



**FACULTAD DE POSTGRADO
TESIS DE POSTGRADO**

**IMPLEMENTACIÓN EN LA NUBE DE
UN MOTOR ESTADÍSTICO PARA
DETECCIÓN TEMPRANA DE
ESTUDIANTES EN RIESGO DE
DESERCIÓN**

SUSTENTADO POR:

UAYEB CABALLERO RODRÍGUEZ

**PREVIA INVESTIDURA AL TÍTULO DE
MÁSTER EN GESTIÓN DE TECNOLOGÍAS DE
INFORMACIÓN**

TEGUCIGALPA, FM,

HONDURAS, C.A.

ABRIL 2019

UNIVERSIDAD TECNOLÓGICA CENTROAMERICANA

UNITEC

FACULTAD DE POSTGRADO

AUTORIDADES UNIVERSITARIAS

RECTOR

MARLON BREVÉ REYES

SECRETARIO GENERAL

RÓGER MARTÍNEZ MIRALDA

VICERRECTORA ACADÉMICA

DESIRE TEJADA CALVO

DECANA DE LA FACULTAD DE POSTGRADO

CLAUDIA MARÍA CASTRO VALLE

**TRABAJO PRESENTADO EN CUMPLIMIENTO DE LOS
REQUISITOS EXIGIDOS PARA OPTAR AL TÍTULO DE**

**MÁSTER EN GESTIÓN DE TECNOLOGÍAS DE
INFORMACIÓN**

ASESOR
JORGE RAÚL MARADIAGA CHIRINOS

MIEMBROS DE LA TERNA:

CARLOS PEREZ
ALFONSO ALFONSO
RAFAEL ALBERTI

DERECHOS DE AUTOR

© Copyright 2019
Uayeb Caballero Rodríguez

Todos los derechos son reservados.

**AUTORIZACIÓN DEL AUTOR(ES) PARA LA CONSULTA,
REPRODUCCIÓN PARCIAL O TOTAL Y PUBLICACIÓN
ELECTRÓNICA DEL TEXTO COMPLETO DE TESIS DE POSTGRADO**

Señores

**CENTRO DE RECURSOS PARA EL APRENDIZAJE Y LA INVESTIGACIÓN (CAP)
UNIVERSIDAD TECNOLÓGICA CENTROAMERICANA (UNITEC)**

Estimados Señores:

Yo, Uayeb Caballero, de Tegucigalpa autor del trabajo de postgrado titulado: IMPREMENTACIÓN EN LA NUBE DE UN MOTOR ESTADÍSTICO PARA DETECCIÓN TEMPRANA DE ESTUDIANTES EN RIESGO DE DESERCIÓN, presentado y aprobado en Abril 2019, como requisito previo para optar al título de máster en GESTIÓN DE TECNOLOGÍAS DE INFORMACIÓN y reconociendo que la presentación del presente documento forma parte de los requerimientos establecidos del programa de maestrías de la Universidad Tecnológica Centroamericana (UNITEC), por este medio autorizo a las Bibliotecas de los Centros de Recursos para el Aprendizaje y la Investigación (CAP) de UNITEC, para que con fines académicos puedan libremente registrar, copiar o utilizar la información contenida en él, con fines educativos, investigativos o sociales de la siguiente manera:

- 1) Los usuarios puedan consultar el contenido de este trabajo en las salas de estudio de la biblioteca y/o la página Web de la Universidad.
- 2) Permita la consulta y/o la reproducción a los usuarios interesados en el contenido de este trabajo, para todos los usos que tengan finalidad académica, ya sea en formato CD o digital desde Internet, Intranet, etc., y en general en cualquier otro formato conocido o por conocer.

De conformidad con lo establecido en los artículos 9.2, 18, 19, 35 y 62 de la Ley de Derechos de Autor y de los Derechos Conexos; los derechos morales pertenecen al autor y son personalísimos, irrenunciables, imprescriptibles e inalienables. Asimismo, el autor cede de forma ilimitada y exclusiva a UNITEC la titularidad de los derechos patrimoniales. Es entendido que cualquier copia o reproducción del presente documento con fines de lucro no está permitida sin previa autorización

por escrito de parte de UNITEC.

En fe de lo cual se suscribe el presente documento en la ciudad de Tegucigalpa, a los 1 días del mes de abril del año 2019

Uayeb Caballero Rodríguez

11443024

* La autorización firmada se encuentra adjunta a mí expediente



FACULTAD DE POSTGRADO

IMPLEMENTACIÓN EN LA NUBE DE UN MOTOR ESTADÍSTICO PARA DETECCIÓN TEMPRANA DE ESTUDIANTES EN RIESGO DE DESERCIÓN

UAYEB CABALLERO

RESUMEN

Esta investigación es para todos aquellos lectores que estén interesados en cómo realizar implementaciones en la nube relacionadas con las tecnologías de la información. El problema que estamos abordando es sobre cómo identificar los factores de riesgo para los estudiantes con una probabilidad de abandonar el próximo término o período académico. UNITEC, como universidad responsable de la educación profesional a nivel de maestría, tiene que identificar a todos aquellos estudiantes con perfiles en los que corren el riesgo de omitir al menos un período académico. Es importante que los jefes de carrera puedan identificar rápidamente a todos aquellos conductores que clasifique a un estudiante como un riesgo de abandono escolar y aproveche esta información para mejorar la toma de decisiones. Mi enfoque se centra en cómo la inteligencia empresarial debe utilizar temas como la ciencia de datos, la minería de datos, el análisis exploratorio de datos como parte de una solución analítica. Es por eso que este mismo documento será una guía clara sobre cómo realizar una implementación correcta de la solución de análisis avanzado. Incluirá un fuerte componente estadístico porque las nuevas tendencias de implementación inteligente utilizan mucho aprendizaje automático y, antes de realizar una implementación de esta, debemos realizar una completa gestión y limpieza de datos. Además, este documento propone cómo la ciencia de datos se relaciona rápidamente con los métodos de investigación tradicionales. Trataremos paso a paso todos los niveles sobre cómo producir nuevos conocimientos y cómo la información tecnológica juega un papel importante en el nivel de la solución de la aplicación.

