UNIVERSIDAD PERUANA DE CIENCIAS APLICADAS

Facultad de Ingeniería

Carrera de Ingeniería de Sistemas de Información

**Modelo Predictivo para el diagnóstico de la Diabetes
Mellitus Tipo 2 soportado por SAP Predictive Analytics**

TESIS

Para optar el título profesional de: Ingeniero de Sistemas de Información

AUTORES

Diego Alberto Ordóñez Barrios (0000-0002-7639-9025)

Erick Raphael Vizcarra Infantes (0000-0002-9137-9743)

Asesor de tesis

Chumpitaz Avendaño, Max Raul (0000-0003-4155-40109)

Lima, 31 de Julio 2018

Resumen Ejecutivo

El presente proyecto se centra en el desarrollo de un modelo predictivo que permite pronosticar el diagnóstico de la diabetes mellitus tipo 2, siendo soportado por la herramienta SAP Predictive Analytics. Tiene como propósito el definir un modelo predictivo cuya implementación permita la optimización del proceso de diagnóstico de la Diabetes Mellitus tipo 2, además permitiendo que el resultado pueda brindar indicios sobre las acciones que una institución prestadora de servicios de cobertura de salud (tanto pública como privada) puede tomar por cada paciente en beneficio del mismo.

Para lograr el propósito del proyecto, se ha realizado una investigación donde hemos alineado las 10 metas mundiales planteadas por la Organización Mundial de la Salud (OMS) a las 4 agrupaciones de enfermedades crónicas de mayor impacto económico, con lo que se ha identificado a la diabetes como la enfermedad crónica de mayor impacto para el Perú debido al creciente factor de incidencia en el país, causado principalmente por serias deficiencias en las costumbres diarias de alimentación y ejercicio en la población peruana, además de ser una enfermedad cuya propagación es alta en países en vías de desarrollo como el Perú debido a que no es mitigada adecuadamente por falta de prevención, desconocimiento o por motivos tan diversos como los económicos. Seguidamente, se realiza un benchmarking de herramientas de Predictive Analytics y las capacidades disponibles de las mismas para identificar cuál de ellas brinda el mejor soporte al modelo predictivo planteado, según el contexto identificado.

Palabras clave: diabetes mellitus; análisis predictivo; Auto Classification algorithm.

Abstract

The project is focused on the development of a predictive model that enables prediction of the development of type 2 diabetes mellitus supported by SAP Predictive Analytics. Its main purpose is the definition of a predictive model that allows institutions that offer health coverage (both public and private) to optimize their diagnostic process and also enables the use of the prediction result in order to determine which actions could be taken, based on medical recommendations, on behalf of the patients benefit.

To achieve the purpose of the project, an investigation has been done where there was an alignment between the 4 main chronic diseases based on their economic impact and the 10 global goals set by the World Health Organization, identifying diabetes as the chronic disease of the biggest impact for Peru due to the growing incidence factor in the country, caused mainly because of serious deficiencies in daily nutritional habits and a lack of workout culture, along with being a disease that has the most incidence in developing countries such as Peru since it is not mitigated accordingly because of lack of prevention, knowledge or economic motives. There has also been a benchmarking of Predictive Analytics tools in order to see which one complies the best with the requirements of both the chronic disease and the Peruvian context.

Keywords: diabetes mellitus; predictive analytics; Auto Classification algorithm.

Tabla de Contenidos

CAPÍTULO 1. DESCRIPCION DEL PROYECTO	12
OBJETO DE ESTUDIO	13
DOMINIO DE LA NECESIDAD	14
PLANTEAMIENTO DE LA SOLUCION	14
OBJETIVOS DEL PROYECTO.....	15
<i>Objetivo General</i>	15
<i>Objetivos Específicos</i>	15
INDICADORES DE ÉXITO	15
PLANIFICACION DEL PROYECTO	16
<i>Alcance</i>	16
<i>Plan de Gestión del Tiempo</i>	16
<i>Plan de Gestión de Recursos Humanos</i>	17
<i>Plan de Comunicaciones</i>	20
<i>Plan de Gestión de Riesgos</i>	22
CAPÍTULO 2. LOGROS DE LOS STUDENT OUTCOMES	27
STUDENT OUTCOME (A).....	28
STUDENT OUTCOME (B).....	29
STUDENT OUTCOME (C).....	29
STUDENT OUTCOME (D)	30
STUDENT OUTCOME (E).....	31
STUDENT OUTCOME (F).....	32
STUDENT OUTCOME (G)	33
STUDENT OUTCOME (H)	35
STUDENT OUTCOME (I).....	36
STUDENT OUTCOME (J).....	37
STUDENT OUTCOME (K).....	37
STUDENT OUTCOME (J) - CAC	38
CAPÍTULO 3. MARCO TEORICO	40
DATOS, INFORMACION Y CONOCIMIENTO	41
DESCUBRIMIENTO DE CONOCIMIENTO EN BASES DE DATOS	42
PREDICTIVE ANALYTICS.....	43
APRENDIZAJE SUPERVISADO Y MACHINE LEARNING	43
MODELOS PREDICTIVOS	43
VALIDACION Y LIMPIEZA DE DATOS	44
DISCRETIZACION	44

ENFERMEDADES CRONICAS	45
DIABETES	45
CAPÍTULO 4. ESTADO DEL ARTE.....	47
REVISION DE LA LITERATURA	48
<i>Redes Neuronales</i>	48
<i>Árboles</i>	49
<i>K-vecinos más cercanos</i>	49
<i>Redes bayesianas</i>	50
<i>LEM2</i>	50
<i>Jerarquía difusa</i>	50
<i>Algoritmos distribuidos</i>	50
MODELOS PREDICTIVOS EN SALUD	51
<i>Desarrollo de un Modelo Predictivo para la Diabetes Mellitus de Tipo 2 usando Data Clínica y Genética</i>	51
<i>Predicción de Diabetes con Algoritmos de Aprendizaje Supervisado de Redes Neuronales Artificiales</i>	53
<i>Mejorando el Ratio de Predicción de Diagnósticos de Diabetes usando Lógica Difusa y Redes Neuronales</i>	55
<i>ARIC: Modelo Predictivo de Riesgo de Diabetes Tipo 2</i>	57
<i>Evaluación de Algoritmos de Redes Neuronales para Modelos de Predicción de Diabetes Tipo 2</i>	60
<i>Sistema de Recomendación Confiable y Objetivo para el Diagnóstico de Enfermedades Crónica</i>	62
<i>Modelo predictivo de diabetes con algoritmos de Percepción Multicapa, Redes Bayesianas y J48</i>	64
<i>Modelo de predicción de diabetes con el algoritmo J48 mejorado</i>	66
<i>Predicción de Diabetes usando una Red Bayesiana</i>	68
<i>Diagnóstico de Diabetes utilizando Técnicas de Clasificación y Minería</i>	69
<i>H2RM: Un Modelo de Razonamiento Híbrido para la Predicción y Gestión de la Diabetes Mellitus</i> ...	70
<i>Detección Temprana de Diabetes Mellitus usando un Modelo Jerárquico Difuso</i>	72
<i>Modelo predictivo de Diabetes en Hadoop/Map reduce</i>	73
<i>Modelo predictivo de puntuación de Riesgo de Salud</i>	76
<i>Modelo predictivo escalable con RapidMiner</i>	80
CAPÍTULO 5. MODELO PREDICTIVO PARA EL DIAGNOSTICO DE LA DIABETES MELLITUS TIPO 2	83
ANALISIS DE ENFERMEDADES CRONICAS.....	84
<i>Definición</i>	84
<i>Objetivo</i>	84
<i>Alcance</i>	84
<i>Contexto</i>	84
<i>Identificación de Enfermedades</i>	85
<i>Relación entre las Enfermedades Crónicas y las Metas Mundiales de la OMS</i>	88

<i>Cifras de Enfermedades Crónicas.....</i>	<i>90</i>
<i>Resultado del análisis.....</i>	<i>94</i>
BENCHMARKING DE HERRAMIENTAS DE PREDICTIVE ANALYTICS	95
<i>Definición</i>	<i>95</i>
<i>Objetivo.....</i>	<i>95</i>
<i>Alcance.....</i>	<i>95</i>
<i>Contexto.....</i>	<i>96</i>
<i>Herramientas y Criterios de Evaluación</i>	<i>99</i>
<i>Análisis de las herramientas</i>	<i>100</i>
<i>Benchmarking resultante.....</i>	<i>124</i>
<i>Análisis de Algoritmos.....</i>	<i>125</i>
<i>Conclusiones del Análisis de Algoritmos</i>	<i>129</i>
ELABORACION DEL MODELO PREDICTIVO	130
<i>Definición</i>	<i>130</i>
<i>Objetivo.....</i>	<i>130</i>
<i>Alcance.....</i>	<i>130</i>
<i>Contexto.....</i>	<i>130</i>
<i>Conjuntos de Datos.....</i>	<i>131</i>
<i>Clasificación de Datos</i>	<i>131</i>
<i>Recopilación de Datos de Prueba.....</i>	<i>132</i>
<i>Preparación de Repositorio de Datos.....</i>	<i>133</i>
<i>Preparación de la Herramienta.....</i>	<i>134</i>
<i>Definición del Proceso de Aplicación del Modelo Predictivo</i>	<i>135</i>
<i>Propuesta de Continuidad.....</i>	<i>136</i>
<i>Conclusiones del Modelo Predictivo.....</i>	<i>139</i>
CAPÍTULO 6. RESULTADOS Y VALIDACION DEL MODELO PREDICTIVO	140
<i>EVALUACION Y DEFINICION DE ALGORITMOS</i>	<i>141</i>
<i>DEFINICION DE PROCEDIMIENTO DE INGRESO DE INFORMACION</i>	<i>141</i>
<i>MODELO PREDICTIVO RESULTANTE.....</i>	<i>142</i>
CAPÍTULO 7. GESTION DEL PROYECTO.....	145
<i>PRODUCTO FINAL</i>	<i>146</i>
<i>GESTION DEL TIEMPO</i>	<i>146</i>
<i>GESTION DE LOS RECURSOS HUMANOS</i>	<i>147</i>
<i>GESTION DE LAS COMUNICACIONES.....</i>	<i>149</i>
<i>GESTION DE LOS RIESGOS</i>	<i>149</i>
<i>LECCIONES APRENDIDAS</i>	<i>151</i>
REFERENCIAS BIBLIOGRAFICAS	154

REFERENCIAS ELECTRONICAS.....	156
ANEXO 1: ACTA DE VALIDACION – EXPERTO EN MEDICINA.....	166
ANEXO 2: ACTA DE VALIDACION – EXPERTO EN TECNOLOGIA	168
ANEXO 3: ACTA DE VALIDACION – CLIENTE DEL PROYECTO	170

Índice de Tablas

Tabla 1: Hitos del Proyecto	16
Tabla 2: Roles y miembros del equipo	18
Tabla 3: Matriz de Comunicaciones del Proyecto	20
Tabla 4: Evaluación del Riesgo	23
Tabla 5: Roles y responsabilidades	24
Tabla 6: Mitigación de riesgos	25
Tabla 7: Coeficiente de correlación por algoritmo	55
Tabla 8: Comparación de las funciones predictivas multivariantes publicadas con las estimadas localmente.....	58
Tabla 9: Comparación de las características del área bajo la curva del funcionamiento del receptor de varios modelos predictivos usando NHS-1992.....	59
Tabla 10: Calidad de calibración de varios modelos predictivos usando NHS-1992	59
Tabla 11: Atributos del conjunto de datos de Diabetes	60
Tabla 12: Clasificación de la precisión de distintos algoritmos	61
Tabla 13: Atributos del conjunto de datos Pima Indian	64
Tabla 14: Datos estadísticos de todas las características	65
Tabla 15: Clasificación de Resultados.....	65
Tabla 16: Comparación de desempeño de algoritmos.....	67
Tabla 17: Resultados de diagnóstico de diabetes por clase	69
Tabla 18: Conjunto de datos vinculados usados en el estudio.	76
Tabla 19: Características de la muestra	77
Tabla 20: Puntuaciones de riesgo correlativos	77
Tabla 21: Asociación entre las puntuaciones de riesgos y nuevos diagnósticos de enfermedades en años subsecuentes.....	78
Tabla 22: Curvas de supervivencia Kaplan-Meier para la diabetes	79
Tabla 23: Modelos Cox de riesgos proporcionales para el diagnóstico y la mortalidad por enfermedad, por quintiles de puntuación de riesgo.	80
Tabla 24 Resultados de pruebas de algoritmos	141
Tabla 25 Hitos del proyecto.....	146
Tabla 26 Roles del proyecto	147
Tabla 27 Analisis de riesgos	149

Índice de Ilustraciones

Ilustración 1: Organigrama del proyecto	17
Ilustración 2: Clasificación de riesgos	22
Ilustración 3: Matriz de probabilidades de impacto	23
Ilustración 4 Representación gráfica del Student Outcome C	30
Ilustración 5 Representación gráfica del Student Outcome D	31
Ilustración 6 Representación gráfica del Student Outcome E	32
Ilustración 7 Representación gráfica del Student Outcome E	32
Ilustración 8 Representación gráfica del Student Outcome F	33
Ilustración 9 Representación gráfica del Student Outcome G	35
Ilustración 10 Representación gráfica del Student Outcome H	36
Ilustración 11 Representación gráfica del Student Outcome J - CAC	39
Ilustración 12: El modelo de Datos-Información-Conocimiento (Aamodt, Nygard 1995)	41
Ilustración 13: Proceso de Descubrimiento de Conocimiento (Maimon, Rokach 2005) ..	42
Ilustración 14: Metodos de discretización	45
Ilustración 15: Dispersión entre triglicéridos, índice de masa corporal e incidencia de diabetes mellitus tipo 2	52
Ilustración 16: Medición de desempeño de algoritmos	54
Ilustración 17: Dispersión de resultados de los algoritmos frente al coeficiente de correlación R	55
Ilustración 18: Ecuación de cálculo de métrica MAE	63
Ilustración 19: Flujo del algoritmo propuesto	66
Ilustración 20: Ecuaciones de evaluación	67
Ilustración 21: Clasificación de clases de diagnóstico de diabetes	68
Ilustración 22: Arquitectura del Sistema de Análisis Predictivo en aplicaciones de Salud	73
Ilustración 23: Sistema Predictivo – Coincidencia de Patrones	74
Ilustración 24: Proceso básico de construcción automática, optimización de la parametrización y evaluación de múltiples modelos predictivos en RapidMiner.	81
Ilustración 25: Conjunto de los métodos de aprendizaje como se muestra en RapidMiner	81
Ilustración 26: Mortalidad Proporcional de ENT	92
Ilustración 27: Tasa de Mortalidad de ENT	93

Ilustración 28: Evolución de la mortalidad por cáncer.....	93
Ilustración 29: Tendencias en la diabetes	94
Ilustración 30: Arquitectura de Plataforma Angoss	102
Ilustración 31: Arquitectura de Plataforma Alpine Data	104
Ilustración 32: Arquitectura de Plataforma FICO	110
Ilustración 33: Proceso de Analítica Predictiva.....	113
Ilustración 34 Cuadro de aplicación del modelo predictivo.	135
Ilustración 35 Plan de continuidad del modelo predictivo.	138
Ilustración 36 Modelo Predictivo.	143

Introducción

El proyecto de investigación tiene como propósito desarrollar un modelo predictivo para el diagnóstico de la diabetes mellitus tipo 2, soportado por SAP Predictive Analytics. Es un proyecto definido y aplicado para el contexto peruano, enfocado directamente a las instituciones prestadoras de servicios de cobertura de salud tanto públicas como privadas y se basa en el análisis sintomático del paciente, identificando aquellos factores que intervienen en el desarrollo y desenlace de la diabetes mellitus tipo 2, con el propósito de brindar a entidades prestadoras de servicios de salud información que les sea de utilidad para determinar perfiles de riesgo por cliente, además de ser aplicable también para mejorar la situación de salud pública en el país al poderse usar como base para investigaciones científicas a futuro que motiven la concientización del riesgo real que se tiene de desarrollar la enfermedad. Con el fin de brindar un modelo de trascendencia para la realidad peruana, se desarrolla una investigación sobre la enfermedad crónica de mayor factor epidemiológico en el Perú, identificándola como la diabetes y además se identificó que sus causas podían ser identificadas o deducidas por ser propias de los hábitos diarios. Asimismo, se desarrolla un benchmarking de herramientas de Predictive Analytics para identificar cuál supone el mejor soporte al modelo planteado y a la realidad de las instituciones identificadas como foco.

CAPÍTULO 1. Descripción del Proyecto

En este capítulo se describe todo lo que respecta al proyecto y los datos del mismo, en el presente capítulo se van a definir cuáles son los objetivos del proyecto, el problema a resolver, la motivación de la investigación, los indicadores de éxito y la planificación del proyecto, entre otras cosas.

OBJETO DE ESTUDIO

El proyecto tiene como título: *“Modelo Predictivo para el diagnóstico de la Diabetes Mellitus Tipo 2 soportado por SAP Predictive Analytics”* el cual plantea realizar un análisis sobre la incidencia de enfermedades crónicas en el Perú, identificando que se trata de la diabetes debido a que es la enfermedad de mayor trascendencia según su proyección de propagación y el impacto que tiene a nivel económico, así como hacer un benchmarking entre herramientas de análisis de Predictive Analytics que provean un marco adecuado para el soporte del modelo predictivo desarrollado, considerando las variables propias de la herramienta tales como costo, escalabilidad y catálogo de algoritmos disponible, identificando esta solución como SAP Predictive Analytics. En base a ambas mediciones, se define la necesidad de diseñar un modelo predictivo que permita diagnosticar la diabetes mellitus tipo 2, y a través del uso del modelo se plantea que las entidades prestadoras de servicios de salud públicas y privadas lo implementen en un contexto comercial, utilizando las habilidades predictivas del modelo en beneficio del diagnóstico del cliente y la optimización de procesos de consulta. Finalmente, se busca desarrollar un modelo con continuidad asegurada, que actúe en beneficio de organizaciones públicas como privadas que se encuentren en el rubro de salud en el Perú. Es necesario resaltar que el modelo predictivo no está basado en el desarrollo de un modelo matemático ni de algoritmos para el desarrollo de la predicción, sino en el uso de algoritmos disponibles en la herramienta identificada.

A continuación, se presentan las definiciones principales que sustentan el proyecto a realizar.

Predictive Analytics

Es un conjunto de tecnologías de inteligencia de negocios que descubren relaciones y patrones en largos volúmenes de datos que pueden ser usados para predecir comportamiento y eventos. A diferencia de otras tecnologías de inteligencia de negocios, predictive analytics ofrece una vista hacia el futuro, usando eventos pasados para predecirlo.¹

Enfermedades Crónicas

Las enfermedades crónicas son enfermedades de larga duración y por lo general de progresión lenta. Las enfermedades cardíacas, los infartos, el cáncer, las

¹ Cfr. The Data Warehousing Institute 2007

enfermedades respiratorias y la diabetes, son las principales causas de mortalidad en el mundo, siendo responsables del 63% de las muertes. En 2008, 36 millones de personas murieron de una enfermedad crónica, de las cuales la mitad era de sexo femenino y el 29% era de menos de 60 años de edad.²

DOMINIO DE LA NECESIDAD

Necesidad

Necesidad de optimización del diagnóstico de la diabetes mellitus tipo 2 a través de un proceso de evaluación de rasgos sintomáticos y hábitos cotidianos empleando técnicas de Predictive Analytics.

PLANTEAMIENTO DE LA SOLUCION

La solución propuesta es el *modelo predictivo que permita diagnosticar la diabetes mellitus tipo 2, soportado por SAP Predictive Analytics*.

El propósito del modelo predictivo es brindar un análisis predictivo sobre el diagnóstico de la diabetes y que permita a las organizaciones prestadoras de servicios de cobertura de salud acceder a información de diagnóstico de diabetes mellitus tipo 2 de manera rápida, a través del análisis sintomático y hábitos cotidianos del paciente.

Se utiliza un algoritmo de predicción, el cual no será desarrollado en este proyecto. El algoritmo será de la herramienta de SAP Predictive Analytics, y es elegido según se adecúe mejor al análisis predictivo de enfermedades, por lo cual dentro del Benchmarking de herramientas de Predictive Analytics, se realiza un análisis sobre las capacidades técnicas/algoritmos disponibles en cada herramienta que suponía una adecuada opción para el soporte del modelo.

Como parte de la gestión del proyecto se usa la metodología PMBOK y para el desarrollo del proyecto se usa la metodología Ad-Hoc, por lo cual no se generan entregables de una metodología en específico, solo hay entregables propios de PMBOK, sino que al usar Ad-Hoc se generan entregables propios del proyecto, los cuales son definidos en el cronograma del proyecto.

² Cfr. OMS 2016

OBJETIVOS DEL PROYECTO

Objetivo General

Elaborar un modelo predictivo que permita pronosticar el diagnóstico de la diabetes mellitus tipo 2 soportado por SAP Predictive Analytics.

Objetivos Específicos

OE1: Analizar los principales tipos de enfermedades crónicas y las herramientas de Predictive Analytics que permitan la implementación de un modelo predictivo de enfermedades crónicas para el contexto de instituciones del rubro de salud públicas y privadas en el Perú.

OE2: Diseñar un modelo predictivo que permita pronosticar el diagnóstico de la Diabetes Mellitus tipo 2, siendo soportado por SAP Predictive Analytics.

OE3: Validar el modelo predictivo a través del juicio de un experto en medicina.

OE4: Asegurar la continuidad del modelo permitiendo diagnosticar el desarrollo de la diabetes mellitus tipo 2 a través de un modelo dinámico, escalable y de mejora continua.

INDICADORES DE ÉXITO

Para OE1:

IE1: Obtención de la constancia de calidad de QS sobre el análisis e identificación de la enfermedad crónica de mayor factor epidemiológico en el Perú y la plataforma de Predictive Analytics adecuada para el contexto peruano, además de validar la calidad del entregable.

Para OE2:

IE2: El modelo diseñado debe ser aprobado por un especialista en medicina y por el cliente del proyecto.

Para OE3:

IE3: Aprobación por parte del experto en medicina sobre la implementación del modelo predictivo diseñado en un ambiente controlado con data recopilada.

Para OE4:

IE4: Conformidad por parte de un experto en tecnologías de Predictive Analytics sobre el modelo diseñado, implementado y la continuidad del modelo.

PLANIFICACION DEL PROYECTO

Alcance

El proyecto se desarrolla en base al análisis de la identificación de la enfermedad crónica de mayor epidemiología en Perú, considerando además las variables que intervienen en la obtención de la misma, tales como factores características del paciente y hábitos cotidianos. Se realiza un benchmarking en herramientas de Predictive Analytics considerando la aplicabilidad de la herramienta en base a la enfermedad identificada y modelo diseñado. El proyecto no contempla el desarrollo de algoritmos de predicción de modelos, por el contrario, plantea usar los algoritmos disponibles en la herramienta elegida (siendo esta una variable a considerar en el benchmarking a realizar).

Plan de Gestión del Tiempo

El proyecto se desarrolla en los ciclos 2016-01 y 2016-02, para lo cual los hitos y entregables son los siguientes:

Tabla 1: Hitos del Proyecto

Hito del proyecto	Fecha Estimada	Entregables incluidos	Prioridad
Definición del proyecto	Semana 3 2016-1	<ul style="list-style-type: none">• Project Chárter• Documentos de planes de Gestión	Alta
Análisis de enfermedades crónicas	Semana 6 2016-1	<ul style="list-style-type: none">• Informe de análisis de enfermedades crónicas	Media
Análisis de herramientas de Predictive Analytics (Benchmarking)	Semana 8 2016-1	<ul style="list-style-type: none">• Benchmarking de herramientas de Predictive Analytics.	Alta
Diseño de modelo predictivo	Semana 6 2016-2	<ul style="list-style-type: none">• Modelo predictivo que permita diagnosticar la Diabetes Mellitus Tipo 2 soportado por SAP Predictive Analytics	Alta
Implementación del modelo predictivo	Semana 13 2016-2	<ul style="list-style-type: none">• Prueba del modelo predictivo en ambiente controlado	Alta

		<ul style="list-style-type: none"> • Informe de resultado de implementación • Memoria de tesis • Paper de proyecto 	
--	--	---	--

Fuente: Elaboración propia.

Plan de Gestión de Recursos Humanos

El equipo del proyecto está organizado según el siguiente organigrama, en él estamos mostrando claramente las líneas de reporte de cada miembro del equipo.

Ilustración 1: Organigrama del proyecto

Fuente: Elaboración propia

A continuación, se detalla los roles y miembros de cada uno:

Tabla 2: Roles y miembros del equipo

Rol	Miembro	Responsabilidades
Comité de Proyecto	Rosario Villalta Jimmy Armas Oscar Gómez Pedro Shiguihara	<ul style="list-style-type: none"> - Dar los lineamientos generales. - Controlar el cumplimiento del proyecto. - Verificar el cumplimiento de los Student Outcomes.
Profesor Gerente IT Consulting	Max Chumpitaz	<ul style="list-style-type: none"> - Dar seguimiento y control al proyecto. - Brindar asesoramiento en la planificación de proyectos.
Alumno Gerente	Eduardo Ulloa	<ul style="list-style-type: none"> - Dar seguimiento y control al proyecto.
Asesor	Jimmy Armas	<ul style="list-style-type: none"> - Brindar asesoramiento en el desarrollo y gestión del proyecto.
Profesor Revisor	Andrés Melgar	<ul style="list-style-type: none"> - Brindar asesoramiento en el desarrollo y control de la memoria del proyecto.
Jefe de Proyecto	Diego Ordóñez Barrios	<ul style="list-style-type: none"> - Desarrollar el análisis de enfermedades crónicas y su epidemiología. - Desarrollar el benchmarking de herramientas de Predictive Analytics, identificando la más apropiada para el propósito del proyecto. - Diseñar un modelo predictivo del desarrollo de una enfermedad crónica soportado por SAP Predictive Analytics.

		<ul style="list-style-type: none"> - Implementar el modelo predictivo en un ambiente controlado para simular el correcto desarrollo del mismo. - Asegurar la calidad de todos los entregables a través de la duración del proyecto. - Asegurar la continuidad del modelo predictivo planteado en un ámbito real de las Instituciones prestadoras de servicios de cobertura de salud.
Jefe de Proyecto	Erick Vizcarra Infantes	<ul style="list-style-type: none"> - Desarrollar el análisis de enfermedades crónicas y su epidemiología. - Desarrollar el benchmarking de herramientas de Predictive Analytics, identificando la más apropiada para el propósito del proyecto. - Diseñar un modelo predictivo del desarrollo de una enfermedad crónica soportado por SAP Predictive Analytics. - Implementar el modelo predictivo en un ambiente controlado para simular el correcto desarrollo del mismo. - Asegurar la calidad de todos los entregables a través de la duración del proyecto. - Asegurar la continuidad del modelo predictivo planteado en un ámbito real de las Instituciones prestadoras de servicios de cobertura de salud.
Recurso QA	Recurso QS	<ul style="list-style-type: none"> - Validar la calidad de los artefactos - Validar la trazabilidad de los entregables y/o artefactos.

Fuente: Elaboración propia.

Plan de Comunicaciones

Para asegurarnos de que todas las necesidades de comunicación de los stakeholders a lo largo del proyecto sean contempladas se ha realizado el siguiente análisis de stakeholders y luego de él se ha plasmado cada una de sus necesidades en la matriz. A continuación, se muestra la matriz de comunicaciones definidas para el proyecto.

Tabla 3: Matriz de Comunicaciones del Proyecto

Información requerida	Contenido	Respons. de elaborarlo	Para su entrega a los Stakeholders	Método de comunicación a utilizar	Frecuencia	Código de EDT asociado
Inicio del Proyecto	Elaboración y aprobación del Project Charter	Jefes de Proyecto	- Profesor Cliente - Profesor Gerente - Miembro del Comité	- Documento Físico	- Única vez	1.1.4
Planificación del Proyecto	Elaboración de los Documentos de Gestión del Proyecto	Jefes de Proyecto	- Profesor Cliente - Profesor Gerente	- Documento Físico - Documento Digital	- Única vez	2.2.3
Planificación del Proyecto	Elaboración de documentación de seguimiento de proyecto	Jefes de Proyecto	- Profesor Cliente - Profesor Gerente	- Documento Físico - Documento Digital	- Semanalmente	2.1.2
Ejecución del Proyecto	Elaboración de los entregables de Análisis del Proyecto	Jefes de Proyecto	- Profesor Cliente - Profesor Gerente - Comité del Proyecto	- Documento Físico - Documento Digital	- Semanalmente	3.3.5 3.5.4
Ejecución del Proyecto	Elaboración de los entregables de la fase de ejecución del Proyecto	Jefes de Proyecto	- Profesor Cliente - Profesor Gerente - Comité del Proyecto	- Documento Físico - Documento Digital	- Semanalmente	3.15.2 3.18.2 3.19.2

Cierre del Proyecto	Gestión de conformidades, actas de aceptación y cierre del proyecto	Jefes de Proyecto	<ul style="list-style-type: none"> - Profesor Cliente - Profesor Gerente - Comité del Proyecto 	<ul style="list-style-type: none"> - Documento Físico - Documento Digital 	- Única vez	5.1.2 5.1.3 5.1.4
---------------------	---	-------------------	---	---	-------------	-------------------------

Fuente: Elaboración propia

Plan de Gestión de Riesgos

Se hará uso de la siguiente RBS para la categorización de los riesgos:

Ilustración 2: Clasificación de riesgos

Fuente: Elaboración propia.

- Escala de probabilidad: Para medir la probabilidad de ocurrencia del riesgo se usará la siguiente escala lineal:
 - 0.1 muy improbable
 - 0.3 improbable
 - 0.5 moderado
 - 0.7 probable
 - 0.9 casi certeza

- Escala de impacto: Para medir el impacto de la ocurrencia del riesgo se usará la siguiente escala lineal:
 - 0.1 muy bajo
 - 0.3 bajo
 - 0.5 moderado
 - 0.7 alto
 - 0.9 muy alto

- Matriz de Probabilidad e impacto: en ella se ilustran los criterios para la definición de escalas de la severidad de los riesgos. Para este proyecto se ha definido la siguiente escala de severidad del riesgo:
 - > 0.25 - Criticidad Alta - **Rojo**
 - $0.07 > x < 0.25$ – Criticidad Moderada - **Amarillo**
 - ≤ 0.07 – Criticidad Baja – **Verde**

En función a estas escalas se ha establecido la siguiente matriz:

Ilustración 3: Matriz de probabilidades de impacto

Probabilidad	Amenazas					Oportunidades				
1	0,10	0,30	0,50	0,70	1,00	1,00	0,70	0,50	0,30	0,10
0.7	0,07	0,21	0,35	0,49	0,70	0,70	0,49	0,35	0,21	0,07
0.50	0,05	0,15	0,25	0,35	0,50	0,50	0,35	0,25	0,15	0,05
0.30	0,03	0,09	0,15	0,21	0,30	0,30	0,21	0,15	0,09	0,03
0.10	0,01	0,03	0,05	0,07	0,10	0,10	0,07	0,05	0,03	0,01
	0.10	0.30	0.50	0.7	1	1	0.7	0.50	0.30	0.10

Impacto (escala de relación) sobre un objetivo (por ejemplo, coste, tiempo, alcance o calidad).
Cada riesgo es clasificado de acuerdo con su probabilidad de ocurrencia y el impacto sobre un objetivo en caso de que ocurra.
Los umbrales de la organización para riesgos bajos moderados o altos se muestran en la matriz y determinan si el riesgo es calificado como alto o bajo para ese objetivo.

Fuente: Elaboración propia.

- Los umbrales para la definición del impacto para los cuatro objetivos (costo, alcance, tiempo y calidad) del proyecto se encuentran en el cuadro que se presenta a continuación:

Tabla 4: Evaluación del Riesgo

Condiciones Definidas para Escalas de Impacto de un Riesgo sobre los Principales Objetivos del Proyecto (Sólo se muestran ejemplos para impactos negativos)					
Objetivo del Proyecto	Se muestran escalas relativas o numéricas				
	Muy bajo / 0.10	Bajo / 0.30	Moderado / 0.50	Alto / 0.70	Muy alto / 0.90
Costo	Aumento de costo insignificante	Aumento del costo < 10%	Aumento del costo del 10-20%	Aumento del costo del 20-40%	Aumento del costo > 40%
Tiempo	Aumento de tiempo insignificante	Aumento del tiempo < 5%	Aumento del tiempo del 5-10%	Aumento del tiempo del 10 - 20%	Aumento del tiempo > 20%
Alcance	Disminución del alcance apenas apreciable	Áreas de alcance secundarias afectadas	Áreas de alcance principales afectadas	Reducción del alcance inaceptable para el patrocinador	El elemento terminado del proyecto es efectivamente inservible
Calidad	Degradación de la calidad apenas perceptible	Sólo las aplicaciones muy exigentes se ven afectadas	La reducción de la calidad requiere la aprobación del patrocinador	Reducción de la calidad inaceptable para el patrocinador	El elemento terminado del proyecto es efectivamente inservible
<p>Esta tabla presenta ejemplos de definiciones de impacto de los riesgos para cuatro objetivos del proyecto diferentes. Estos deben adaptarse al proyecto individual y a los umbrales de riesgo de la organización en el proceso Planificación de la Gestión de Riesgos. Las definiciones del impacto pueden desarrollarse para las oportunidades de forma similar.</p>					

Fuente: Elaboración propia

Roles y Responsabilidades

Tabla 5: Roles y responsabilidades

Roles para el manejo de riesgos	Miembro	Responsabilidades
Jefes de Proyecto	Diego Ordóñez Erick Vizcarra	<ul style="list-style-type: none"> ✓ Identificar los riesgos principales del proyecto según el alcance planteado y evaluar sus causas. ✓ Evaluar de los riesgos a través de un análisis de impacto y probabilidad. ✓ Identificar y analizar estrategias de gestión para el manejo de riesgos. ✓ Establecer acciones adecuadas contra los riegos identificados, según la identificación sobre la actitud a adoptar por cada riesgo. ✓ Establecer controles y procedimientos para una adecuada gestión del riesgo. ✓ Monitorear constantemente los riesgos identificados y buscar nuevos factores de riesgo. ✓ Establecer planes de contingencia ✓ Identificar y analizar controles/estrategias aplicables a los riesgos registrados. ✓ Validar la eficacia de los controles/estrategias diseñados
Gerente Alumno IT Consulting	Eduardo Ulloa	<ul style="list-style-type: none"> ✓ Monitorear constantemente los riesgos de la cartera de proyectos de IT Consulting. ✓ Monitorea el desempeño de los proyectos IT Consulting.
Profesor Gerente de IT Consulting	Max Chumpitaz	<ul style="list-style-type: none"> ✓ Monitorear constantemente los riesgos del proyecto. ✓ Brindar asesorías sobre estrategias que ayuden a minimizar los riesgos y su impacto. ✓ Contribuir con el diseño de controles que sean aplicables al proyecto de acuerdo a su naturaleza.
Profesor Cliente	Jimmy Armas	<ul style="list-style-type: none"> ✓ Monitoreo constantemente el desarrollo adecuado del proyecto. ✓ Contribuir con el diseño de controles que sean aplicables al proyecto de acuerdo a su naturaleza.
Comité de Proyectos	Rosario Villalta Jimmy Armas Oscar Gómez	<ul style="list-style-type: none"> ✓ Monitoreo del desarrollo adecuado del proyecto y de los entregables finales validados. ✓ Constatar la calidad de los entregables finales realizados y validar el alcance de los logros propuestos.

	Pedro Shiguihara	
Recurso Quality Services	Recurso QS	✓ Validación de la concordancia y correcto desarrollo de los distintos entregables del proyecto.
Revisor	Andrés Melgar	✓ Apoyo y asesoría para la elaboración de la memoria.

Fuente: Elaboración propia

Mitigación de riesgos

- En la siguiente tabla estarán los riesgos que suponemos estarán presentes en el proyecto y su estrategia de mitigación. Estos riesgos estarán registrados en el registro de riesgos en caso lleguen a suceder en el transcurso del proyecto

Tabla 6: Mitigación de riesgos

Cód. Riesgo	Descripción del Riesgo	Probabilidad	Impacto	Tipo de Estrategia	Descripción de la estrategia planteada
IMPECDB-R001	El recurso de QS no cumpla con las funciones que se le delega.	30% Baja	2 Baja	Evitar	<i>Notificar al líder de línea del recurso para exigir que cumpla con sus responsabilidades mediante seguimiento continuo.</i>
IMPECDB-R002	Insuficientes fuentes disponibles para investigación requerida para el desarrollo del proyecto	10% Muy baja	4 Alta	Aceptar - Activamente	<i>Realizar una búsqueda rigurosa de fuentes físicas o digitales con prudente tiempo de anticipación.</i>
IMPECDB-R003	Escasa disponibilidad del Cliente para las reuniones pactadas.	30% Baja	2 Baja	Aceptar - Activamente	<i>Pactar anticipadamente las fechas de las reuniones y/o en caso ocurra un imprevisto reprogramar instantáneamente la reunión.</i>

IMPECDB-R004	Constante cambio del alcance del proyecto.	30% Baja	4 Alta	Mitigar	<i>Mantener comunicación con el cliente y la gerencia del desarrollo del proyecto para que estén alineados al alcance y objetivos del proyecto y aprueben cualquier cambio que pudiera ocurrir. Ejecutar procedimiento de cambios de alcance en el proyecto.</i>
IMPECDB-R005	Pérdida de información digital del proyecto.	10% Muy baja	5 Muy alta	Evitar	<i>Realizar copias digitales periódicamente de la información como estrategia de contingencia, en diversas plataformas Cloud y ambientes físicos.</i>
IMPECDB-R006	Retraso en la recepción de observaciones del revisor	50% Media	1 Muy Baja	Mitigar	<i>Coordinar los plazos definidos por la escuela con el revisor y enviar la documentación a tiempo para que se pueda cumplir con ellos.</i>

Fuente: Elaboración propia

Seguimiento y control de riesgos

- Los jefes de proyecto son los únicos responsables del desarrollo adecuado del proyecto, realizando una constante identificación de riesgos y definiendo y planes de contingencia en caso se concreten riesgos que impacten en el proyecto o supongan una mayor exposición del proyecto.
- Los jefes de proyecto tienen la responsabilidad de actuar ante un riesgo materializado de manera inmediata, siguiendo las estrategias definidas por cada riesgo identificado.
- Los jefes de proyecto deben informar inmediatamente al cliente e involucrados directos al proyecto en caso se materialicen los riesgos de modo que puedan abordarse en un debido tiempo colaborando esfuerzos para mitigarlos.