Palabras claves: UNITEC / Maquinas de aprendizaje / nube / Tecnologías de la Información / Riesgo de deserción.



GRADUATE SCHOOL

IMPLEMENTATION IN THE CLOUD OF A STATISTICAL ENGINE FOR EARLY DETECTION OF STUDENTS AT RISK OF DROPOUT

UAYEB CABALLERO

ABSTRACT

This research is for all those readers that are interested in how to do implementations into the cloud related to information technologies. The problem that we are addressing is about how to identify risk factors for students with a likelihood of dropout the next academic term or period. UNITEC as university responsible for educating professional at a master level has to identify all those students with profiles where they are at risk to skip at least one academic period, it is an important thing that career heads will be able to identify quickly all those drivers that classify a student as a risk of dropout and take advantage of this information to improve decision making. The focus of this one is on how business intelligence should be using topics as data science, data mining, exploratory data analysis as a part of an analytical solution. That's why this same document will be a clear guideline in how to do a right roadmap implementation of advanced analytics solution. It will include a strong statistical component because the new trends of smart implementation use a lot of machine learning and before to do an implementation of this one, we must do a complete data wrangling and cleaning. As well this document proposed how data science is quickly related to traditional research methods. We will be addressing step by step the whole levels in how to produce new knowledge and how technology information plays an important role in the application solution level.

Key words: UNITEC / Machine Learning / cloud / Information Technology / Dropout risk.

DEDICATORIA

Dedico este trabajo a mi eterna inspiración la sonrisa de mi madre y mi padre y el ejemplo que han sido mis hermanos en mi vida. Todos mis logros son el fruto de esas noches de desvelo de mis padres, eternamente agradecido con ellos.

AGRADECIMIENTO

Agradezco en primer lugar a Dios y la Virgen maría por darme la fuerza de terminar y alcanzar mis metas. También agradezco mucho a todas las personas que me apoyaron para terminar este arduo trabajo. Mis padres que siempre confían y se preocupan por mí. Mi hermana por acompañarme todas esas noches de desvelo mientras trabajábamos los dos. Mi hermano por ser mi inspiración y un ejemplo sobre lo que es dedicación y a mi novia por esta conmigo en todos esos cafés mientras trabajaba y se tomaba el tiempo de escuchar todas mis locas ideas.

Un agradecimiento especial a las autoridades de UNITEC por compartir los datos de los estudiantes de posgrado y realizar este análisis y Laureate Education Inc. Por prestarme los espacios tecnológicos y poder realizar esta pequeña investigación.

ÍNDICE DE CONTENIDO

CAPÍTULO I. PLANTEAMIENTO DE LA INVESTIGACIÓN	1
1.1 Introducción	1
1.2 Antecedentes	2
1.3 Definición del problema.....	3
1.3.1 Enunciado del problema	3
1.3.2 Formulación del problema.....	3
1.3.3 Preguntas de investigación	4
1.4 Objetivos del proyecto.....	4
1.5 Justificación	5
CAPÍTULO II. MARCO TEÓRICO	7
2.1 La importancia de la retención en los negocios de educación.....	7
2.1.1 Los negocios hoy en día	7
2.1.2 La importancia de la retención en la educación superior	9
2.1.3 El diseño de las intervenciones mediante el uso de los factores de riesgo	11
2.1.3.1 Características Individuales	13
2.1.3.2 Características Académicas	13
2.1.3.3 Características Socioeconómicas.....	15
2.1.3.4 Características institucionales	16
2.2 Data-Driven	16
2.2.1 Democratización de la información	20
2.2.2 Como saber si se manejan las mejores prácticas de data-Driven	21
2.3 Ciencia de datos en instituciones educativas.....	22
2.3.1 Que es ciencia de datos	23
2.3.2 Que hace un científico de datos y por qué son tan necesarios hoy en día	23
2.3.3 El manejo y definición de las métricas	25
2.3.4 Como está compuesto un proyecto de ciencia de datos	27
2.4 Análisis Exploratorio de datos.....	31
2.4.1 Limpieza de datos.	31
2.4.2 Ingeniería de variables	35
2.4.2.1 Transformaciones.....	35
2.4.2.2 Imputaciones.....	36

2.4.2.3 Agregaciones	36
2.4.2.4 Estandarizaciones.....	38
2.5 Modelado estadístico	42
2.5.1 Selección de variables	42
2.5.1.1 La mejor selección de subconjuntos	43
2.5.1.2 Procedimientos de selección de modelos paso a paso	43
2.5.2 Selección de modelos estadísticos.....	45
2.5.2.1 Máquinas de aprendizaje supervisadas	45
2.6 Como llevar un modelo estadístico a la nube	48
2.6.1 Software como un servicio (SaaS)	48
2.6.2 Solución basada en SaaS	49
2.6.3 Microsoft Azure Machine Learning Services	51
2.6.4 Azure Data Factory Versión 2.....	51
2.6.5 Azure Databricks	52
2.6.6 Docker administración de contenedores	52
CAPÍTULO III. METODOLOGÍA.....	54
3.1 Planteamiento de la investigación	54
3.2 Operatividad de variables	56
3.3 Enfoque de la investigación	59
3.4 Alcance de la investigación	59
3.5 Diseño de la investigación.....	60
3.6 Herramientas e Instrumentos	61
CAPÍTULO IV. RESULTADOS	71
4.1 Análisis exploratorio de datos de registro	71
4.2 Análisis de información financiera	88
4.3 Análisis exploratorio información de tickets	103
4.4 Análisis Exploratorio de la plataforma de Blackboard	108
4.5 Elaboración de modelos estadísticos.....	114
4.5.1 Modelo para nuevos estudiantes en semana dos.....	115
4.5.2 Modelo para nuevos estudiantes semana tres a cinco.....	118
4.5.3 Modelo para nuevos estudiantes semana tres a cinco.....	120
4.5.4 Resumen de todos los modelos.	122