CAPÍTULO 2. Logros de los Student Outcomes

En este capítulo se describe y detalla el logro de cada uno de los 12 Students Outcomes estipulados por la acreditadora ABET para la carrera de Ingeniería de Sistemas de Información. El objetivo principal del programa establecido por ABET es evidenciar el cumplimiento de las habilidades, conocimientos y comportamientos que se adquieren en el desarrollo del proyecto.

STUDENT OUTCOME (A)

Aplica conocimientos de matemáticas, ciencias, computación e ingeniería.

El cumplimiento del Student Outcome (A) se ve reflejado a lo largo del proyecto a través de diversas etapas e hitos del mismo, que se ven plasmados mediante diversos entregables y documentos y configuraciones, entre ellos:

1. En el ***Benchmarking de Herramientas de Predictive Analytics*** se realiza un análisis de diversas plataformas de Predictive Analytics para hacer una matriz de evaluación de herramientas según criterios definidos por la consultora Forrester. En la evaluación se realizaron modificaciones a los pesos de acuerdo a nuestra necesidad de uso de la herramienta, se tomó como base a Forrester. Una vez elegido la mejor herramienta para nuestras necesidades se realizó una evaluación de los algoritmos del SAP Predictive Analytics. Se analizaron las familias de algoritmos para determinar cual tiene aplicabilidad para nuestro modelo predictivo, siendo las familias de Árboles de Decisiones, Regresión y Redes neuronales/clasificación las familias que brindan resultados coherentes para la predicción de la diabetes, la comparación de dichas familias se realizó estadísticamente para determinar la precisión de los algoritmos según los resultados de la predicción y los resultados esperados, realizando ejecuciones del mismo modelo para distintos grupos de datos como manera de garantizar la validez de la predicción.
2. Como parte del diseño del modelo predictivo, se realizó el ***Diseño de Base de datos*** necesario para almacenar la data para el modelo, este diseño incluyó la creación de procesos de transformación y carga, stored procedures y funciones para realizar transformaciones de la información, necesarias para poder realizar la predicción. Así también se realizó el ***diseño de la estructura de datos*** de la base de datos que soporte la entrada y almacenamiento de información luego de la predicción. En síntesis, en cuanto a la solución a nivel de repositorio de datos se realizó la definición estructural completa necesaria para definir como recibir la información que alimenta el modelo y como guardarla de tal manera que sea útil para el uso que se propone del modelo.
3. Dentro del plan de continuidad del proyecto se diseñó y construyó una ***Web de captación de información*** para hacer crecer el volumen de información de entrada para el modelo, a mayor información se tendrá predicciones más precisas. La web se construyó empleando Wordpress como motor y requirió conocimiento sobre despliegues web y la gestión adecuada de formularios según el diseño que se buscaba tener. Además, fue desplegada de manera pública y cuenta con hosting público.
4. Se realizó una ***Clasificación de datos*** (discretización) necesaria previa al ingreso de la data al modelo, debido a que la información se obtuvo mediante encuestas se tenía registros con información en un formato distinto al formato de entrada o simplemente se debería realizar transformaciones como por ejemplo para el cálculo del IMC, así también se realizó una agrupación por rangos de respuestas para categorizar de una mejor manera el input de información.
5. Se definió de forma matemática el cálculo de variables mínimas necesarias para la validez del modelo, además de trabajar en proporciones correctas de datos aleatorios para las fases de entrenamiento y validación del modelo realizado.

STUDENT OUTCOME (B)

Diseña y conduce experimentos en base al análisis e interpretación de datos.

El cumplimiento del Student Outcome (B) se ve reflejado claramente en los siguientes entregables o documentos del proyecto, entre ellos:

1. Elaboración y ejecución de **Encuestas de captación de información**, las cuales fueron construidas luego del análisis de factores de riesgo determinantes y significativos, tomando en cuenta estudios previos para la predicción de la diabetes y asesoría con un especialista médico, estas variables no requieren de algún análisis de sangre o intrusivo, son variables que un paciente puede conocer sin ningún problema. Como primera instancia se distribuyó esta encuesta para recopilar información, luego como parte de la continuidad del proyecto se diseñó una web para recopilar esa información y almacenarla en una base de datos.
2. Como parte de la implementación del modelo, se realizaron **Pruebas de algoritmos** para evaluar los resultados, y determinar si lo predicho es correcto o no, por lo que se hizo un análisis de los resultados de acuerdo a la precisión de los algoritmos a través de una prueba para conjuntos diferentes de datos con el propósito de garantizar la precisión del algoritmo elegido a través de dos fases: la primera fase siendo la de entrenamiento del modelo y la segunda siendo la validación real del mismo a través de la ejecución del entrenamiento con otro conjunto de datos validando sus resultados.
3. Con la información recopilada, se tuvo una base de información que no estaba lista para entrar al modelo, tuvo que pasar por un proceso de pre procesamiento para quedar listo como input al modelo predictivo, luego se **analizó la información, transformó y clasificó la data (discretización)**, para finalmente diseñar la **estructura de datos** que soporta el modelo predictivo.

STUDENT OUTCOME (C)

Diseña sistemas, componentes o procesos para encontrar soluciones en la atención de necesidades teniendo en cuenta restricciones económicas, sociales, políticas, éticas, de salud y seguridad y otras propias del entorno empresarial.

El cumplimiento del Student Outcome (C) se ve reflejado el **modelo predictivo para el diagnóstico de la Diabetes Mellitus tipo 2**, que fue diseñado en base al análisis de las variables sintomáticas que pueden indicar el desarrollo de la enfermedad y el impacto que esta genera en la sociedad, siendo el modelo aplicado en beneficio de la misma en una instancia que permita la optimización del proceso de diagnóstico de la enfermedad, se adapte al proceso natural de obtención de citas médicas en las entidades prestadoras de servicios de salud y aproveche el lapso entre la obtención de la cita y el momento en que esta se lleva a cabo, recopilando datos de manera anónima para el entrenamiento y validación del modelo y recopilando datos del documento de identidad del paciente cuando se encuentre en producción, sólo en caso lo autorice el paciente, permitiendo además de la optimización del proceso, la investigación

científica con la información obtenida sobre la relevancia e impacto de determinadas variables en el diagnóstico de la enfermedad. Con el fin de desarrollar un modelo predictivo cuya implementación sea de impacto inmediato y que no requiera del input de un profesional médico, se determina que el modelo emplee exclusivamente variables sintomáticas de pleno conocimiento por parte del paciente, permitiendo un fácil y rápido diagnóstico preliminar de la diabetes.

Ilustración 4 Representación gráfica del Student Outcome C

Fuente: Elaboración Propia.

STUDENT OUTCOME (D)

Participa en equipos multidisciplinarios desarrollando sus tareas eficientemente con profesionales de diferentes especialidades o dominios de aplicación.

Si bien el cumplimiento del Student Outcome (D) se evidenció a lo largo de todo el proyecto mediante la constante interacción de los diversos interesados e involucrados del proyecto, resaltando las interacciones más claras con equipos multidisciplinarios:

1. Durante el **diseño, elaboración y validación del modelo** se trabajó en conjunto con especialistas médicos como el Dr. Juan Carlos Bartolo y el Dr. Antonio Felices y especialistas en análisis predictivo como el Dr. Andrés Melgar, con la finalidad de garantizar que el modelo utilice variables de riesgos médicas que sean significativas y que se empleen técnicas de machine learning y de predicción de acuerdo al contexto, garantizando además la viabilidad de implementación de la propuesta del modelo predictivo en un ámbito de diagnóstico médico real.
2. A lo largo del proyecto se interactuó con distintos roles como el Profesor Cliente, Gerente Profesor, Gerente Alumno, Revisor de Tesis, Co autor de Paper, Jefe de Línea de QS y recursos QS, los cuales cumplían con tener un background tecnológico que ayudo a validar el procedimiento seguido respecto a la implementación tecnológica que implicaba el desarrollo del modelo, validando con ellos la propuesta de implementación del modelo como su escalabilidad y propuesta de continuidad.

Ilustración 5 Representación gráfica del Student Outcome D

Fuente: Elaboración Propia.

STUDENT OUTCOME (E)

Identifica, formula y resuelve problemas de ingeniería.

El proyecto se sustenta sobre la necesidad identificada en la Ilustración 6, de esta manera el modelo se orienta hacia la optimización del proceso de diagnóstico de la diabetes mellitus tipo 2 por parte de instituciones prestadoras de servicios de salud a con el fin de contribuir al ámbito científico empleando variables que pueden ser identificadas por el paciente por su naturaleza cotidiana. Es de esta manera que el modelo predictivo propuesto utiliza variables sintomáticas identificadas como críticas para el diagnóstico de la diabetes mellitus tipo 2, y además se propone un modelo escalable que permite la incorporación de nuevas variables gracias a la flexibilidad del formulario de ingreso de datos y la estructura de datos subyacente que soporta al modelo, tanto para input como para output. En la Ilustración 7 se puede apreciar la etapa de planificación del modelo predictivo en el que se identifica el objetivo del modelo y como es que se va a obtener la información para la implementación de modelo. Las variables empleadas fueron corroboradas por un especialista en medicina y la escalabilidad/aplicabilidad del proyecto validada por un experto en tecnología.

Ilustración 6 Representación gráfica del Student Outcome E

Fuente: Elaboración Propia.

Ilustración 7 Representación gráfica del Student Outcome E

Fuente: Elaboración Propia.

STUDENT OUTCOME (F)

Propone soluciones a problemas de ingeniería con responsabilidad profesional y ética.

El proyecto desarrollado se plantea en el sector de salud, para entidades prestadoras de servicios de salud, cumpliendo con la responsabilidad social y ética al recopilar datos de los pacientes sin atentar contra su privacidad y derecho a la protección de información: esto quiere decir que en la información recopilada de entrenamiento y

validación del modelo, no se almacenaron datos personales como nombre, apellidos, dirección, teléfonos y demás datos de dicha índole. Asimismo, la recopilación de la información se dio con pleno consentimiento de los entrevistados a los cuales se les explico el propósito del estudio. Como parte de la continuidad del proyecto se ha construido una web para seguir recopilando información, en la cual se explica el propósito del proyecto y donde se recopila el documento de identidad del paciente y los datos acordes a las variables identificadas, solo en caso el paciente acceda a brindar dicha información.

Es así que el proyecto desarrollado mantiene la confidencialidad e integridad de la información brindada por los encuestados cuyas respuestas fueron usadas como sustento del entrenamiento y validación del modelo predictivo. De la misma manera los resultados obtenidos en la predicción fueron constatados y evidenciados con especialistas de Sistemas de Información en las *Actas de validación del modelo*. Adicionalmente, el entrenamiento del modelo se realizó con un conjunto de datos aleatorio y distinto al empleado para la validación del mismo, garantizando la imparcialidad frente a los resultados obtenidos en la ejecución de la predicción.

Ilustración 8 Representación gráfica del Student Outcome F

Fuente: Elaboración Propia.

STUDENT OUTCOME (G)

Comunica ideas o resultados de manera oral o escrita con claridad y efectividad.

El cumplimiento del Student Outcome (G) se ve reflejado a lo largo del desarrollo del proyecto y en distintos escenarios, la Ilustración 9 muestra los frentes de comunicación presentados.

1. **Gestión del proyecto**, se realizaron reuniones coordinadas semanalmente con el profesor cliente y profesor gerente, así como también coordinaciones vía correos electrónicos y/o mensajería instantánea. Estas reuniones tenían como objetivo hacer seguimiento al avance del proyecto y/o solventar algunas dudas del mismo. La comunicación con el revisor de tesis se realizó mediante reuniones

presenciales y virtuales y por vía correo electrónico para el envío de avances y revisiones.

2. **Calidad del proyecto**, se realizaron coordinaciones vía correo electrónico con la empresa virtual Quality Services, para el envío de solicitudes de recursos con la Gerente Alumna y para el envío de paquetes de entregables para su revisión con los Recursos QS.
3. **Expertos**, se coordinaron reuniones presenciales con expertos de medicina para solventar dudas y validar las propuestas presentadas. De igual manera con expertos en analítica para incrementar el conocimiento y tener mejores propuestas.
4. **Asesoría Paper**, se tuvo sesiones semanales con el Asesor del paper para obtener un feedback de los avances y/o tareas asignadas para poder elaborar y mejorar el paper.
5. **Comité Evaluador**, se realizaron presentaciones a lo largo del proyecto para exponer los avances y resultados a la fecha, evidenciando el cumplimiento de objetivos con la propuesta innovadora del proyecto ante una necesidad identificada.

A lo largo de todo el proyecto también se ve plasmada la comunicación escrita, ya sea a través de los **entregables de gestión, entregables propios del proyecto y/o actas de reunión**.

Ilustración 9 Representación gráfica del Student Outcome G

Fuente: Elaboración Propia.

STUDENT OUTCOME (H)

Identifica el impacto de las soluciones de Ingeniería de sistemas de información en el contexto global, económico y del entorno de la sociedad.

La solución propuesta, tras la implementación del modelo como se muestra en la ilustración 10, tiene impacto para cada contexto (global, económico y social) para las empresas prestadoras de servicios de salud, puesto que mediante este modelo es posible obtener los siguientes beneficios:

1. **Investigación científica**, beneficia al contexto nacional y global ya que permitirá a las organizaciones obtener información para la realización de estudios propios sobre la influencia de los factores sintomáticos en el desarrollo de la enfermedad.
2. **Nuevos programas de salud**, beneficia al contexto social ya que con el análisis de los resultados obtenidos en el modelo se podrán crear nuevos programas de salud en beneficio de la sociedad.
3. **Identificación de oportunidades de negocios**, beneficia al contexto económico de las organizaciones al permitir identificar potenciales nuevos productos que no solo benefician a la sociedad, si no también que son rentables para las entidades prestadoras de servicios de salud.

Ilustración 10 Representación gráfica del Student Outcome H

Fuente: Elaboración Propia.

STUDENT OUTCOME (I)

Reconoce la necesidad de mantener sus conocimientos actualizados.

El cumplimiento del Student Outcome (I) se ve reflejado a lo largo del desarrollo del proyecto y en distintos entregables o documentos:

1. Dentro del **Benchmarking de Herramientas**, se realizó un análisis comparativo de herramientas de Predictive Analytics, si bien es cierto, el proyecto se puede realizar con cualquiera de las herramientas debido a que la función de las mismas es predecir, el aprendizaje se da en el uso de cada herramienta para la gestión del modelo, ya que cada una la gestiona de manera distinta. Asimismo para las pruebas con determinados algoritmos, se debe conocer los parámetros de configuración de los algoritmos para una correcta predicción. Finalmente, la elección de la herramienta se basó en su totalidad en el análisis de sus características según criterios actuales de medición de herramientas de predicción, definido por entidades líderes como Gartner y Forrester.
2. En el **análisis de enfermedades crónicas** se realizó un estudio de las principales enfermedades crónicas y su impacto en el país, identificando tasas de mortalidad por géneros, así como la relación entre estas enfermedades, en algunos casos contar con una de las enfermedades analizadas estaba ligado a contraer otras de

las enfermedades crónicas y en otros casos, las agrupaciones de enfermedades crónicas hacían que el enfoque de estudio hacia una determinada familia de enfermedades no sea posible por todas las variedades posibles de enfermedades, como por ejemplo las cardiovasculares. El análisis de enfermedades crónicas permitió además identificar que la mejor manera de conocer que enfermedades eran las más relevantes para el objeto de estudio a través del análisis de las metas globales de la OMS, las cuales nos permiten alinear el proyecto a las necesidades mundiales de salud.

3. Como parte de la continuidad el proyecto, se diseñó una web para capturar información para el crecimiento de la base de datos del modelo, de esa manera se atraviesa un flujo continuo de aprendizaje del modelo y se afinan las predicciones.

STUDENT OUTCOME (J)

Analiza hechos del mundo contemporáneo identificando el impacto en el desempeño profesional de Ingeniería.

El cumplimiento del Student Outcome (J) se ve reflejado a lo largo del desarrollo del proyecto y en distintos entregables o documentos:

1. En el **análisis de enfermedades crónicas** se realizó un estudio de las principales enfermedades crónicas y su impacto en el país, identificando variables que nos ayudaron a determinar a la diabetes como la enfermedad de mayor incidencia e impacto proyectado mundialmente, siendo de esta manera identificada como el punto de enfoque para desarrollar un proyecto de alto impacto
2. Dentro del **Benchmarking de herramientas** se realizó la comparación de las herramientas, identificando variables de calificación según estándares definidos por entidades líderes en el ámbito tecnológico mundial, las cuales de acuerdo a nuestras necesidades permitieron calificar e identificar a la herramienta de Predictive Analytics más adecuada.

STUDENT OUTCOME (K)

Utiliza técnicas, herramientas y metodologías necesarias para la práctica de la ingeniería.

Dentro de la gestión del proyecto, se utilizó la metodología **PMBOK**, como una base y guía fundamental para la elaboración de los documentos de gestión, siendo estos entregables validados por el profesor gerente de la empresa IT Consulting y la calidad de los documentos con los recursos de la empresa QS.

El uso de herramientas para la realización de la analítica predictiva es fundamental, por lo que con el resultado del **Benchmarking de herramientas** se identificó a **SAP Predictive Analytics** como la herramienta que soporte el proyecto, esta herramienta es intuitiva y permite gestionar el modelo. Cabe resaltar que el uso que esta herramienta no es de uso exclusivo del modelo, esto quiere decir que el modelo no está construido en base a la herramienta, pero si es soportado por la misma. La metodología seguida

para el diseño del modelo se realizó de tal manera que se trabaje con repositorios no propios de la herramienta, sino que sean bases de datos convencionales, de tal forma que se puedan emplear otras herramientas y además se pueda explotar el conocimiento que la herramienta genera.

STUDENT OUTCOME (J) - CAC

Comprende y brinda soporte para el uso, entrega y gestión de sistemas de información dentro de un entorno de sistemas de información.

El **informe de la implementación del proyecto** muestra paso a paso las configuraciones y etapas a seguir en la implementación del modelo. Este documento muestra el cómo se determinaron las transformaciones de información para que las soporte el modelo predictivo, las pruebas realizadas con los algoritmos para determinar el algoritmo que entregara mejores resultados.

La estrategia de **Continuidad y soporte del Modelo Predictivo** tiene como objetivo asegurar a través del tiempo la validez del modelo, cumpliendo con las necesidades. De esta manera se definen 3 fases:

1. Post Implementación: En esta primera fase el modelo atraviesa un proceso de enriquecimiento de información a través de su despliegue en el ciclo natural de diagnóstico de la enfermedad, permitiendo realizar nuevos grupos de entrenamiento, aprendizaje y validación de resultados, afinando la precisión lograda.
2. Adaptabilidad del modelo: La adaptabilidad radica en la posibilidad de adopción de nuevas variables en el modelo, soportado en su totalidad por el modelo planteado debido a la flexibilidad con la que se desarrolló la estructura subyacente de datos que son procesados por el modelo. Adicionalmente, en esta fase se llevará a cabo el descubrimiento de patrones, observaciones y consideraciones que van a permitir profundizar en el estudio de los factores de riesgo que desencadenan en la enfermedad.
3. La tercera fase es la evolución del modelo producto del despliegue del mismo: a través de este despliegue y en el corto-mediano plazo se verán notorias mejoras en el proceso de diagnóstico de la enfermedad, agilizándolo y afinándolo, permitiendo que el profesional médico a través de un juicio de experto ágil producto de la visibilidad que tiene de los datos del paciente, tome una decisión sobre el mejor proceder en cuanto a la salud del paciente. Adicionalmente, en esta fase, producto de la base nutrida de datos que se recopilará, se podrá realizar la transición a un entorno cloud para el ámbito de alojamiento y procesamiento, eximiendo de esta manera a la organización de requerir alto potencial de procesamiento in-house.

Ilustración 11 Representación gráfica del Student Outcome J - CAC

Fuente: Elaboración Propia.

CAPÍTULO 3. Marco Teórico

El presente capítulo presentará la definición de todos los términos y conceptos esenciales en el ámbito de predicción de enfermedades para permitir que el lector se familiarice con el tema del proyecto. Dentro de los conceptos a desarrollar, se encontrarán conceptos básicos, complementarios y específicos al contexto de enfermedades.

DATOS, INFORMACION Y CONOCIMIENTO

Los datos son una representación de hechos, conceptos o instrucciones plasmados de manera formal para la comunicación, interpretación o procesamiento por humanos o modos automáticos (Hicks, 1993). Los datos por si solo son inútiles: deben ser cambiados a una forma usable y ser ubicados en un contexto para que tenga valor. Los datos se vuelven información cuando son transformados para comunicar significado o conocimiento, ideas o conclusiones (Senn, 1990). El conocimiento es la información obtenida, incorporada en los recursos de razonamiento de un agente y lista para ser usada en un proceso de decisión; es el resultado de un proceso de aprendizaje (Aamodt, Nygard, 1995)

Ilustración 12: El modelo de Datos-Información-Conocimiento (Aamodt, Nygard 1995)

Anexando los conceptos anteriores con el proyecto desarrollado, podemos identificar los datos como aquella información que ingresa de manera puntual al modelo predictivo desarrollado y que sirve para alimentarlo: datos propios de pacientes como estatura, peso, niveles de hemoglobina, etc. La información es lo que se obtiene de la transformación de datos a través de una serie de relaciones lógicas, lo cual en el caso del proyecto sería la consolidación de los rasgos del paciente, siendo necesario destacar que en la creación de información no se obtiene directamente conocimiento, sino que es necesario que esta atraviese una serie de recursos que permitan razonar y aprender de la misma, logrando así conocer que es lo que la información brinda en base a los datos ingresados. En este sentido,

existe una similitud entre el concepto de conocimiento y el resultado que muestra el modelo debido a que se emplea el procesamiento a través de algoritmos para obtener aprendizaje.

DESCUBRIMIENTO DE CONOCIMIENTO EN BASES DE DATOS

El descubrimiento del conocimiento es la extracción no trivial de información previamente desconocida, implícita y potencialmente útil a partir de un conjunto de datos. En base a un conjunto de datos F , un lenguaje L y una medida de certeza C , se define un patrón S en L que describe las relaciones entre un conjunto de datos F_s de F con una certeza C , para que S sea más simple que la enumeración de todos los hechos en F_s . El resultado de un programa que monitorea el conjunto de hechos en una base de datos y produce patrones, de esta manera, es conocimiento descubierto (Frawley, Piatetsky-Shapiro, Matheus, 1992)

El descubrimiento de conocimiento en bases de datos es la agrupación de todos los procesos, automatizados y no automatizados, que permiten la exploración de bases de datos, largas y pequeñas, para extraer conocimiento potencial. (Fayyad, Piatetsky-Shapiro, Smyth, 1996).

Ilustración 13: Proceso de Descubrimiento de Conocimiento (Maimon, Rokach 2005)

La habilidad de descubrir conocimiento en bases de datos, empleando tecnologías vigentes actualmente brinda nuevas posibilidades a la gran mayoría de industrias de ser aplicada correctamente. Es necesario sin embargo indicar que resulta elemental identificar que técnicas son las que se deben aplicar para realizarlo, debido a que cada una presenta una serie de ventajas que las hacen más adecuadas para determinados casos de negocio. Para fines del proyecto, empleando la premisa de la predicción inmediata se procederá a desarrollar el concepto de predictive analytics.

PREDICTIVE ANALYTICS

Predictive Analytics es un conjunto de tecnologías de inteligencia de negocios que descubren relaciones y patrones en largos volúmenes de datos que pueden ser usados para predecir comportamiento y eventos. A diferencia de otras tecnologías de inteligencia de negocios, predictive analytics ofrece una vista hacia el futuro, usando eventos pasados para predecirlo (Eckerson 2007). Predictive Analytics incluye modelos estadísticos y otros métodos empíricos destinados a la creación de predicciones empíricas (a diferencia de las predicciones basadas exclusivamente en teoría), además de métodos que identifican la calidad de las predicciones en la práctica (Dubin, 1969).

Es importante entender las limitaciones de predictive analytics. Primero, no se puede obtener progreso sin un conjunto de datos que sirva de entrenamiento y que sea de un tamaño y calidad adecuado. Segundo, es crucial tener una definición clara del concepto que va a ser predicho y tener ejemplos históricos del ejemplo. (Elkan, 2013).

APRENDIZAJE SUPERVISADO Y MACHINE LEARNING

El objetivo de un algoritmo de aprendizaje supervisado es obtener una clasificación a través de lo aprendido de ejemplos de entrenamiento. Esta clasificación es algo que puede ser usado para hacer predicciones en ejemplos de prueba. (Elkan, 2013).

Dentro del contexto de aprendizaje supervisado nace el término de Machine Learning o aprendizaje de máquina, el cual hace referencia a la detección de manera automática de patrones significantes en datos. Una característica de la aplicación de machine learning es que se aplica en casos donde los patrones a identificar son de alta complejidad, por lo que un ser humano no cuenta con la pericia suficiente para aprender de esto de manera específica (Shalev-Shwartz, Ben-David, 2014).

MODELOS PREDICTIVOS

Modelo predictivo es un nombre dado a una colección de técnicas matemáticas con la meta en común de encontrar una relación matemática entre un objetivo, respuesta o variable dependiente y varios factores de predicción y variables independientes, con el objetivo de medir valores futuros de estos factores de predicción e insertándolos en la relación matemática para predecir valores futuros de la variable objetivo. Como estas relaciones nunca son perfectas en la práctica, es deseable dar una medición de dudeza en las predicciones. (Dickey, 2012)

Para identificar aquellos factores que interceden en la predicción, se pueden agrupar en tres categorías: aquellos que tienen poca posibilidad de afectar el resultado, aquellos con cierta certeza para afectar los resultados y deben ser considerados en el modelo y los que se encuentran en el medio, que pueden influir o no en el resultado final, siendo necesario identificar a través de una serie de técnicas si deben ser incluidas en el modelo (Dickey, 2012).

VALIDACION Y LIMPIEZA DE DATOS

La validación de datos significa confirmar que se trata de datos confiables, mientras que la limpieza significa corregir errores en ella. No es inusual tener dificultades en identificar si un valor es correcto o no, y si no es correcto, es imposible encontrar el valor correcto. Sin embargo, es posible adquirir seguridad razonable de que ciertos datos son propensos a ser correctos o incorrectos, debido a que aprueban o no aprueban una serie de controles. Cuando los datos son muy propensos a ser erróneos, es mejor descartarlo, dedicando mayor esfuerzo a la validación de datos que a la reparación de los mismos (Elkan, 2013).

Dentro de los tipos de limpieza de datos, existe el mezclar registros que hacen referencia a la misma entidad y no deben estar separados (Elkan, 2013), siendo esto algo cotidiano si se trata de un conjunto de información proveniente de diversas fuentes.

Asimismo, no es inusual que, según el algoritmo empleado para realizar la predicción, se deba hacer una transformación de las variables hacia valores discretos o hacia valores binarios que sean soportados por el algoritmo. Resulta cotidiano tener intervención humana en esta fase debido a la complejidad de la información con la que se puede trabajar. (Elkan, 2013).

DISCRETIZACION

La discretización puede transformar atributos numéricos hacia discretos, la selección de características puede eliminar algunos atributos irrelevantes (Liu & Setiono, 1995).

Los métodos de discretización se dividen en dos tipos, métodos supervisados y métodos no supervisados. Los métodos supervisados si la clase de los ejemplos es considerada en la discretización como atributo de decisión y los métodos no supervisados si la clase no es considerada o es tratada como un atributo normal (Giráldez, Aguilar-Ruiz & Riquelme, 2001).

Hay tres ejes diferentes mediante el cual se pueden clasificar los métodos de discretización: global vs locales, supervisados vs no supervisado y estático vs dinámico (Dougherty, Kohavi & Sahami, 1995).

Ilustración 14: Metodos de discretización

	Global	Local
Supervised	1RD (Holte) Adaptive Quantizers ChiMerge (Kerber) D-2 (Catlett) Fayyad and Irani / Ting Supervised MCC Predictive Value Max.	Vector Quantization Hierarchical Maximum Entropy Fayyad and Irani C4.5
Unsupervised	Equal width interval Equal freq. interval Unsupervised MCC	k-means clustering

Fuente: Dougherty et al. (1995)

ENFERMEDADES CRONICAS

También conocidas como enfermedades no transmisibles (ENT) son condiciones médicas crónicas o enfermedades que no son infecciosas. Entre los ejemplos más comunes están apoplejía, ataques al corazón, diabetes, cáncer, asma y depresión. Algunas de las ENT están precedidas por conductas poco saludables, seguido de la aparición de factores de riesgo metabólico y de enfermedad. Los factores de riesgo asociados a las enfermedades crónicas son, sobrepeso y obesidad, presión arterial elevada, aumento de los niveles de glucosa en la sangre y niveles no óptimos de colesterol en la sangre. La mayoría de estos factores de riesgo son considerados modificables mediante cambios en los comportamientos o con medicación. Los comportamientos clave que podrían reducir los factores de riesgo de enfermedades crónicas se dan mediante una dieta saludable, participación en actividad física regular, no consumir tabaco, y evitar el uso nocivo del alcohol (Bradshaw, Steyn, Levitt & Nojilana).

DIABETES

La diabetes mellitus es un grupo de enfermedades metabólicas caracterizadas por hiperglicemia resultante de defectos de segregación de insulina, la acción de la insulina o ambos. La hiperglucemia crónica de la diabetes se asocia con daño a largo plazo, disfunción e insuficiencia de varios órganos, especialmente los ojos, riñones, nervios, corazón y vasos sanguíneos. La mayoría de diagnósticos/casos de diabetes se divide en dos categorías etiopatogénicos. En la diabetes tipo 1, la causa es una deficiencia absoluta de la secreción de insulina. En la diabetes tipo 2, mucho más frecuente, la causa es una combinación de resistencia a la acción de la insulina y una respuesta secretora inadecuada de la insulina

compensatoria. En este tipo de diabetes, se puede presentar un grado de hiperglucemia suficiente para causar cambios patológicos y funcionales en diferentes tejidos afectados, pero sin síntomas clínicos, pudiendo estar presente por varios años antes de que se detecte la diabetes. (Mellitus, 2005).

CAPÍTULO 4. Estado del arte

Este capítulo expone los trabajos científicos realizados previamente referente a la predicción de la diabetes mellitus tipo 2, en estos trabajos se realizaron diversas pruebas para comparar los resultados obtenidos con el uso de algoritmos de predicción.

Revisión de la Literatura

La revisión de la literatura muestra numerosos resultados en cuanto a la predicción de la diabetes, usando distintos algoritmos y obteniendo distintos resultados de acuerdo a los parámetros tomados en cuenta para la predicción.

La capacidad predictiva es la capacidad de construir y evaluar modelos destinados a generar predicciones exactas de nuevas observaciones, donde nuevas puede ser interpretado como temporal y/o como sección transversal (Wang, Kung, & Byrd, 2016).

En Wessler (2013) se define la capacidad predictiva como el proceso de usar un conjunto de sofisticadas herramientas estadísticas para desarrollar modelos y estimaciones de lo que el entorno hará en el futuro. Para crear la capacidad de predicción, las organizaciones tienen que depender de una plataforma de análisis predictivo que incorporan almacenes de datos, algoritmos de análisis predictivo (por ejemplo, análisis de regresión, aprendizaje automático, y redes neuronales), cuadros de mando y presentación de informes que proporcionan las decisiones óptimas para los usuarios.

Redes Neuronales

En los estudios de Sapon, Ismail, Zainudin & Ping (2011), Anand, Kirar & Burse (2012), Thirugnanam, Kumar, Srivatsan & Nerlesh (2012) con relación a los algoritmos de redes neuronales se encuentran puntos por mejorar en la predicción de la diabetes. En Sapon et al. (2011) se muestra un estudio de un modelo de predicción de la diabetes con diferentes algoritmos de aprendizaje supervisado de red neuronal artificial. El algoritmo Scaled Conjugate Gradient obtuvo un mejor rendimiento con una mejor correlación entre los objetivos y resultados previstos. En Anand et al. (2012) se menciona que es necesario un sistema de apoyo de diagnóstico médico para doctores en el proceso del diagnóstico. Por eso realizaron un estudio de varios modelos predictivos de algoritmos de redes neuronales para el diagnóstico de la diabetes. Obteniendo que al pre-procesar en distintos algoritmos y extrayendo algunas características de los modelos se puede mejorar la exactitud. En Thirugnanam et al. (2012) se identifican que todos los modelos de predicción de diabetes tienen una falencia por lo que presentan un enfoque de diagnóstico para prevenir la diabetes aplicando técnicas de redes neuronales lógica difusa y razonamiento basado en casos.

Obteniendo un enfoque para el diagnóstico de la diabetes usando un nuevo diseño influenciado de parámetros de entrada.

Árboles

En los estudios de Hussein, Omer, Li & Ati (2012), Kaur & Chhabra (2014), Iyer, Jeyalatha, & Sumbaly (2015), Van Poucke et al. (2016) relacionados a los algoritmos de árboles, identifican que la diabetes viene desarrollando un crecimiento en su presencia mundial en los últimos años.

En el estudio de Hussein et al. (2012) mencionan que con el crecimiento de las enfermedades crónicas es necesario un mayor control. Los autores proponen un modelo de Chronic Disease Diagnosis (CDD) usando varios algoritmos de árbol de decisiones. Por otro lado en Kaur & Chhabra (2014) muestran que la diabetes es una enfermedad que genera un gran riesgo para tener otras enfermedades por lo que es necesario realizar analítica en ella. Es así que proponen un modelo predictivo para el diagnóstico de la diabetes usando el algoritmo J48 y mejorándolo. Con la modificación del algoritmo lograron una precisión del 99.87%. En Iyer et al. (2015) mencionan que la diabetes viene creciendo en el mundo y que se estima que para el 2025 habrán 380 millones de personas con diabetes, por esto buscan un diagnóstico más rápido y eficiente al analizar patrones encontrados en el análisis de clasificación al emplear algoritmos de árboles de decisiones y redes bayesianas. Finalmente, en Van Poucke et al. (2016) observaron que la gran cantidad de información relacionada a salud puede usarse para predecir, prevenir y personalizar la medicina, por lo que desarrollan un marco que soporte un análisis predictivo escalable usando herramientas visuales, esto permitió desarrollar un robusto proceso de creación automática, optimización de parámetros y evaluación de varios modelos predictivos con el algoritmo que dio mejores resultados Random Forest.

K-vecinos más cercanos

En el estudio de Lee et al. (2011) observan que hay variables predictivas que tienen incidencia para el desarrollo de la Diabetes Mellitus Tipo 2. Comparan los diversos algoritmos, concluyendo que el mejor es K-vecinos más cercanos. Es así que diseñan un

modelo predictivo basado en varios algoritmos estadísticos, comparando su modelo con otros obtuvieron un ratio de error mayor al esperado.

Redes bayesianas

En Kumari, Vohra & Arora (2014) identifican a la diabetes como una enfermedad crónica de mayor reto en la salud mundial, y utilizan un modelo de redes bayesianas para predecir la diabetes, obteniendo resultados con 99,51% de precisión. Por otro lado, en Koklu & Unal (2013) mencionan que el análisis manual de información en estos tiempos se ha vuelto inútil con la gran cantidad de información, por lo que es necesario un análisis computarizado. De tal manera que desarrollan un modelo de diagnóstico de diabetes usando algoritmos de redes bayesianas, percepción multicapa y J48. Obteniendo distintos resultados de precisión para cada algoritmo, siendo el de redes bayesianas el de mejor resultado.

LEM2

En Ali, Hussain, Siddiqi, Hussain & Lee (2015) mencionan que hay numerosos modelos que proponen resolver los problemas de diagnóstico y el manejo de la diabetes, ellos proponen un modelo híbrido (H2RM), usando el algoritmo LEM2, que resuelve el problema del manejo de la diabetes para los tipos 1 y 2, obteniendo en promedio y una precisión equilibrada del 95,9%.