4.5.5 Análisis de estudiantes de primer año	123
4.6 Migración de modelo a la nube.....	125
CAPITULO V CONCLUSIONES DE INVESTIGACIÓN	129
5.1 Conclusiones.....	129
5.2 Recomendaciones.....	130
BIBLIOGRAFÍA.....	132
GLOSARIO DE TÉRMINOS.....	134

ÍNDICE DE FIGURAS

Figura 1 Modelo DMBOOK.....	18
Figura 2 Epicycles of Analysis.....	28
Figura 3 Tipos de variables.....	31
Figura 4 Ejemplo Boxplot.....	33
Figura 5 Ejemplo de transformación.....	35
Figura 6 ejemplo de lista de números.....	39
Figura 7 Resumen de una variable numérica.....	39
Figura 8 Boxplot variable numérica sin estandarizar.....	40
Figura 9 Información percentiles.....	40
Figura 10 Resumen de variable estandarizada.....	41
Figura 11 Boxplot variable estandarizada.....	41
Figura 12 Algorithm Best subset selection.....	43
Figura 13 Forward stepwise (Adelante paso a paso).....	44
Figura 14 Backward stepwise (Hacia atrás paso a paso).....	44
Figura 15 Arquitectura usando componentes en la nube.....	51
Figura 16 Distribución Estudiantes por carrera.....	72
Figura 17 Carreras excluida.....	73
Figura 18 Porcentaje de deserción por periodo académico.....	73
Figura 19 Niveles por variable categóricas.....	74
Figura 20 Distribución Departamentos de Honduras sin limpiar.....	75
Figura 21 Deserción por departamento (Valores limpios).....	75
Figura 22 Deserción por zona geográfica del país.....	76
Figura 23 Número de estudiantes sin clasificación de zona geográfica.....	76
Figura 24 Deserción por zona geográfica - datos limpios.....	77
Figura 25 Porcentaje por tipo de estudiantes.....	77
Figura 26 Porcentaje de deserción por tipo de estudiantes.....	78
Figura 27 Distribución de estudiantes agrupados.....	79
Figura 28 Porcentaje de deserción por tipo de estudiantes agrupados.....	79
Figura 29 Distribución mostrada desde Microsoft Azure.....	80
Figura 30 ETL Entrenamiento para nuevos estudiantes Inicio del Periodo.....	81
Figura 31 Validación del modelo de inicio de ciclo.....	82
Figura 32 Variables más importantes modelo inicio de semestre.....	83
Figura 33 Distribución predicción modelo inicio de semestre.....	84
Figura 34 Validación por deciles modelo inicio de semestre.....	85
Figura 35 Distribución modelo para estudiantes continuos inicio de semestre.....	86
Figura 36 Validaciones modelo inicio de semestre, estudiantes continuos.....	87
Figura 37 Validación por decil modelos inicio de semestre estudiantes continuos.....	87
Figura 38 Modelo de datos origina de información financiera.....	88
Figura 39 Comportamiento de pago semana 1	89
Figura 40 Comportamiento de pago semana 4	90

Figura 41 Comportamiento de pago semana 8	91
Figura 42 Comportamiento de pago semana 11.	92
Figura 43 Modelo de datos de información financiera transformada.	93
Figura 44 Boxplot número de pagos.....	93
Figura 45 Boxplot número de pagos tratado.....	94
Figura 46 Nuevo resumen de variables financieras tratadas.....	94
Figura 47 ETL entrenamiento de modelo financiero.....	96
Figura 48 Ajuste de potenciación mediante arboles de decisiones.....	97
Figura 49 Validación modelo de información financiera.	98
Figura 50 Modelo financiero utilizado en otras semanas.	99
Figura 51 Modelo financiero utilizado en otras semanas.	100
Figura 52 Precisión modelo financiero con variable de respuesta semana 4.....	101
Figura 53 Precisión modelo financiero con variable de respuesta semana 8.....	102
Figura 54 Precisión modelo financiero con variable de respuesta semana 10.....	103
Figura 55 Deserción por campus de estudiantes con tickets.....	104
Figura 56 Tickets no contestados por semana y deserción.	105
Figura 57 Comportamiento de deserción estudiantes sin respuesta a tickets.	105
Figura 58 Deserción por clasificación de tickets modulo uno.....	106
Figura 59 Deserción por clasificación de tickets modulo dos.	107
Figura 60 Deserción por número de tickets.	107
Figura 61 Suma de logins por semana.	108
Figura 62 Mediana de logins por semana.	109
Figura 63 Mediana de logins por semana y desertores.	110
Figura 64 Mediana de logins por semana y desertores acumulado.	111
Figura 65 Mediana de accesos por semana y desertores acumulado, modulo uno.....	112
Figura 66 Mediana de accesos por semana y desertores acumulado, modulo cuatro.....	113
Figura 67 Comportamiento de estudiantes que desertan con sus interacciones.....	113
Figura 68 Propuesta final de modelos a construir.....	114
Figura 69 Curva ROC Nuevos estudiantes semana dos.....	116
Figura 70 Pesos, modelo semana dos estudiantes nuevos.	117
Figura 71 Precisión mediante deciles, modelo semana dos estudiantes nuevos.....	117
Figura 72 variables importantes modelo semana dos estudiantes nuevos.	118
Figura 73 Precisión mediante deciles, modelo semana tres a cinco estudiantes nuevos.	119
Figura 74 variables importantes modelo semana tres a cinco estudiantes nuevos.	120
Figura 75 Curva ROC modelo de fin de ciclo estudiantes nuevos.	121
Figura 76 validación deciles modelo fin de semestre estudiantes nuevos.	122
Figura 77 Explicación de factores de riesgo académicos para estudiantes nuevos	124
Figura 78 Comportamiento financiero de estudiantes que tienen bajo compromiso.....	124
Figura 79 Estudiantes con buen compromiso y alto promedio de respuesta de tickets.....	125
Figura 80 Modelos estadísticos exportados en Microsoft Azure.....	126
Figura 81 Configuración de web service.	127
Figura 82 Fábrica de datos.....	127

ÍNDICE DE TABLAS

Tabla 1 Niveles de madurez de la gobernabilidad de la información (Saffirio, 2017)	20
Tabla 2 funciones de los científicos de datos.....	23
Tabla 3 Ejemplo HAVE para definir KPIs	27
Tabla 4 Distribución de Estudiantes por carrera.....	36
Tabla 5 Tabla agregada en base a una regla de negocio.....	37
Tabla 6 Ejemplo de correlación con variable de respuesta.....	37
Tabla 7 Operación de variables- Estudiantes nuevos	56
Tabla 8 Operación de variables - factores de riesgo	57
Tabla 9 Operación de variables - Calculo de deciles	58

CAPÍTULO I. PLANTEAMIENTO DE LA INVESTIGACIÓN

1.1 Introducción

El presente trabajo de tesis condensa todo el fundamento científico y tecnológico para poder implementar y poner en producción un motor estadístico empleando tecnologías en la nube integrándose con distintos tipos de servicios web brindando soluciones a problemas específicos de negocios.

UNITEC es una de las instituciones tecnológicas más importantes a nivel centro americano, esta posee estudiantes de distintos niveles ofreciendo conocimiento aplicable al mundo laboral, siendo un productor de mentes constructivas e intelectuales que proveen soluciones a cualquier problema de negocio aportando valor tanto en la operación como en la toma de decisiones.

Formar a los estudiantes en distintas áreas sociales y laborales es una clara misión de muchas instituciones, para que la universidad pueda alcanzar esto debe de fomentar la resiliencia, identificar factores de riesgo por el cual se presentan casos de deserción y diseñar procesos clave a nivel de marketing, planes de desarrollo estudiantil, financieros, innovación y mejorar la experiencia universitaria.

La identificación temprana de riesgos puede salvar y retener al estudiante por lo que en esta tesis se demostrará como mediante la aplicación de modelos estadísticos se extraerá provecho a la información histórica aplicada en postgrado, explicando por qué se modelan distintas técnicas estadísticas y por qué a estas mismas se le consideran máquinas de aprendizaje.

Finalmente se explicará cómo se puede hacer operativo mediante las tecnologías de información en la nube un servicio netamente estadístico para proveer y cooperar con los equipos

destinados a la retención. Es muy importante señalar que esta investigación trata uno de los temas que se deben de considerar en cualquier organización; y este tópico es ciencia de datos y analítica avanzada, ya que se considera que una empresa que no opere sus tomas de decisiones mediante esto pierde mucho mercado y suma a la desventaja competitiva.

1.2 Antecedentes

Laureate Education, Inc. (NASDAQ: LAUR) es una compañía de capital privado que presta servicios de educación superior, su sede principal está en Baltimore, Maryland, EE.UU. A través de su filial Laureate International Universities reúne 62 instituciones de educación superior en 20 países del mundo, con más de 500.000 estudiantes entre sus 100 campus. (Inc, 2019)

UNITEC de Honduras formando parte de la red de universidades de Laureate ofrece una amplia movilidad académica internacional para estudiantes y docentes, así como Intercambios académicos, programas de doble titulación, pasantías de alumnos en los parques de Walt Disney World y los cursos de verano en universidades altamente prestigiosas dentro y fuera de la red Laureate en destinos como EE.UU., España, Taiwán, México, Argentina, Chile, Perú y los países de la región centroamericana.

Considerando lo mucho que aporta UNITEC formando profesionales a nivel nacional e internacional se ha decidido y evaluado mediante el equipo de operaciones que el desarrollo de un motor para detectar factores de riesgo que aportan a la deserción estudiantil es un objetivo muy importante para seguir aportando en la mejora continua de la experiencia académica.

1.3 Definición del problema

1.3.1 Enunciado del problema

UNITEC en Honduras ofrecen programas de estudio a 22,000 estudiantes de Pregrado (Tradicional, Adulto Trabajador) y Posgrado. Actualmente buscan reducir su tasa de deserción neta y ofrecer una experiencia estudiantil superior y de alta calidad. Así mismo, el equipo de retención cuenta con un sistema de alerta para identificar a estudiantes de alto riesgo en base a su rendimiento académico a nivel de postgrado.

En el año 2017 se implementó un motor estadístico que permite identificar con tiempo esos estudiantes de pregrado que están en alto peligro de deserción mediante un análisis de factores que aportan a la calificación de riesgo, los modelos diseñados fueron configurados en la nube tanto para UNITEC como CEUTEC.

Mejorar la alarma temprana a través de un motor estadístico en la nube para postgrado que permita una mayor precisión en la identificación de estudiantes de alto riesgo y los factores que influyen a su deserción aportará mucho valor en la planeación estratégica de retención ya que no se cuenta con ella de manera inmediata.

1.3.2 Formulación del problema

¿Es posible la construcción de un motor estadístico en la nube para la detección temprana de estudiantes en riesgo de deserción?

1.3.3 Preguntas de investigación

¿Quiénes son los estudiantes de primer año de posgrado en riesgo de abandonar sus estudios en el siguiente ciclo?