Jerarquía difusa

Por su parte en Lukmanto & Irwansyah (2015) mencionan que para el 2035 la pre-valencia de la diabetes crecerá en 55% con respecto al 2013, es así que proponen una inteligencia computacional usando un modelo de jerarquía difusa que permite una temprana detección de la diabetes. Ellos compararon la información del modelo con la decisión del doctor. Teniendo una similitud del 87,46% con la declaración de los médicos.

Algoritmos distribuidos

Finalmente, en Eswari, Sampath & Lavanya (2015) indican que con el crecimiento de Big Data y la gran cantidad de información de salud es necesario estructurar y enfatizar el tamaño

de esta información para una posible solución, por lo que usan un algoritmo predictivo analítico para predecir la pre-valencia de los tipos de diabetes y complicaciones asociadas con el tipo de tratamiento brindado.

Modelos Predictivos en Salud

Desarrollo de un Modelo Predictivo para la Diabetes Mellitus de Tipo 2 usando Data Clínica y Genética

Lee et al. (2011) indica que existen estudios genéticos recientes que proveen evidencia convincente que la ubicación de un gen y el polimorfismo de un grupo de moléculas están directamente asociados con el riesgo de desarrollar diabetes mellitus de tipo 2. Es bajo esta premisa que se plantea desarrollar un modelo predictivo de diabetes mellitus tipo 2 usando data clínica y genética.

Para el desarrollo del modelo, se empleó un método estadístico de muestreo de 684 participantes, donde 472 fueron sujetos aleatorios y 212 fueron sujetos con diabetes mellitus de tipo 2. La edad promedio era de 64.1 años para los aleatorios y 64.5 años para los pacientes diagnosticados con diabetes, por lo que las edades no eran significativamente diferentes. Se sometieron a los 684 participantes a un muestreo de 499 grupos de moléculas asociadas previamente con la diabetes mellitus de tipo 2. Se consideró además factores que pueden interceder en el desarrollo de la enfermedad, tal como el índice de masa corporal y el conteo de triglicéridos con el fin de afinar el resultado determinando un ratio de clasificación incorrecta de diabetes mellitus tipo 2 de 27.98 ± 2.76 para todos los algoritmos probados.

Para mejorar los resultados de la clasificación, se empleó un modelo de ajuste multi escenario, seleccionando factores de riesgo que permitan agrupaciones más reducidas según características en común para realizar un análisis de reconocimiento de patrón. En base a esto, se determinaron tres grupos: el primero contenía factores asociados con la obesidad (índice de masa corporal, peso y ratio entre cintura y cadera), en el segundo se contenía los ratios más altos de glucosa, resistencia a la insulina, colesterol y hemoglobina y en el tercero se incluían los valores más altos de índice de masa corporal, peso y ratio de trabajo del corazón.

Se identificó a través del uso de modelos estadísticos de decisión de árbol que la prevalencia de la diabetes mellitus tipo 2, en hombres y mujeres, era similar en los grupos

2 y 3 pero menor en el grupo 1. A continuación se muestra un cuadro de dispersión entre la relación de triglicéridos, índice de masa corporal e incidencia de la diabetes mellitus tipo 2:

Ilustración 15: Dispersión entre triglicéridos, índice de masa corporal e incidencia de diabetes mellitus tipo 2

Fuente: Lee et al. (2011)

En los clusters 1-3, 2-3 y 3-3 podemos identificar a los pacientes que sufren de la enfermedad y notamos una tendencia hacia el alto nivel de triglicéridos e índice de masa corporal para estos clusters.

En base al análisis estadístico y clasificación de factores, se determinaron los siguientes factores de riesgo para desarrollar la enfermedad:

- Información clínica: talla, índice de masa corporal, circunferencia de la cintura, presión sanguínea, colesterol, conteo de triglicéridos y nivel de glucosa.
- Historia familiar y personal: hipertensión, diabetes mellitus de tipo 2, enfermedades cerebrovasculares y otras enfermedades vasculares
- Polimorfismo de grupos de moléculas (SNP): se identificaron 499 SNPs en 87 genes candidatos asociados a la diabetes mellitus tipo 2.

Para construir un modelo predictivo óptimo se optó inicialmente por emplear el método estadístico de chi cuadrado, sin embargo se obtuvieron ratios de error muy altos (entre 28.24

y 52.69%) debido a la distribución desigual de tipos de genotipificación (dentro de los 499 SNPs hubo un 70-75% de genotipos dominantes, un 5-15% recesivos y el resto eran heterocigóticos, además, la data clínica era variable mientras la data de grupos de moléculas era relativamente estática, por lo que al comparar ambos factores un factor de riesgo sumamente importante podría ser desestimado. Los modelos construidos empleando árboles de decisión y métodos de machine-learning no mostraron muchas diferencias por los mismos factores. Para obtener un mejor resultado, resultaría necesario cuantificar adecuadamente los SNPs para clasificar inicialmente a los pacientes en grupos de alto riesgo y bajo riesgo. El análisis concluye que los ratios de error eran más altos es lo esperado, por lo que se recomienda emplear muestreos mayores y estrategias de análisis mejores para integrar la genotipificación para cuantificar objetivamente los SNPs de alto riesgo.

Predicción de Diabetes con Algoritmos de Aprendizaje Supervisado de Redes Neuronales Artificiales

Akmal et al. (2011) proponen emplear redes neuronales multicapas para obtener predicciones exitosas de diabetes frente a métodos tradicionales, basándose en que hay algoritmos de redes neuronales que han sido probado que son útiles para diagnosticar males cardíacos, tuberculosis, neumonía y cáncer al pulmón.

Se empleó un muestreo de 250 pacientes de diabetes dentro de las cuales consideran 27 variables tales como presión sanguínea, acidez de la orina, creatinina y otros identificados como valores numéricos.

Los autores definen al aprendizaje supervisado como un enfoque para identificar la relación de input-output en base al entrenamiento empleando un conjunto de datos, aprendiendo del output en base a la señal de error que se emita según el diagnóstico, continuando hasta identificar el mínimo valor de error permitido.

Se emplearon algoritmos disponibles en las herramientas de redes neuronales de MATLAB, empleando los siguientes:

- Método del gradiente conjugado de Fletcher-Powell (a)
- Método del gradiente conjugado de Polak-Ribière (b)
- Método del gradiente conjugado a escala.(c)

Para identificar la relación entre resultados deseados y resultados obtenidos se empleó un análisis de regresión donde se obtenía un coeficiente de correlación R, siendo 1 un resultado ideal. A continuación, se muestran los desempeños de los algoritmos

Ilustración 16: Medición de desempeño de algoritmos

(a)

(b)

(c)

Fuente: Akmal et al. (2011)

En base al test de regresión realizado obtenemos el siguiente bosquejo. En las figuras a continuación, identificamos al coeficiente de correlación R correspondiente a cada algoritmo respectivamente, reflejado por la línea punteada:

Ilustración 17: Dispersión de resultados de los algoritmos frente al coeficiente de correlación R

Fuente: Akmal et al. (2011)

Podemos identificar que el algoritmo de gradiente conjugado a escala casi sobre lapa la línea de referencia, indicando un resultado cercano entre lo esperado y lo predicho. El objetivo del estudio es determinar el algoritmo que tiene el mejor desempeño con la data recolectada. Los resultados del coeficiente R son los siguientes:

Tabla 7: Coeficiente de correlación por algoritmo

Algorithm	Epochs	Correlation Coefficient, R
Fletcher-Powell Conjugate Gradient	39	0.097219
Polak-Ribière Conjugate Gradient	52	0.056466
Scaled Conjugate Gradient	108	0.88026

Fuente: Akmal et al. (2011)

Identificando que el algoritmo de gradiente conjugado a escala es lejanamente el de mejor desempeño en cuanto a la predicción de diabetes frente a las pobres performances de los algoritmos de Polak-Ribière y Fletcher-Powell.

Mejorando el Ratio de Predicción de Diagnósticos de Diabetes usando Lógica Difusa y Redes Neuronales

Thirugnanam et al. (2012) indican que se han hecho muchos enfoques en predecir la diabetes pero que nunca se ha tenido un enfoque especial en la precisión de los modelos desarrollados.

Los sistemas existentes son limitados y proveen resultados exactos si se usan conjuntos de datos de mucha calidad, estandarizados o se usan pocos conjuntos de datos. Es bajo esta premisa que el paper presenta un enfoque nuevo para el diagnóstico de diabetes empleando dos fases para predecir el estado de la diabetes: la primera es emplear lógica difusa, redes neuronales y razonamiento en base a casos como un enfoque individual y la segunda es aplicar un algoritmo a los valores obtenidos en la primera fase, resultando en una predicción con mayor objetividad.

Se propone un marco de trabajo en el cual un usuario ingresa a un sistema y se ve forzado a ingresar datos sobre sus características que pueden demostrar disposición al desarrollo de diabetes, para luego analizar estos parámetros. Se normalizan y se aplican los tres enfoques mencionados en la fase inicial: lógica difusa, redes neuronales y razonamiento en base a casos guardando estos resultados y finalmente aplicando un algoritmo para derivar los resultados, generando un reporte de diagnóstico final. Los pasos seguidos para el análisis de datos son los siguientes:

- Recolección de información
 - Datos tales como edad, género, información de enfermedades familiares, medicinas para presión alta, consumo de tabaco, dieta, actividad física, índice de masa corporal, frecuencia de orina, hambre incrementada, nivel de sed, pobre curación de heridas, estilo de vida y picazón.
 - Análisis de información a través de técnicas de redes neuronales:
 - Usando el algoritmo de backpropagación: se alimenta la información a la capa de input que consiste de un numero n de neuronas, las actividades de la capa oculta dependen de la capa de input y consiste de $n+1$ neuronas, y la capa de output representa 0 o 1, siendo 1 cuando el paciente es afectado por diabetes
 - Usando conceptos de lógica difusa: se define el conjunto de datos según lo indicado anteriormente y se le otorga un valor basado en la comparación entre un paciente enfermo de diabetes y uno sano, estando en un rango entre 0 y 2. Luego se definen las reglas de decisión según cada parámetro: por ejemplo, para edades menores a 35 podemos determinar un conjunto difuso “bajo” por la baja propensidad a desarrollar la enfermedad. De esta manera y empleando conectores Y y O para evaluar las condiciones del paciente y el resultado según la ponderación de cada variable.
 - Aplicando conceptos de razonamiento en base a casos: se definen dos fases, una es pre procesamiento de datos y una de procesamiento de consulta. En la

primera se adquiere el conocimiento y los datos y en la segunda se procesa en base a un conjunto de datos ya definido para identificar la propensidad del paciente a desarrollar la enfermedad.

- Aplicación del algoritmo basado en reglas:
 - En base a la aplicación del enfoque de análisis de información planteado anteriormente, se emplea un algoritmo basado en reglas para discernir entre información certera y dudosa, asegurando la predicción final.

Se concluye que el enfoque planteado a través de los 3 frentes de análisis de información es más útil que los enfoques tradicionales debido a la reducción de errores a través de la mezcla de análisis de información y la aplicación de un algoritmo basado en reglas. Asimismo, el marco de trabajo resultaría beneficioso debido a que permite al paciente obtener información de manera rápida (en caso se implemente esto a un sistema) y que pueda actuar de manera que combata aquellas variables que lo vuelven propenso al desarrollo de la enfermedad.

ARIC: Modelo Predictivo de Riesgo de Diabetes Tipo 2

Woon-Loong et al. (2012) indican que la identificación de individuos de alto riesgo es crucial para la efectiva implementación de los programas de prevención de la diabetes mellitus tipo 2. Numerosos estudios muestran que las funciones predictivas multivariables se desempeñan tan bien como la exposición posterior a 2 horas de glucosa en identificar a los individuos de alto riesgo de desarrollar diabetes. Sin embargo, no se había podido aprovechar la utilidad de las funciones predictivas porque no habían estudios que evalúen la habilidad de las funciones de predecir con precisión el riesgo de desarrollar diabetes tipo 2.

Es por ese motivo que los autores evalúan el desempeño de 3 modelos predictivos de determinación de riesgo de diabetes tipo 2 para determinar la viabilidad de adoptar el modelo en la población Asiática. Los 3 modelos evaluados fueron escogidos porque tenían excelentes cualidades de calibración y discriminación, siendo los modelos: *The San Antonio Health Study (SAHS) model*, *The Atherosclerosis Risk in Communities Study (ARIC) model* y *The Framingham Offspring Study (FRAM) model*.

Se utilizó información de Chinos de Singapur, malayos e indios asiáticos para poder validar los modelos dentro del contexto Asiático, la información se obtuvo de un estudio de 1992 (Singapore Prospective Study Program, en National Health Survey - NHS.1992).

Los autores analizaron los 3 modelos predictivos publicados comparándolos con el modelo predictivo local, el cual tiene la información del NHS 1992.

Tabla 8: Comparación de las funciones predictivas multivariantes publicadas con las estimadas localmente.

Risk factor at baseline	SAHS Model		ARIC Model		Framingham Model*	
	locally-estimated (95% CI)	<u>Published</u>	locally-estimated (95% CI)	Published	locally-estimated (95%CI)	Published
Age (years)	-0.004 (-0.026; 0.018)	0.028	-0.023 (-0.047; 0.000)	0.0173	NA	NA
Gender	0.098 (-0.391; 0.588)	0.661	NA	NA	NA	NA
Ethnicity	0.646 (0.186; 1.107)	0.412	0.677 (0.214; 1.140)	0.4433	NA	NA
FPG (mg/dl)	0.119 (0.090; 0.148)	0.079	0.120 (0.090; 0.149)	0.0880	NA	NA
IFG	NA	NA	NA	NA	1.795 (1.259;2.331)	1.98
SBP (mmHg)	0.015 (0.000; 0.030)	0.018	0.015 (0.000; 0.029)	0.0111	0.843 (0.372;1.315)	0.50
Triglyceride(mg)	NA	NA	0.000 (-0.001;0.001)	0.0027	0.695 (0.232;1.158)	0.58
HDL (mg/dl)	-0.033 (-0.058;-0.008)	-0.039	-0.027 (-0.053;-0.001)	-0.0122	0.498 (0.032;0.965)	0.94
BMI (kg/m ²)	0.154 (0.097; 0.212)	0.070	NA	NA	NA	NA
BMI (overweight)	NA	NA	NA	NA	1.206 (0.728;1.683)	0.30
BMI (obese)	NA	NA	NA	NA	1.574 (0.881;2.267)	0.92
Waist (cm)	NA	NA	0.071 (0.046; 0.096)	0.0273		NA
Height (cm)	NA	NA	-0.044 (-0.072;-0.016)	-0.0326		NA
Family history of T2DM	0.447 (-0.008;0.903)	0.481	0.473 (0.018; 0.928)	0.4981	0.548 (0.111;0.986)	0.57

Bold and italicized fonts: Estimates from local models whose 95% confidence intervals do not overlap with the published estimates, i.e. we assumed that the published estimates were the true values in the hypothesis testing

*Categorical variables are used for the Framingham model with the following rule: fasting glucose level: (1 = 100 - 125 mg/dL, 0 ≤ 100 mg/dL), overweight (1 = BMI ≥ 25 & BMI < 30, 0 = otherwise), obese (1 = BMI ≥ 30, 0 = otherwise), low high density lipoprotein (HDL) (1 = HDL < 40 mg/dL in men or < 50 mg/dL in women, 0 = otherwise), high triglyceride level (1 ≥ 150 mg/dL, 0 = otherwise), elevated blood pressure (systolic blood pressure ≥ 130 or diastolic blood pressure ≥ 85 mm Hg, 0 = otherwise)For the SAHS model, the 95% confidence intervals of the coefficients estimated from the NHS-92 cohort included the published SAHS coefficient for all but four risk factors: age, gender, FPG at baseline and BMI; the coefficients for age and gender were greater for the published SAHS model than the locally-estimated coefficients. In contrast, the locally estimated coefficients for FPG and BMI were larger than the published coefficients.

Fuente: Woon-Loong et al. (2012)

En la Tabla se muestran los resultados de los coeficientes de los modelos publicados de SAHS, ARIC y Framingham, frente a los estimados localmente con la información del NHS 1992. Para el modelo ARIC, los intervalos de confianza del 95% de los siguientes factores de riesgo no estaban incluidos en el modelo publicado: la edad, el ayuno del FPG al inicio del estudio, la circunferencia de la cintura y los triglicéridos. Para el modelo Framingham, todo lo localmente estimado concuerda con el modelo publicado, excepto por el efecto del peso, que se encontró significativamente mayor en el modelo local.

Tabla 9: Comparación de las características del área bajo la curva del funcionamiento del receptor de varios modelos predictivos usando NHS-1992

Model 1	AUC (95% C.I.)	Model 2	AUC (95% C.I.)	P value for comparison
SAHS-local	0.857 (0.821;0.894)	SAHS-published	0.839 (0.803;0.874)	0.028
ARIC-local	0.864 (0.829;0.899)	ARIC-published	0.847 (0.812;0.883)	0.065
FRAM-local	0.826 (0.784;0.867)	FRAM-published	0.805 (0.762;0.849)	0.076
SAHS-published	0.839 (0.803;0.874)	ARIC-published	0.847 (0.812;0.883)	0.230
SAHS-published	0.839 (0.803;0.874)	FRAM-published	0.805 (0.762;0.849)	0.028
ARIC-published	0.847 (0.812;0.883)	FRAM-published	0.805 (0.762;0.849)	0.007
FPG-local	0.782 (0.736;0.828)	SAHS-local	0.857 (0.821;0.894)	< 0.001
		ARIC-local	0.864 (0.829;0.899)	< 0.001
		FRAM-local	0.826 (0.784;0.867)	0.047
OGTT-local	0.778 (0.727;0.828)	SAHS-local	0.857 (0.821;0.894)	0.004
		ARIC-local	0.864 (0.829;0.899)	0.002
		FRAM-local	0.826 (0.784;0.867)	0.110
FPG-local	0.782 (0.736;0.828)	SAHS-published	0.839 (0.803;0.874)	< 0.001
		ARIC-published	0.847 (0.812;0.883)	< 0.001
		FRAM-published	0.805 (0.762;0.849)	0.323
OGTT-local	0.778 (0.727;0.828)	SAHS-published	0.839 (0.803;0.874)	0.021
		ARIC-published	0.847 (0.812;0.883)	0.011
		FRAM-published	0.805 (0.762;0.849)	0.371

Fuente: Woon-Loong et al. (2012)

La tabla anterior compara el área bajo la curva de varios modelos predictivos. A pesar que la curva de los 3 modelos localmente estimados es ligeramente mayores que sus contrapartes publicadas, solo en el caso del modelo SAHS la diferencia alcanza significancia estadística. Todos los modelos multivariados estimados localmente logran un mejor poder de discriminación en comparación con el modelo que utiliza solo FPG (todos $p < 0,001$), mientras que el modelo Framingham estimado localmente es el único que no fue estadísticamente mejor que el modelo que solo utiliza 2hPG ($P = 0,110$).

Fuera de los tres modelos publicados, el modelo ARIC tuvo un mayor poder de discriminación ($AUC = 0,847$), seguido del modelo SAHS ($AUC = 0,839$) y el modelo Framingham ($AUC = 0,805$). Los modelos publicados de SAHS y ARIC son estadísticamente mejores en discriminar los casos de diabetes tipo 2 de los no casos en comparación el modelo local que solamente utiliza FPG ($P < 0,001$) o sólo 2hPG ($P = 0,021$ y $0,011$ respectivamente).

Tabla 10: Calidad de calibración de varios modelos predictivos usando NHS-1992

Model	HL-statistic	Predicted incidence (%)
SAHS-local	3.52	7.8
ARIC-local	2.28	7.8
FRAM-local	1.45	7.8
FPG-local	3.08	7.8
OGTT-local	3.90	7.8
SAHS-published	48.33	13.5
ARIC-published	14.71	9.8
FRAM-published	264.89	2.0
<i>Recalibrated</i> SAHS-published	23.19	11.2
<i>Recalibrated</i> ARIC-published	10.31	8.5
<i>Recalibrated</i> FRAM-published	103.15	16.9

Fuente: Woon-Loong et al. (2012)

La estadística del índice de reclasificación neta revela que en general el modelo publicado de ARIC es mejor solamente marginando que el modelo publicado SAHS. Los modelos publicados muestran pobre calibración en comparación con los modelos locales que tienen buenas propiedades de calibración. El modelo Framingham muestra el peor resultado en cuando a recalibración, ARIC y SAHS tienen los mejor resultados, pero SAHS muestra un sobreestimado de casos con diabetes tipo 2 en comparación al modelo ARIC. El análisis realizado por los autores sugiere que el mejor modelo de los 3 evaluados es el Modelo ARIC.

Evaluación de Algoritmos de Redes Neuronales para Modelos de Predicción de Diabetes Tipo 2

Anand et al. (2012) mencionan que la diabetes es una enfermedad de mayor crecimiento siendo un problema alrededor del mundo. El diagnóstico de la diabetes tipo 2 en etapas tempranas es un reto complicado de alcanzar debido a las complejas interdependencias de varios factores. Es por tal razón que indican que es necesario el desarrollo de un sistema de soporte de diagnóstico el cual sea beneficioso para practicantes de la medicina en su proceso de diagnóstico.

Los autores refieren que si el conjunto de data en entrenamiento contiene atributos de clasificación irrelevantes el análisis puede producir un resultado menos preciso. Es por eso que el pre procesamiento de la data es requerido para mejorar la precisión del análisis, sin el pre procesamiento las redes neuronales podrían ser muy lentas.

Tabla 11: Atributos del conjunto de datos de Diabetes

Table 1. Attributes of Diabetes Data Set

No	Attribute	Description	Missing Value
1.	Pregnant	A record of the number of times the woman pregnant	110
2	Plasma glucose	Plasma glucose concentration measured using two hours oral glucose tolerance test (mm Hg)	5
3.	Diastolic BP	Diastolic blood pressure	35
4.	Triceps SFT	Triceps skin fold thickness (mm)	227
5.	Serum-Insulin	Two hours serum insulin (muU/ml).	374
6.	BMI	Body mass index (weight Kg/height in (mm) ²)	11
7.	DPF	Diabetes pedigree function	0
8.	Age	Age of patient(year)	0
9.	Class	Diabetes on set within five year	0

Fuente: Anand et al. (2012)

Usaron el conjunto de datos de The Pima Indian del repositorio de machine learning UCI.

El conjunto de datos consta de 9 atributos.

Tabla 12: Clasificación de la precisión de distintos algoritmos

Method	Classification Accuracy	Author
BP	75.2	Statlog
SNB	75.4	Ster et al.
NB	75.5	Ster et al.
kNN	75.5	Grudzinski
MML	75.5	Zarndt
RBF	75.7	Statlog
LVQ	75.8	Ster et al.
Semi-Naïve Bayes (5xCV)	76	Friedman
MLP + BP	76.4	Ster et al.
FDA	76.5	Ster et al.
ASI	76.6	Ster et al.
SMART	76.8	Statlog
GTO DT (5xCV)	76.8	Bennet et al.
BFGS quasi Newton	77.08	Yildirim et al.
LM	77.08	Yildirim et al.
LDA	77.5	Statlog et al.
GD	77.6	Yildirim et al.
SVM (5xCV)	77.6	Bennet et al.
GDA-LS-SVM	79.16	Polat et al.
GRNN	80.21	Yildirim et al.
LDA-MWSVM	89.74	Calisir et al.
MI-MCS-SVM	93.58	Giveki et al.

Fuente: Anand et al. (2012)

Los autores realizaron un comparativo de la tasa de precisión de N algoritmos, de los cuales algunos algoritmos como kNN, PCA, LDA, GDA permiten pre procesar la data, esto permite que la información faltante sea completada de acuerdo a patrones que detecta cada algoritmo, cada uno realizando su tarea de manera distinta al otro. Ellos muestran cómo es que el algoritmo LDA Morlet Wavelet Support Vector Machine Classifier o LDA-MWSVM el cual tiene como algoritmo de pre procesamiento a Linear Discriminant Analysis o LDA, mejora significativamente su ratio de precisión.

Concluyen que el pre procesamiento de la data y característica de extracción en conjunto, con modelos de Redes Neuronales mejora la clasificación de la precisión.

Sistema de Recomendación Confiable y Objetivo para el Diagnóstico de Enfermedades Crónica

Hussein et al. (2012) proponen un sistema de diagnóstico de enfermedades crónicas que sea exacto y confiable para brindar monitoreo remoto 24/7. En el planteo de su hipótesis, se indica que si se predice el riesgo de enfermedades crónicas de un paciente con alta precisión se debe esperar una mejora de las condiciones de salud del paciente y de ajustes en su estilo de vida, reduciendo los costos de cobertura de salud.

Para aumentar la precisión, es necesario reducir los falsos positivos y falsos negativos y es por esto que se considera el origen y la complejidad de la data médica recolectada (entendiendo por complejidad a la data no estandarizada, con alto número de parámetros o faltante). Se indica que la información médica es generada por diversas fuentes y por diversas personas, haciéndola propensa a distintos tipos de errores como data faltante o errónea, siendo este un reto sumamente importante de cara a la efectividad del diagnóstico. En el paper se plantea un alcance de clasificación de árbol de decisiones para aumentar la efectividad del diagnóstico, con resultados demostrados sobre data médica. Asimismo, para la recolección de información de un sistema de diagnóstico de enfermedades crónicas resulta necesario clasificar la información que será registrada:

- Data de entrenamiento: es información de pacientes diabéticos previos y consta de 935 registros de hospitales de Oman, dentro de los cuales existen atributos, valores y diagnósticos del doctor, empleando una escala de 1 a 10 donde 1-2 representan excelente prognosis y 9-10 representan condición crítica.
- Data demográfica: datos de edad, nombre, nivel de educación, tipo de dispositivo de conexión al sistema y tipo de conectividad.
- Data médica del paciente: nivel de azúcar en la sangre, presión sanguínea y peso.

Dentro de la clasificación de árbol de decisiones identificada, se emplea el método de clasificación de Random Forest (RF) para el análisis y segmentación de la información ingresada: se distingue entre información relacionada entre si en base a las escalas de diagnóstico, utilizando estas relaciones para generar un modelo predictivo para determinar la clase de diagnóstico del paciente.

En base al análisis y clasificación de estos datos, se brinda un valor numérico que representa el diagnóstico de riesgo de una enfermedad del paciente. Además, se toma el diagnóstico y se vuelve a analizar para ahora compararlo frente a diagnósticos de otros pacientes, identificando coincidencias adicionales.

Según el flujo planteado, concluido este análisis se envía información a los proveedores de salud quienes aprobarían la exactitud del diagnóstico y actualizarían el perfil del paciente, enviándole acciones de mitigación frente a los factores que suponen riesgo de desarrollo de la enfermedad y enviando la información del diagnóstico al sistema para mejorar la manera en que este aprende y brinda diagnósticos a futuro.

El modelo predictivo planteado es comparado frente a otros modelos de árbol de decisiones como J48, Decision Stump y Arbol REP. La información analizada es de pacientes en hospitales de Omán y consta de 935 pacientes y 13 atributos (incluyendo la escala de diagnóstico del doctor). Para la evaluación del modelo RF frente a los demás se empleó la métrica MAE (medición del error absoluto) dictada por la siguiente fórmula:

Ilustración 18: Ecuación de cálculo de métrica MAE

$$MAE = \frac{1}{n} \sum_{i=1}^n |f_i - y_i| = \frac{1}{n} \sum_{i=1}^n |e_i|$$

Fuente: Hussein et al. (2012)

Donde MAE representa el promedio de errores absolutos e_i , donde $e_i = f_i - y_i$, donde f_i es el diagnóstico y y_i es el diagnóstico real.

En base a la evaluación de resultados se identificó que:

- El error promedio de 10 árboles RF es 0.2877
- RF ha demostrado correctamente un 99.7% de casos correctamente clasificados
- RF mejora la precisión manejando la data internamente a través de las fases anteriormente mencionadas de análisis de sus predicciones una vez que estas ya han sido emitidas: este enfoque de aprender enriquece al modelo y estimula su exactitud.

Se concluye que el método de RF generará predicción efectiva de riesgo de enfermedades crónicas de pacientes que los sufren, debido a que aprende de sus propios resultados y enriquece sus predicciones futuras.

Modelo predictivo de diabetes con algoritmos de Percepción Multicapa, Redes Bayesianas y J48

Koklu & Unal (2013) observan que existe mucha información médica almacenada en bases de datos para investigaciones o análisis para ayudar a tomar decisiones médica, el análisis manual se ha vuelto ineficiente, es así que métodos de análisis computarizados son esenciales. La minería de datos se presenta como una de las tecnologías actualmente disponibles, ha demostrado ser válida, sensitiva, y un método confiable para descubrir patrones y relaciones. Por lo tanto aplicar esa información necesitada de análisis con la tecnología es necesaria para desarrollar un sistema de ayuda de decisiones médicas para practicantes médicos. Los autores usan el conjunto de datos de The Pima Indians Diabetes Data Set del repositorio de Machine Learning UCI. Y usan la herramienta Open Source Weka, la cual tiene una colección de algoritmos de machine learning para minería de datos.

Tabla 13: Atributos del conjunto de datos Pima Indian

No	Name	Type
1	Number of times pregnant	Numeric
2	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Numeric
3	Diastolic blood pressure (mm Hg)	Numeric
4	Triceps skin fold thickness (mm)	Numeric
5	2-Hour Serum insulin (mu U/ml)	Numeric
6	Body mass index (weight in kg/(height in m) ²)	Numeric
7	Diabetes pedigree function	Numeric
8	Age (years)	Numeric

Fuente: Koklu & Unal (2013)

El conjunto de datos comprende 8 atributos, todos los atributos son numéricos.

Tabla 14: Datos estadísticos de todas las características

Attribute	Min value	Max value	Mean Value	Standard Dev.
1	0	17	3.845	3.37
2	0	199	120.895	31.973
3	0	122	69.105	19.356
4	0	99	20.536	15.952
5	0	846	79.799	115.244
6	0	67.1	31.993	7.884
7	0.078	2.42	0.472	0,331
8	21	81	33.241	11.76

Fuente: Koklu & Unal (2013)

Cada atributo tiene datos estadísticos, valor mínimo, valor máximo, valor medio, desviación estándar.

Para la clasificación tuvieron como problema identificar a cuál del conjunto de categorías pertenece una nueva observación, sobre la base de un conjunto de datos de entrenamiento que contienen observaciones cuya categoría de miembro es conocida. La técnica de análisis de cluster es mayormente usada para objetos agregados dentro de grupos de acuerdo a la similitud de sus medidas.

Para realizar la clasificación de la información se usaron los 3 algoritmos para determinar cuál es el algoritmo que tiene una mejor precisión.

Tabla 15: Clasificación de Resultados

No	Classification Types	Accuracy	Time(sec.)
1	MLP	75.130	1.13
2	J48	73.828	0.08
3	Naïve Bayes	76.302	0.01

Fuente: Koklu & Unal (2013)

Los resultados muestran que el mejor ratio de precisión la tiene el algoritmo de Redes Bayesianas con 76,302%, este algoritmo también tuvo el mejor desempeño con el conjunto de datos de The Pima Indian.

Modelo de predicción de diabetes con el algoritmo J48 mejorado

Kaur & Chhabra (2014) mencionan que la diabetes tiene gran incidencia con otras enfermedades, siendo esto un gran factor de riesgo. Por lo que los autores deciden minar la información de The Pima Indian Diabetes, se realiza una comparación de algoritmos y se mejora uno de ellos para mejorar los resultados, todo realizado desde la herramienta Weka como API de Matlab.

Se realizaron los siguientes pasos dentro de Weka para realizar la minería de datos:

- Pre-procesamiento de la data y visualización
- Selección de atributos
- Clasificación (Arboles de decisiones)
- Predicción (Vecindad más cercana)
- Evaluación del modelo
- Clustering (Cobweb, K-means)
- Reglas de asociación

Ilustración 19: Flujo del algoritmo propuesto

Fuente: Kaur & Chhabra (2014)

El algoritmo propuesto usa la herramienta Weka como API de Matlab, este software permite la transformación de los datos de un lado a otro y utilizar la mayoría de las funciones disponibles en Weka, como clasificadores de entrenamiento. Al hacer esto se mejora la precisión del algoritmo J48 en comparación con el algoritmo original. El algoritmo propuesto funciona de la siguiente forma: El archivo de datos arff se carga desde WEKA en MATLAB. A continuación, se realiza la refinación del conjunto de datos. Posteriormente se

aplica el clasificador J48. Al final se obtienen los resultados, la precisión y se calcula la tasa de errores.

Tabla 16: Comparación de desempeño de algoritmos

Algorithm	Accuracy	Error
NaiveBayes	76.3021	23.6979
MultilayerPreception	73.3906	26.6094
MultiClassClassifier	77.2135	22.7865
RandomTree	68.099	31.901
REPTree	68.099	31.901
LADTree	74.0885	25.9115
BFTree	73.5677	26.4323
ADTree	72.9167	27.0833
RandomForest	73.9583	26.0417
J48	73.8281	26.1719
Proposed Algorithm	99.8700	0.1300

Fuente: Kaur & Chhabra (2014)

También realizan una evaluación de algoritmos para determinar el de mejor precisión. Como era de esperarse luego del mejoramiento del algoritmo J48, este sigue siendo el de mejor desempeño por lejos.

Se utilizaron las siguientes ecuaciones para la evaluación de los algoritmos:

Ilustración 20: Ecuaciones de evaluación

$$TA = \frac{(TP+TN)}{TP+TN+FP+FN}$$

$$RA = \frac{(TP+FP)*(TN+FN)+(FN+TP)*(FP+TP)}{(Total*Total)}$$

Fuente: Kaur & Chhabra (2014)

TA significa la precisión total, TP es el verdadero positivo, TN es el verdadero negativo, FP es el falso positivo, FN es el falso negativo. RA representa la precisión aleatoria.

Los resultados experimentales mostraron una mejora significativa sobre el algoritmo J48 existente. Y se probó que el algoritmo propuesto puede alcanzar una precisión del 99.87%.

Predicción de Diabetes usando una Red Bayesiana

Kumari et al. (2014) proponen la predicción de diabetes empleando técnicas de minería de datos, empleándola para identificar patrones en los conjuntos de datos a analizar. Se define a la minería de datos como un campo interdisciplinario de las ciencias de la computación enfocado al proceso de descubrimiento de patrones en largos conjuntos de datos, empleando métodos de inteligencia artificial, machine learning, estadística y sistemas de bases de datos. Para la predicción, se utiliza el modelo de red bayesiana que es un modelo gráfico que muestra relaciones de probabilidad entre variables de interés, y al mezclar esta información con técnicas estadísticas, los resultados son sumamente beneficiosas debido a que al ser una red, se gestionan las situaciones donde existe falta de información, además, permite identificar relaciones causales e incrementar el entendimiento del problema. Finalmente, la red bayesiana es útil porque el modelo planteado de enfermedad crónica tiene semánticas causales y de probabilidad, por lo que es un caso ideal para mezclar conocimiento previo y conocimiento nuevo.

Para el análisis predictivo se tomó un conjunto de datos de pacientes con variables tales como presión sanguínea, índice de masa corporal, edad, grosor de doblado de la piel de los tríceps, número de veces embarazado, ratio de tolerancia a la glucosa y nivel de insulina. Seguidamente se plantearon las siguientes tres clases posibles de diagnóstico:

Ilustración 21: Clasificación de clases de diagnóstico de diabetes

Fuente: Kumari et al. (2014)

Donde la clase 0 determinaba que no había riesgo, la clase 1 que había riesgo de desarrollar diabetes y la clase 2 que había diabetes.

Para la realización de la predicción se priorizaron las características primarias que reportan la incidencia de diabetes, por ejemplo si los valores de glucosa en plasma era menor a 100mg/dl y el valor del test de tolerancia a la glucosa era menor a 140 mg/dl, automáticamente se descartaba la diabetes del paciente. Se empleó una herramienta llamada

Waikato Environment for Knowledge Analysis de la Universidad de Waikato, en Nueva Zelanda y se obtuvieron los siguientes resultados:

Tabla 17: Resultados de diagnóstico de diabetes por clase

CLASS	ENTRIES	% OF PERSONS
NO	65	31.55
PRE	51	24.75
YES	90	43.68

Fuente: Kumari et al. (2014)

Para el análisis de error se empleó una matriz de confusión donde se identificaron los verdaderos negativos (TN), verdaderos positivos (TP), falsos positivos (FP) y falsos negativos (FN), calculando el error empleando la formula $(TP + TN) / (P+N)$ donde $P = TP + FN$ y $N = FP + TN$, obteniendo un resultado de precisión muy alto de 99.51.

Se identifica que la red bayesiana provee una respuesta eficiente para el diagnóstico de diabetes con un margen de error bastante reducido. Sin embargo, se considera que la cantidad de información usada puede no ser suficiente y los atributos no haber sido los totalmente correctos, por lo que se debe enfatizar en identificar aquellos que son de suma trascendencia para la predicción de la enfermedad.

Diagnóstico de Diabetes utilizando Técnicas de Clasificación y Minería

Iyer et al. (2015) presentan la creación de un modelo de minería de datos basado en dos algoritmos de clasificación para proveer una solución al diagnóstico de diabetes en mujeres. El primer algoritmo estudiado es el de árboles de decisión, el cual es usado para clasificar y luego predecir usando nodos y nodos internos. Los nodos internos son los casos de prueba usados para separar las instancias de características distintas y luego habiendo nodos de hoja, los cuales denotan la variable de la clase. El árbol de decisiones provee una técnica poderosa para la clasificación y predicción de la diabetes.