¿Cuáles son los principales factores o variables significativas que influyen en la deserción estudiantil?

¿Quiénes son los estudiantes con mayor probabilidad de abandonar estudios en el siguiente ciclo académico?

¿Cuáles son las estrategias de intervención de mayor eficacia para mejorar la tasa de retención ciclo a ciclo?

1.4 Objetivos del proyecto

Objetivo General:

- Desarrolla un motor estadístico en la nube que ayude a identificar los factores de riesgo más importantes por cada uno de los estudiantes de posgrado con mayor riesgo de deserción.

Objetivos Específicos:

- Elaborar un reporte al equipo de retención estudiantil con el listado de los estudiantes con mayor riesgo de deserción.
- Construir un motor estadístico en la nube usando Microsoft Azure para procesar, transformar y calcular la calificación de riesgo de los estudiantes.
- Utilizar la información de CAP para describir el comportamiento de estudiantes que han desertado y utilizarlo como información predictora en los modelos estadísticos.

- Proporcionar un resumen de todos los modelos construidos con la validación de las variables más importantes y todas las métricas para evaluar la precisión y calidad de los modelos construidos.

1.5 Justificación

Entender el comportamiento exacto de nuestros clientes es una de las necesidades más grande que presenta hoy en día cualquier organización, antes podíamos notar como se construían servicios que cumplían una necesidad o una demanda de forma colectiva, ahora esto es algo que ha evolucionado mucho ya que se requieren productos a la medida y que tengan una alta disponibilidad.

Esto representa un gran reto tecnológico ya que se necesitan piezas de software más sofisticadas y a su vez más especializadas, como toda área de estudio el requerir algo más especializado requiere de un conocimiento científico difícil de adquirir y encontrar, esto abre muchas oportunidades a la investigación científica y apostar más en la inversión de proyectos de innovación en la nube.

Hoy en día una de las fuentes más grandes que generan ideas de negocio son dadas gracias a la analítica avanzada y como esta se complementa mediante la ciencia de datos, tener la habilidad de poder encontrar patrones específicos en la información es uno de los retos más frecuentes ya que estos son los que nos dan esas aperturas de especializar y ajustar nuestros productos a la medida.

Combinar estas soluciones de alto desempeño matemático mediante las máquinas de aprendizaje y las tecnologías en la nube provee una alta oportunidad para ser un ente diferenciador

tanto en innovación tecnológica como en servicios de negocios, claramente lo que se busca son sistemas expertos que nos ayuden a tomar decisiones más certeras y aumenten la fidelidad de la marca.

Siendo UNITEC una de las empresas de educación que, apuesta por el avance tecnológico, el usar servicios en la nube para automatizar la detección de estudiantes de postgrado con una alta probabilidad de deserción es una clara oportunidad de mejora que ayudara a los miembros del equipo de retención no solo a saber quiénes son ellos, si no, mediante la correcta lectura de sus factores de riesgo se podrá diseñar planes operativos para mejorar de la calidad de vida y experiencia estudiantil.

CAPÍTULO II. MARCO TEÓRICO

El presente capítulo muestra el sustento teórico para un entendimiento sobre el desarrollo de esta investigación, se tocarán tanto temas como de negocios, matemática y tecnología, como aplicando una correcta interacción entre ellos se puede desarrollar un producto que ayuda a la toma de decisiones, este capítulo es altamente importante ya que aquí se encontrara justificaciones sobre la aplicación y selección de tecnologías como a su vez los distintos tipo de algoritmos matemáticos para trabajar en análisis de ciencia de datos.

2.1 La importancia de la retención en los negocios de educación

2.1.1 Los negocios hoy en día

Las industrias hoy en día están creciendo de manera acelerada, y las necesidades que deben cumplir son más exigentes, los clientes requieren productos más personalizados. La gran pregunta que podemos hacernos sería la siguiente: ¿Cómo puedo alinear mi organización con esta demanda? ¡Y la respuesta a esta pregunta podrás encontrarla en los DATOS!

Uno de los activos más críticos de las organizaciones son los datos almacenados y generados gracias a ellos; Podemos descubrir diferentes tipos de patrones que nos ayudarán a comprender mejor a nuestros clientes. Pero qué tipo de datos es lo que tenemos que analizar, cómo puedo juntarlos para medir qué es correcto y qué tipo de información es lo que tengo dentro de mi organización; son una de las primeras preguntas que debe hacerse.

¿Sabías que podemos tener dos grandes tipos de información? La información externa es una de ellas, y son generadas por el consumo de nuestros clientes, la otra es información interna,

todas son datos generados en el proceso de operación cuando estamos produciendo el producto o el servicio, por lo tanto, necesitamos estratégicamente relacionarlos para tener una comprensión completa de todas nuestras fuentes de datos.

Ahora tome un segundo e imagine todos los datos que se están generando en ese momento. Ahora trate de evaluar cuánto tiempo necesitará si analiza toda esa información sin un objetivo específico. Big Data ahora es un concepto popular y muchas organizaciones están tratando de alinear sus objetivos corporativos utilizando este importante. Big Data tiene dos áreas interesantes. La primera es cómo podemos almacenar una gran cantidad de información y la segunda es la parte de análisis. Una vez que almacenamos todos esos datos, necesitamos usar herramientas estadísticas para encontrar valor en ellos.

Una encuesta realizada por LogLogic encuentra que el 49% de las organizaciones están algo o muy preocupados por la gestión de Big Data, pero que el 38% no entiende lo que es, y el 27% dice que tiene una comprensión parcial; además, la encuesta encontró que 59% de las organizaciones carecen de las herramientas necesarias para gestionar los datos de sus sistemas de TI. Según la encuesta, el 62% mencionó que ya había gestionado más de un terabyte de datos; eso indica que el volumen de datos en el mundo está aumentando a un ritmo casi incomprensible. (Camargo-Vega, Camargo-Ortega, & Joyanes-Aguilar, 2015)

La lectura continua de nuestras métricas y KPI nos dirá cómo va la estrategia de negocios, es por eso por lo que es esencial definir lo que queremos lograr y alinearnos con los objetivos corporativos de la empresa, y aquí es como comienza nuestra aventura, Proponer como se deben de leer estas métricas y con qué frecuencia ayuda a la estrategia de las áreas para garantizar la mejora continua.

2.1.2 La importancia de la retención en la educación superior

La educación superior hoy en día atraviesa grandes cambios y grandes retos la tecnología ha hecho que la información y el conocimiento esté al alcance de un clic por lo que el realizar investigaciones ya no es una tarea tan tediosa y difícil de elaborar, esto a su vez genera un potencial riesgo, la educación informal crece cada día más ofreciendo conocimiento a la mano y aplicable; educación superior tiene una fuerte competencia para seguir innovando y ofrecer ese valor de que esos recursos extra sean un complemento más para agudizar el perfil de estudiante.

El tener este tipo de reto o riesgo hace que la universidad este más pendiente de cómo atiende al estudiante y lo evalúa constantemente para que cualquier aspecto de su vida estudiantil este bajo control y constante chequeo. Distintos motivos se pueden presentar para que el estudiante se encuentre en riesgo de abandonar su carrera profesional o decida moverse a otra institución generando como resultado una pérdida de la formación de un futuro profesional.

A nivel mundial se trata de abarcar este problema de distintas maneras, entre ellas se podrían mencionar dos tipos de modelos heurísticos y estadísticos, los modelos heurísticos son aquellos que en base simple reglas se puede establecer un patrón de que estudiantes necesitan esa atención para poderlos retener, (Pineda Báez & Pedraza Ortiz, 2009) publicaron que ellos para la universidad de Católica del norte de Colombia aplicaron un par de encuestas y complementaron con información universitaria donde perfilaban el estudiante y en base a eso pudieron identificar como poder intervenir a los estudiantes con determinado perfil, lo cual es una excelente práctica pero se manejan riesgos de implementar esas encuestas y no poder abarcar a todo el estudiantado.

La segunda es mediante modelos estadísticos, es esta se someten distintos tipos de algoritmos de predicción o clasificación, métodos de validación cruzada y selección de variables,

(Nespereira, Fernández Vilas, & Díaz Redondo, 2015) proponen un modelo utilizando información de la plataforma de aprendizaje en línea, toman en cuenta toda la actividad que ellos tienen y mediante series de tiempos proponen un modelo lineal que siendo esta evaluada retorna una probabilidad de éxito o fracaso del curos, esto fue aplicado en una universidad de Portugal.

La situación que se presenta en universidades como UNITEC y CEUTEC los docentes pueden saber el desempeño académico de los estudiantes basado en reglas personalizadas viendo la actividad y motivación en los cursos y hacer sus propias intervenciones, de la misma manera, esta institución cuenta con su propia área de retención, donde cuentan con un motor estadístico construido por Laureate para detectar a esos estudiantes en un alto riesgo de deserción para los niveles académicos de grado. Para posgrado por los momentos no se cuenta con una herramienta que identifique a estos estudiantes y las razones de sus calificaciones de riesgo altas.

En que debe de invertir la universidad hoy en día para garantizar que el estudiante estará monitorizado en distintas áreas que complementan la experiencia de la vida estudiantil, se considera que la evaluación de las características demográficas si aportan gran valor al análisis del perfil estudiantil, de la misma manera tomar en cuenta el comportamiento financiero nos puede ayudar para diseñar planes de financiamiento en los que los estudiantes puedan autofinanciarse o en el caso de que sean becas saber que tanto el estudiante aprovecha una de ellas.

El rendimiento académico se puede considerar que es uno de los temas más complejos ya que de aquí se pueden tomar tantas características a evaluar para determinar el desempeño del estudiante; entre ellas alertas de éxito estudiantil, probabilidades de que el estudiante no complete un curso con éxito o termine rotando de carreas por el mal asesoramiento.

Cuál es la correcta intervención que se le debe de hacer a los estudiantes de primer año ya que en ellos se encuentran los porcentajes más altos de deserción estudiantil, el tener un staff de retención que identifique todos estos riesgos es lo que se quiere alcanzar, para poder lograr este tipo de objetivo primero se debe establecer cuando es estratégico aplicar una mediación, debería ser esta directa al estudiante o escalarla a los equipos operativos indicados, como puede ser el caso de finanzas o tutorías.

Retención es uno de los términos más utilizados por la institución ya que en esta palabra se puede sustentar el éxito de culminar una carrera (Ferrer-Rodriguez, 2015) define este concepto como: “aquellos estudiantes que permanecen en la misma institución donde comenzaron hasta que terminaron el programa. Una parte importante de las funciones de la administración es encargarse de los procesos de matrícula juntamente con el mercadeo y el reclutamiento, para aumentar la cantidad de estudiantes admitidos y para retenerlos” (p. 29). Y si es correcto señalar que esos procesos de reinscripción son claves por que se deben de contactar aquellos estudiantes con más posibilidad de regresar a la universidad.

2.1.3 El diseño de las intervenciones mediante el uso de los factores de riesgo

Se ha definido lo vital e importante que es retener al estudiante en la institución, esta es una tarea de ambos tanto de la universidad como del estudiante. El alumno debe de cumplir con sus asignaciones, cumplir con el plan de carrera, seguir las recomendaciones de sus docentes, sacar provecho de las instalaciones de la universidad y los múltiples servicios que este ofrece para facilitar la experiencia estudiantil (p).