El segundo algoritmo es el de Naive Bayes que es un algoritmo de probabilidad que sigue los pasos de ejecución, clasificación, estimación y predicción. Se trata de un algoritmo que no es afectado por problemas que afectan a las herramientas existentes de minería de datos: se puede aplicar a un conjunto de datos largo en tiempo real y se pueden realizar estimaciones complejas del parámetro analizado sin problema.

Se consideraron las variables de número de veces embarazada, concentración de glucosa en plasma, la presión sanguínea, el grosor de la piel de tríceps, la insulina, el índice de masa corporal, edad y el factor hereditario. Se siguieron pasos de transformación de información reemplazando data faltante y normalizando data, pasando todos los atributos a tener valores

de 0 a 1. Los datos usados fueron la base de datos de diabetes en Indios Pima, del Instituto Nacional de Enfermedades Diabéticas, Digestivas y de Riñón.

Se empleó la herramienta WEKA (Waikato Environment for Knowledge Analysis de la Universidad de Waikato, en Nueva Zelanda) y se obtuvo una clasificación correcta de instancias de 74.8698% en el caso del algoritmo de árbol de decisiones

Debido a que se empleó una técnica de 70-30 en la distribución de información para análisis (70) y entrenamiento (30), se volvió a realizar la ejecución de la predicción para el primer algoritmo, aumentando las instancias clasificadas correctamente para el caso del árbol de decisiones a un 76.9565%.

Se aplicó la misma metodología para la evaluación del modelo Naive Bayes y se obtuvo un ratio de instancias clasificadas correctamente de 79.5652%, significativamente mayor que el del primer algoritmo, identificando que es sumamente trascendente emplear información que sirva de entrenamiento para el modelo, con el fin de aumentar el ratio de detección correcto. Se concluye que ambos modelos son eficientes en el diagnóstico de la diabetes, pero se requiere mayor número de datos de entrenamiento y que el modelo aprenda de sus predicciones para poder reducir este ratio de error. Asimismo, se resalta que es necesario adoptar información de todas partes del mundo para ahondar en la precisión del modelo predictivo y además, se plantea que el trabajo realizado puede ser utilizado para su implementación en un sistema que brinde diagnósticos automáticos.

H2RM: Un Modelo de Razonamiento Híbrido para la Predicción y Gestión de la Diabetes Mellitus

Ali et al. (2015) proponen un modelo de razonamiento híbrido para la predicción y gestión de la diabetes mellitus tipo 1 y 2. Se determinaron un conjunto de reglas de predicción en base a la data de 50 pacientes de diabetes de un hospital coreano entre los años 2008 a 2011. En la investigación se diseñó el modelo de razonamiento híbrido empleando:

- Conjuntos de resultados clínicos de pacientes recolectado por doctores en visitas a hospitales y guías virtuales para gestionar anormalidades en el tratamiento de diabetes. Esto constituyó las fuentes de conocimiento para la predicción y gestión de la enfermedad. (PCG)
- Conjunto de métodos y procedimientos empleados para traducir las tablas clínicas y guías virtuales para que sigan un formato estructurado de datos y hagan referencia a un rango determinado. (CGT)

- Un conjunto de técnicas matemáticas y de inteligencia artificial que permiten asignar atributos de valores continuos a valores discretos, emplea técnicas de identificación de atributos esenciales y permiten identificar reglas empleando el algoritmo de LEM2. (RKA)
- Conjuntos de bases de conocimiento de reglas para predecir la diabetes tipo 1 y 2. (KB)
- Metodología de razonamiento basado en reglas para la predicción de la diabetes y provee un rango de razonamiento para definir reglas sobre observaciones anormales. (HRBR)
- Un conjunto de modelos estadísticos, tales como análisis de regresión y de tendencias para i identificar observaciones anormales y predecir tendencias futuras. (CTA)

Se atravesó una etapa de depuración de información redundante por parte de un experto. Para tratar el caso de información faltante, se emplearon tres criterios:

- Cese de tratamiento del atributo
- Valor medio/frecuente
- Valor inmediato previo/siguiente

Luego, se realizó la normalización de la información empleando métodos de conversión a valores discretos, para definir intervalos, pero con suma cautela de retener la significancia de cada rango de categorías, debido a que si no se hace esto se perdería la información real a evaluar.

La investigación desarrollada integra los componentes anteriormente mencionados de la siguiente manera:

- PCG: fuentes de conocimiento
- CGT: traducción y extracción de conocimiento
- RKA: proceso de minería de reglas
- KB: repositorio de conocimiento
- HRBR: proceso de predicción y clasificación en tiempo real
- CTA: análisis de tendencia futura.

Inicialmente el modelo aplica un proceso manual de depuración de tablas clínicas y luego emplea la conversión de valores a valores discretos según sus rangos, para luego emplear el algoritmo LEM2 en el análisis de reglas de decisión. El modelo obtuvo un desempeño de 95.91% en clasificar los tipos de diabetes, permitiendo además en base al uso de análisis de tendencias correlacionadas brindar una vista más detallada de las condiciones del paciente a un doctor de manera adecuada, asistiéndolos en controlar los factores de riesgos. El modelo híbrido se desempeña adecuadamente, pero es necesario llevar mejoras acabo, inclusive

sugiriendo implementar la funcionalidad de tratamiento a seguir en base a las condiciones del paciente.

Detección Temprana de Diabetes Mellitus usando un Modelo Jerárquico Difuso

Lukmanto & Irwansyah (2015) proponen un método para la detección temprana de la diabetes basándose en actividades de encuesta y adquisición de conocimiento.

Se empleó como data inicial información de encuestas en un laboratorio de hospital de Yakarta, Indonesia y a través de reuniones con dos especialistas médicos para adquirir conocimiento. Se trabajó con un muestreo sobre los resultados de los exámenes de glucosa de los pacientes donde también se incluye la esas, la medición de la glucosa sin ingerir alimento por más de 8 horas, pero no más de 14 y 2 horas después que el paciente consumió, además de realizar la evaluación HbA1c para identificar el nivel de glucosa promedio en los últimos 120 días. Además, se identificó que existen tres características que pueden contribuir a la detección temprana: orina incrementada, niveles de sed altos y hambre repentina.

Para determinar si una persona tiene el potencial para desarrollar diabetes, se considera que debe tener 2 de los 3 síntomas anteriormente mencionados y un factor de riesgo resultante en uno de los análisis de glucosa, comparando además esto con la edad del paciente y determinando si tiene potencial bajo, medio o alto a desarrollar la enfermedad. La información y el conocimiento adquirido fueron cuantificados a través de rangos que permitieron valorizar cada uno e identificar el riesgo que supone.

Se propone el siguiente flujo:

- Variable de input: valores de polidipsia (exceso de sed), poliuria (exceso de orina), polifagia (exceso de hambre), resultados de exámenes de glucosa y edad. Estas variables atravesaron un proceso de conversión a variables difusas, las cuales luego serán normalizadas para mostrar el resultado final.
- Creación de variable temporal: se crea una variable denominada riesgo oral obtenida a través de la inferencia difusa de la polidipsia y la polifagia, luego evaluando este riesgo oral con el riesgo de glucosa en base al procesamiento difuso de los resultados de la glucosa actual con la glucosa 2 horas después de ingerir alimentos y determinando a través de un nuevo procesamiento inferencial difuso estos dos factores de riesgo.
- Variable de output: es la variable que provee el resultado final, la cual determina si el paciente tiene o no riesgo de desarrollar la diabetes

El proceso de evaluación del riesgo de desarrollo temprano de la diabetes es realizado a través de una simulación del modelo jerárquico difuso en MATLAB y se realiza la validación de resultados obtenidos por el modelo con los resultados obtenidos por un experto médico analizando la semejanza entre el diagnóstico del modelo y el diagnóstico del doctor.

Se concluye que la aplicación de inteligencia computacional a través del uso de modelos jerárquicos difusos con la habilidad de detectar de manera temprana la diabetes e identificar el riesgo del paciente a desarrollar la enfermedad es de suma utilidad debido a que se obtuvo un 87.46% de efectividad en la predicción desarrollada.

Modelo predictivo de Diabetes en Hadoop/Map reduce

Eswari et al. (2015) mencionan que en recientes años se genera una gran cantidad de datos médicos, y con el crecimiento de Big Data es necesario estructurar y enfatizar el tamaño de dicha información para una solución posible. Identifican a la diabetes como una de las enfermedades de mayor incidencia en países en desarrollo como la India. Con la ayuda de los desarrollos tecnológicos, es necesario combinar la robusta información compartida de diabetes y los sistemas de comunicación electrónica pueden facilitar el acceso a los servicios médicos a todos los niveles de pacientes. El desarrollo de un Intercambio de Información de Salud puede extraer información de muchos repositorios e integrar la data dentro de un registro único de pacientes, al que los proveedores de atención puedan acceder de forma segura. Es así que los autores usan un algoritmo de análisis predictivo en Hadoop/Map Reduce para predecir la prevalencia de la diabetes, y sus complicaciones asociadas y el tipo de tratamiento para el paciente.

Ilustración 22: Arquitectura del Sistema de Análisis Predictivo en aplicaciones de Salud

Fuente: Saravana et al. (2015)

La arquitectura incluye las fases de Colección de datos, Data Warehousing o almacenamiento de datos, análisis predictivo, informes de procesamiento analizados. La colección de datos comprende los Electronic Health Records (EHR)/ Patient Health Records (PHR), sistemas clínicos, recursos externos, archivos en varios formatos.

El análisis predictivo permite a los proveedores de salud precisamente esperar y responder a las necesidades del paciente, proveyendo la habilidad de tomar mejores decisiones financieras y clínicas basadas en predicciones hechas por el sistema. El sistema usa Hadoop/Map Reduce para predecir y clasificar el tipo de diabetes y sus complicaciones.

El descubrimiento de patrones de análisis predictivo debe incluir lo siguiente:

- Minería de reglas de asociación
- Clustering
- Clasificación
- Uso de estadísticas
- Aplicación de reglas deductivas predefinidas a través de la data

Ilustración 23: Sistema Predictivo – Coincidencia de Patrones

Fuente: Saravana et al. (2015)

Cada vez que el Data Warehouse envía conjunto de datos al sistema Hadoop, el Mapa reduce las tareas distribuyéndolas en nodos. Se tiene un Nodo maestro, el NN y un nodo buscador de trabajo, el JT, el que realiza la tarea de reducir el trabajo usando el mapa. Los nodos trabajadores o esclavos reciben la información del nodo maestro, procesan la tarea de coincidencia de patrones con la data de diabetes con ayuda del nodo de data (DN) y el Buscador de tareas (TT). La coincidencia de predicción es el proceso de comparar el valor del umbral analizado con el valor obtenido. Si el proceso de coincidencia de patrones es completado por todos los nodos trabajadores según el requerimiento, se almacenan en discos intermediarios. Si la tarea de reducción la inicio el Nodo Maestro, todos los Nodos Trabajadores asignados leen la data procesada de los discos intermediarios. Según la consulta recibida del cliente a través del Nodo Maestro, la reducción de tareas es desempeñada por los Nodos Trabajadores. Finalizando con la distribución en varios servidores de los resultados obtenidos en la fase de reducción.

Luego del análisis la información llega a varios servidores y es replicada en varios nodos. Al emplear correctamente la tecnología de comunicación electrónica para intercambiar la información de pacientes sobre los centros de salud para que reciban el tratamiento en el tiempo correcto en locaciones remotas a bajo costo.

La implementación de Big Data Analytics con Hadoop provee una forma sistemática para alcanzar mejores resultados como disponibilidad y bajos costos en salud para la población. Al transformar la información clínica sobre diabetes de varios pacientes para su análisis a resultados útiles, este análisis puede hacer entender al paciente las complicaciones que se presenten.

Modelo predictivo de puntuación de Riesgo de Salud

Hamad et al. (2015) indican que los investigadores siempre tienen problemas para encontrar la mejor forma de capturar el estado de salud en general de cada individuo con medidas subjetivas y objetivas. Con la creciente disponibilidad de fuentes de Big Data en forma de solicitudes digitalizadas vinculables y registros administrativos, epidemiólogos e investigadores de salud tiene ahora la oportunidad de realizar estudios utilizando grandes conjuntos de datos longitudinales, como los de Medicare.

Los algoritmos de predicción o “puntuación de riesgo” valiosos en entornos corporativos, fueron inicialmente usados por actuarios y aseguradoras para crear algoritmos de predicción para pronosticar los gastos de salud y por los Centros de Servicios Medicare y Medicaid para determinar los pagos a las organizaciones de mantenimiento de salud.

Los autores utilizaron el software DxGC Intelligence de Verisk Health. La muestra con la que demuestran las propiedades y valor predictivo de la puntuación de riesgo es una cohorte de trabajadores de manufactura de Alcoa, un empleador de USA.

Tabla 18: Conjunto de datos vinculados usados en el estudio.

Dataset	Contents
Personnel	Age
	Race
	Gender
	Employment status (e.g., active, on leave)*
Claims	International Classification of Diseases codes
	Current Procedural Terminology codes
	Dates of healthcare encounters
National Death Index	Date of death
Eligibility	Insurance status

* This variable was used to determine which employees to include in our sample, i.e., those who were actively employed on January 1, 1996.

Fuente: Hamad et al. (2015)

Utilizaron los datos de las solicitudes para identificar casos de enfermedades como la diabetes, hipertensión, asma/enfermedad pulmonar obstructiva crónica, depresión y enfermedad cardíaca isquémica. Para examinar la mortalidad, enlazaron el conjunto de datos con el National Death Index para obtener la fecha de muerte de cada individuo en la muestra que haya muerto, incluyendo a los que dejaron la empresa en algún momento del periodo de estudio.

Realizaron varios análisis para ilustrar las propiedades de la puntuación de riesgo, examinaron sus correlatos demográficos, y para demostrar su relación con una variedad de resultados de salud y de no salud.

Para valorar la puntuación de riesgo asociados con el desarrollo de enfermedades a largo plazo, realizaron un análisis de tiempo hasta el evento para cada una de las condiciones de la enfermedad. Los análisis les dieron los siguientes resultados.

Tabla 19: Características de la muestra

Sociodemographic characteristics.	
Female (%)	10.1
Age in 1996 (mean $\pm$ SD)	46.7 $\pm$ 8.7
Race (%)	
White	87.1
Black	8.5
Hispanic	3.7
Other	0.7
Health characteristics	
Deaths during 1996–2011 (%)	8.2
New disease diagnoses during 1996–2011 (%)	15.9
Diabetes	39.6
Hypertension	6.6
Asthma/COPD	4.7
Depression	14.6
Risk score properties (1996)	
Mean $\pm$ SD	1.12 $\pm$ 1.36
Median	0.79
Min, Max	0.23, 33.19
Quintiles	
Q1	0.23, 0.53
Q2	0.53, 0.69
Q3	0.69, 0.93
Q4	0.93, 1.38
Q5	1.38, 33.19

Inclusion criteria: Employed on January 1, 1996 with at least one risk score in the period 1996–2011 (N = 14,161). COPD = chronic obstructive pulmonary disease.

Fuente: Hamad et al. (2015)

El mayor puntaje de riesgo para la muestra es de 1.12 con una desviación estándar de 1.36. Los puntajes de riesgo son bastante estables en el tiempo

Tabla 20: Puntuaciones de riesgo correlativos

Variable	Beta	[95% CI]
Age (per 10-year increment)	0.51**	[0.48, 0.54]
Female	0.12**	[0.03, 0.20]
Race (Ref = White)		
Black	0.45**	[0.31, 0.59]
Hispanic	0.078	[-0.085, 0.24]
Other	-0.14	[-0.51, 0.22]
Year	0.025**	[0.021, 0.030]
Constant	-51.74**	[-61.06, -42.41]
Observations	151,931	
Individuals	13,880	

* $p < 0.05$,

** $p < 0.01$.

Note: Sample includes individuals employed at the firm on January 1, 1996. Analysis conducted using multivariable linear regression with individual-level random effects. Robust standard errors are clustered at the individual level.

Fuente: Hamad et al. (2015)

Los factores de riesgo para mayores puntuaciones de riesgo que determinaron incluyen la edad ($\beta = 0.51$ por 10 años de incremento, $p < 0,001$), ser mujer ($\beta = 0,12$, $p = 0,005$), y ser de raza negra ($\beta = 0,45$, $p < 0,001$). Después de un control por estas covariables, observaron un aumento anual en el promedio de riesgo de 0,025 unidades durante el período de estudio ($p < 0,001$). Dado el envejecimiento de la muestra y estas tendencias a largo plazo, la puntuación de riesgo media en el año 2011 fue de 1.83 con una desviación estándar de 2,72.

Tabla 21: Asociación entre las puntuaciones de riesgos y nuevos diagnósticos de enfermedades en años subsecuentes.

	Coefficient [95% CI]				
	Asthma	Depression	Diabetes	Hypertension	Ischemic heart disease
Previous year risk score	0.00041** [0.00026, 0.00057]	0.00021** [0.000097, 0.00033]	0.00047** [0.00026, 0.00068]	-0.000077 [-0.00037, 0.00022]	0.00041** [0.00019, 0.00063]
Age	0.000024 [-0.000015, 0.000063]	-0.00017** [-0.00020, -0.00013]	0.00018** [0.00013, 0.00023]	0.00027** [0.00018, 0.00035]	0.00050** [0.00045, 0.00055]
Female	0.0023** [0.00098, 0.0036]	0.0025** [0.0013, 0.0037]	-0.0030** [-0.0044, -0.0017]	-0.0030** [-0.0052, -0.00076]	-0.0047** [-0.0058, -0.0037]
Race (ref white)					
Black	-0.00072 [-0.0019, 0.00048]	-0.00093* [-0.0018, -0.000030]	0.0067** [0.0045, 0.0089]	0.010** [0.0073, 0.013]	-0.0010 [-0.0027, 0.00064]
Hispanic	-0.0026** [-0.0038, -0.0014]	0.000030 [-0.0015, 0.0015]	0.0052** [0.0022, 0.0082]	0.0023 [-0.0014, 0.0060]	-0.00071 [-0.0030, 0.0016]
Other	-0.0021 [-0.0055, 0.0012]	-0.0010 [-0.0043, 0.0023]	0.00090 [-0.0056, 0.0074]	-0.0062 [-0.014, 0.0020]	-0.0017 [-0.0069, 0.0034]
Observations	143,822	144,392	139,633	127,321	141,706
Individuals	13,681	13,736	13,293	12,191	13,441

* $p < 0.05$,

** $p < 0.01$.

Note: Sample includes individuals employed at the firm on January 1, 1996. Analyses are conducted using linear probability models with individual-level random effects, in which an individual's risk score in one year predicts their likelihood of a new diagnosis of disease in the following year. Standard errors are clustered at the individual level. To be considered a new diagnosis, the individual must have been free of the disease for the first two years of the study. For each of these conditions, individuals with one or more inpatient claims or two or more outpatient claims with a relevant ICD diagnosis code in a 365-day period are considered to have the disease in question. Each model includes dummy variables for year to control for secular trends.

Fuente: Hamad et al. (2015)

Cada incremento de 1 en la escala de riesgo se asocia, en promedio, con una mayor probabilidad de recibir un nuevo diagnóstico de asma (0,04%, $p < 0,001$), depresión (0,02%, $p < 0,001$), la diabetes (0,05%, $p < 0,001$), y la CI (0,04%, $p < 0,001$) en el año siguiente.

Tabla 22: Curvas de supervivencia Kaplan-Meier para la diabetes

Fuente: Hamad et al. (2015)

Examinaron las capacidades de predicción a largo plazo de la puntuación de riesgo, utilizando el valor de la línea de base de la puntuación de riesgo en 1996 para examinar el

tiempo de diagnóstico para cada condición. Las curvas de supervivencia Kaplan-Meier demuestran un aumento monótonico en la probabilidad de diagnóstico para cada condición con los quintiles de mayor puntuación de riesgo.

Tabla 23: Modelos Cox de riesgos proporcionales para el diagnóstico y la mortalidad por enfermedad, por quintiles de puntuación de riesgo.

	Hazard Ratio [95% CI]					
	Asthma	Depression	Diabetes	Hypertension	Ischemic heart disease	Mortality
Risk score quintile (ref = Q1)						
Q2	1.50*	1.39	1.18	1.31**	1.43**	1.32
	[1.09, 2.07]	[0.98, 1.95]	[0.97, 1.45]	[1.15, 1.48]	[1.12, 1.84]	[0.92, 1.91]
Q3	1.91**	1.96**	1.37**	1.20**	1.67**	1.08
	[1.39, 2.62]	[1.40, 2.74]	[1.12, 1.68]	[1.05, 1.36]	[1.31, 2.14]	[0.74, 1.56]
Q4	2.32**	2.28**	1.70**	1.39**	2.35**	1.35
	[1.69, 3.20]	[1.61, 3.22]	[1.38, 2.08]	[1.22, 1.59]	[1.85, 2.99]	[0.94, 1.94]
Q5	2.75**	2.18**	1.91**	1.43**	2.73**	2.24**
	[1.98, 3.81]	[1.50, 3.17]	[1.54, 2.37]	[1.24, 1.65]	[2.13, 3.49]	[1.57, 3.19]
Female	1.27	1.33*	0.56**	0.78**	0.31**	0.63**
	[0.99, 1.62]	[1.02, 1.74]	[0.45, 0.69]	[0.68, 0.88]	[0.23, 0.43]	[0.46, 0.85]
Race (ref white)						
Black	0.75	0.84	1.81**	1.86**	0.95	1.29*
	[0.54, 1.06]	[0.56, 1.26]	[1.51, 2.18]	[1.63, 2.12]	[0.76, 1.19]	[1.02, 1.63]
Hispanic	0.38**	1.07	1.92**	1.00	0.87	1.05
	[0.20, 0.76]	[0.64, 1.80]	[1.48, 2.49]	[0.81, 1.23]	[0.63, 1.21]	[0.70, 1.57]
Other	0.67	0.85	1.06	0.88	1.44	1.03
	[0.17, 2.70]	[0.21, 3.43]	[0.50, 2.25]	[0.48, 1.58]	[0.64, 3.21]	[0.39, 2.77]
Observations	8,841	8,886	8,522	7,641	8,665	9,012

* p < 0.05,

** p < 0.01.

Note: Sample includes individuals employed at the firm on January 1, 1996. To be considered a new diagnosis, the individual must have been free of the disease for the first two years of the study. For each of these conditions, individuals with one or more inpatient claims or two or more outpatient claims with a relevant ICD diagnosis code in a 365-day period are considered to have the disease in question. Models include dummy variables to control for age group at baseline (20–30 years old, 30–40 years old, etc.). Individuals were censored at the last date that they were active at the firm based on the personnel dataset. For mortality, individuals were censored at September 1, 2011, after which we do not have data on mortality.

Fuente: Hamad et al. (2015)

El uso de los quintiles de puntuación como predictor de riesgo en modelos de Cox confirma la relación observada en las curvas de Kaplan-Meier, es decir, hay un aumento monótonico en la probabilidad de diagnóstico con un riesgo más alto para la mayoría de los quintiles de puntuación de las condiciones de salud.

Modelo predictivo escalable con RapidMiner

Van Poucke et al. (2016) mencionan que la acumulación de data clínica ha sobrepasado la capacidad para la agregación y el análisis efectivo para mejorar la calidad clínica, seguridad del paciente, y un cuidado integrado del paciente. Un análisis inteligente de la data trae consigo una representación efectiva de las complejas relaciones entre los síntomas, enfermedades y los tratamientos. Pero como es común solo se utilizan las plataformas de análisis para monitorear e integrar la data, dejando de lado el análisis y la toma de decisiones. La ventaja de la aplicación de análisis predictivo es poder brindar un apoyo en decisiones médicas validado, y guiar hacia la Predicción, Prevención y Medicina Personalizada

(PPPM). PPPM es un concepto integrador en el cuidado de la salud que permite predecir las predisposiciones individuales antes del inicio de la enfermedad, proporcionar medidas preventivas específicas, y crear algoritmos de tratamiento adaptados a la persona.

Los autores realizaron una pequeña comparación entre las herramientas que puedan cubrir sus necesidades eligiendo la herramienta RapidMiner. Para realizar sus pruebas utilizaron la base de datos clínica de The MIMIC II (Version 2.6).

Realizaron un Modelo automático de creación, optimización de la parametrización y la evaluación.

Ilustración 24: Proceso básico de construcción automática, optimización de la parametrización y evaluación de múltiples modelos predictivos en RapidMiner.

Fuente: Van Poucke et al. (2016)

Definieron 17 sobprocesos conteniendo un algoritmo predictivo, la estructura del proceso permite la parametrización si es necesario. Implementaron un algoritmo evolucionario para la optimización de parámetros SVM.

Ilustración 25: Conjunto de los métodos de aprendizaje como se muestra en RapidMiner

Fuente: Van Poucke et al. (2016)

Adicionalmente desarrollaron 2 procesos como extensión para permitir la característica de selección automática, construcción de modelos y evaluación bajo varias características de selección de esquemas.

Lograron la integración de MIMIC II en RapidMiner, permitiendo el análisis predictivo escalable de datos clínicos de los pacientes en estado crítico, a través de un entorno visual de código libre. Uno de los mayores beneficios de la integración es la manipulación sin problemas, la extracción de datos, el pre-procesamiento y análisis predictivo de grandes cantidades de datos, utilizando herramientas visuales sin la necesidad de realizar una sola línea de código.

CAPÍTULO 5. Modelo Predictivo para el diagnostico de la diabetes mellitus tipo 2

El siguiente capítulo muestra el detalle del trabajo realizado con la finalidad de plantear la estructura del esquema del modelo predictivo. De tal manera, se muestra la planificación, inputs, procesamiento de la información y las pruebas del modelo para el desarrollo del modelo predictivo.

ANALISIS DE ENFERMEDADES CRONICAS

Definición

El análisis de enfermedades crónicas es desarrollado en la fase previa del diseño del modelo predictivo, siendo el primero de dos entregables realizados como etapa preliminar al modelo. El propósito es la identificación de la enfermedad crónica de mayor impacto social y epidemiológico en el Perú, con el fin de usarla como base para el desarrollo del modelo predictivo.

Objetivo

Analizar las enfermedades crónicas de mayor impacto para el contexto peruano e identificar cuál es aquella en la que se deba enfocar el modelo predictivo, según su relevancia.

Alcance

El análisis de enfermedades crónicas tiene como fin determinar mediante un análisis, la enfermedad crónica que presenta un mayor impacto, factor social y epidemiológico para su desarrollo y evolución en el escenario peruano, con el propósito de identificar aquel enfoque que se tendrá para el desarrollo del modelo predictivo como resultado final del presente proyecto de investigación.

Contexto

En un mundo de ritmos acelerados y una adopción general de pobres costumbres de cuidado de salud, por la misma celeridad que nos exige el siglo XXI, resulta común sentir que el campo de enfermedades y su propagación han crecido de una manera estrepitosa: vemos la esperanza de vida incrementar y vemos en las noticias regularmente celebraciones de las personas más longevas del mundo, sin embargo no resulta algo ajeno a nosotros el enterarse de muertes cercanas a edades tempranas y por factores no propios de la degeneración natural del cuerpo humano, sino por enfermedades al corazón, cáncer o complicaciones derivadas de la diabetes.

Vivimos en una vorágine de exigencias familiares, académicas y laborales, por lo cual resulta natural descuidar nuestro cuerpo, aunque esto termine siendo altamente perjudicial. No es sorpresa identificar que las muertes por enfermedades crónicas estén proyectadas a aumentar en un 72% (de 29 millones a 50 millones de muertes) para el año 2020 según una publicación

del American Journal of Public Health ³. Es bajo esta premisa que identificamos a las enfermedades crónicas como un foco de alto potencial sobre el cual se pueden desarrollar proyectos de análisis como el propuesto, puesto que se trata de un fenómeno inherente a la modernidad y cuyo pico aún estamos por experimentar.

Identificación de Enfermedades

Para continuar con el análisis es necesario partir de una definición formal del concepto estudiado. Según la OMS, las enfermedades crónicas o no transmisibles están definidas como aquellas enfermedades de larga duración, no contagiosas y por lo general de lenta progresión, siendo la principal causa de mortalidad en el mundo, siendo responsables del 70% de las causas de muerte. Las principales enfermedades crónicas son las enfermedades cardiovasculares, diabetes, cáncer, obesidad, artritis y enfermedades respiratorias.⁴

Tratándose de una cantidad de enfermedades representativa, surge la necesidad de ir identificando cuáles serán las que se emplearán como principales objetos de estudio. Debido a la alta trascendencia económica derivada de las enfermedades crónicas y su tratamiento, se empleará el criterio monetario inicialmente.

A través de la cuantificación económica, identificamos a los 4 tipos principales de enfermedades crónicas de acuerdo al gasto que incurren en los EE.UU son las siguientes:

1. Enfermedades cardiovasculares
2. Diabetes
3. Cáncer
4. Obesidad

Debido a que estas 4 agrupaciones son las de mayor impacto económico, se tomaran como punto de partida las mismas para realizar el siguiente análisis referente a su descripción y sintomatología:

Enfermedades cardiovasculares

Las enfermedades cardiovasculares (ECV) son un grupo de desórdenes o trastornos del corazón y los vasos sanguíneos, entre estos se encuentran:

- ✓ **La cardiopatía coronaria:** enfermedad en los vasos sanguíneos que irrigan el corazón.
- ✓ **Las enfermedades cerebrovasculares:** enfermedades en los vasos sanguíneos que irrigan el cerebro.

³ Cfr. American Journal of Public Health

⁴ Cfr. OMS 2016 Enfermedades Crónicas

- ✓ **Las artropatías periféricas:** enfermedades en los vasos sanguíneos que irrigan los miembros superiores e inferiores.
- ✓ **La cardiopatía reumática:** lesiones del corazón y de las válvulas cardíacas de debidas a la fiebre reumática, enfermedad causada por las bacterias estreptococos.
- ✓ **Las cardiopatías congénitas:** malformaciones del corazón presentes desde el nacimiento.
- ✓ **Las trombosis venosas profundas y embolias pulmonares:** coágulos de sangre (trombos) en las venas de las piernas, que pueden desprenderse (émbolos) y situarse en los vasos de los pulmones o del corazón.

Los ataques al corazón y los accidentes vasculares cerebrales (AVC) suelen ser fenómenos agudos que se originan, en su mayoría, por obstrucciones que impiden el fluido de la sangre hacia el cerebro o el corazón.⁵

Diabetes

La diabetes es una enfermedad crónica que aparece cuando el páncreas no produce suficiente insulina o cuando el organismo no utiliza eficazmente la insulina que se produce. La insulina es una hormona que regula el azúcar en la sangre. El efecto de la diabetes no controlada es la hiperglucemia (aumento de azúcar en la sangre), que con el tiempo desencadena fallas en muchos órganos y sistemas, especialmente los sistemas nervioso y sanguíneo.

- ✓ **Diabetes Tipo 1:** Es también llamada insulín dependiente, juvenil o de inicio en la infancia, se da por una producción insuficiente de insulina, por lo cual requiere de la administración diaria de la hormona. La causa de la diabetes tipo 1 se desconoce hasta el momento y no se puede prevenir con el conocimiento actual.
- ✓ **Diabetes Tipo 2:** Es también llamada no insulín dependiente o de inicio en la edad adulta, se caracteriza por una utilización ineficaz de la insulina. La diabetes tipo 2 representa el 90% de casos a nivel mundial.
Hasta hace poco este tipo de diabetes solo se mostraba en adultos, pero en la actualidad se da también en niños.
- ✓ **Diabetes Gestacional:** Se caracteriza por hiperglucemia (aumento del azúcar en la sangre) que aparece en la etapa gestante y alcanza valores por encima de los normales pero inferiores a los establecidos para diagnosticar diabetes.⁶

Cáncer

El cáncer es un término genérico para agrupar un conjunto de enfermedades que pueden afectar a cualquier parte del organismo; también se habla de tumores malignos o neoplasias malignas. Una característica del cáncer es la multiplicación rápida de las

⁵ Cfr. OMS 2016 Enfermedades Cardiovasculares

⁶ Cfr. OMS 2016 Diabetes

células malignas o anormales que se entienden por encima de sus límites habituales, llegando hasta invadir partes adyacentes del cuerpo o propagarse a otros órganos, este proceso se le conoce como metástasis.⁷

Los principales tipos de cáncer, por cantidad de defunciones mundiales (2012), son:

- ✓ **Cáncer pulmonar (1,59 millones)**
- ✓ **Cáncer hepático (745 000)**
- ✓ **Cáncer gástrico (723 000)**
- ✓ **Cáncer colorrectal (694 000)**
- ✓ **Cáncer mamario (521 000)**
- ✓ **Cáncer de esófago (400 000)**

Obesidad

La obesidad, al igual que el sobrepeso, se define como la acumulación anormal o excesiva de tejido adiposo o grasa, siendo perjudicial para la salud.

El Índice de Masa Corporal o IMC es un indicador simple de la relación entre el peso y la talla que se utiliza comúnmente para identificar la obesidad, al igual que el sobrepeso, en los adultos, se calcula mediante la división del peso de una persona en Kilogramos entre su talla en Metro al cuadrado; tener un IMC igual o mayor a 30 determina obesidad.⁸

Los expertos advierten que sus efectos más negativos actúan como un agente que determina y agrava a corto plazo patologías graves como la **diabetes**, algunos tipos de **cáncer**, hipertensión, complicaciones **cardiovasculares**.⁹

A partir de estas 4 enfermedades no transmisibles se realiza el análisis de las misma para la elección de la enfermedad que se utilizada para el modelo predictivo para el diagnóstico de la enfermedad.

⁷ Cfr. OMS 2016 Cáncer

⁸ Cfr. OMS 2016 Obesidad

⁹ Cfr. DMedicina 2016

Relación entre las Enfermedades Crónicas y las Metas Mundiales de la OMS

En el 2014 la Organización Mundial de la Salud emite un informe sobre las enfermedades crónicas, en dicho informe se expresan metas mundiales para la reducción de dichas enfermedades, a continuación, mostraremos las metas que están relacionadas con las 4 principales enfermedades crónicas:

Meta mundial 1: Reducción relativa de la mortalidad general por enfermedades **cardiovasculares, cáncer, diabetes** o enfermedades respiratorias crónicas en un 25% para 2025.

Meta mundial 3: Reducción relativa de la prevalencia de la actividad física insuficiente en un 10% para 2025. *(La actividad física regular reduce el riesgo de **cardiopatía** isquémica, accidente **cerebrovascular**, **diabetes**, y **cáncer de mama y de colon**).*

Meta mundial 4: Reducción relativa de la ingesta poblacional media de sal o sodio en un 30% para 2025. *(El consumo excesivo de sodio en la dieta se ha asociado al aumento del riesgo de hipertensión y **enfermedades cardiovasculares**).*

Meta mundial 5: Reducción relativa de la prevalencia del consumo actual de tabaco en un 30% en las personas de 15 años o más para 2025. *(Desarrollo de **enfermedades cardiovasculares, diabetes, cáncer**.)*

Meta mundial 6: Reducción relativa de la prevalencia de la hipertensión en un 25%, o contención de la prevalencia de la hipertensión, en función de las circunstancias del país, para 2025. *(Si no se controla, la hipertensión causa **accidentes cerebrovasculares, infarto de miocardio, insuficiencia cardíaca, demencia, insuficiencia renal y ceguera**.)*

Meta mundial 7: Detención del aumento de la **diabetes y la obesidad** para 2025.

Meta mundial 8: Tratamiento farmacológico y asesoramiento (incluido el control de la **glucemia**) de al menos un 50% de las personas que lo necesitan para prevenir ataques **cardíacos y accidentes cerebrovasculares** para 2025.

Meta mundial 9: 80% de disponibilidad de tecnologías básicas y medicamentos esenciales asequibles, incluidos los genéricos, necesarios para tratar las principales enfermedades no transmisibles, en centros tanto públicos como privados para 2025. *(Esta meta se refiere a la necesidad básica de tecnologías y medicamentos para llevar a cabo intervenciones costo eficaces de atención primaria a fin de hacer frente a las **enfermedades cardiovasculares, la diabetes y el asma**).*

Para las enfermedades cardiovasculares se tiene 7 de las 9 metas relacionadas. La diabetes está relacionada con 6 de 9 metas. Para el caso del cáncer 3 de 9 metas están relacionadas a

esta enfermedad (considerando que dos de las metas solo menciona algunos tipos de cáncer), y para la obesidad hay 2 de 9 metas relacionadas. Evidenciado esto, se puede inferir que, dentro de las metas mundiales, no solo de la OMS, sino también de los países comprometidos con la salud, que hay una mayor responsabilidad para afrontar 2 de las 4 enfermedades crónicas más importantes, la diabetes y las enfermedades cardiovasculares, con un 77.78% y un 66.67% de representación de cada enfermedad en las metas respectivamente.

Estas metas solo se enfocan en un compromiso por los países para mejorar la situación frente a las enfermedades crónicas, hay algunos países que están haciendo progresos en este aspecto, pero la mayoría no está en camino de cumplir las metas mundiales.

Es así que por un lado tenemos el compromiso mundial y por el otro la realidad, el cumplimiento o el camino al cumplimiento de estas metas, la OMS indica que los países deben pasar del compromiso a la acción priorizando intervenciones de gran impacto y definir metas nacionales relativas a las enfermedades crónicas y hacerse responsables de su cumplimiento.

Es de suma importancia resaltar la alta incidencia de enfermedades crónicas en países en vías de desarrollo. El 79% de muertes atribuidas a enfermedades crónicas ocurren en estos países, siendo los factores principales aquellos correspondiente al comportamiento de la persona: uso de tabaco, dieta poco saludable y un alto índice de sedentarismo¹⁰

El Ministerio de Salud del Perú indica que, en el país, las enfermedades crónicas son la causa del 58,5% del total de muertes. La probabilidad de morir entre los 30 – 70 años por cualquiera de las 4 principales enfermedades crónicas es del 11%¹¹. Por su parte GFK, indica que el 61% de los peruanos se considera sedentario, siendo este un factor de incidencia para el desarrollo y/o evolución de una o más enfermedades crónicas¹². Asimismo, es necesario resaltar la alta deficiencia de la dieta peruana en cuanto a la contribución de vitaminas y nutrientes esenciales para mantener un adecuado estado de salud, tanto por rasgos propios de las costumbres alimenticias como por desconocimiento generalizado sobre una dieta balanceada¹³.

¹⁰ Cfr. OMS 2007 Prevención y control integral de enfermedades crónicas

¹¹ Cfr. MINSA 2012

¹² Cfr. Perú21 2013

¹³ Cfr. Datum Internacional 2013

Cifras de Enfermedades Crónicas

Enfermedades Cardiovasculares¹⁴:

- Son la principal causa de muerte en el mundo.
- En el 2012, murieron 17,5 millones de personas, representando el 31% de las muertes en el mundo. De estas muertes, 7,4 millones se debieron a cardiopatía coronaria y 6,7 millones a los accidentes vasculares cerebrales.
- Más de las tres cuartas partes de las muertes se produce en países de ingresos bajos y medios.
- La mayoría de las enfermedades cardiovasculares se podrían prevenir actuando sobre factores como el consumo de tabaco, las dietas malsanas, la obesidad, inactividad física o el consumo nocivo de alcohol.
- En el Perú, en el 2016, las muertes por enfermedades cardiovasculares representaron el 22%. ¹⁵
- Entre el 20%/30% de los casos de infartos se da en personas sin síntoma alguno (infarto silente o silencioso).¹⁶

Diabetes¹⁷:

- En el 2012, fallecieron 1,5 millones de personas como consecuencia directa de la diabetes.
- En el mundo hay 422 millones de personas con diabetes.
- Más del 80% de las muertes por diabetes se da en países de ingresos bajos y medios.¹⁸
- Según proyecciones de la OMS, en el 2030, la diabetes será la séptima causa de mortalidad en el mundo, con un crecimiento 2x de muertes entre 2005 – 2030.
- La prevención de la diabetes tipo 2 o el retraso de su aparición se puede dar mediante una dieta saludable, actividad física regular, manteniendo un peso corporal normal y evitando el consumo de tabaco.
- En el Perú, en el 2014, las muertes por diabetes representaron un 2% de la mortalidad total de enfermedades.

Cáncer¹⁹:

- En el 2012, hubo 14 millones de nuevos casos de cáncer y 8,2 millones de muertes relacionadas con el cáncer.
- Es una de las principales causas de morbilidad y mortalidad en todo el mundo.

¹⁴ Cfr. OMS 2016 Enfermedades Cardiovasculares

¹⁵ Cfr. OMS 2016 Perfiles de Países: Perú 2016

¹⁶ Cfr. Infobae 2016 Infarto asintomático

¹⁷ Cfr. OMS 2016 Diabetes

¹⁸ Cfr. OMS 2016 Diabetes: Día Mundial de la Salud 2016: Vence a la diabetes

¹⁹ Cfr. OMS 2016 Cáncer

- Se prevé que el número de nuevos casos se eleven en aproximadamente 70% en los próximos 20 años.
- En el 2012, en los hombres se diagnosticaron con más frecuencia los cánceres de pulmón, próstata, colon y recto, estómago e hígado. En las mujeres fueron los de mama, colon y recto, pulmón, cuello uterino y estómago.
- Aproximadamente un 30% de las muertes por cáncer se debe a cinco factores de riesgo conductuales y dietéticos, como el IMC elevado, ingesta reducida de frutas y vegetales, carencia de actividad física, consumo de tabaco y alcohol.
- Más del 60% de los nuevos casos anuales totales del mundo se producen en África, Asia, América Central y Sudamérica, estas regiones representan el 70% de las muertes por cáncer en el mundo.
- En el Perú, en el 2014, se tiene un 20% de mortalidad por cáncer.²⁰

Obesidad²¹:

- Desde el 1980, la obesidad se ha más que duplicado en el mundo.
- En 2013, más de 42 millones de niños menores de 5 años tenían sobrepeso.²²
- En 2014, más de 1900 millones personas mayores de 18 años tenían sobrepeso, de las cuales 600 millones eran obesos.
- El 39% de las personas adultas de 18 años o más tenían sobrepeso, y el 13% eran obesas.

Luego de haber realizado este análisis de cifras mundiales, podemos decir que las 4 enfermedades no transmisibles (ENT) son mortales, algunas más que otras, pero lo que se debe resaltar es que tanto la diabetes, las enfermedades cardiovasculares y la obesidad se pueden prevenir, en cuanto al cáncer se puede disminuir su impacto si se consideran los factores de riesgo, al igual que con las otras 3 enfermedades estos factores de riesgos previenen que se desarrollen dichas enfermedades.

Si bien se tienen porcentajes distintos entre la mortalidad de la diabetes (2%), cáncer (20%) y enfermedades cardiovasculares (22%) en el Perú, entre estas 3 se tiene una representación del 44%. Estas 3 enfermedades tienen como factor de riesgo la obesidad, y las 4 enfermedades analizadas concuerdan en tener como factor de riesgo dietas malsanas, índice de masa corporal elevada, e insuficiente actividad física.

²⁰ Cfr. OMS 2014 Perfil Peruano

²¹ Cfr. OMS 2016 Obesidad

²² Cfr. OMS 2016 10 datos sobre la obesidad

Ilustración 26: Mortalidad Proporcional de ENT

Fuente: Organización Mundial de la Salud (2014)

En cuanto a la mortalidad del cáncer y las enfermedades cardiovasculares (ECV) tienen un punto a favor representando un mayor número de muertes en relación con la diabetes.

Otro factor importante a analizar es la diversificación de las enfermedades, las ECV agrupan un conjunto de varias enfermedades, el cáncer es también una agrupación de enfermedades, la diabetes tiene solo tres variantes explicadas anteriormente y la obesidad la consideraremos como factor de riesgo clave para las tres enfermedades, que ahora en adelante son las que quedan en carrera en este análisis. Siendo el cáncer y las ECV conjuntos de enfermedades y la diabetes una enfermedad única con tres variantes, destacamos esto como punto a favor de la diabetes para la realización de un modelo predictivo.

La tasa de mortalidad por edades separada por sexo, muestran diferencias entre los hombres y mujeres en las ECV y el cáncer, caso contrario, la diabetes tiene una mínima diferencia entre las mortalidades por sexo, esto quiere decir, que, si bien en el cáncer y en las ECV existen diferencias marcadas en la manera que afecta una enfermedad a un sexo u al otro, en el caso de la diabetes, la enfermedad casi no hace diferencia entre sexo. Este aspecto nos es favorable para un modelo predictivo, siendo un punto a favor para la diabetes.

Ilustración 27: Tasa de Mortalidad de ENT

Fuente: Organización Mundial de la Salud (2014)

Si bien la tasa de mortalidad para los cánceres viene disminuyendo, en el caso de las ECV se muestra un escenario abrupto, con subidas y bajadas, en cuanto a la diabetes podemos observar que se viene manteniendo en el tiempo. Ahora mostraremos dos gráficas distintas, una para el cáncer y otra para la diabetes, ya que tienen la menor diferencia entre los sexos. En las gráficas del cáncer (evolución de la mortalidad) se encuentra abierto por tipo de cáncer y una gráfica por sexo, se puede observar una disminución gradual de la mortalidad.

Ilustración 28: Evolución de la mortalidad por cáncer

Fuente: Organización Mundial de la Salud (2014)

En la gráfica de la diabetes (tendencia en la diabetes) podemos observar que el porcentaje de población que cuenta con diabetes viene en aumento.

Ilustración 29: Tendencias en la diabetes

Fuente: Organización Mundial de la Salud (2014)

Por un lado, se tiene que la mortalidad en el cáncer viene disminuyendo en el Perú y por el otro vemos que la incidencia de diabetes en la población está en crecimiento. Este es otro punto a favor de la diabetes, por su crecimiento y por la no discriminación en cuanto a sexo.

Resultado del análisis

Es así que la diabetes, según el análisis realizado, tiene 3 puntos a favor frente a 1 para las enfermedades cardiovasculares. Tomando en consideración el hecho de que las ECV y el cáncer agrupan un conjunto de varias enfermedades, nuestro análisis apunta a que la diabetes

es la enfermedad crónica adecuada para ser escogida para el desarrollo de nuestro modelo debido a su gran trascendencia e impacto a nivel salud y económico en el país.

Esto viene reforzado con la proyección de la OMS, que indica que la incidencia de diabetes en el mundo se verá duplicada, dentro de los próximos 20 años. Asimismo, siendo la diabetes una enfermedad especialmente crítica para países en vías de desarrollo, se adecúa totalmente al contexto planteado para la realización del proyecto. Adicionalmente, la diabetes como enfermedad crónica es la que requiere de un menor número de factores ajenos a las costumbres propias para desarrollarla: es decir, mientras que las enfermedades cardiovasculares son en gran medida hereditarias o males congénitos o el cáncer requiere de una serie de factores externos para desarrollarse (por ejemplo, alto nivel de contaminación del aire), la diabetes requiere principalmente una serie de malas costumbres que resultan más cotidianas, tales como deficiencias de alimentación o de actividad física.

Finalmente, en base a lo anteriormente mencionado, la diabetes será nuestro enfoque para el desarrollo del modelo predictivo propuesto como resultado del proyecto.

BENCHMARKING DE HERRAMIENTAS DE PREDICTIVE ANALYTICS

Definición

El benchmarking de herramientas de predictive analytics es el segundo entregable desarrollado en la fase previa del diseño del modelo predictivo. El propósito del documento es la identificación de la herramienta de Predictive Analytics más adecuada para utilizar como soporte al desarrollo del modelo predictivo propuesto.

Objetivo

El benchmarking tiene como objetivo analizar las herramientas de predictive analytics disponibles en el mercado e identificar cuál es aquella en la que se deba basar el modelo predictivo, según su relevancia.

Alcance

Determinar la herramienta más adecuada para el desarrollo del modelo predictivo, considerando criterios de Forrester.

Contexto

Gartner ha definido para el año 2016 diez principales tendencias tecnológicas que intervendrán en los negocios de hoy en día. A través de la revisión de estas corrientes, identificamos dos grandes grupos de necesidades: la necesidad de integrar el ambiente real del ser humano con la faceta digital del mismo y el desear aprender de la información con la que contamos de manera ágil e innovadora. Esta urgencia por aprender sobre la información de la cual disponemos afronta diversos retos:

- Volumen hace referencia a la gran cantidad de información que se genera día a día a cada segundo. Considera la información que se genera en la web, en redes sociales como Facebook, Twitter, todos los mensajes, fotos, comentarios que contienen las redes sociales. Esto no solo representan Terabytes, ni Zetabytes de información, más bien representan Brontobytes ($1e^9$ Terabytes) de data. Este incremento de conjuntos de datos requiere de una gran cantidad de almacenamiento y un análisis tradicional de la data. Con la tecnología Big Data de las herramientas de analítica predictiva, se puede almacenar y analizar estos conjuntos de datos.
- Velocidad hace referencia a la rapidez en la que la data es generada y la rapidez en la que esta data se mueve. Pensemos en los mensajes virales que se esparcen por las redes sociales o los videos virales en YouTube o en la rapidez en la que la tarjeta de crédito es verificada para evitar un fraude. Las herramientas de Predictive Analytics tienen la capacidad de analizar esta información mientras es generada sin almacenarla en una base de datos.
- Variedad se refiere a los diferentes tipos de información que se puede utilizar. Antiguamente se ponía foco a data estructurada que se almacenaba en tablas o bases de datos relacionales, tal como la data financiera. Es un hecho, que el 80% de la data del mundo es data no estructurada, la cual no es sencilla de almacenar en tablas o bases de datos relacionales, un ejemplo de esa data son las fotos, los videos, mensajes en Facebook, comentarios de fotos, actualizaciones de estado. Las herramientas de Predictive Analytics cuentan con funcionalidades destinadas a la estructuración de información y clasificación de la misma, afín de incrementar la precisión de sus pronósticos y permitiendo el uso de data estructurada y no estructurada.
- Veracidad se refiere a al desorden o confiabilidad de la data. Con las diversas formas y fuentes de información, es menos controlable la calidad o exactitud, por ejemplo en Twitter se usan tweets, mensajes directos, #hashtags, abreviaciones, mensaje coloquial. La tecnología de analítica permite analizar estos tipos de data. Los volúmenes a menudo compensan la falta de calidad o exactitud de la información pero se debe tener suma cautela, debido a que los volúmenes pueden también distorsionar la calidad de la predicción.

- Valor se refiere a la habilidad de transformar la data en valor. Este quinto aspecto es el más esencial, ya que es un conjunto del Volumen, Velocidad, Variedad y Veracidad de la data. Sin la correcta gestión de estas características no se generara el valor esperado por la organización y los esfuerzos por impulsar tecnologías de Predictive Analytics en las organizaciones serán inservibles. El valor que puede generar la utilización de la tecnología es lo que genera un mayor impacto.

Bajo los retos anteriormente planteados, resulta necesario analizar que herramientas de Predictive Analytics nos ofrece actualmente el mercado basándonos en la investigación de las consultoras Gartner y Forrest Wave.

El reporte de Cuadrante Mágico de Plataformas de Analíticas Avanzadas está actualizado a febrero del 2016, en este informe se analizan 16 proveedores de plataformas de Predictive Analytics las cuales se categorizan en los 4 cuadrantes:

1. **SAS** – Q1
2. **IBM** – Q1
3. **KNIME** – Q1
4. **RapidMiner** – Q1
5. **Dell** – Q1
6. **SAP** – Q2
7. **Angoss** – Q2
8. **Microsoft** – Q3
9. **Alteryx** – Q3
10. **Predixion Software** – Q3
11. **Alpine Data** – Q3
12. **FICO** – Q4
13. **Lavastorm** – Q4
14. **Megaputer** – Q4
15. **Prognoz** – Q4
16. **Accenture** – Q4

La definición de los 4 cuadrantes mágicos de Gartner es la siguiente:

- **Líderes (Q1)** se desenvuelven bien frente a su actual visión y están bien posicionados para lo que vendrá mañana.

- **Desafiantes** (Q2) se desenvuelven bien en el presente o dominan un gran segmento de mercado, pero no demuestran un completo entendimiento de la dirección del mercado.
- **Visionarios** (Q3) entienden hacia donde apunta el mercado o tienen la visión de hacer cambiar las reglas del mercado, pero no se desenvuelven de la mejor manera.
- **Jugadores de nicho** (Q4) enfocados a pequeños segmentos, o están fuera de foco y no innovan o superan a otros.

Por otro lado, en el reporte de Forrester actualizado al Q2 del 2015, en este se evalúan 13 proveedores de plataformas de Predictive Analytics las cuales obtuvieron la siguiente categorización:

1. **SAS** – Líder
2. **IBM** – Líder
3. **SAP** – Líder
4. **Alteryx** – Fuerte Desempeño
5. **Angoss** – Fuerte Desempeño
6. **Dell** – Fuerte Desempeño
7. **FICO** – Fuerte Desempeño
8. **KNIME** – Fuerte Desempeño
9. Oracle – Fuerte Desempeño
10. **RapidMiner** – Fuerte Desempeño
11. **Predixion Software** – Contendiente
12. **Microsoft** – Contendiente

El resumen de las 3 categorías es la siguiente:

- **Líderes.** IBM y SAS tienen amplitud y profundidad sin igual en sus soluciones. Ambos tienen productos extremadamente maduros y ninguna se ha quedado de los brazos cruzados. Ambos tienen altas puntuaciones en cada categoría. SAP por su lado es un líder en crecimiento con miras a enfrentar a IBM y SAS con una continua y agresiva inversión en sus capacidades de analítica predictiva.
- **Fuertes intérpretes.** Alpine Data Labs, Alteryx, Angoss, Dell, FICO, KNIME, Oracle y RapidMiner son fuertes intérpretes. Todos ellos tienen algo que los convierte en las mejores soluciones para las empresas. Con mejores calificaciones estratégicas se tiene a Alteryx, Angoss, FICO, Oracle y RapidMiner, quienes habrían podido ser líderes.
- **Contendientes.** Microsoft y Predixion Software son contendientes. Ambos empezaron de un nicho pero tienen mucho espacio por crecer y un valor único para las empresas. Microsoft es puro Cloud y Predixion Software empodera a Excel para realizar sus capacidades de analítica predictiva en Cloud.

Los Criterios de Evaluación de Forrester está agrupada en 3 categorías las cuales en su totalidad suman 45 criterios, las 3 categorías de alto nivel son las siguientes:

- **Oferta actual.** Se evaluó la arquitectura de cada solución, la seguridad, adquisición de datos y problemas con los datos, preparación de los data sets, algoritmos soportados y los métodos, capacidades de evaluación, usabilidad de la herramienta, aplicación en el negocio, y otras características para establecer las capacidades de la oferta actual del vendedor. Todos los productos evaluados deben haber estado a disposición del público antes del 1 de octubre del 2014.
- **Estrategia.** Revisaron la estrategia de cada vendedor para evaluar como sus planes cumplen con la demanda de los clientes. Los criterios más importantes en esta categoría son la adquisición y el precio de las soluciones, el soporte para la implementación del vendedor, el roadmap y la estrategia de salida al mercado.
- **Presencia en el mercado.** En este grupo se encuentran entre los criterios más importantes las finanzas del vendedor, sus ganancias globales, industrias atendidas, debilidad del mercado, tecnología, socios de servicios, todos estos criterios sirven para evaluar el peso de la presencia del vendedor en el mercado.

Herramientas y Criterios de Evaluación

Habiendo recabado esta información hemos decidido por hacer un Benchmarking con 12 proveedores de plataformas de Predictive Analytics, ya que haciendo el cruce entre las investigaciones de Gartner y Forrester 12 de los proveedores coinciden, siendo en el caso de Gartner estos 12 proveedores como los 12 mejores ubicados en los cuadrantes (análisis a febrero del 2016) y en el caso de Forrester son casi la totalidad de proveedores evaluados (análisis a Q2 del 2015). Esto evidencia que se tienen casi la misma cantidad de proveedores significativos en el mercado, mejorando en ciertos aspectos para destacar del resto. Consideramos este punto válido para el inicio de nuestro benchmarking. Por consiguiente, las 12 herramientas a analizar serán:

1. **Angoss** – Angoss Predictive Analytics
2. **Alpine Data Labs** – Alpine Chorus
3. **Alteryx** – Alteryx Analytics
4. **Dell** – Statistica
5. **FICO** – FICO Predictive Analytics
6. **IBM** – IBM Predictive Analytics
7. **KNIME** – KNIME Analytics Platform
8. **Microsoft** – Microsoft Azure Machine Learning
9. **Predixion Software** – Predixion Insight
10. **RapidMiner** – RapidMiner Studio
11. **SAP** – SAP Predictive Analytics
12. **SAS** – SAS Enterprise Miner

Puesto que nuestro análisis no requiere de criterios tan técnicos, sino más bien un análisis de los puntos claves que debe tener la herramienta, el estado actual del vendedor dueño de la herramienta, la estrategia, la presencia en el mercado, hemos decidido usar los 15 mejores criterios propuestos por Forrester en su momento. Los cuales están agrupados en 3 categorías como se muestra a continuación:

- ✓ Oferta actual
 1. Arquitectura
 2. Seguridad
 3. Data
 4. Análisis
 5. Gestión del modelo
 6. Usabilidad de la herramienta
 7. Aplicaciones de negocio
- ✓ Estrategia
 8. Adquisición y precio
 9. Capacidad de ejecución
 10. Soporte de implementación
 11. Roadmap
 12. Ratio de salida al mercado
- ✓ Situación del vendedor
 13. Situación Financiera
 14. Base de Clientes
 15. Socios de Negocio

En cada uno de los criterios se utilizará una calificación de 0 – 5, siendo:

- ✓ 0 – Cumplimiento débil
- ✓ 1 – Cumplimiento parcial
- ✓ 2 – Cumplimiento normal
- ✓ 3 – Cumplimiento por encima del normal
- ✓ 4 – Cumplimiento sobresaliente
- ✓ 5 – Cumplimiento excelente

Análisis de las herramientas

Angoss – Angoss Predictive Analytics

Angoss es una empresa con Sede Principal en Toronto, Canadá, con oficinas en Estados Unidos, Reino Unido y Asia.

El portafolio Angoss provee de una suite de escritorio, cliente servidor y software de Big Data Analytics y soluciones Cloud, entre otros. La Suite de Predictive Analytics incluye

KnowledgeSEEKER, KnowledgeSTUDIO y KnowledgeREADER, todo manejado desde su cliente Cloud KnowledgeCLOUD.

KnowledgeSEEKER tiene capacidades para Data Mining y Predictive Analytics, entre sus funciones incluye importación/exportación de data, preparación de la data, perfilado de data, transformación de data, particionamiento de la data, muestreo de árboles de decisión, árboles de estrategia, análisis market basket, soporte para R, evaluación del modelo, despliegue y automatización del flujo de trabajo. La plataforma soporta el ciclo de vida completo del desarrollo. Los árboles de decisiones son altamente gráficos y son amoldables. La plataforma cuenta también con Árboles de Estrategia únicos, los cuales integran reglas de negocio con modelos predictivos en un solo gráfico de árbol.

La plataforma mejora la eficiencia del Data Mining en un 40% y minimiza la curva de aprendizaje, lo cual permite construir más modelos en un tiempo menor de codificación. Mejora la transparencia de la información, mostrándola en cada paso del proceso de Data Mining, realizando una rápida evaluación de la calidad de la información, detectando patrones y tendencias, y desempeña una variable de reducción. La plataforma se puede integrar a un software de otro proveedor para brindar un mejor análisis visual.

Angoss es reconocida por reconocidos analistas de la industria como la mejor herramienta de Árboles de decisiones, sus árboles de decisiones son amigables para los usuarios, permiten un crecimiento manual y automático, y muestran los resultados visualmente para una mejor interpretación. La plataforma puede conectarse a distintas bases de datos y permite el uso de distintos formatos, se pueden integrar fácilmente con otros ambientes analíticos permitiendo la importación y exportación de data a o desde Texto, Excel, SAS, SPSS, R, Hadoop, y a bases de datos vía ODBC.

Las capacidades de analítica predictiva como la preparación de la data, perfilado, visualización, árboles de decisiones y diseño de la estrategia y de despliegue, permiten a los usuarios transformar los insights (conocimiento predicho) en ideas reales.

KnowledgeSTUDIO permite el modelado y desarrollo de cuadro de mandos, automatización del flujo de trabajo, regresión lineal y logística, árboles de decisión, redes neuronales, principales componentes de análisis, también cuenta con técnicas de aprendizaje sin supervisión como análisis de cluster, análisis market basket (reglas de asociación), cuadros de mando y provee soporte para R y modelos de bases de datos en ejecución. La herramienta también provee de todas las capacidades de avanzadas de modelos de predicción. Automatiza la tediosa tarea de coarse classing (binning) con peso flexible e inteligente del

optimizador de evidencia (WOE) y reduce el tiempo dedicado a esta tarea manual de hasta 50%.

Con los reconocidos árboles de decisiones permiten un crecimiento manual y automático, muestra los resultados para una fácil interpretación. El conjunto de árboles, construidos con Random Forest, bagging, y algoritmos de impulso, mejoran aún más la experiencia del usuario permitiendo la creación de múltiples árboles de una sola vez y combinándolas en un solo conjunto modelo de árboles.

Al igual que KnowledgeSEEKER, la plataforma KnowledgeSTUDIO mejora la eficiencia del Data Mining en un 40% y minimiza la curva de aprendizaje, lo cual permite construir más modelos en un tiempo menor de codificación. También permite la interconexión con distintas bases de datos y permite la importación y exportación de data a o desde Texto, Excel, SAS, SPSS, R, Hadoop, y a bases de datos vía ODBC.

KnowledgeREADER es una plataforma para el análisis de texto ideal para el análisis temático y de sentimiento. Es compatible con los árboles de decisión y estrategia, y se puede emplear para el análisis de tendencias, descubrimiento de asociación, exploración de documentos y otras formas de análisis de texto. Enriquece el proceso de descubrimiento de texto con una robusta funcionalidad de minado que provee acceso al Procesado Natural de Lenguaje (NLP), análisis de sentimiento, dashboard de descubrimiento de texto, mapas de descubrimiento de asociación, clasificación y extracción de temática y entidad.

La plataforma tiene el mejor NLP proporcionando los insights de datos más significativos al descomponer fragmentos de texto en elementos gramaticales para interpretar la vaguedad de la expresión del cliente.

Fuente: Butler Analytics

Alpine Data Labs – Alpine Chorus

Alpine Data provee de una plataforma de analítica que permite a los científicos de datos y a los usuarios colaborar con analítica avanzada a través de una simple interfaz web. Alpine Chorus es instalado sobre un ligero web server y se conecta directamente cualquier fuente de información (base de datos o Hadoop) por lo que no requiere de movimiento de data para analizar la data.

Alpine Chorus es plataforma comprensiva de Analítica Avanzada que provee el completo ciclo de vida en un solo ambiente, y permite construir, desplegar, y consumir aplicaciones analíticas e insights de una manera ágil y colaborativa. Entre sus características están la prueba de hipótesis, modelado predictivo usando funcionalidades estadísticas incluidos los análisis Serie-Tiempo, Clasificación, regresión, árboles de decisiones y otros, la transformación de datos, creación y construcción de modelos, también repara y apalanca data, modelos, y asegura los resultados para evitar los silos de datos y provee un flujo de trabajo que cubre de inicio a fin la extracción, transformación, modelado y evaluación.

Alpine Chorus cuenta con una interfaz visual de Drag and Drop que permite a los usuarios y a los científicos de datos analizar su información sin tener que realizar consultas SQL o escribiendo en MapReduce o en R. La interface contiene operadores específicos para funciones como reemplazo o filtrado, permitiendo crear complejos flujos de trabajo para limpiar, mezclar, transformar y preparar su proceso de data. La plataforma está diseñada para operar en ambiente Hadoop, Greenplum, Oracle, PostgreSQL y bases de dato Netezza.

Una de sus características resaltantes es la capacidad para analizar data donde sea que esta se ubique remotamente, permitiendo a los usuarios crear cajas de arena de provisión, y crear los modelos sin necesidad de mover su información. La plataforma indexa dinámicamente la metadata de cada proyecto, creando un diccionario viviente que es accesible por cualquier Stakeholder en el proceso. Si se tiene algún percance y se necesita ayuda, Alpine Data brinda soporte bajo demanda.

Los algoritmos de Alpine Chorus están optimizados para impulsar la analítica desde cualquier fuente de análisis, cuando se ejecutan flujos de trabajo, el Motor de Ejecución Distribuido de Alpine Chorus almacena el flujo de trabajo lógico y envía instrucciones a múltiples sistemas de bases de datos automáticamente. Esta capacidad permite a los analistas y científicos ejecutar algoritmos a escala sin mover su data u optimizar sus algoritmos basado en la lógica de su base de datos.

La herramienta ha sido diseñada para utilizar ampliamente clusters de cómputo en paralelo para ejecutar analíticas eficientemente. El flujo de trabajo basado en Hadoop o en base de datos, tomando ventaja de MapReduce, Spark o SQL según convenga. El Motor de Ejecución Distribuido impulsa el código en la plataforma de datos en paralelo a través de todo el cluster.

La plataforma contiene también, una colección comprensiva de minería de datos predictivos, y algoritmos de modelado que enriquece el negocio para manipular, modelar, y apalancar Big Data en el ciclo de vida del negocio.

Los usuarios pueden compartir sus espacios de trabajo o cajas de aprovisionamiento con cualquier Stakeholder. Cada recurso involucrado al proyecto puede ser adjuntado al espacio de trabajo incluido los flujos de trabajo, documentos de definición de datos, y planes de proyecto. El flujo de trabajo cuenta con un historial que provee un resumen cronológico de cada versión del flujo de trabajo dentro de cualquier espacio de trabajo, permitiendo el manejo de versiones, dicho de otra manera el usuario puede revertir versiones anteriores de flujo de trabajo.

La plataforma también permite colaboración entre el equipo de trabajo, los Stakeholders pueden interactuar con la data y hacer comentarios en el estado del proyecto. Permitiendo también ver el estado actual del proyecto. Los usuarios pueden desarrollar modelos predictivos y ponerlos en producción rápidamente vía PMML dentro de la plataforma. Los flujos de trabajo y scripts que son desarrollados en R y SQL son almacenados para uso futuro. La plataforma también ofrece una API para conectarse con otras plataformas.

Ilustración 31: Arquitectura de Plataforma Alpine Data

Fuente: Alpine Data

Alteryx – Alteryx Analytics

Alteryx Analytics incluye en su portafolio: Alteryx Designer, Alteryx Server y Alteryx Analytics Gallery. Alteryx Designer permite mezclar data interna, de terceros y basada en la nube, construir poderosas aplicaciones predictivas y analíticas espaciales basadas en R sin ninguna codificación, compartir la data de insights con los tomadores de decisiones. Tiene las técnicas de modelado predictivo, como regresión lógica, árboles de decisiones, técnicas de clustering, como K-centroid clustering y análisis de componentes principales, técnicas de investigación de data, como scatter plots y análisis de asociación, todas las técnicas pueden ser incluidas sin la necesidad de programar. Alteryx Server, escala los flujos de trabajo de análisis crítico para cumplir con los requisitos de datos y análisis, programar múltiples flujos de trabajo, y publicar y compartir aplicaciones analíticas de la organización. Alteryx Analytics Gallery deja consumir, compartir y publicar aplicaciones analíticas y macros.

Alteryx Designer cuenta con 4 funciones especiales: Preparar y mezclar la data, Analítica Predictiva, Analítica Espacial y Compartir Insights. Alteryx toma un enfoque distinto para preparar y mezclar la data desde cualquier fuente de información y lo hace a través de una interfaz de Drag and Drop intuitiva. Alteryx Designer conecta y limpia los datos de Data Warehouses, aplicaciones Cloud, hojas de cálculo y otras fuentes. Crea el conjunto de datos correcto para el análisis y visualización usando herramientas de calidad de datos, integración y transformación. Mezcla y brinda archivos de data espacial y luego se unen fácilmente con datos de terceros, como datos demográficos. Acelera el proceso de crear analítica predictiva con herramientas especialmente diseñadas. Alteryx hace la analítica predictiva accesible a

cualquier analista brindando un flujo de trabajo repetible que ofrece capacidades de análisis de datos de autoservicio necesarios para el análisis predictivo, junto con herramientas de Drag and Drop para poder crear un modelo sin la necesidad de programar. Con más de 30 herramientas de Drag and Drop todas basadas en el lenguaje R y personalizables para cumplir con necesidades específicas. Se elimina el tradicional reporte estático, brindando visualización interactiva para validar los resultados del modelo. Se puede integrar codificación directa de R al flujo de trabajo y ser compartida con los analistas por un científico de datos. Alteryx Designer permite trabajar con data de geolocalizaciones y procesarla desde archivos planos o ESRI hacia Big Data, las herramientas espaciales permiten el análisis a través de técnicas como drive time, trade área, spatial matching y creation. Los insights pueden ser compartidos con aplicaciones de terceros como herramientas de mapeo o visualización. La plataforma permite compartir los insights en una variedad de formatos como Excel, ESRI, XML, PDF. También permite la creación de reportes personalizados que pueden ser actualizados cuando se requiera o programados. Reduce el tiempo necesario para crear el conjunto de datos perfecto y añadir analíticas avanzadas para la visualización. Permite expandir el valor de los flujos de trabajo añadiendo una simple interfaz e interconexión que puede ser consumida por los ambientes de Alteryx Server o Alteryx Analytics Gallery.

Alteryx Server acelera el tiempo de los insights analíticos y faculta a los analistas y usuarios de negocios en toda la organización para tomar las mejores decisiones, más informadas, basadas en datos. Al proporcionar una plataforma escalable para desplegar y compartir análisis, permite a los equipos colaborar en decisiones críticas de negocio rápida y sencillamente. Alteryx Server brinda una arquitectura flexible que puede ser desplegada bajo demanda o en Cloud así que se puede escalar hacia arriba o abajo dependiendo de la necesidad del negocio. Utiliza capacidades de programación y automatización para programar o ejecutar los flujos de trabajo analíticos, para generar reportes y actualizar los conjuntos de datos en un servidor centralizado para que todos puedan acceder a los datos en cualquier momento y lugar. También ejecutar los flujos de trabajo simultáneamente aplicando procesos de tolerancia a fallas y de seguridad.

Permite el control de versiones en un repositorio centralizado con una colaboración eficiente y segura.

Alteryx Analytics Gallery es la primera plataforma de analítica Cloud que reconoce que los tomadores de decisiones esperan que sus aplicaciones sean atractivas como las de casa. Es por ello que Analytics Gallery es una fuerza disruptiva, brindando una poderosa experiencia

analítica basada en el consumidor que permite a las organizaciones llegar al valor de Big Data exponencialmente más rápido. Sus dos principales funciones son, Publicar Apps Analítica y Personalizar Apps Analíticas.

Permite guardar los flujos de trabajo en un repositorio centralizado con control de versiones para una colaboración efectiva y segura entre los analistas de datos, publicar la data, macros y flujo de trabajo analítico en un estudio privado, dar y revocar acceso a aplicaciones, macros o a colecciones enteras por usuario.

Alteryx Analytics Gallery cuenta con una galería de aplicaciones analíticas en la cual se puede navegar para buscar la que cumpla las necesidades de la organización. Ejecutar estas aplicaciones y crear apps analíticas para los dispositivos de la organización o para servicios Cloud.

Dell – Statistica

Statistica es un software propietario desarrollado por StatSoft, que fue comprada por Dell. Statistica brinda soluciones para análisis de datos, gestión de datos, visualización de datos y minería de datos con modelado predictivo, clustering, clasificación y técnicas de exploración en una sola plataforma.

Statistica provee de una plataforma de minería de datos/textos y de analítica predictiva robusta y comprensiva que puede ofrecer soluciones que van desde la detección del fraude en los seguros y/o banca hasta la optimización de procesos en el sector manufactura.

La plataforma cuenta con 5 tecnologías, Statistica Decisioning Platform, Statistica Enterprise, Statistica Data Miner, Statistica Text Miner, Statistica Lives Score.

Statistica permite el descubrimiento y visualización de la data en tiempo real, permite a los analistas, trabajadores del conocimiento, administradores y ejecutivos colaborar y explorar la información a través de dashboards interactivos. Statistica optimiza el desempeño adoptando matemática a los datos, mejorando su rendimiento donde sea, permite crear y conseguir modelos predictivos en la base de datos, como en SQL Server o directamente dentro de Hadoop. Brinda manuales y templates fáciles de usar para la minería de datos, análisis predictivo, aprendizaje automático, la previsión y la minería de texto en Big Data. Combina tecnologías como Hadoop para el desempeño y escalabilidad, con búsquedas Lucene/SOLR, Mahout machine learning y procesadores avanzados de lenguaje natural. Permite la analítica avanzada a cualquier dato, sea estructurado, no estructurado o semi estructurado y desde cualquier fuente de datos, como Cloud, bases de datos SQL o no SQL. El despliegue o implementación de Statistica es agilizado permitiendo el soporte para Cloud, en el cual se puede realizar el despliegue en minutos, siendo costo-efectivo, totalmente

integrado y configurable. Usa el estándar de la industria de lenguajes script, estando totalmente integrado con R de manera nativa. Permite el despliegue de modelos con código PMML, así como SAS, Java, Teradata y SQL. Reduce la carga en TI, ofreciendo una arquitectura agnóstica, flexible, adaptable e integrable que complementa las inversiones existentes.

Permite un fácil uso a los usuarios de negocio y a los analistas. Agiliza la mezcla de la data, el descubrimiento interactivo, la visualización y el despliegue. Brinda una plataforma completa que soporta el ciclo de vida de la analítica, incluyendo gestión del modelo, combinando reglas de negocio y modelos analíticos para su calificación.

Proporciona una recuperación de la información con un motor rico en búsqueda y visualización de la data combinando análisis y búsqueda en un entorno unificado. Mejora la agilidad operativa y brinda una ventaja competitiva con capacidades de extracción y análisis fuera de la caja, capacidades que descubren inteligencia procesable en grandes cantidades de información variada. Permite la creación de flujos de trabajo de procesamiento de contenido con la función de Drag-and-Drop. Proporciona sugerencias automáticas basadas en metadatos extraídos.