Ahora en el caso de la institución debe de garantizar que cada uno de los distintos servicios que se ofrece al estudiante complemente de una manera idónea y oportuna las necesidades del estudiante por lo que debemos hacernos la siguiente interrogante, Como puede la universidad identificar en que esta experimentando dificultades el estudiante. Normalmente a este proceso se le denomina identificación de factores de riesgo.

De los factores de riesgo de más interés son el comportamiento financiero y académico de los estudiantes, estas son de las intervenciones más fuertes en las que la institución trabaja e invierte muchas horas de planeación para poder abarcar e identificar el correcto segmento de los estudiantes. De estas planeaciones resultan estrategias de tutorías y financiamientos. Otras universidades centran sus investigaciones en alertas de éxito estudiantes, estas consisten en identificar si un estudiante puede perder ya sea el curso o la motivación de continuar sus estudios universitarios.

(Vergara Morales, Eva Boj, Barriga, & Díaz Larenas, 2017) Explica los distintos tipos de factores de riesgo como:

Según el tiempo: i) deserción precoz: entendida como el abandono de un programa antes de comenzar, habiendo sido aceptado; ii) deserción temprana: referida al abandono del programa durante los primeros cuatro semestres; y iii) Deserción tardía: entendida como abandono desde el quinto semestre en adelante. (p. 611)

Según el espacio: i) cambio de programa dentro de una misma institución; ii) cambio de institución educativa; y iii) salida del sistema educativo, donde existe la posibilidad de reingreso en un futuro (a la misma institución o a otra). (p. 611)

Es muy importante reconocer estos distintos tipos de deserción ya que en base a esto se puede diseñar nuestra variable de respuesta o variable dependiente, haciendo una introducción rápida de esto, los factores de riesgo normalmente son las variables con mayor correlación con la variable a predecir para el caso de esta investigación aquellos estudiantes que desertaran al

siguiente periodo académico, la deserción hecha en el mismo periodo académico es considerada como un factor de riesgo a ser evaluada.

La deserción dada en el mismo periodo académico puede estar dada en ocasiones por factores externos del análisis y dominio de datos de la universidad, sin embargo, se puede encontrar una correlación bastante fuerte de que estudiantes que salen no regresan inmediatamente al siguiente periodo, por lo que se podría recomendar fuertemente el diseño de intervenciones especiales para este tipo de estudiantes.

2.1.3.1 Características Individuales

Una de las características individuales que influye significativamente en la deserción de los estudiantes corresponde a las expectativas sobre las instituciones y sobre las condiciones de vida universitaria (Patriarca, 2013). Es decir, en la medida que los estudiantes satisfacen sus expectativas y perciben de manera favorable las condiciones de adaptación universitaria, deciden permanecer en la universidad. Por otro lado, (Vries, 2011) identificaron que los problemas de vocación constituyen un aspecto que influye en la decisión de abandonar una carrera universitaria.

2.1.3.2 Características Académicas

De las variables comunes analizadas en esta categoría podemos encontrar número de créditos matriculados, números de cursos matriculados, número de cursos aprobadas, número de cursos reprobados, número de cursos retirados, adicional a esto se puede analizar el detalle de cada uno de estos cursos como saber las notas individuales de cada una de ellas, la cantidad de horas invertidas en caso de que sean clases en línea y otra cantidad de variables a revisar. La combinación entre ellas puede ayudar a mejorar la correlacionar con la variable de respuesta de deserción.

Estas a su vez generan intervenciones para saber el índice de esfuerzo y compromiso en caso de que sean clases en línea y se pueda tener acceso a ellas, aplicando algoritmos como análisis de componentes principales se puede extraer de la varianza capturada de la primera dimensión un promedio del desempeño de este, cruzando esta misma variable con las notas se podrían identificar por cuadrantes un tipo de perfil académico, que a su vez esta podría ser de gran ayuda para el motor estadístico. En el caso de esta investigación no se llega a tal tipo de análisis sin embargo si se deja como posible manera de poder mejoras los modelos para modalidades en línea.

Para el caso de tipo de estudiante este análisis puede variar ya que una variable de número de cursos aprobados históricamente en un primer periodo para un estudiante nuevo esta será de cero por lo que combinar esta información con los estudiantes de retorno no es buena práctica ya que se podría generar un sesgo por interpretación incorrecta de la información, sin embargo, cosas de las que se pueden analizar exclusivamente de los estudiantes de nuevo ingreso son pruebas de aptitud, matemáticas, lenguas, entre otras. Esto es muy importante para estudiantes de pregrado sin embargo como nuestro estudio está enfocado para los estudiantes de posgrado esta información no es útil y podría ser de mejor ayuda información como de qué universidad realizaron su grado, información socioeconómica, número de maestrías cursadas entre otras.

No toda la información propuesta es accesible de lado de la institución por lo que se deja como recomendación y buena práctica aplicar encuestas de que capturen un poco más el perfil del aspirante a entrar, la pregunta clave seria, se puede utilizar esta información desde el día uno para poder hacer un tipo de intervención, la respuesta es sí, empírica o heurísticamente se pueden establecer reglas de un perfil básico del aspirante y está en base a la experiencia del negocio podría ser un patrón de cómo y a quien se debe de contactar. De igual manera estas van sirviendo desde

día uno para ir construyendo una base de datos histórica que será de mucha utilidad en futuras recalibraciones.

2.1.3.3 Características Socioeconómicas

El análisis socio económico es uno de los pilares de análisis de mercado, ya que nos da pautas muy marcadas y propias de los estudiantes en la mayoría de los casos, cuales son una de las variables más importantes para analizar, definitivamente el género no es una de ellas, ya que esta variable está sobrevalorada especialmente en un ámbito académico, ya que no se puede inferir comportamientos o rendimientos. Y el querer generar intervenciones dependiendo de genero se podría considerar una mala práctica hoy en día.