Statistica Decisioning Platform ayuda a las organizaciones a tomar decisiones más efectivas usando analítica predictiva. Mediante la aplicación de análisis predictivo para determinar los patrones de los datos históricos, una empresa comercial puede afinar mejor y lograr sus objetivos para la retención de clientes, captación de clientes, rendimiento de los empleados, la disminución del riesgo, y una mayor rentabilidad. Cuenta con 7 atributos clave:

1. Reglas de decisión: gestión y ejecución de reglas de negocio basadas en el contexto del negocio, regulaciones, interpretación de los resultados del análisis predictivo y qué hacer con ese resultado.
2. Modelado Predictivo: utilización de los datos históricos de la organización para discriminar, segmentar, clusterizar y pronosticar efectivamente usando las últimas técnicas en Statistica Data Miner.
3. Gestión del modelo: el despliegue, gestión y monitoreo efectivo de modelos predictivos mediante Statistica Enterprise Server.
4. Minería de texto: utilización de data no estructurada con datos numéricos.
5. Scoring Server batch y ejecución en tiempo real: emplear modelos predictivos en batch o en tiempo real usando la escalabilidad de Statistica Live Score.
6. Arquitectura abierta y automatización: capacidad de integrar sistemas existentes, usando estándares de la industria, con la API de Statistica.
7. Visualización de datos: entendimiento de lo que hacen los modelos predictivos y por qué lo hacen a través de un monitor gráfico.

Statistica Enterprise Server es altamente escalable y de clase empresarial, completamente habilitado para el análisis de datos de la web y de bases de datos, basándose en la tecnología de procesamiento distribuido y soporta configuraciones de arquitectura multi capa de Cliente-Servidor.

Dentro de la suite brindada por Statistica Data Mining tenemos los siguientes algoritmos:

Redes bayesianas, Random Forest, Boosted Trees, Support Vector Machines, K-nearest Neighbors, Multivariate Adaptive Regression Splines, clasificación, árboles de regresión, entre otros.

Se proporcionan métodos para la evaluación competitiva automática de modelos, para calcular la predicción media a través de modelos. Todos los procedimientos en el software Statistica pueden ser codificados vía Statistica Visual Basic, el cual permite el lenguaje R y también se pueden incluir otros lenguajes de programación.

FICO – FICO Predictive Analytics

FICO Big Data Analyzer es un entorno de análisis especialmente diseñado para usuarios de negocios, analistas y científicos de datos para obtener información valiosa a partir de la exploración y el análisis de cualquier tipo y tamaño de los datos de Hadoop. Esto hace que grandes volúmenes de datos accesibles sean enmascarados en la complejidad de Hadoop, lo que permite a todos los usuarios a manejar con más valor de negocio con cualquier data. Mejora la productividad y el tiempo para los insights. Al hacer Big Data más accesible a los analistas de negocios y científicos de datos, que les permite resolver problemas de negocios que presiona de forma rápida y colaborativa. El Analytics Hub permite a los usuarios almacenar de forma segura, buscar, volver a usar y manejar los activos de análisis creados en cualquier parte de la organización. Esto elimina la necesidad de replicación de datos, reduciendo el coste total de propiedad Hadoop. Combinado con FICO Analytic Modeler y otros componentes de la suite de gestión de decisiones FICO, FICO Big Data Analyzer está diseñada para el ciclo de vida entero del Big Data Analytics, que abarca la exploración de datos, construcción de modelos y la aplicación de decisiones operativas.

Provee de una interfaz intuitiva que soporta queries para el análisis de datos y brinda una interfaz gráfica para los analistas del negocio. Permite a las organizaciones agregar Big Data preservando la integridad de los datos. El Analytics Hub permite a los usuarios busca, gestión y reusar activos analíticos. Es compatible con la mayoría de versiones y distribuciones de Hadoop, incluso con Cloudera y HortonWorks. Acceso rápido y listo a queries avanzados, incluyendo más de 250 pre cargados para analítica. Los modelos existentes construidos en

prácticamente cualquier herramienta pueden ser importados para uso colaborativo de Big Data en Hadoop. La solución es integrada nativamente con Hadoop.

FICO Model Builder es compatible con el proceso completo de desarrollo del modelo, permitiendo a los usuarios acceder, explorar y analizar los datos, calcular las variables, documentar los pasos clave de modelado, y construir, evaluar y desplegar modelos. Empleado en conjunto con otros programas, tales como el FICO FICO Blaze Advisor. FICO Model Builder ayuda a las organizaciones a adaptarse rápidamente modelos predictivos y reglas de negocio a cambios rápidos en las condiciones.

Es compatible con todo el ciclo de vida de modelado, desde la exploración de datos y de la construcción de modelos hasta su despliegue. Contiene una amplia variedad de algoritmos, incluyendo árboles de decisión y modelos segmentados Ensemble, que aprenden a reconocer patrones complejos dentro de los datos valiosos y masivos. La automatización inteligente acelera el modelado, mientras que permite a los analistas asumir el control y refinar los modelos para satisfacer sus necesidades de toma de decisiones de negocio. Permite una fácil colaboración usando lógica consistente, metodologías comunes, librerías reusables y plantillas, y un modelo robusto de documentación. La interfaz de usuario está basada en un enfoque hacia el usuario para facilitar el compromiso con los miembros del equipo. Ofrece integración con plataformas analíticas para usar las herramientas en otro nivel, integración con Hadoop, MapReduce, SAS y R.

FICO Blaze Advisory permite trabajar en colaboración mediante el intercambio y la reutilización de normas, conjuntos de reglas, flujos de reglas y modelos de objetos. Fácil de escribir, editar y pruebas de reglas utilizando un entorno de desarrollo integrado.

La FICO Decision Management Suite, junto con el FICO Analytic Cloud, ofrecen un entorno completo que hace que sea rápido y fácil de espigar puntos de vista de los datos, y desarrollar modelos analíticos y servicios de decisión que hacen operativos esos puntos de vista. Integra sin problemas de análisis, decisiones, optimización, visualización de datos y exploración, desarrollo rápido de aplicaciones y otras capacidades para proporcionar una solución completa y ágil de toma de decisiones.

Ilustración 32: Arquitectura de Plataforma FICO

Fuente: FICO

IBM – IBM Predictive Analytics

IBM ofrece una manera fácil de usar productos y soluciones de análisis predictivo, que cumplen con las necesidades de diferentes usuarios y nivel de experiencia desde analistas principiantes hasta avanzados. IBM Predictive Analytics ayuda al negocio a transformar la data en insights predictivos para guiar la toma de decisiones e interacciones. El portafolio de IBM Predictive Analytics incluye IBM SPSS Modeler, IBM SPSS Analytic Server, IBM SPSS Statistics, SPSS Predictive Analytics Enterprise, IBM Analytical Decision Management, IBM SPSS Data Collection e IBM Social Media Analytics.

IBM SPSS Modeler es una herramienta de analítica predictiva con un rango de algoritmos y técnicas avanzados que incluyen análisis de texto, análisis de entidades, gestión y optimización de decisiones que es diseñado para brindar inteligencia predictiva a las decisiones hechas por los individuos, grupos, sistemas y por la organización. Esta plataforma está diseñada para mejorar las decisiones y resultados, ayudar a extraer el valor de los datos. Integrarse con los sistemas existentes. Automatiza y optimiza decisiones transaccionales al

combinar el análisis predictivo, reglas y calificación para dar recomendaciones en tiempo real. Usa una gran variedad de enfoques de modelado de una sola corrida y compara los resultados de los diferentes métodos de modelados. La plataforma puede analizar cualquier tipo de información estructurada, no estructurada. También realiza análisis de las relaciones en las redes sociales entre las entidades sociales y la implicación de estas relaciones con el comportamiento de individuos, pudiendo construir modelos predictivos en base al comportamiento de grupos de personas. Entre los algoritmos con los que cuenta el IBM SPSS Modeler están:

- Detección anómala
- A priori
- Redes Bayesianas
- Árboles de decisiones, árboles interactivos
- Regresión Cox
- Lista de decisiones
- Support Vector Machine
- KNN
- Regresión logística
- Redes Neuronales
- Otros

IBM SPSS Analytic Server permite que la plataforma de análisis predictivo de IBM utilice datos procedentes de distribuciones Hadoop para mejorar las decisiones y los resultados con el uso de IBM SPSS Modeler. La plataforma provee de una arquitectura abierta, integrada y centrada en Big Data. Compatibilidad con distribuciones de IBM, Cloudera, Hortonworks y Apache Hadoop. Interfaz interactiva que permite centrarse en el análisis de datos, también permite la incorporación de nuevos algoritmos estadísticos para obtener datos. La herramienta es escalable.

IBM SPSS Statistics es una solución estadística integrada que se centra en el proceso analítico. La plataforma ayuda a entender la data, identificar tendencias y producir estimaciones precisas. Ayuda rápidamente a entender y predice desde conjuntos de datos en cualquier formato usando procedimientos estadísticos avanzados, asegurando alta precisión para guiar un proceso de toma de decisiones de calidad. Incrementa el poder analítico, productividad y flexibilidad a través de opciones de programación como comandos y lenguajes externos de programación como R, Python. Aborda el análisis no estándar y cuestiones como la calidad de los datos y la automatización con las características tales como la preparación automática de datos, árboles de decisión y la previsión. Soporta negocios empresariales que necesitan técnicas avanzadas tales como el modelado de ecuaciones

estructurales (SEM), la evaluación de muestreo y prueba, y una pequeña muestra y análisis de rara ocurrencia.

IBM SPSS Predictive Analytics Enterprise presenta una amplia y profundo análisis descriptivo y predictivo, preparación y automatización de datos y brinda analíticas de datos estructurados y no estructurados de cualquier fuente. Incluye una infraestructura de despliegue para integrar la inteligencia predictiva con las reglas de negocio y optimizar las decisiones operativas.

Ilustración 33: Proceso de Analítica Predictiva

Fuente: IBM

KNIME – KNIME Analytics Platform

KNIME Analytics Platform es una plataforma de código abierto enfocada al análisis de datos para descubrir potencial en los mismos, trabajarlos para obtener vistas de las tendencias actuales o predecir tendencias futuras. Al ser una herramienta de código abierto, es una elección sumamente popular en los círculos académicos enfocados al uso de analítica, lo cual a su vez genera que se tenga una comunidad masiva de profesionales de información que permiten satisfacer las dudas que se puedan presentar.

El código de la herramienta está escrito en Java y basado en Eclipse, aprovechando sus capacidades a través de la disponibilidad de plug-ins que soportan la integración con

métodos de minería de texto, minería de imágenes y análisis de series de tiempo. Asimismo, al ser una herramienta de código abierto, se encuentra totalmente abierta a la integración con otros proyectos de código abierto, demostrado por la inclusión de algoritmos de machine learning de Weka, R y JFreeChart.

Dentro de las principales funcionalidades de la herramienta, destaca la capacidad de agrupación de datos: la herramienta soporta el agrupar datos de texto simples, bases de datos, documentos, imágenes, redes e inclusive data basada en Hadoop dentro de un mismo flujo de trabajo, aumentando así las posibilidades de información con la que puede trabajar. Asimismo, destaca la capacidad de agrupación de herramientas: KNIME Analytics Platform soporta scripts de los siguientes 4 lenguajes: Python, R, SQL y Java con el propósito que los usuarios puedan reusar el expertise de otros profesionales en el área de ciencias de datos e incentivando el conocimiento colectivo en el ámbito científico de información. Además, ofrece una interfaz gráfica sencilla que brinda la opción de trabajar bajo un ambiente de desarrollador (con acceso a código) o bajo un enfoque netamente amigable y visual.

La herramienta provee una combinación de acceso a datos, transformación de datos, investigación, análisis predictivo y visualización del conocimiento creado, soportando también la automatización del procesamiento predictivo para que se integre a sistemas en producción de la organización que use la herramienta.

KNIME Analytics Platform incorpora cientos de nodos de procesamiento de data de entrada y salida, soportando además las fases de pre procesamiento y limpieza, modelado, análisis y minería de datos además de vistas interactivas como graficas de dispersión y coordinadas paralelas. Además, emplea algoritmos de minería de datos que se encuentran alineados al estado del arte de dicho ámbito, tales como clustering, inducción de reglas, árbol de decisiones, reglas de asociaciones, algoritmo de naive bayes, redes neuronales y soporte de máquinas de vectores para lograr un mejor entendimiento de la información de la que se dispone.

Adicionalmente a las funcionalidades que ofrece KNIME Analytics Platform, el proveedor también cuenta con los siguientes productos comerciales que complementan a la solución principal:

- KNIME Personal Productivity: Se trata de un producto que provee una manera eficiente para construir y mantener flujos de trabajo de KNIME, permitiendo el reuso y gestión de porciones de código y meta nodos.
- KNIME Partner Productivity: Provee a las organizaciones que brindan servicios de consultoría la habilidad de encriptar y cifrar meta nodos encapsulados para que se puedan compartir con clientes y a su vez protegiendo su propiedad intelectual, siendo

esto clave para una herramienta de código abierto donde la delimitación de lo propio y lo público resulta difuso.

- KNIME Team Space: Mejora los esfuerzos colaborativos del equipo al proveer un ambiente de almacenaje de flujos de datos y flujos de trabajo de análisis con la capacidad de ser compartido y trabajado colaborativamente por miembros del equipo.
- KNIME Server Lite: Permite la gestión avanzada de las capacidades colaborativas de la herramienta, como por ejemplo autenticación de usuarios, derechos de usuarios, ejecución programada remota, generación de reportes, espacio de datos comparativo, acceso a los repositorios de flujos de trabajo y meta nodos.
- KNIME Server: Ofrece mejoras a la extensión Server Lite, con un mayor enfoque en el aspecto de autenticación de usuario y accesos de usuario, incluyendo también soporte a Web Services, versionamiento de flujos de trabajo y soporte comercial.
- KNIME Big Data Extension: Provee nodos para acceder a información guardada en Sistemas de Archivos Distribuidos de Hadoop.
- KNIME Cluster Execution: Provee una capa de conexión entre KNIME y un cluster para mejorar la performance optimizando el uso de cluster con KNIME.

KNIME Analytics Platform es una herramienta de código abierto que opera bajo la licencia general pública de GNU versión 3, por lo que es de libre descarga y operación. Sin embargo, el uso de las extensiones comerciales si tienen un costo y pueden ser adquiridas directamente de KNIME. El soporte a la herramienta generalmente se realiza a través de publicaciones en la comunidad de KNIME, habiendo soporte adicional en caso se adquieran las licencias de pago o teniendo la posibilidad de adquirir soporte dedicado por parte de KNIME a través de un contrato de servicios.

Microsoft – Microsoft Azure Machine Learning

Microsoft Azure es la solución de Microsoft de tipo Platform as a Service (PaaS) e Infrastructure as a Service (IaaS). Bajo estos conceptos, Azure ofrece espacios virtuales que permiten alojar aplicaciones y archivos, realizar procesamiento muy complejos para ser realizados por infraestructuras in-house, desarrollar colaborativamente y desplegar aplicaciones webs y móviles, acceder a servicios de analítica en tiempo real, analítica de big data, machine learning y datawarehousing, además de poder crear redes virtuales para gestionar conexiones dedicadas y tráfico web y gestionar la seguridad propia de todas las funcionalidades anteriormente mencionadas.

Bajo este catálogo de funcionalidades y servicios surge Microsoft Azure Machine Learning. Se trata de un servicio de analítica predictiva basada en Cloud dentro del marco de servicios

de Azure que fue lanzado en Febrero 2015, siendo considerado un serio contrincante en el ámbito de analítica predictiva por el respaldo de una empresa líder como Microsoft.

El servicio tiene como principal enfoque brindar una experiencia simplificada pero poderosa de gestión, transformación y aprendizaje de información, aprovechando la infraestructura cloud que la soporta y los web services propios de la plataforma Azure. Azure Machine Learning permite crear, probar, operar y gestionar soluciones de analítica predictiva en la nube a través de un explorador web. Haciendo frente a las tendencias de lenguajes de gestión y predicción de datos, Azure provee integración con R y Python.

Dentro de las facilidades de Microsoft Azure Machine Learning, destacamos la posibilidad de desarrollar modelos predictivos a través de funciones simples e intuitivas tales como el Drag and Drop, empleando además una nutrida galería de módulos y plantillas existentes para simplificar el desarrollo de tareas de machine learning, garantizando la rapidez y simpleza.

La herramienta provee un ambiente de trabajo llamado Machine Learning Studio basado en web donde se encontrará el catálogo de módulos disponibles para simplificar el proceso de creación del modelo predictivo. Además, Microsoft Azure Machine Learning cuenta con un API llamado Machine Learning API usado para desplegar modelos predictivos como web services tolerantes a los errores y escalables, ofreciendo comunicación entre aplicaciones externas y el modelo de analítica predictiva.

Machine Learning Studio soporta tres formatos de ingreso de datos: carga a través de un archivo local que contiene la información a analizar, usando un módulo para importar data de servicios de datos en cloud o accediendo al conjunto de datos de un modelo previo. Los módulos del Studio soportan conjuntos de datos de hasta 10 GB y en caso sea necesario trabajar con una cantidad mayor de información, podemos realizar muestreos usando Hive o la base de datos SQL de Azure. Los siguientes módulos se encuentran disponibles:

- Módulos de recomendación
- Módulo de SMOTE para replicar los casos de uso en el conjunto de datos de manera balanceada
- Módulos de Scripting: R, Python, SQL
- Módulos de unión de información
- Módulos de verificación cruzada, regresión ordinal, agrupación de parámetros cuando se tratan de muchas iteraciones.

La herramienta soporta la ejecución de 4 módulos en paralelo para un experimento. Además, se puede visualizar la data de manera interactiva y obtener estadísticas del modelo trabajado.

El catálogo de algoritmos de Machine Learning Studio se basa en algoritmos vigentes según las tendencias tecnológicas actuales, tales como árboles de decisión escalables, sistemas de recomendación bayesiana, redes neuronales profundas y junglas de decisión desarrolladas en Microsoft. Asimismo, la herramienta soporta paquetes de código abierto y escalable de Machine Learning como Vowpal Wabbit. Además, la herramienta provee resultados comparativos por cada algoritmo, con el propósito de identificar el más adecuado según el propósito de nuestro modelo.

Azure provee la posibilidad de monitorear el modelo predictivo desarrollado a través de su panel de control, habiendo un dashboard por cada modelo actualizado en tiempo real. Asimismo, se garantiza una disponibilidad del 99.9% de transacciones de la API generada. Finalmente, el aspecto de costos es determinado por variables dependientes del uso de la herramienta. El uso de Machine Learning requiere una suscripción a Azure y se considera una suscripción mensual al studio y se cobra por cada hora de uso, además de realizarse cobros por horas de procesamiento de API de producción y por cantidad de transacciones en el mismo. Microsoft ofrece soporte técnico a través de su servicio de soporte técnico de Azure con un costo mensual y se ofrece complementariamente asesoría gratuita para los servicios de facturación y administración de suscripciones. En caso se opte por una versión de prueba, Microsoft cuenta con una comunidad de usuarios que podrán asesorar las dudas que se puedan presentar en el modelo.

Predixion Software – Predixion Insight

Predixion Software, una empresa de desarrollo de software localizada en California y enfocada en la industria de analítica predictiva. Dentro de su catálogo de soluciones, se encuentra Predixion Insight. Se trata de una plataforma de analítica predictiva basada en cloud que permite el despliegue de la analítica predictiva en diversos ambientes de producción. El enfoque más importante de la herramienta es llevar los insights que identifica a la línea principal del negocio, con el propósito de asesorar la toma de decisiones, aprovechando el ingreso de información propio del uso del concepto de internet de las cosas (IoT) y explotando las posibilidades de uso de dispositivos móviles.

La herramienta se basa en un modelo semántico de Machine Learning y es únicamente adecuada para implementaciones en organizaciones donde se trabaje con una infraestructura tecnológica que se base en los conceptos de IoT, lo cual puede ser visto como una limitante considerando que no es una tendencia aun común pero es de suma utilidad ya que la

información que se analiza es aquella que se genera en tiempo real, sin necesidad de tomar muestreos de situaciones anteriores.

Predixion Insight tiene la posibilidad de ofrecer un ambiente de creación y despliegue aplicaciones web desarrolladas en Java, R, C#, SQL y Javascript para acceder instantáneamente y de manera sencilla a la información que se predice. Este es considerado el principal beneficio de la solución debido a que se pueden tomar acciones inmediatas según transcurran los hechos y se puede diseñar a medida lo que verdaderamente necesita el negocio analizar. Asimismo, el asistente de despliegue de la herramienta genera el código fuente para facilitar el despliegue del modelo predictivo al dispositivo que se necesite o a la nube.

Al ser una herramienta que se enfoca netamente en el aspecto de IoT y enfrentar el reto de normalizar y estructurar la data proveniente de diversos dispositivos, se cuenta con funcionalidades que permite la combinación de datos de múltiples fuentes para un posterior modelado, integración y construcción de modelo, además de contar con automatizaciones de buenas prácticas para preparar, normalizar, limpiar y muestrear un conjunto de datos, permitiendo que el científico de datos se dedique a modelar, colaborar e iterar en resultados en lugar de desempeñarse en tareas redundantes. Asimismo, la herramienta cuenta con algoritmos especialmente dedicados a la gestión de datos provenientes de fuentes diversas, propio de IoT.

En el catálogo de algoritmos disponibles se definen las siguientes clasificaciones:

- Algoritmos de clasificación: asignación de una nueva observación a una de varias clases pre determinadas, analizando información sobre la información.
 - o Bosque de decisión
 - o Análisis del componente principal
 - o Múltiples árboles de decisión
 - o Regresión logística con regularización L1
 - o Reglas de clasificación MS/R
 - o Árboles de decisión MS/R
 - o Regresión logística MS/R
 - o Naive Bayes MS/R
 - o Redes Neuronales MS/R
- Algoritmos de estimación: usados para la predicción de un parámetro basado en valores observados de otros parámetros.
 - o Árboles gaussianos MS/R
 - o Regresión lineal MS/R
 - o Red neuronal MS/R
 - o Regresión logística MS
 - o Árboles de regresión MS

- Algoritmos de segmentación: separación de grupos en sub-grupos (segmentos y clusters) en base a características compartidas y no conocidas.
 - o Clustering de probabilidades MS/R
 - o Clustering de inputs continuos y nominales MS/R
- Algoritmos de asociación: es una técnica de minería de datos que identifica ítems o eventos que ocurren juntos.
 - o Reglas de asociación MS/R
 - o Árboles de asociación MS/R
- Algoritmos de predicción: es una técnica usada para predecir valores futuros en base a observaciones anteriores.
 - o ARIMA R
 - o ARTP MS
 - o Mezcla de ARIMA y ARTP MS
 - o ARIMA / R Auto ARIMA MS.

Para la medición y comparación de desempeños de algoritmos, se muestran cuadros comparativos de resultados y desempeño de modelos para determinar cuál es el más adecuado para la necesidad específica del negocio.

RapidMiner – RapidMiner Studio

RapidMiner Studio es la principal plataforma de analítica predictiva de código abierto, con un enfoque sumamente importante en ser de fácil uso y aplicación para cualquier proceso de negocio, permitiendo identificar oportunidades y riesgos de manera inmediata.

La herramienta, en sincronía con las tendencias en el mundo de la ciencia de datos, soporta código desarrollado en R y Python. Además, cuenta con múltiples métodos de carga, transformación y visualización de datos proveniente de diversas fuentes como Excel, Access, Oracle, IBM DB2, Microsoft SQL, Netezza, Teradata, MySQL, PostgreSQL y SPSS. Al ser una solución de código abierto, no se restringe las funcionalidades que se puedan crear a partir de los APIs que provee.

La solución no presenta restricciones de tamaño de información, empleando analítica de diversos tipos para satisfacer todo tipo de requerimientos de procesamiento de información: cuenta con analítica in-memory para fuentes pequeñas de información y analítica basada en Hadoop para soportar conjuntos de datos de grandes tamaños.

Como principales ventajas destaca el tener un catálogo de algoritmos de machine learning existente casi en su totalidad como parte de sus funcionalidades, aprovechando además el hecho de ser una herramienta completa para usarla tanto como para la preparación del modelo de datos como para operar y validar los resultados del modelo. Asimismo, la interfaz

gráfica del aplicativo es sumamente sencillo: existe un enfoque especial en el diseño visual intuitivo con recomendaciones automáticas, ampliando el uso de la herramienta tanto para científicos de datos como para no profesionales en la materia.

RapidMiner Studio respeta dentro de su estructura funcional las mejores prácticas y estándares de trabajo en el ámbito de predicción y en el ambiente de trabajo colaborativo: cuenta con numerosas plantillas y procesos predeterminados además de soportar el trabajo en paralelo y la subsecuente integración de lo desarrollado. Asimismo, permite ejecutar los modelos predictivos de manera automática o causada por algún evento definido, simulando así las acciones humanas acorde a algún resultado.

Al ser una herramienta de código abierto, la comunidad y el ámbito colaborativo resultan esenciales para el desarrollo de nuevas capacidades de la herramienta o del modelo. Sin embargo, las recomendaciones y asesoría dedicada por partes de expertos de la empresa RapidMiner requieren adquirir el servicio profesional.

Funcionalmente, la versión gratuita difiere de la versión profesional en las fuentes de datos disponibles: en la versión gratuita no se soporta fuentes de bases de datos comerciales, de nube, Hive (Hadoop), NoSQL ni fuentes como SPSS. Asimismo, sólo se ofrece 20 MB de repositorio y 8 GB de capacidad de procesamiento en caso se registre una versión gratuita, mientras que si se adquiere el servicio se tendrá 5 GB de repositorio y 60 GB de capacidad de procesamiento. Finalmente, los aspectos de colaboración, computación Hadoop, procesamiento de alta performance y despliegue sólo se encuentran disponibles en la versión profesional. Las cotizaciones se realizan acorde al perfil de cliente que adquiera la solución y el uso que se le dará a la herramienta, lo cual se evalúa a través de un cuestionario y una sesión con agentes de venta.

SAP – SAP Predictive Analytics

La empresa alemana SAP participa en el mercado de análisis masivo de información con la herramienta SAP Predictive Analytics. La herramienta tiene como propósito agilizar y simplificar el proceso de modelado predictivo, permitiendo la automatización del mismo.

SAP destaca que el proceso de depuración de información inicial es el más tedioso y propenso a errores en la elaboración de un modelo predictivo, por lo que ofrece una solución que automatiza este proceso a través de la definición, por parte de los usuarios del negocio, de componentes reusables del negocio llamados registros analíticos, los cuales son

empleados para crear conjuntos de datos para modelar, produciendo resultados de manera más rápida y precisa.

La herramienta presenta soporte para conectarse a fuentes de información de casi cualquier formato, desde archivos estructurados hasta formatos de archivos propietarios como SAS y SPSS. Además, permite trabajar tanto con datos nominales u ordinales, además de poder lidiar óptimamente con valores faltantes o incorrectos.

Dentro de la necesidad de brindar una herramienta accesible a usuarios de todo nivel, existe la posibilidad de ser guiado a través de la construcción de un modelo predictivo sofisticado en un lapso pequeño de tiempo empleando algoritmos que permitan análisis de asociación, de regresión, de cluster, de clasificación, de evolución a través del tiempo, de distribución de probabilidad, identificación de vínculos y funciones estadísticas con soporte univariable y multivariable. Además, permite que los científicos más especializados desarrollen funciones acorde a sus conocimientos y necesidades a través del uso del lenguaje R como del uso de algoritmos predictivos disponibles en la nube, con el uso de la plataforma SAP HANA. A través de la funcionalidad de análisis de experto, se simplifica el uso de algoritmos de código abierto basados en R con una interfaz de Drag and Drop para agilizar la construcción de modelos que empleen estos algoritmos. Asimismo, maneja una interfaz web para ingresar a la herramienta que está diseñada principalmente para el ámbito de un profesional de negocio y no de información, además de actualizar los modelos en tiempo real según lo que se vaya aprendiendo a través de las predicciones que realice.

SAP Predictive Analytics destaca por alta capacidad de procesamiento, permitiendo manejar más de 15000 atributos en paralelo de manera automática y empleando este análisis masivo en alimentar el modelo diseñado o propuesto y así generar mejores resultados a través del tiempo.

Dentro de las opciones de visualización de información se incluyen:

- Tamaño de clusters
- Densidad
- Distancias
- Comparación de variables
- Vista de árbol de decisiones
- Probabilidad de resultados
- Cuadro de evolución de información a través del tiempo
- Cuadro de coordenadas paralelas para variables numéricas y categóricas, incluyendo selección de variable, matriz de dispersión etc.

Asimismo, se ofrece la posibilidad de trabajar en un entorno colaborativo para compartir, complementar y publicar modelos predictivos con un mínimo esfuerzo.

La solución cuenta con la capacidad de identificar aquellas variables que son particularmente trascendentes para una pregunta del negocio, permitiendo visualizar aquello que es verdaderamente relevante y que no lo es o que cosa es redundante con otros atributos. Además, es posible generar cuadros de puntaje predictivo que, gracias a la interfaz abierta de SAP Predictive Analytics, pueden ser transmitidos a las plataformas que más lo necesiten: bases de datos, procesos de negocios o ámbitos propios del negocio, como por ejemplo call centers.

Adicionalmente, SAP Predictive Analytics provee una solución para el análisis de vinculaciones para aprovechar al máximo la influencia social, identificando vínculos entre el cliente y su entorno, identificando potencialidad de riesgo y fraude.

SAP Predictive Analytics provee un marco sobre el cual se puede sustentar el flujo total de la construcción de un modelo predictivo, permitiendo además la alimentación del mismo a través de diversas fuentes y aprovechando un extenso catálogo de algoritmos y acceso a funcionalidades de alta performance que permiten el análisis masivo de información sin sacrificar costo ni desempeño.

SAS – SAS Enterprise Miner

SAS Enterprise Miner es una herramienta de minería de datos destinada al procesamiento de los mismos y la creación de modelos predictivos y descriptivos de gran precisión, basándose en el análisis de grandes cantidades de información disponibles en la organización.

La solución está enfocada para ser usable por cualquier persona dentro de la organización, sin embargo se tiene un principal enfoque en aquellos roles que cumplen una labor principalmente científico-analítica más que de negocio.

En la fase de preparación, resumen y exploración de información, SAS Enterprise Miner permite acceder e integrar información estructurada e no estructurada, filtrar aquellos datos que no deberían formar parte del objeto de estudio, muestreo y particionamiento de información e importación de diversas fuentes.

Como parte del desarrollo del modelo predictivo y el enfoque de un modelo que aprenda de sus resultados (machine learning), tiene definidos los siguientes métodos para acceder y analizar información:

- Clustering
- Mapas de auto organización
- Regresión linear y logística
- Árboles de decisión
- Redes neuronales

- Razonamiento basado en memoria
- Minería basada en series de tiempo
- Algoritmos de alta performance: partición de datos, transformación, selección de variables, árboles, bosques, redes neuronales, regresión

Asimismo, la herramienta soporta el lenguaje de programación R para permitir la integración de otros algoritmos dentro del modelo desarrollado, incluyendo además la posibilidad de realizar evaluaciones de modelos basadas en R para medir el desempeño de cada uno y determinar cuál es el que mejor se aplica a la solución deseada.

Dentro de las funciones más resaltantes de la herramienta, existe la capacidad de asesoría para la elaboración del modelo a través de la comparación de los resultados del mismo con diferentes algoritmos empleando diagnóstico estadístico y métricas de retorno de inversión, además de la identificación de puntos de corte del modelo que pueden servir para identificar nuevas posibilidades o nuevas oportunidades de negocio. Asimismo, integra la minería de texto para el análisis correcto de data estructurada y no estructurada.

SAS Enterprise Miner soporta la generación de código de puntajes en SAS; C, Java y PMML, permitiendo la exportación de los resultados a IBM DB2, Oracle y Teradata. Asimismo, la solución brindada ofrece un amplio espectro de escalabilidad al ofrecer una arquitectura que soporta desde soluciones de usuario hasta soluciones necesarias para empresas multinacionales, sustentado por la opción de realizar procesamiento masivo o en memoria según sea necesario y empleando procesamiento en paralelo.

Benchmarking resultante

Criterios / Herramientas		Angoss – Angoss Predictive Analytics	Alpine Data Labs – Alpine Chorus	Alteryx – Alteryx Analytics	Dell – Statistica	FICO – FICO Predictive Analytics	IBM – IBM Predictive Analytics	KNIME – KNIME Analytics	Microsoft – Microsoft Azure Machine Learning	Predixion Software – Predixion Insight	RapidMiner – RapidMiner Studio	SAP – SAP Predictive Analytics	SAS – SAS Enterprise Miner
Oferta actual	55%	1.95	1.675	2.15	2.1	2.125	2.5	1.95	2.225	2.05	1.975	2.525	1.825
1. Arquitectura	10%	3	2.5	3.5	3	3	4	3	4	4	3	4	4
2. Seguridad	5%	3	3	4	4	4	4	2	4	3	3.5	4	3
3. Data	5%	4	3.5	4	4	4	5	4	3	5	4	5	3
4. Análisis	15%	4	3	4	4	4	5	4	4	4	4	5	3.5
5. Gestión del modelo	10%	3	3	4	4	4	4.5	4	4	4	4	4.5	3
6. Usabilidad de la herramienta	5%	4	4	4	4	4	4	3	5	3	4	4.5	3
7. Aplicaciones de negocio	5%	4	3	4	4	4.5	5	4	4.5	2	2	5	3
Estrategia	40%	1.35	1.15	1.4	1.3	1.3	1.54	1.3	1.225	1.175	1.2	1.6	1.225
8. Adquisición y precio	10%	3	3	3	3	3	3	3.5	3	2	3	3	2
9. Capacidad de ejecución	10%	4	3	4	4	4	4.5	4	3	3	3	4	4
10. Soporte de implementación	10%	4	3	4	3	3	4	3	3	3.5	3	5	3
11. Roadmap	5%	3	3	3	3	3	3	2	3.5	3.5	3	4	3.5
12. Ratio de salida al mercado	5%	2	2	3	3	3	4.8	3	3	3	3	4	3
Situación del vendedor	5%	0.15	0.15	0.15	0.15	0.15	0.225	0.18	0.2	0.165	0.185	0.25	0.21
13. Situación Financiera	2%	3	3	3	3	3	4.5	4	4	3.5	4	5	5
14. Base de Clientes	1%	3	3	3	3	3	4.5	4	3	3.5	3.5	5	3
15. Socios de Negocio	2%	3	3	3	3	3	4.5	3	4.5	3	3.5	5	4
Puntaje Total	100%	3.45	2.975	3.7	3.55	3.575	4.265	3.43	3.65	3.39	3.36	4.375	3.26

Fuente: Elaboración Propia, Basado en Gualtieri, M. & Curran, R. (2015)

En base al benchmarking realizado identificamos que la herramienta de mayor puntuación según los criterios establecidos es SAP Predictive Analytics, destacando la robustez de la herramienta en cuanto a la oferta que brinda a los usuarios en aspectos de funcionalidades, soporte y alineamiento a las últimas tendencias tecnológicas. Al ser una herramienta que puede desempeñarse en ambientes Cloud, su implementación no conlleva inversión en infraestructura ni en recursos humanos en la medida que lo demandaría una solución in-house, lo cual resulta particularmente atractivo para las empresas aseguradoras que se basan en estructuras tecnológicas y organizacionales sumamente complejas.

El respaldo que tiene la herramienta elegida es también un factor a considerar. El impacto de un modelo predictivo en una organización es de suma importancia por lo que es necesario garantizar el adecuado funcionamiento de la misma y la asistencia técnica cuando sea necesario, sobre todo si se trata de una herramienta en la nube.

Se realizó una simulación de uso de la herramienta y se identificó otro beneficio con el que cuenta la herramienta, una interfaz simple e intuitiva, lo cual garantiza que la curva de aprendizaje para dominarla y desarrollar el modelo predictivo será sustancialmente menor a lo que supondría una herramienta de mayor complejidad de uso.

El catálogo de funcionalidades y capacidades de la herramienta es el más completo de las herramientas analizadas e identificadas como trascendentes en el ambiente de análisis de datos. La vasta disponibilidad de modelos, algoritmos, capacidad de acceso, tratamiento y conversión de información es además un punto sumamente atractivo en beneficio de SAP Predictive Analytics, principalmente porque la herramienta no es restrictiva, permitiendo el desarrollo de modelos más acorde a necesidades específicas del proyecto y con soporte a la escalabilidad que se le quiere brindar al modelo desarrollado.

Finalmente, para propósitos del proyecto, resulta beneficioso el convenio que la Universidad tiene con SAP, lo cual permite el fácil y libre acceso a la licencia de la herramienta sin costo alguno, contando además con soporte especializado en caso sea necesario.

Análisis de Algoritmos

El éxito de un modelo predictivo radica en su capacidad de permitir identificar relaciones basadas en reglas matemáticas que conlleven a predecir un resultado futuro. El proceso de identificación de relaciones se basa en las reglas determinadas en un algoritmo. Por definición, un algoritmo es un conjunto de pasos para lograr un objetivo, siendo este objetivo la predicción cuando se trata de un algoritmo destinado a pronosticar el futuro. De esta manera, podemos identificar que el éxito del modelo propuesto se basa no solo en la

identificación correcta de variables relevantes para el pronóstico deseado, sino además en la elección adecuada del paso a paso que se seguirá para realizarlo.

Según la identificación previa de la herramienta a emplear, se propondrá a continuación un análisis de las familias de algoritmos disponibles en la herramienta de forma *out of the box*: es decir, aquellos que vengan por defecto en la instalación inicial de la herramienta. Luego, se identifican aquellos algoritmos que cumplen con las condiciones necesarias según el contexto de la predicción a realizar y se realizarán pruebas, con el propósito de identificar aquel algoritmo que sea el más cercano a la realidad.

Es necesario introducir el contexto del modelo predictivo a realizar: Se propone desarrollar un modelo predictivo destinado al diagnóstico de la Diabetes Mellitus Tipo 2 según un análisis sintomático y de factores cotidianos: es decir, basándose en respuestas ya recopiladas sobre las costumbres o rasgos del paciente.

La herramienta elegida trae por defecto 24 algoritmos, clasificados en 7 familias según su naturaleza. Estas 7 familias serán explicadas a continuación, identificando si son aplicables al contexto del modelo o no.

Algoritmos de serie de tiempo.

Una serie de tiempo está definida como una secuencia de datos medidos a través de puntos en el tiempo, siendo estos puntos generalmente equidistantes al tratarse muchas veces de intervalos ya definidos (por ejemplo, medición periódica de la presión sanguínea en intervalos de 30 minutos y evaluando como esta fluctúa según la ingesta de alimentos). SAP Predictive Analytics cuenta con 3 algoritmos en esta familia:

- Single Exponential Smoothing
- Double Exponential Smoothing
- Triple Exponential Smoothing

Este tipo de algoritmos son además destinados para pronosticar datos sin alguna tendencia o patrón estacional. Sin embargo, al basar la predicción en una línea de tiempo, resulta un conjunto de algoritmos no aplicable para el modelo propuesto debido a que la información recopilada de los pacientes se basa exclusivamente en un cuestionario a responder en el momento en que se realiza, no en un cuestionario desplegado a través de varias instancias de vida del paciente.

Algoritmos de regresión.

Los algoritmos de regresión tienen como propósito la predicción de variables continuas empleando otras variables disponibles en el conjunto de datos. Este tipo de algoritmos

buscan modelar la relación entre variables y es refinada de manera iterativa, empleando como medición de error las predicciones hechas por el modelo.

SAP Predictive Analytics cuenta con 8 algoritmos de este tipo:

- Regresión Lineal
- Regresión Exponencial
- Regresión Geométrica
- Regresión Logarítmica
- Regresión Linear Múltiple
- Regresión Polinomial
- Regresión Logística
- Regresión Automática

El último algoritmo (regresión automática) es una característica particular de la herramienta frente a otras que también cuentan con un catálogo similar de posibilidades de regresión. Se trata de una funcionalidad que emplea una técnica llamada Minimización Estructural del Riesgo y construye un modelo polinomial con capacidad de procesamiento de múltiples atributos. La familia de algoritmos de regresión se aplica al contexto del modelo a desarrollar, sin embargo resulta necesario discernir correctamente en la evaluación de estos algoritmos sobre cuales verdaderamente cumplen con los requisitos de predicción estipulados por el proyecto.

Algoritmos de asociación.

Los algoritmos de asociación están destinados a realizar análisis de asociaciones para descubrir relaciones escondidas en largos conjuntos de datos. Las relaciones que se descubren son de gran utilidad para la aplicación en casi toda industria imaginable: podemos identificar patrones de comportamiento en compras por ejemplo y mejorar la distribución en un local de abarrotes, alentando la compra de determinados productos (como por ejemplo el famoso caso de la compra de cerveza y pañales). La herramienta cuenta con dos algoritmos de asociación:

- A priori
- A prioriLite

Al ser algoritmos destinados a la identificación de relaciones entre datos, no son aplicables al proyecto debido a que el modelo propuesto busca predecir la diabetes en el grupo de estudio, no analizar los factores que la determinan. Sin embargo se trata de una familia de algoritmos de alta utilidad por el potencial de aprendizaje que puede brindar.

Algoritmos de agrupamiento.

Los algoritmos de agrupamiento tienen como propósito agrupar características del conjunto de datos en clusters según su similaridad, permitiendo identificar patrones según los distintos grupos resultantes. La herramienta provee un algoritmo de agrupamiento:

- K-Means Clustering

Al ser un proyecto destinado al diagnóstico de la enfermedad mencionada anteriormente, el algoritmo de agrupamiento no es aplicable debido a que no cumple con el propósito del modelo, sin embargo es sumamente útil al permitir identificar agrupaciones de rasgos en común y ser usable para otro tipo de análisis relacionado también a la enfermedad.

Algoritmos de árboles de decisión.

Los algoritmos de árboles de decisión son de los métodos más usados para inferir resultados. El planteamiento de este tipo de algoritmos está basado en reglas de decisiones: si se cumple con determinada condición, entonces se tiene determinada característica. Son ampliamente usados para predicciones y tienen particular relevancia en el ámbito de medicina.

La herramienta provee tres algoritmos de árboles de decisión:

- HANA C 4.5
- R-CNR Tree
- CHAID

Los cuales son aplicables al contexto del proyecto al identificar relaciones y permitir cumplir el objetivo de diagnosticar la enfermedad al grupo de datos estudiado, por lo que estos algoritmos serán evaluados en la implementación del modelo.

Algoritmos de detección de partes aisladas.

Los algoritmos de detección de partes aisladas están destinados a la identificación de observaciones que no cumplen una secuencia lógica con las observaciones analizadas, siendo tan distintas que parecieran haber sido generadas por algún otro mecanismo. Está destinado principalmente para la detección de fraudes a través del análisis de actividades fraudulentas o la identificación de la ocurrencia de enfermedades no epidémicas con una misma causa. La herramienta provee cuatro tipos de identificación de partes aisladas:

- Inter Quartile Range
- Nearest Neighbor Outlier
- Anomaly Detection
- Variance Test

Al ser algoritmos destinados a identificar observaciones anormales a las del grupo de estudio y esto no ser parte del objetivo del proyecto, no se considerará el uso de este tipo de algoritmos al estar destinado para otro propósito.

Algoritmos de redes neuronales y clasificación.

Los algoritmos de redes neuronales se basan en el concepto de neuronas en el cerebro. La característica más resaltante de este tipo de algoritmos es que no sigue un camino lineal: es procesada de manera colectiva, paralelamente a través de nodos (o neuronas) y con la habilidad de adecuar su estructura interna según la información con la que se trabaje. Generalmente, son empleados para pronosticar, clasificar e identificar patrones estadísticos. La herramienta elegida cuenta con dos algoritmos de redes neuronales:

- R-NNet Neural Network
- R-MONMLP Neural Network

Dentro de esta familia, destaca además un algoritmo propio de SAP:

- Auto Classification

El algoritmo mencionado anteriormente está destinado a la clasificación binaria/categorica, identificando el tipo de modelo y el algoritmo de clasificación más adecuado para el resultado buscado, reduciendo en gran medida el esfuerzo invertido en preparar datos y probar el modelo con distintos algoritmos. Además, permite la creación de conjuntos de datos de entrenamiento y validación para evaluar los resultados y mejorar su predicción.

El presente grupo de algoritmos será considerado para el desarrollo del proyecto al ser aplicable para el contexto de una predicción binaria (es decir, aplicable al diagnóstico de la Diabetes Mellitus tipo 2).

Conclusiones del Análisis de Algoritmos

En base al análisis realizado, se emplearán algoritmos de los siguientes tres tipos de familias:

- Árboles de Decisión
- Regresión
- Redes neuronales/Clasificación

Como parte del proyecto, se realiza la evaluación de todos los algoritmos aplicables de las tres familias mencionadas anteriormente, con el propósito de definir cuál es el que provee el mejor resultado en cuanto a la predicción del diagnóstico de la Diabetes Mellitus tipo 2. El criterio a emplear para determinar la aplicabilidad del algoritmo se basará en los datos de input que se utilizarán en el modelo y la aplicabilidad de la funcionalidad de predicción de

cada uno, mientras que la eficacia será medida en base a las predicciones correctas del conjunto de datos de prueba.

ELABORACION DEL MODELO PREDICTIVO

Definición

El entregable final del proyecto profesional es un modelo predictivo para el diagnóstico de la diabetes mellitus a través del análisis sintomático/de hábitos cotidianos del paciente.

Objetivo

El modelo elaborado tiene como propósito describir las actividades desarrolladas en el marco de la elaboración de un modelo predictivo destinado al diagnóstico de la diabetes mellitus tipo 2 a través del análisis sintomático/de hábitos cotidianos del paciente.

Alcance

El alcance del modelo contempla su elaboración basándose en análisis sintomático y de hábitos cotidianos del paciente, evaluando los algoritmos disponibles en la herramienta elegida y empleando el que resulte más preciso según el análisis de los resultados de las predicciones ejecutadas con conjuntos de datos de prueba.

Contexto

El principal reto del desarrollo de un modelo predictivo radica en cómo volverlo aplicable a un proceso de índole comercial sin interrumpir su normal ejecución, siendo este uno de los principales motivos por el cual aún no están desplegados de manera masiva en la fase de diagnóstico de pacientes atendidos en entidades prestadoras de servicios de salud. Bajo esta premisa, el reto principal del modelo elaborado radica en cómo desarrollarlo de tal manera que se inserte de forma natural en el proceso de una cita o revisión médica rutinaria.

Es en base a este reto que se plantea la necesidad de desarrollar un modelo que reciba información de conocimiento común del paciente, de tal manera que no dependa de los resultados de un análisis médico (que puedan demorar) y cuya información pueda ser recopilada a través de una encuesta sencilla desde la comodidad del hogar de paciente, de forma virtual, optimizando así el tiempo destinado a la cita médica.

Conjuntos de Datos

La definición de variables a emplear como base para el modelo predictivo se realizó a través de la investigación de los principales factores de conocimiento común que pueden brindar indicios sobre el desarrollo de la diabetes mellitus tipo 2. Al ser una enfermedad caracterizada por el exceso de azúcar en la sangre²³ es posible identificar que las costumbres alimenticias son indicadores directos sobre la posibilidad de desarrollar la enfermedad. Asimismo, identificamos que el índice de masa corporal, determinado por la relación entre la altura y el peso del paciente es derivado directamente de la dieta del paciente, por lo que son también importantes para considerar.

Partiendo de la premisa que la alimentación es un indicador de suma importancia para el diagnóstico de la enfermedad, es necesario también considerar que variables pueden brindar indicios de problemas en la alimentación y calidad de vida: naturalmente se infiere que los problemas cardiovasculares y de colesterol son indicadores de una alimentación deficiente en la gran mayoría de casos²⁴ y al ser males que suelen ser de pleno conocimiento de los pacientes, resultan opciones ideales para considerar en el modelo predictivo.

El desarrollo de la diabetes mellitus tipo 2 es favorecido de manera directa por costumbres deficientes en cuanto al ejercicio diario y por la adopción de un estilo de vida sedentario²⁵, por lo que la actividad física pasa a ser un factor crucial en cuanto al análisis del desarrollo de la enfermedad.

Adicionalmente, la diabetes mellitus tipo 2 tiene un componente genético de suma importancia, necesario de considerar si se busca predecir su incidencia, teniendo especial consideración en cuanto a cómo se clasifica: cuanto más cercano sea el parentesco de alguna persona con la enfermedad, mayor será el riesgo de desarrollarla²⁶.

Clasificación de Datos

El siguiente paso posterior a la determinación de variables a emplear es como clasificarlas de tal manera que se pueda estandarizar su ingreso al modelo que procesará la predicción. Según las variables definidas, podemos identificar tres tipos de clasificaciones necesarias de

²³ American Diabetes Association (2016)

²⁴ Kanter, Kris-Etherton, Fernandez, Vickers and Katz (2012)

²⁵ New England Journal of Medicine (2016)

²⁶ Arslanian, Bacha, Saad and Gungor (2004)

realizar: aquellas que se determinan en rangos simples (si/no), aquellas que se clasifican en rangos de mayor complejidad, como por ejemplo la variable de parentesco con alguna persona que tenga la enfermedad y finalmente aquellas que pueden tener un valor dentro de un rango no parametrizado. En el caso de este último tipo de clasificación, se procederá a transformar las variables de tal manera que encajen en un rango definido.

Inicialmente determinamos las variables de clasificación simple por su naturaleza como las siguientes:

- Diagnóstico de la Diabetes Mellitus Tipo 2
- Problemas de Glucosa
- Problemas de Colesterol
- Problemas de Presión
- Actividad Física (superior a las 3 horas semanales, según la Asociación Americana del corazón²⁷)
- Dieta balanceada (tomando como referencia una dieta balanceada en granos, vegetales y frutas como acuerdo común por toda organización ligada a la salud)

Por consiguiente, determinamos las siguientes variables como aquellas cuya clasificación es de mayor complejidad, determinada de la siguiente manera:

- Parentesco familiar con algún paciente con la enfermedad
 - Definir rangos y sustento
- Índice de Masa Corporal
 - Definir rangos y sustento

Finalmente, las variables que no son clasificadas en un rango son aquellas numéricas. En el caso particular de las dos variables del modelo que son de esta naturaleza, son transformadas en el índice de Masa Corporal y empleadas bajo este concepto:

- Altura
- Peso

Recopilación de Datos de Prueba

La recopilación inicial de datos se realizó basándose en un cuestionario virtual alojado en Google Forms con preguntas necesarias para recopilar las variables definidas anteriormente.

²⁷ Heart.org (2016)

Por un tema de estricta confidencialidad, no se recopilaban datos personales de los pacientes más allá de su edad y sexo.

La distribución del formulario desarrollado se desarrolló a través de 3 vías: distribución virtual directa por el equipo del proyecto, distribución virtual por personas a nuestro alrededor (familiares, compañeros y amigos) y finalmente, con el fin de recopilar información sobre pacientes con la enfermedad, se realizaron encuestas presenciales en el Centro de Atención Integral en Diabetes e Hipertensión (CEDHI) de EsSalud, teniendo como resultado total la recopilación de 656 pacientes, de los cuales 47 contaban con la enfermedad (un 7%). Esta muestra inicial nos permite sentar la línea base de cara al aprendizaje de predicción del modelo desarrollado, puesto que los toma como datos de prueba referenciales sobre los cuáles basar la predicción de nuevos valores que serán recopilados a través del formulario que será incluido como parte del modelo predictivo desarrollado.

Preparación de Repositorio de Datos

El repositorio de datos del modelo predictivo desarrollado está basado en una base de datos de motor SQL Server 2012, la cual ha sido conectada a la herramienta utilizada empleando drivers JDBC a través de la interfaz gráfica de preparación de conexiones. El repositorio de información almacena datos en dos instancias a modo de tablas detalladas a continuación:

- Previa al procesamiento de información: contiene la información recopilada a través del formulario virtual creado. Los datos presentes en esta tabla se encuentran transformados a un formato numérico que sea legible para la ejecución posterior de la predicción.
- Posterior al procesamiento de información: contiene información de los resultados de procesamiento, incluyendo los datos empleados para la predicción y los resultados de la misma.

El potencial de mantener la información actualizada en un repositorio no solo permite la consulta del resultado de la predicción en el momento de la cita médica, sino permite además recopilar una fuente de información sobre la precisión del algoritmo, la influencia de determinadas variables en el diagnóstico de la enfermedad y demás casos prácticos que pueden desprenderse de un repositorio centralizado de información previa al procesamiento como posterior al mismo.

Preparación de la Herramienta

El proceso de preparación de la herramienta se realizó con relativa simpleza producto de la alta usabilidad de la herramienta. De manera inicial, una vez definida la estructura del repositorio de datos con los tipos de datos acorde a cada variable, se procede con la instalación de los drivers JDBC para la conexión de la herramienta con la base de datos. Los componentes de conexión usados en la preparación de la herramienta fueron los de lectura y escritura de datos.

Definición del Proceso de Aplicación del Modelo Predictivo

En base a lo definido anteriormente se define el proceso de implementación del modelo predictivo en un entorno de una entidad prestadora de servicios de salud de la siguiente manera:

Ilustración 34 Cuadro de aplicación del modelo predictivo.

Fuente: Elaboración propia.

Propuesta de Continuidad

La propuesta de continuidad del proyecto se divide en 3 fases siguientes a la implementación del modelo, categorizadas de la siguiente manera:

Post Implementación:

De manera inmediata a la puesta en producción del modelo, se apreciará el enriquecimiento del mismo a través de la recopilación de información de pacientes. La consecuencia directa de esto será la posibilidad de emplear datos recopilados con la salvedad que sean utilizados de forma anónima para el posterior entrenamiento del modelo que permite afinar la precisión del diagnóstico de la enfermedad.

Adaptabilidad del modelo:

La adaptabilidad del modelo es el siguiente paso inmediato en cuanto a la continuidad del proyecto. Se divide en dos etapas: identificación de nuevas variables de estudio y descubrimiento de patrones, observaciones y consideraciones para el diagnóstico de la enfermedad. La identificación de nuevas variables de estudio se realizará en base a la evolución del modelo: como tal, el modelo recopila en su fase de input de datos variables adicionales que no son usadas en la predicción (edad y sexo), sin embargo estas pueden incluirse en el modelo sin afectar el despliegue del mismo, siendo necesario sencillamente la selección de estas nuevas variables dentro de la fase de entrenamiento de esta nueva versión del modelo predictivo y en la puesta en producción del mismo. El descubrimiento de patrones, observaciones y consideraciones para el diagnóstico de la Diabetes Mellitus tipo 2 es un resultado natural de la implementación del mismo y no depende de la identificación de nuevas variables, sino simplemente depende de la recopilación de datos de pacientes y la precisión de diagnósticos realizados por la herramienta, corroborando los resultados de la misma de manera posterior a la evaluación médica sobre si el paciente tiene la enfermedad o no, permitiendo de esta forma desarrollar un repositorio que permita la extracción de conocimiento sobre la incidencia de determinados factores en el desarrollo de la enfermedad, identificación de factores y motivos de sesgo en el diagnóstico de la enfermedad y demás patrones que puedan identificarse producto de la base de datos resultante como consecuencia de la puesta en producción del modelo.

Evolución del modelo:

La tercera fase de la propuesta de continuidad del modelo radica en 2 etapas: la mejora de procesos de pronóstico de enfermedades crónicas y la transición de procesamiento *in-house* a un entorno *Cloud* empleando SAP Hana. La primera etapa se llevará a cabo como consecuencia directa de la implementación del modelo, ya que esta tendrá como resultado final la optimización de procesos de diagnóstico aprovechando el tiempo resultante entre la solicitud de una cita médica y el momento en que se lleva a cabo, además de permitir al profesional médico acceder al análisis sintomático de manera directa sin necesidad de realizar el cuestionario nuevamente e interpretar las respuestas del paciente para brindar una asesoría personalizada sobre las acciones a tomar para mejorar su salud, además de afinar el criterio usado para determinar cuando un paciente debe atravesar un examen de glucosa para confirmar o descartar la diabetes mellitus tipo 2. La segunda etapa se basa en la transición de un esquema *in-house* a un entorno *cloud* soportado por SAP Hana como consecuencia directa de la necesidad de procesamiento adicional por parte de la herramienta, considerando la magnitud de información recopilada. La posibilidad de migrar a un entorno en la nube evidencia la escalabilidad y adaptabilidad del modelo desarrollado frente a lo que el futuro le depara al modelo desarrollado. Este ámbito en la nube sería administrado en cuanto a hardware por SAP pero la gestión de la herramienta se mantendrá de manera local, manteniendo las mismas capacidades que cuando el procesamiento y despliegue era totalmente *in-house*.

Ilustración 35 Plan de continuidad del modelo predictivo.

Fuente: Elaboración propia.

Conclusiones del Modelo Predictivo

El principal reto que supone la elaboración de un modelo predictivo radica en su aplicabilidad a un entorno organizacional en donde no genere un retraso en el normal desarrollo de las actividades de la misma. Bajo esta premisa, el planteamiento del modelo radicó en aprovechar el lapso de tiempo entre la obtención de la cita médica y que esta ocurra. De esta manera, y aprovechando las tecnologías disponibles en el mercado, se propuso un modelo integral que permita agilizar el diagnóstico de la diabetes mellitus tipo 2 en pacientes que soliciten una cita médica a través de la recopilación de datos de conocimiento del paciente que sean ingresados en un formulario virtual y a su vez alimenten una base de datos usada para la predicción.

Es necesario resaltar además que la proyección del proyecto como tal supone la escalabilidad para la inclusión de otras variables, además de basarse en una herramienta que puede ser desplegada en la nube según la demanda de procesamiento. Consideramos además que la aplicación de una herramienta altamente usable contribuye a la facilidad de despliegue del modelo en cualquier entorno administrado no necesariamente por un profesional en la ciencia de datos.

Finalmente, el modelo propuesto permite la recopilación de una base de datos escalable cuyos resultados almacenados, tanto a nivel de respuestas recopiladas por los pacientes como en base a las predicciones realizadas, pueden ser usados para el estudio de los factores de riesgo de la enfermedad, permitiendo además afinar el procedimiento de recomendaciones de estilo de vida brindadas al paciente en base a las respuestas recopiladas y también siendo un formulario aplicable a personas con la enfermedad diagnosticada, de tal manera que sus respuestas puedan brindar una vista clara sobre las acciones que dicha persona está tomando día a día y como estas pueden afectar el desarrollo de la enfermedad).

CAPÍTULO 6. Resultados y validación del Modelo predictivo

El siguiente capítulo muestra los resultados del trabajo realizado. De tal manera, se muestra los resultados de las pruebas del modelo para el desarrollo del modelo predictivo y su validación.

EVALUACION Y DEFINICION DE ALGORITMOS

La evaluación de algoritmos se realizó a través de dos instancias de predicción detallados a continuación:

- Entrenamiento del algoritmo a través de la ejecución de la predicción con un conjunto de 525 registros totalmente independientes, elegidos de manera aleatoria.
- Ejecución del modelo entrenado empleando los modelos entrenados y un conjunto de 132 datos totalmente distinto a lo usado para el entrenamiento previo.

La definición de la precisión de los algoritmos se determinó según la cantidad de diagnósticos correctos (positivos o negativos) de la enfermedad sobre el total de registros, obteniendo los siguientes resultados:

Tabla 24 Resultados de pruebas de algoritmos

Algoritmo	Precisión Entrenamiento	Precisión Validación
R-CNR Tree	98.1%	90.8%
Auto Classification	95.4%	91.7%
R-Nnet Neural Network	94.9%	90.1%

Fuente: Elaboración propia.

Tomando en cuenta los resultados de la precisión validación se determinó que el algoritmo de Auto Classification brindaba mayor precisión en el diagnóstico, identificando correctamente los diagnósticos positivos y negativos con una precisión del 91.7%.

DEFINICION DE PROCEDIMIENTO DE INGRESO DE INFORMACION

El modelo predictivo supuso como reto principal la identificación de una oportunidad de negocio bajo la cual se pueda implementar en el proceso natural de diagnóstico de la enfermedad sin dilatarla, al contrario, contribuyendo a la optimización de este proceso. Bajo esta premisa, se identifica la posibilidad de desarrollar un proceso de ingreso de información que se desarrolle en la ventana de tiempo existente entre la solicitud de cita médica y el momento en que esta se lleva acabo, el cual se basaba en el desarrollo de un formulario virtual que se distribuye al paciente solicitante de la cita médica a través de un correo desplegado diariamente donde se indica el vínculo para llenarlo.

El formulario propuesto tiene como preguntas aquellas que les dan respuesta a las necesidades de información según las variables definidas, además de un campo de registro que permite identificar, a través de un documento de identidad, al paciente que contestó

determinadas características en el formulario, de tal manera que se pueda visualizar en la cita médica. De manera complementaria, el procedimiento de ingreso de información incluye el grabado de las respuestas del paciente en el repositorio de información, siendo además visible por el especialista médico quien puede brindar consejos para llevar una vida más saludable en el momento de la cita médica.

Según las necesidades plasmadas y los recursos disponibles para el proyecto, se determinó la viabilidad de desarrollar un formulario basado en Wordpress y alojado de manera pública, a través del cual se pueda recopilar la información solicitada. Asimismo, el desarrollo del formulario a través de Wordpress permite la adaptabilidad del mismo a distintas interfaces de escritorio como móviles, permitiendo de esta manera desarrollar un *input* de datos que se mantenga en vigencia de cara a las tendencias tecnológicas que apuntan hacia el incremento del uso de dispositivos móviles y el descenso en el uso de computadoras.

El formulario desarrollado se basa en un plugin de Wordpress de nombre FormidableForms, que permite además la fácil exportación de las respuestas recopiladas a un formato de Excel, el cuál luego atravesará un proceso de transformación hacia la base de datos y finalmente se cargará en la tabla de ingreso al modelo, la cual realizará procesamientos diarios y guardará los resultados de los mismos en el repositorio de consulta.

MODELO PREDICTIVO RESULTANTE

El modelo resultante es el mostrado a continuación:

Ilustración 36 Modelo Predictivo.

Fuente: Elaboración propia.

Como todo modelo predictivo, atraviesa tres fases: ingreso de información, procesamiento (complementado por analítica) y resultado de predicción.

La primera fase está basada en el formulario virtual de ingreso de información donde se recopilarán los resultados de las preguntas realizadas al paciente, mediante las cuales se recopilarán las variables definidas anteriormente y estas se alojarán en una base de datos soportada por el motor SQL Server 2012. La exportación de los datos recopilados en el formulario hacia la base de datos se hará de manera diaria a través de una rutina automática de extracción, transformación y carga (proceso ETL) que además ejecutará la transformación de los datos hacia los rangos ya definidos por cada uno.

La segunda fase es la de procesamiento, dentro de ella tenemos cuatro procesos: el primero es la depuración de datos no considerados, tales como la edad y el sexo, los cuales han sido desestimados en el primer alcance del proyecto (a pesar que se han recopilado, de tal manera que si se considera oportuno se pueden considerar). A continuación, se realiza la transformación de las variables de altura y peso hacia la variable de índice de masa corporal, que indica la medición de la grasa del cuerpo en función de la altura (siendo de suma importancia para el diagnóstico de la enfermedad). En la fase de procesamiento, además, se realiza el entrenamiento del modelo sobre los perfiles de riesgo de diagnóstico de la enfermedad a través de la predicción de información recopilada. Dicho entrenamiento no debe ser en cada ejecución del modelo predictivo, pero si resulta necesario hacerlo de manera periódica, con el propósito que se afine el diagnóstico de la enfermedad con mayor precisión. Finalmente, el cuarto paso de procesamiento es la ejecución de la predicción empleando un algoritmo de redes neuronales basado en R, el cual brindó la mayor precisión en cuanto al diagnóstico de la enfermedad según las pruebas realizadas a través de las fases de entrenamiento y simulación. Asimismo, el procesamiento se realizará diariamente, una vez recopilada la información del día.

Finalmente, la fase de salida de información se ejecutará de manera diaria, posterior al procesamiento de los datos diarios, y se basará en el almacenamiento de la información trabajada en una base de datos del motor SQL Server 2012, la cual será consultada por el especialista médico en el momento de cita médica del paciente. Adicionalmente, en caso el paciente tenga la enfermedad, puede llenar el formulario y el especialista médico podrá brindar recomendaciones sobre costumbres diarias que tiene el paciente que pueden contribuir al empeoramiento de la enfermedad.

CAPÍTULO 7. Gestión del Proyecto

Este capítulo aborda la gestión realizada a lo largo del proyecto, centrándose específicamente en la comparación de las actividades planificadas y lo que se realizó finalmente, sea entorno a hitos, entregables, roles, riesgos, comunicaciones y otros aspectos diversos. Permite identificar los aspectos críticos, inconvenientes o situaciones insatisfactorias que condicionaron el desarrollo del proyecto.

PRODUCTO FINAL

Las primeras actividades dentro de la realización del modelo predictivo se tenía identificación la enfermedad a predecir y la herramienta que soporte el modelo, mediante la investigación se logró determinar que la enfermedad de mayor impacto dentro de nuestro análisis es la diabetes; por otro lado, se realizó un benchmarking de herramientas top en cuanto a predictive analytics, optando por la herramienta de SAP Predictive Analytics. Para obtener data a predecir y data de aprendizaje se realizó una encuesta en la que se preguntaban por los factores de riesgo determinantes dentro del modelo predictivo. Se pudo obtener más de 600 encuestados (se tienen dos tipos de información, la data necesaria para aprender de ella, encuestados diabéticos, y la data de encuestados no diabéticos, con la que se realizaran las pruebas). Para la implementación del modelo predictivo se realizaron pruebas con algoritmos para determinar el que brinda los mejores resultados de predicción, resultado el algoritmo de Redes Neuronales basado en R el de mejor precisión para la predicción del diagnóstico de la Diabetes Mellitus Tipo 2.

GESTION DEL TIEMPO

Con respecto al cronograma, presentado por hitos, se tuvo un pequeño retraso y debido a la ampliación del alcance del benchmarking de herramientas, se tuvo listo para la semana 8, y su aprobación se retrasó por temas de coordinación de horarios con el cliente, estando aprobado para la semana 10, así mismo se incluirá dentro de ese mismo entregable el análisis de algoritmos a usar con la herramienta SAP Predictive Analytics, esta actualización del entregable está programada para la semana 4 del 2016-02.

Tabla 25 Hitos del proyecto

Hito del proyecto	Fecha Estimada	Entregables incluidos	Prioridad
Definición del proyecto	Semana 3 2016-1	<ul style="list-style-type: none">• Project Chárter• Documentos de planes de Gestión	Alta
Análisis de enfermedades crónicas	Semana 6 2016-1	<ul style="list-style-type: none">• Informe de análisis de enfermedades crónicas	Media
Análisis de herramientas de Predictive Analytics	Semana 8 2016-1	<ul style="list-style-type: none">• Benchmarking de herramientas de Predictive	Alta

(Benchmarking)		Analytics.	
Diseño de modelo predictivo	Semana 6 2016-2	<ul style="list-style-type: none"> Modelo predictivo que permita diagnosticar la Diabetes Mellitus Tipo 2 soportado por SAP Predictive Analytics 	Alta
Implementación del modelo predictivo	Semana 13 2016-2	<ul style="list-style-type: none"> Prueba del modelo predictivo en ambiente controlado Informe de resultado de implementación Memoria de tesis Paper de proyecto 	Alta

Fuente: Elaboración propia.

GESTION DE LOS RECURSOS HUMANOS

En la siguiente tabla se especifican las funciones desempeñadas por los roles, en algunos casos han variado de acuerdo a la planificación.

Tabla 26 Roles del proyecto

Rol	Miembro	Funciones desempeñadas
Comité de Proyecto	Rosario Villalta Jimmy Armas Oscar Gómez Pedro Shiguihara	<ul style="list-style-type: none"> - Dar los lineamientos generales. - Controlar el cumplimiento del proyecto. - Revisión de presentación parcial y final.
Profesor Gerente IT Consulting	Max Chumpitaz	<ul style="list-style-type: none"> - Dar seguimiento y control al proyecto. - Asesoramiento en la planificación de proyectos. - Asesoramiento frente a posibles problemas.
Alumno Gerente	Eduardo Ulloa	<ul style="list-style-type: none"> - Dar seguimiento y control al proyecto.

Asesor	Jimmy Armas	<ul style="list-style-type: none"> - Asesoramiento en el desarrollo y gestión del proyecto. - Gestionar contactos de importancia para el proyecto.
Profesor Revisor	Andrés Melgar	<ul style="list-style-type: none"> - Asesoramiento en el desarrollo y control de la memoria del proyecto. - Asesoramiento de temas importantes en el proyecto.
Jefe de Proyecto	Diego Ordóñez Barrios	<ul style="list-style-type: none"> - Desarrollar el análisis de enfermedades crónicas y su epidemiología. - Desarrollar el benchmarking de herramientas de Predictive Analytics, identificando la más apropiada para el propósito del proyecto. - Diseño de encuestas para la obtención de información. - Diseño de modelo predictivo del desarrollo de una enfermedad crónica soportado por SAP Predictive Analytics. - Implementar el modelo predictivo en un ambiente controlado para simular el correcto desarrollo del mismo. - Asegurar la calidad de todos los entregables a través de la duración del proyecto. - Asegurar la continuidad del modelo predictivo planteado en un ámbito real de las Instituciones prestadoras de servicios de cobertura de salud.
Jefe de Proyecto	Erick Vizcarra Infantes	<ul style="list-style-type: none"> - Desarrollar de análisis de enfermedades crónicas y su epidemiología. - Desarrollo de benchmarking de herramientas de Predictive Analytics, identificando la más apropiada para el propósito del proyecto. - Diseño de encuestas para la obtención de información. - Diseño un modelo predictivo del desarrollo de una enfermedad crónica soportado por SAP Predictive Analytics.

		<ul style="list-style-type: none"> - Implementar el modelo predictivo en un ambiente controlado para simular el correcto desarrollo del mismo. - Asegurar la calidad de todos los entregables a través de la duración del proyecto. - Asegurar la continuidad del modelo predictivo planteado en un ámbito real de las Instituciones prestadoras de servicios de cobertura de salud.
Recurso QA	Recurso QS	<ul style="list-style-type: none"> - Validación de la calidad de los artefactos. - Validación de la trazabilidad de los entregables y/o artefactos.

Fuente: Elaboración propia.

Se presentaron inconvenientes a lo largo del ciclo 2016-01 con 2 roles, el primero fue el recurso de QS, se nos asignó en primera instancia a uno pero por desempeño este fue cambiado de un momento para otro, esto significó un retraso de 2 días con respecto a nuestra planificación; el segundo inconveniente fue con el revisor de la memoria, se tuvieron problemas de coordinación en las primeras semanas, luego de una reunión se pudieron solucionar los inconvenientes.

GESTION DE LAS COMUNICACIONES

Dentro de la gestión de las comunicaciones se presentó un inconveniente de comunicación con QS, ya que solicitaban distintos requerimientos para la solicitud de recursos que los expresados por el alumno gerente.

La comunicación con todos los roles del proyecto fue en dos tipos, vía correo electrónico, vía telefónica y presencialmente. Solo para los casos del cliente, profesor gerente, alumno gerente y revisor se realizaron gestiones presenciales, vía correo electrónico, vía telefónica y presencial, para los demás roles se realizó una comunicación escrita vía correo electrónico.

GESTION DE LOS RIESGOS

En la siguiente tabla se muestran los riesgos planteados, los nuevos riesgos presentados, identificando cuáles de ellos se materializaron y la estrategia de mitigación ejecutada:

Tabla 27 Analisis de riesgos

Identificación		Análisis		Estrategia de respuesta	
Cód. Riesgo	Descripción del Riesgo	Probabilidad	Materialización	Tipo de Estrategia	Acciones realizadas
IMPECDB-R001	El recurso de QS no cumpla con las funciones que se le delega.	30% Baja	Si	Evitar	<i>Se notificó al líder de línea del recurso para exigir que cumpla con sus responsabilidades mediante seguimiento previo a la fecha de cierre del plazo.</i>
IMPECDB-R002	Insuficientes fuentes disponibles para investigación requerida para el desarrollo del proyecto	10% Muy baja	No	Aceptar - Activamente	<i>No aplica</i>
IMPECDB-R003	Escasa disponibilidad del Cliente para las reuniones pactadas.	30% Baja	No	Aceptar - Activamente	<i>No aplica</i>
IMPECDB-R004	Constante cambio del alcance del proyecto.	30% Baja	Si	Mitigar	<i>Numerosas reuniones con el cliente y distintos especialistas para validar el alcance inicial del proyecto, dando como resultado delimitar el alcance.</i>
IMPECDB-R005	Pérdida de información digital del proyecto.	10% Muy baja	No	Evitar	<i>No aplica</i>
IMPECDB-R006	Retraso en la recepción de observaciones del revisor	50% Media	Si	Mitigar	<i>Al inicio de ciclo se presentaron problemas en la recepción de observaciones de la memoria por parte del revisor, se realizó una reunión presencial para subsanar los inconvenientes.</i>

Fuente: Elaboración propia.

LECCIONES APRENDIDAS

- Tener una mejor coordinación entre los roles del proyecto.
- Todo cambio o modificación en el cronograma debe ser comunicado entre todas las partes involucradas en cumplir o monitorear el mismo.
- Hacer seguimiento constante a los involucrados del proyecto para cumplir con el cronograma.
- Frente a un inconveniente que no tenga solución en primera instancia, escalarlo de manera inmediata con el fin que se logre un común acuerdo.
- Toda reunión o comunicación debe estar debidamente evidenciada y notificadas a las partes interesadas con el fin de hacer notar cualquier discrepancia.
- Tener planes de respaldo en caso la primera opción de solución no sea efectiva o no pueda darse por razones ajenas, esto podría retrasar la ejecución del proyecto.

Conclusiones

- Se desarrolló un análisis profundo sobre las principales enfermedades crónicas aplicadas al contexto peruano y se identificó que la Diabetes Mellitus tipo 2 tenía mayor representatividad por el impacto socioeconómico de la misma, además de ser la enfermedad crónica específica de mayor incidencia y con mayor proyección de crecimiento en el futuro.
- Se identificó a SAP Predictive Analytics como la herramienta más adecuada para soportar un modelo predictivo aplicado al contexto de entidades prestadoras de servicios de salud públicas y privadas a través del análisis de sus características, divididas en aspectos de: Oferta Actual, Estrategia y Presencia en el mercado.
- Se diseñó un modelo predictivo destinado al diagnóstico de la Diabetes Mellitus tipo 2, soportado por SAP Predictive Analytics, el cual es aplicable al contexto de una entidad prestadora de servicios de salud a través de su implementación en el proceso de evaluación médica del paciente, interviniendo desde la fase de obtención de la cita médica hasta la fase de evaluación del paciente durante el desarrollo de la misma. La implementación planteada fue validada a través del juicio de experto del Subdirector del Centro Nacional de Telemedicina – EsSalud, el Dr. Juan Carlos Bartolo Kato quien verificó la viabilidad de la propuesta planteada.
- El modelo planteado tiene el potencial de escalar a un esquema Cloud según la oferta propuesta por SAP Predictive Analytics dentro de sus funcionalidades, dependiendo de la necesidad de procesamiento y alojamiento de datos y es la evolución natural que tendrá el modelo conforme vaya recopilando una mayor cantidad de información.
- Se cumplieron los objetivos del proyecto satisfactoriamente, brindando un modelo predictivo junto con una propuesta para la implementación del mismo a un contexto real, de entidades prestadoras de servicios de cobertura de salud públicas y privadas en el Perú cuya viabilidad ha sido comprobada a través del juicio de experto de un profesional en la materia de medicina y telemedicina.

Recomendaciones

- El presente proyecto permite proponer investigaciones futuras basadas en el desarrollo de algoritmos en lenguaje R que permitan afinar el diagnóstico de la Diabetes Mellitus tipo 2 empleando las mismas o un mayor número de variables, con el fin de brindar mayor precisión que a su vez sea usada en beneficio no solo del diagnóstico inmediato de la enfermedad, sino del estudio de los factores que intervienen en la misma.
- Asimismo, es necesario que cuando se desarrolle un proyecto ligado a la medicina y a la tecnología se cuente con conformidades sobre la viabilidad del mismo por parte de expertos en ambas materias, lo cual garantiza el futuro éxito de la implementación desde ambos frentes.
- Recomendamos plantear, en el desarrollo de modelos predictivos futuros, una estructura de información que pueda ser llenada a través de un proceso implementado de forma real, el cual no requiera desarrollar transformaciones de datos adicionales y además sea soportado por un motor de datos robusto y escalable, garantizando de esta manera la implementación adecuada del proyecto.

Bibliografía

Referencias bibliográficas

- Ali, R., Hussain, J., Siddiqi, M. H., Hussain, M., & Lee, S. (2015). H2RM: A hybrid rough set reasoning model for pre-diction and management of diabetes mellitus. *Sensors*, 15(7), 15921-15951.
- Anand, R., Kirar, V. P. S., & Burse, K. (2012). Data pre-processing and neural network algorithms for diagnosis of type ii diabetes: a survey. *International Journal of Engineer-ing and Advanced Technology (IJEAT)* ISSN, 2249-8958.
- Arslanian, S., Bacha, F., Saad, R. and Gungor, N. (2004). Family History of Type 2 Diabetes Is Associated With Decreased Insulin Sensitivity and an Impaired Balance Between Insulin Sensitivity and Insulin Secretion in White Youth. *Diabetes Care*, 28(1), pp.115-119.
- Bradshaw, D., Steyn, K., Levitt, N., & Nojilana, B. NON-COMMUNICABLE DISEASES. *Women*, 60, 72.
- Chin, C. W. L., San Chia, E. H., Ma, S., Heng, D., Tan, M., Lee, J., ... & Salim, A. (2012). The ARIC predictive model re-liably predicted risk of type II diabetes in Asian populations. *BMC medical research methodology*, 12(1), 1.
- Dougherty, J., Kohavi, R., & Sahami, M. (1995, July). Supervised and unsupervised discretization of continuous features. In *Machine learning: proceedings of the twelfth international conference* (Vol. 12, pp. 194-202).
- Dubin, R. (1969) *Theory Building*. Estados Unidos: The Free Press.
- Eswari, T., Sampath, P., & Lavanya, S. (2015). Predictive Methodology for Diabetic Data Analysis in Big Data. *Proce-dia Computer Science*, 50, 203-208.
- Fayyad, U.M. Piatetsky-Shapiro, G. Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, (37-54).
- Giráldez, R., Aguilar-Ruiz, J. S., & Riquelme, J. C. (2001). Discretización supervisada no paramétrica orientada a la obtención de reglas de decisión. In *Proceedings of the CAEPIA2001* (pp. 53-62).
- Hicks, J. O (1993). *Management Information Systems: A User Perspective*. Estados Unidos: West Publishing. 668
- Hussein, A. S., Omer, W., Li, X., & Ati, M. (2012). Accurate and Reliable Recommender System for Chronic Disease Di-agnosis. In *GLOBAL HEALTH, The First International Con-ference on Global Health Challenges Venice, Italy*.

- Iyer, A., Jeyalatha, S., & Sumbaly, R. (2015). Diagnosis of diabetes using classification mining techniques. arXiv pre-print arXiv:1502.03774.
- Kanter, M., Kris-Etherton, P., Fernandez, M., Vickers, K. and Katz, D. (2012). Exploring the Factors That Affect Blood Cholesterol and Heart Disease Risk: Is Dietary Cholesterol as Bad for You as History Leads Us to Believe?. *Advances in Nutrition: An International Review Journal*, 3(5), pp.711-717.
- Kaur, G., & Chhabra, A. (2014). Improved J48 classification algorithm for the prediction of diabetes. *International Journal of Computer Applications*, 98(22).
- Koklu, M., & Unal, Y. (2013, August). Analysis of a Population of Diabetic Patients Databases with Classifiers. In *Proceedings of World Academy of Science, Engineering and Technology* (No. 80, p. 792). World Academy of Science, Engineering and Technology (WASET).
- Kumari, M., Vohra, R., & Arora, A. (2014). Prediction of Diabetes Using Bayesian Network.
- Lee, J., Keam, B., Jang, E. J., Park, M. S., Lee, J. Y., Kim, D. B., ... & Kwack, K. B. (2011). Development of a predictive model for type 2 diabetes mellitus using genetic and clinical data. *Osong public health and research perspectives*, 2(2), 75-82.
- Liu, H., & Setiono, R. (1995, November). Chi2: Feature selection and discretization of numeric attributes. In *AI (p. 388)*. IEEE.
- Lukmanto, R. B., & Irwansyah, E. (2015). The Early Detection of Diabetes Mellitus (DM) Using Fuzzy Hierarchical Model. *Procedia Computer Science*, 59, 312-319.
- Maimon, Oded. Rokach, Lior. (2005) *Data Mining And Knowledge Discovery Handbook. Introduction to Knowledge Discovery In Databases*. 1-2.
- Mellitus, D. (2005). Diagnosis and classification of diabetes mellitus. *Diabetes care*, 28, S37.
- Sapon, M., Ismail, K., Zainudin, S., & Ping, C. (2011). Diabetes Prediction with Supervised Learning Algorithms of Artificial Neural Network. In *International Conference on Software and Computer Applications*, Kathmandu, Nepal.
- Senn, J.A. (1990) *Information Systems in Management*. Estados Unidos: Wadsworth Publishing Co.
- Shalev-Shwartz, Shai. Ben-David, Shai. (2014) *Understanding Machine Learning: From Theory to Algorithms*. Estados Unidos: Cambridge University Press. (1-21)
- Thirugnanam, M., Kumar, P., Srivatsan, S. V., & Nerlesh, C. R. (2012). Improving the prediction rate of diabetes diagnosis using fuzzy, neural network, case based (FNC) approach. *Procedia Engineering*, 38, 1709-1718.

- Van Poucke, S., Zhang, Z., Schmitz, M., Vukicevic, M., Van-der Laenen, M., Celi, L. A., & De Deyne, C. (2016). Scalable predictive analysis in critically ill patients using a visual open data analysis platform. *PloS one*, 11(1), e0145791.
- Wang, Y., Kung, L., & Byrd, T. A. (2016). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and So-cial Change*.
- Wessler, M. (2013) *Big Data Analytics for Dummies*.

Referencias electrónicas.

- Aamodt, Agnar & Nygard, Mads. (1995) Different roles and mutual dependencies of data, information and knowledge – an AI perspective on their integration. *Data and Knowledge Engineering* (191-222). Recuperado de: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.42.1967&rep=rep1&type=pdf> (Fecha de consulta: 12 de Mayo 2016).
- Advanced Modeling for High Performance Decision Making | Angoss www.angoss.com. (2016). Angoss.com. Recuperado de: <http://www.angoss.com/predictive-analytics-software/software/knowledgestudio/> (Fecha de consulta: 1 de Mayo 2016).
- Alpine Chorus - In a Nutshell - Butler Analytics. (2014). Butler Analytics. Recuperado de: <http://www.butleranalytics.com/alpine-chorus-in-a-nutshell/> (Fecha de consulta: 2 de Mayo 2016).
- Alpine Chorus - Predictive Analytics Today. (2015). Predictive Analytics Today. Recuperado de: <http://www.predictiveanalyticstoday.com/alpine-chorus/> (Fecha de consulta: 1 de Mayo 2016).
- Alpine Data Labs Chorus 4.0 Predictive Analytics. (2014). Business Intelligence & Advanced Analytics. Recuperado de: <http://www.jenunderwood.com/2014/07/30/alpine-data-labs-chorus-4-0/> (Fecha de consulta: 2 de Mayo 2016).
- American Diabetes Association. (2016). Type 2. Recuperado de: <http://www.diabetes.org/diabetes-basics/type-2/?referrer=https://www.google.com.pe/> (Fecha de consulta: 7 de Noviembre 2016).
- An Architecture for Smarter Decisions (1st ed.). Recuperado de: http://brblog.typepad.com/files/insights_architecture_smarter_decisions_2536wp_en.pdf (Fecha de consulta: 2 de Mayo 2016).

- Angoss KnowledgeSEEKER Software - 2016 Reviews & Demo. (2016). Softwareadvice.com. Recuperado de: <http://www.softwareadvice.com/bi/knowledgeseeker-profile/> (Fecha de consulta: 1 de Mayo 2016).
- Angoss Predictive Analytics - Predictive Analytics Today. (2013). Predictive Analytics Today. Recuperado de: <http://www.predictiveanalyticstoday.com/angoss-predictive-analytics/> (Fecha de consulta: 10 de Mayo 2016).
- Angoss Review - Butler Analytics. (2015). Butler Analytics. Recuperado de: <http://www.butleranalytics.com/angoss-review/> (Fecha de consulta: 10 de Mayo 2016).
- Alteryx Analytics Gallery | Alteryx. (2016). Alteryx.com. Recuperado de: <http://www.alteryx.com/products/alteryx-analytics-gallery> (Fecha de consulta: 2 de Mayo 2016).
- Alteryx Analytics - Predictive Analytics Today. (2014). Predictive Analytics Today. Recuperado de: <http://www.predictiveanalyticstoday.com/alteryx-analytics/> (Fecha de consulta: 2 de Mayo 2016).
- Alteryx Designer | Alteryx. (2016). Alteryx.com. Recuperado de: <http://www.alteryx.com/products/alteryx-designer> (Fecha de consulta: 2 de Mayo 2016).
- Alteryx Server | Alteryx. (2016). Alteryx.com. Recuperado de: <http://www.alteryx.com/products/alteryx-server> (Fecha de consulta: 2 de Mayo 2016).
- Azure Machine Learning Frequently Asked Questions (FAQ): Billing, capabilities, limitations, and support. (2016). Azure.microsoft.com. Recuperado de: <https://azure.microsoft.com/en-us/documentation/articles/machine-learning-faq/> (Fecha de consulta: 3 de Mayo 2016).
- Dickey, David. (2012) Introduction to Predictive Modeling with Examples. SAS Global Forum 2012 (1-2). Recuperado de: <http://support.sas.com/resources/papers/proceedings12/337-2012.pdf> (Fecha de consulta: 12 de Mayo 2016).
- Brownlee, J. (2013). A Tour of Machine Learning Algorithms. Machine Learning Mastery. Recuperado de: <http://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/> (Fecha de consulta: 8 de Setiembre 2016).

Confidently Anticipate and Drive Better Business Outcomes. (2016). SAP. Recuperado de: <http://go.sap.com/documents/2015/05/280754e0-247c-0010-82c7-eda71af511fa.html> (Fecha de consulta: 3 de Mayo 2016).

Data Mining Software, Model Development and Deployment, SAS Enterprise Miner. (2016). Sas.com. Recuperado de: http://www.sas.com/en_id/software/analytics/enterprise-miner.html (Fecha de consulta: 3 de Mayo 2016).

Datum Internacional (2013) Los peruanos no cuentan con la información nutricional necesaria para seguir una dieta saludable. Recuperado de: <http://www.datum.com.pe/pdf/HAS.pdf> (Fecha de consulta: 26 de Abril 2016).

Decision Tree Learning. Chapter 3. Recuperado de: <http://www.cs.princeton.edu/courses/archive/spr07/cos424/papers/mitchell-dectrees.pdf> (Fecha de consulta: 8 de Setiembre 2016).

DÍA MUNDIAL DE LA DIABETES. (2016). Minsa.gob.pe. Retrieved 24 June 2016, from <http://www.minsa.gob.pe/portada/Especiales/2014/diabetes>

Diabetes. (2016). Organización Mundial de la Salud. Re-trieved 23 June 2016, from <http://www.who.int/mediacentre/factsheets/fs312/es/>

DMedicina. (2009). Obesidad | DMedicina. Recuperado de: <http://www.dmedicina.com/enfermedades/digestivas/obesidad.html> (Fecha de consulta: 22 de Abril 2016).

Dmg.org. (2016). PMML 4.1 - Time Series Models. Recuperado de: <http://dmg.org/pmml/v4-1/TimeSeriesModel.html> (Fecha de consulta: 8 de Setiembre 2016).

Eckerson, Wayne. (2007) Predictive Analytics. Extending The Value of Your Data Warehousing Investment. The Data Warehousing Institute (5-6). Recuperado de: <http://hosteddocs.ittoolbox.com/sas-predictive-analytics-062508.pdf> (Fecha de consulta: 12 de Mayo 2016).

Elkan, Charles. (2013) Predictive analytics and data mining. (8-20). Recuperado de: http://cseweb.ucsd.edu/~jmcauley/cse255/files/elkan_dm.pdf (Fecha de consulta: 12 de Mayo 2016).

Elston, S. (2016). Data Science in the Cloud with Microsoft Azure Machine Learning and R - O'Reilly Media Free, Live Events. Oreilly.com. Recuperado de: <http://www.oreilly.com/pub/e/3292> (Fecha de consulta: 3 de Mayo 2016).

- Examining the KNIME open source data analytics platform. (2016). SearchBusinessAnalytics. Recuperado de: <http://searchbusinessanalytics.techtarget.com/feature/Examining-the-KNIME-open-source-data-analytics-platform> (Fecha de consulta: 3 de Mayo 2016).
- FICO Big Data Analyzer - Predictive Analytics Today. (2015). Predictive Analytics Today. Recuperado de: <http://www.predictiveanalyticstoday.com/fico-big-data-analyzer/> (Fecha de consulta: 2 de Mayo 2016).
- FICO® Big Data Analyzer | FICO®. (2014). FICO® | Decisions. Recuperado de: <http://www.fico.com/en/products/fico-big-data-analyzer> (Fecha de consulta: 2 de Mayo 2016).
- FICO® Blaze Advisor® Decision Rules Management System | FICO®. (2014). FICO® | Decisions. Recuperado de: <http://www.fico.com/en/products/fico-blaze-advisor-decision-rules-management-system> (Fecha de consulta: 2 de Mayo 2016).
- FICO® Decision Management Suite 2.0 | FICO®. (2014). FICO® | Decisions. Recuperado de: <http://www.fico.com/en/analytics/decision-management-suite> (Fecha de consulta: 2 de Mayo 2016).
- FICO® Model Builder | FICO®. (2014). FICO® | Decisions. Recuperado de: <http://www.fico.com/en/products/fico-model-builder> (Fecha de consulta: 2 de Mayo 2016).
- Frawley, William J, Piatetsky-Shapiro, Gregory & Matheus, Christopher J. (1992) Knowledge discovery in databases: An overview. AI Magazine (58-59). Recuperado de: <http://aaai.org/journals/ai-magazine/article/viewFile/1011/929> (Fecha de consulta: 12 de Mayo 2016).
- Greenberg, H., Raymond, S. U., & Leeder, S. R. (2011). The prevention of global chronic disease: academic public health's new frontier. American journal of public health, 101(8), 1386-1390. Recuperado de: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3134509/pdf/1386.pdf> (Fecha de consulta: 26 de Abril 2016).
- Gualtieri, M. & Curran, R. (2015). Forrester Wave: Big Data Predictive Analytics Solutions, Q2 2015 (1st ed.). Recuperado de: http://www.sas.com/content/dam/SAS/en_us/doc/analystreport/forrester-wave-predictive-analytics-106811.pdf (Fecha de consulta: 1 de Mayo 2016).
- Heart.org. (2016). American Heart Association Recommendations for Physical Activity in Adults. Recuperado de:

- http://www.heart.org/HEARTORG/HealthyLiving/PhysicalActivity/FitnessBasics/American-Heart-Association-Recommendations-for-Physical-Activity-in-Adults_UCM_307976_Article.jsp (Fecha de consulta: 7 de Noviembre 2016).
- How much does SAS cost?. (2016). Quora. Recuperado de: <https://www.quora.com/How-much-does-SAS-cost> (Fecha de consulta: 3 de Mayo 2016).
- IBM - SPSS Analytic Server. (2016). IBM. Recuperado de: <http://www-03.ibm.com/software/products/es/spss-analytic-server> (Fecha de consulta: 2 de Mayo 2016).
- IBM - SPSS Predictive Analytics Enterprise. (2016). IBM. Recuperado de: <http://www-03.ibm.com/software/products/es/spss-predictive-analytics-enterprise> (Fecha de consulta: 3 de Mayo 2016).
- IBM Predictive Analytics - Predictive Analytics Today. (2013). Predictive Analytics Today. Recuperado de: <http://www.predictiveanalyticstoday.com/ibm-predictive-analytics/> (Fecha de consulta: 2 de Mayo 2016).
- IBM Software Estadístico: SPSS Statistics - España. (2016). IBM. Recuperado de: <http://www-01.ibm.com/software/es/analytics/spss/products/statistics/> (Fecha de consulta: 3 de Mayo 2016).
- IBM SPSS Modeler Features. (2016). IBM. Recuperado de: <http://www-01.ibm.com/software/analytics/spss/products/modeler/features.html> (Fecha de consulta: 2 de Mayo 2016).
- IBM SPSS Statistics. (2016). IBM. Recuperado de: https://www.ibm.com/marketplace/cloud/statistical-analysis-and-reporting/us/en-us?S_TACT=M161007W (Fecha de consulta: 2 de Mayo 2016).
- Infobae (2016). Infarto asintomático, o cuando el corazón falla sin dar señales Recuperado de: <http://www.infobae.com/2016/01/30/1786434-infarto-asintomatico-o-cuando-el-corazon-falla-dar-senales/> (Fecha de consulta: 26 de Abril 2016).
- Insight, P. (2016). Reviews of Predixion Insight : Free Pricing & Demos : Business Intelligence Software. Capterra.com. Recuperado de: <http://www.capterra.com/business-intelligence-software/spotlight/139130/Predixion%20Insight/Predixion%20Software> (Fecha de consulta: 3 de Mayo 2016).
- Jenkyns, Tom. Stephenson, Ben. (2013). Fundamentals of Discrete Math for Computer Science (4): Springer

- Jure Leskovec and Anand Rajaraman. Clustering Algorithms. Stanford University
Recuperado de: <https://web.stanford.edu/class/cs345a/slides/12-clustering.pdf>
(Fecha de consulta: 8 de Setiembre 2016).
- Kart, L., Herschel, G., Linden, A., & Hare, J. (2016). Gartner: Magic Quadrant for Advanced Analytics Platforms (1st ed.). Recuperado de: <https://rapidminer.com/resource/leader-gartner-magic-quadrant-advanced-analytics/>
(Fecha de consulta: 1 de Mayo 2016).
- KNIME Analytics Platform. (2016). Datatransformed.com.au. Recuperado de: <http://www.datatransformed.com.au/knime.htm#.VzjuGpHhCzc> (Fecha de consulta: 3 de Mayo 2016).
- Kriegel Hans-Peter, Kröger Peer, Zimek Arthur. (2010). Outlier Detection Techniques. Ludwig-Maximilians-Universität München. The 2010 SIAM International Conference on Data Mining. Recuperado de: <https://www.siam.org/meetings/sdm10/tutorial3.pdf> (Fecha de consulta: 8 de Setiembre 2016).
- Machine Learning | Microsoft Azure. (2016). Azure.microsoft.com. Recuperado de: <https://azure.microsoft.com/services/machine-learning/> (Fecha de consulta: 3 de Mayo 2016).
- Ministerio de Salud del Perú (2012) Situación de las enfermedades crónicas no transmisibles: problema cáncer. Recuperado de: <http://bvs.minsa.gob.pe/local/MINSA/2283.pdf>
(Fecha de consulta: 26 de Abril 2016).
- Natureofcode.com. (2016). The Nature of Code. Recuperado de: <http://natureofcode.com/book/chapter-10-neural-networks/> (Fecha de consulta: 8 de Setiembre 2016).
- New England Journal of Medicine. (2016). Prevention of Type 2 Diabetes Mellitus by Changes in Lifestyle among Subjects with Impaired Glucose Tolerance “ NEJM. Recuperado de: <http://www.nejm.org/doi/full/10.1056/NEJM200105033441801>
(Fecha de consulta: 7 de Noviembre 2016).
- OMS. (2016). OMS | Enfermedades crónicas. [online] Recuperado de: http://www.who.int/topics/chronic_diseases/es/ (Fecha de consulta: 10 de Marzo 2016).
- OPINIÓN: Epidemia de diabetes mellitus en Perú | Diario Médico Perú. (2016). Diariomedico.pe. Retrieved 24 June 2016, from <http://diariomedico.pe/?p=127>

Organización Mundial de la Salud (2014) Informe Sobre La Situación Mundial De Las Enfermedades No Transmisibles. Recuperado de: http://apps.who.int/iris/bitstream/10665/149296/1/WHO_NMH_NVI_15.1_spa.pdf (Fecha de consulta: 26 de Abril 2016).

Organización Mundial de la Salud (2014) Perfil Peruano. Recuperado de: http://www.who.int/nmh/countries/per_es.pdf (Fecha de consulta: 22 de Abril 2016).

Organización Mundial de la Salud (2014) Perfil Peruano: Cáncer. Recuperado de: http://www.who.int/cancer/country-profiles/per_es.pdf?ua=1 (Fecha de consulta: 26 de Abril 2016).

Organización Mundial de la Salud (2016) Perfiles de Países: Perú 2016. Recuperado de: http://www.who.int/diabetes/country-profiles/per_es.pdf (Fecha de consulta: 22 de Abril 2016).

Organización Mundial de la Salud. (2016). 10 datos sobre la obesidad. Recuperado de: <http://www.who.int/features/factfiles/obesity/facts/es/index2.html> (Fecha de consulta: 22 de Abril 2016).

Organización Mundial de la Salud. (2016). Cáncer. Recuperado de: <http://www.who.int/mediacentre/factsheets/fs297/es/> (Fecha de consulta: 22 de Abril 2016).

Organización Mundial de la Salud. (2016). Diabetes. Recuperado de: <http://www.who.int/mediacentre/factsheets/fs312/es/> (Fecha de consulta: 22 de Abril 2016).

Organización Mundial de la Salud. (2016). Enfermedades cardiovasculares. Recuperado de: <http://www.who.int/mediacentre/factsheets/fs317/es/> (Fecha de consulta: 22 de Abril 2016).

Organización Mundial de la Salud. (2016). Enfermedades crónicas. Recuperado de: http://www.who.int/topics/chronic_diseases/es/ (Fecha de consulta: 22 de Abril 2016).

Organización Mundial de la Salud. (2016). Integrated chronic disease prevention and control. Recuperado de: http://who.int/chp/about/integrated_cd/en/ (Fecha de consulta: 22 de Abril 2016).

Organización Mundial de la Salud. (2016). Obesidad y sobrepeso. Recuperado de: <http://www.who.int/mediacentre/factsheets/fs311/es/> (Fecha de consulta: 22 de Abril 2016).

Organización Mundial de la Salud. (2016). Programa de Diabetes de la OMS. Recuperado de: <http://www.who.int/diabetes/es/> (Fecha de consulta: 22 de Abril 2016).

Otexts.org. (2016). 7.1 Simple exponential smoothing. Recuperado de: <https://www.otexts.org/fpp/7/1> (Fecha de consulta: 8 de Setiembre 2016).

Peru21. (2013). Cifras sobre el sedentarismo en el Perú. Recuperado de: <http://peru21.pe/vida21/cifras-sobre-sedentarismo-peru-2151830> (Fecha de consulta: 22 de Abril 2016).

Powerful, Market-Leading Decision Trees for Data Driven Decisions | Angoss www.angoss.com. (2016). Angoss.com. Recuperado de: <http://www.angoss.com/predictive-analytics-software/software/knowledgeseeker/> (Fecha de consulta: 1 de Mayo 2016).

Precios - Aprendizaje automático | Microsoft Azure. (2016). [Azure.microsoft.com](https://azure.microsoft.com/es-es/pricing/details/machine-learning/). Recuperado de: <https://azure.microsoft.com/es-es/pricing/details/machine-learning/> (Fecha de consulta: 3 de Mayo 2016).

Predictive Analytics - IBM Analytics. (2016). IBM. Recuperado de: <http://www.ibm.com/analytics/us/en/technology/predictive-analytics/> (Fecha de consulta: 3 de Mayo 2016).

Predictive analytics with simplified data mining | Statistica | Dell Software. (2016). [Software.dell.com](http://software.dell.com/products/statistica/). Recuperado de: <http://software.dell.com/products/statistica/> (Fecha de consulta: 2 de Mayo 2016).

Predixion Insight 4.5 Enables Analytics to Run on Device, on Gateway and in the Cloud - Predictive Analytics Today. (2015). Predictive Analytics Today. Recuperado de: <http://www.predictiveanalyticstoday.com/predixion-insight-4-5-enables-analytics-to-run-on-device-on-gateway-and-in-the-cloud/> (Fecha de consulta: 3 de Mayo 2016).

Predixion Insight™. (2016). [Insight.predixionsoftware.com](https://insight.predixionsoftware.com/overview_45). Recuperado de: https://insight.predixionsoftware.com/overview_45 (Fecha de consulta: 3 de Mayo 2016).

PROJECT MANAGEMENT INSTITUTE (2004) Guía de los Fundamentos de la Dirección de Proyectos (Guía del PMBOK®) Tercera Edición. Recuperado de: <http://gio.uniovi.es/documentos/software/GUIAPMBok.pdf> (Fecha de consulta: 10 de Marzo 2016).

RapidMiner | CrunchBase. (2016). Crunchbase.com. Recuperado de: <https://www.crunchbase.com/organization/rapidminer#/entity> (Fecha de consulta: 3 de Mayo 2016).

Revilla, L. (2013). Situación de la vigilancia de diabetes en el Perú, al I semestre de 2013. Bol Epidemiol, 22(39), 825-8.

SAP (2015) SAP Predictive Analytics 2.1. Expert Analytics User Guide. Recuperado de: http://help.sap.com/businessobject/product_guides/pa21/en/pa21_expert_user_en.pdf (Fecha de consulta: 8 de Setiembre 2016).

SAP (2016) SAP Predictive Analytics 3.0. Classification, Regression, Segmentation and Clustering Scenarios. Automated Analytics User Guide. Recuperado de: http://help.sap.com/businessobject/product_guides/pa30/en/pa30_classification_user_en.pdf (Fecha de consulta: 8 de Setiembre 2016).

SAS Enterprise Miner and SAS Factory Miner. (2016). Support.sas.com. Recuperado de: <http://support.sas.com/software/products/miner/> (Fecha de consulta: 3 de Mayo 2016).

SAS Enterprise Miner Key Features. (2016). Support.sas.com. Recuperado de: http://support.sas.com/software/products/miner/miner_key_features.html (Fecha de consulta: 3 de Mayo 2016).

Sas. (2016). What is Big Data and why it matters. Sas.com. Recuperado de: http://www.sas.com/en_us/insights/big-data/what-is-big-data.html (Fecha de consulta: 10 Marzo 2016).

Soft, B. (2013). ¿Que es Big Data? - Business Software. Business Software. Recuperado de: <http://www.businesssoftware.net/que-es-big-data> (Fecha de consulta: 10 de Marzo 2016).

STATISTICA - Predictive Analytics Today. (2013). Predictive Analytics Today. Recuperado de: <http://www.predictiveanalyticstoday.com/statistica/> (Fecha de consulta: 2 de Mayo 2016).

STATISTICA | Data Mining Software. (2016). Statsoft.com. Recuperado de: <http://www.statsoft.com/Products/STATISTICA/Data-Miner> (Fecha de consulta: 2 de Mayo 2016).

STATISTICA | Decisioning Platform. (2016). Statsoft.com. Recuperado de: <http://www.statsoft.com/Products/STATISTICA/Decisioning-Platform> (Fecha de consulta: 2 de Mayo 2016).

- STATISTICA | Predictive Analytics. (2016). Statsoft.com. Recuperado de: <http://www.statsoft.com/Solutions/Cross-Industry/Predictive-Analytics> (Fecha de consulta: 2 de Mayo 2016).
- Tan, P. N., Michael, S., & Kumar, V. (2005). Chapter 6. association analysis: Basic concepts and algorithms. Introduction to Data Mining. Addison-Wesley. Recuperado de: <https://www-users.cs.umn.edu/~kumar/dmbook/ch6.pdf> (Fecha de consulta: 8 de Setiembre 2016).
- The Alpine Data Platform (1st ed.). Recuperado de: <http://alpinedata.com/wp-content/uploads/2016/04/Alpine-Data-Overview.pdf> (Fecha de consulta: 2 de Mayo 2016).
- Unparalleled Voice of the Customer Insights with Text Analytics www.angoss.com. (2016). Angoss.com. Recuperado de: <http://www.angoss.com/predictive-analytics-software/software/knowledgereader/> (Fecha de consulta: 1 de Mayo 2016).
- What is Microsoft Azure (Windows Azure)? - Definition from WhatIs.com. (2016). SearchCloudComputing. Recuperado de: <http://searchcloudcomputing.techtarget.com/definition/Windows-Azure> (Fecha de consulta: 3 de Mayo 2016).
- Why only one of the 5 Vs of big data really matters. (2016). IBM Big Data & Analytics Hub. Recuperado de: <http://www.ibmbigdatahub.com/blog/why-only-one-5-vs-big-data-really-matters> (Fecha de consulta: 1 de Mayo 2016).

Anexos

Anexo 1: Acta de Validación – Experto en Medicina

Proyecto:	Modelo Predictivo para el diagnóstico de la Diabetes Mellitus Tipo 2 soportado por SAP Predictive Analytics		
	Fecha:	14/11/2016	N°:	1

Acta de Reunión - Validación N° 001

Sección 1. Información General

Código y Nombre del Proyecto	Fecha de reunión	Hora inicio	Hora fin
Modelo Predictivo para el diagnóstico de la Diabetes Mellitus Tipo 2 soportado por SAP Predictive Analytics	14/11/2016	9:30 a.m.	10:30 a.m.
Elaborado por:			
Diego Ordoñez Barrios Erick Vizcarra Infantes			
Asistentes	Cargo		Asistió
Diego Ordoñez Barrios	Jefe de Proyecto		No
Erick Vizcarra Infantes	Jefe de Proyecto		Si
Juan Carlos Bartolo Kato	Subdirector de regulación, Iniciativas y Servicios en Telesalud – Centro Nacional de Telemedicina de EsSalud		Si

Sección 2. Agenda

Nro.	Tema
1	Validación del modelo diseñado
2	Aprobación sobre la implementación del modelo predictivo en un ambiente controlado con datos recopilados.

Sección 3. Detalle de lo Tratado

Nro. Tema	Descripción
1	Se presentó el modelo al experto en medicina considerando la viabilidad de la implementación del mismo en un ambiente real.
2	Se presentó el proceso de recopilación de información para el modelo predictivo, las transformaciones y clasificaciones realizadas, las pruebas de predicción y los resultados obtenidos.

Proyecto:	Modelo Predictivo para el diagnóstico de la Diabetes Mellitus Tipo 2 soportado por SAP Predictive Analytics		
	Fecha:	14/11/2016	N°:	1

Sección 4. Acuerdos

Nro. Acuerdo	Acuerdo
1	El experto en medicina aprueba la versión final del modelo, manifestando su conformidad con lo establecido y lo abordado en el mismo.
2	El experto en medicina, expresa su conformidad sobre el procedimiento seguido para la recopilación de información, análisis y predicción de la misma y además la viabilidad de sus resultados.

Sección 5. Tareas

Nro. Tarea	Descripción de la Tarea	Responsable	Fecha
1	No Aplica	-	-

JUAN CARLOS BARTOLO KATO MEDICO CIRUJANO C.M.P. 44802 TELF: 9792-6123 Juan Carlos Bartolo Kato	Erick Vizcarra	Diego Ordoñez
--	--	--

Anexo 2: Acta de Validación – Experto en Tecnología

Proyecto:	Modelo Predictivo para el diagnóstico de la Diabetes Mellitus Tipo 2 soportado por SAP Predictive Analytics		
	Fecha:	23/11/2016	N°:	2

Acta de Reunión - Validación N° 002

Sección 1. Información General

Código y Nombre del Proyecto	Fecha de reunión	Hora inicio	Hora fin
Modelo Predictivo para el diagnóstico de la Diabetes Mellitus Tipo 2 soportado por SAP Predictive Analytics	23/11/2016	1:30 p.m.	2:30 p.m.
Elaborado por:			
Diego Ordoñez Barrios Erick Vizcarra Infantes			
Asistentes	Cargo		Asistió
Diego Ordoñez Barrios	Jefe de Proyecto		Si
Erick Vizcarra Infantes	Jefe de Proyecto		Si
Andres Melgar Sasieta	PhD. en Ingeniería y Gestión del Conocimiento. Universidad Federal de Santa Catarina, Brasil.		Si

Sección 2. Agenda

Nro.	Tema
1	Validación del modelo desarrollado.

Sección 3. Detalle de lo Tratado

Nro. Tema	Descripción
1	Se presentó el modelo al experto en tecnología considerando el diseño del mismo, los experimentos realizados, la viabilidad de la implementación del mismo en un ambiente real y la continuidad/escalabilidad del proyecto.

Sección 4. Acuerdos

Nro. Acuerdo	Acuerdo
1	El experto en tecnología aprueba la versión final del modelo, manifestando su conformidad con lo establecido y lo abordado en el mismo.

Proyecto:	Modelo Predictivo para el diagnóstico de la Diabetes Mellitus Tipo 2 soportado por SAP Predictive Analytics		
	Fecha:	23/11/2016	N°:	2

Sección 5. Tareas

Nro. Tarea	Descripción de la Tarea	Responsable	Fecha
1	No Aplica	-	-

Andres Melgar Sasieta	Erick Vizcarra	Diego Ordoñez

Anexo 3: Acta de Validación – Cliente del Proyecto

Proyecto:	Modelo Predictivo para el diagnóstico de la Diabetes Mellitus Tipo 2 soportado por SAP Predictive Analytics		
	Fecha:	15/11/2016	N°:	3

Acta de Reunión - Validación N° 003

Sección 1. Información General

Código y Nombre del Proyecto	Fecha de reunión	Hora inicio	Hora fin
Modelo Predictivo para el diagnóstico de la Diabetes Mellitus Tipo 2 soportado por SAP Predictive Analytics	15/11/2016	6:00 p.m.	6:30 p.m.
Elaborado por:			
Diego Ordoñez Barrios Erick Vizcarra Infantes			
Asistentes	Cargo		Asistió
Diego Ordoñez Barrios	Jefe de Proyecto		Si
Erick Vizcarra Infantes	Jefe de Proyecto		Si
Jimmy Armas Aguirre	Cliente		Si

Sección 2. Agenda

Nro.	Tema
1	Validación del modelo diseñado

Sección 3. Detalle de lo Tratado

Nro. Tema	Descripción
1	Se presentó el modelo al cliente considerando la viabilidad de la implementación del mismo en un ambiente real.
2	Se presentó el proceso de recopilación de información para el modelo predictivo, las transformaciones y clasificaciones realizadas, las pruebas de predicción y los resultados obtenidos.

Sección 4. Acuerdos

Nro. Acuerdo	Acuerdo
1	El cliente aprueba la versión final del modelo, manifestando su conformidad con lo establecido y lo abordado en el mismo.

Proyecto:	Modelo Predictivo para el diagnóstico de la Diabetes Mellitus Tipo 2 soportado por SAP Predictive Analytics		
	Fecha:	15/11/2016	N°:	3

Sección 5. Tareas

Nro. Tarea	Descripción de la Tarea	Responsable	Fecha
1	No Aplica	-	-

Jimmy Armas Aguirre	Erick Vizcarra	Diego Ordoñez
--	---	--