Otra variables interesantes son el estado actual si el estudiante trabaja, el mismo estudiante financia su carrear u obtiene ayuda ya sea por algún padrino o el gobierno, es un estudiante que tuvo que salir de su casa hogar y vivir en otro lado para tener acceso a educación es otra variable importante a considerar, el reto es como se mantiene actualizada esta información normalmente esta no es frecuentemente actualizada, sin embargo se podrían diseñar maneras indirectas de obtener esta información como por ejemplo empleando uso de gamificación o encuestas a través del aplicativo móvil.

(Guta, 2017) define gamificación como: un proceso para integrar las mecánicas del juego en algo que ya existe para motivar la participación, el compromiso y la lealtad. Esto puede ser casi cualquier cosa, desde su sitio web hasta la presencia en las redes sociales, las operaciones diarias, el compromiso del cliente y más. La gamificación introduce elementos de diseño de juegos en aplicaciones que no son juegos para hacerlos más divertidos y atractivos. Utiliza la competencia,

los puntos, los logros, las reglas de juego, el estado y la autoexpresión para fomentar acciones a través de comentarios positivos.

2.1.3.4 Características institucionales

Analizar el entorno de la universidad es otro aspecto muy importante para considerar los estudiantes deben de sentir que todas sus necesidades investigativas y recreativas son encontradas en la universidad por lo que la continua evaluación lectura de métricas de que tanto uso hacen los estudiantes de los distintos servicios es una obligación. Pero qué tipo de variables podemos encontrar aquí: el número de veces que se visita la biblioteca, mediante Internet de las cosas también se puede capturar información como, por ejemplo: ¿Cuánto tiempo gastan los estudiantes en los gimnasios, cafeterías, aulas? ¿Cuántas veces a la semana entran al campus? ¿Qué tanto uso hace del aplicativo móvil? entre otras.

Internet de las Cosas es el concepto de objetos de todos los días – desde máquinas industriales hasta dispositivos de vestir (wearable devices) – mediante el uso de sensores integrados para recopilar datos y seguir una acción con esos datos a través de una red. De modo que un edificio que utiliza sensores para ajustar automáticamente la calefacción y la iluminación. O bien equipo de producción que alerta al personal de mantenimiento de un fallo inminente. Dicho de manera simple, Internet de las Cosas es el futuro de la tecnología que puede hacer nuestras vidas más eficientes. (know, n.d.)

2.2 Data-Driven

Data-Driven es una metodología que da una forma de operar la toma de decisiones del negocio, normalmente esta opera en base a los datos por lo que mantener estos estables y con un acceso rápido de todo lo que se pueda medir es una de las actividades más críticas en la que las

organizaciones deben enfocarse. Porqué este es uno de los temas más importantes que se deben de considerar antes de comenzar cualquier proyecto de analítica avanzada o ciencia de datos, por la razón de que el insumo más importante para estos proyectos es la información entre más datos juntos más rápido se pueden hacer análisis y encontrar esos patrones que expliquen mejor cualquier pregunta de negocio que se quiera contestar.

Como se pudo observar todos los datos que giran alrededor de la vida estudiantil es de una proporción bastante grande, tenemos muchas fuentes de información que generan muchas variables, operacionalmente tenemos muchos dueños y líderes de servicio, por ejemplo, tenemos un responsable de finanzas, logística, experiencia de aprendizaje en línea, académico, etc. Una de las principales buenas prácticas que comparte Data-Driven es que estos dueños deben de ver la información que ellos generan como una herramienta más, pero los verdaderos dueños de la información deberían de estar centralizado en un solo lugar, esto da paso a lo que denominamos de gobernabilidad de los datos.

La gran cantidad de datos que están entregando los procesos de negocios de las empresas son una fuente de información que puede ayudar a la generación de valor, mediante el apoyo a la toma de decisiones con elementos cuantitativos, confiables y oportunos. Esta necesidad de utilización de los datos para generar información relevante requiere el establecer un conjunto de definiciones, reglas y procesos que regulen como serán tratados los datos, a este conjunto se le denomina Gobernabilidad de Datos. (Saffirio, 2017)

Cuando nos referimos a Gobernabilidad de Datos podemos estar refiriéndonos a:

- Estructuras de la Organización.

- Reglas (políticas, estándares, normas, reglas de negocio).
- Derechos de decisión (establecer cómo decidir y quién decide).
- Responsabilidades.
- Procesos relacionados con la operación de los datos. (Saffirio, 2017)

Una de las mejores opciones para poder manejar todo lo relacionado a gobernabilidad de los datos podría ser a través de Data Management Book of Knowledge (DMBOOK) donde este explica el correcto manejo y administración de los datos alienados a las metas estratégicas de TI y el negocio, de la siguiente manera:



Figura 1 Modelo DMBOOK.

Fuente: (Saffirio, 2017)

Esta es una de las maneras más ordenadas de dimensionar la correcta administración de los datos como se puede observar está compuesto desde la administración diccionario de variables,

seguridad de la información, inteligencia de negocio y todas las formas en las que podemos almacenar información. Este es uno de los puntos más importantes a considerar el momento que se implementa una nueva tecnología de almacenamiento, estos varían desde bases de datos relaciones, archivos como también bases de datos de cache.

Llegar a este nivel de madurez es una tarea que lleva años tanto para empresas grandes como pequeñas ya que centralizar y mapear todo el registro de la información no es una tarea que con solo pagar se podrá tener en un tiempo extremadamente corto, por lo que siempre es muy importante tener a una persona que se responsabilice y mantenga los distintos niveles de madurez de esta.

Tabla 1 Niveles de madurez de la gobernabilidad de la información.

Nivel	Nombre	Características
1	Informal (Piensa Localmente, Actúa Localmente)	<p>Existen unas pocas reglas y políticas relacionadas con la calidad y consistencia de los datos. Hay mucha redundancia en los datos, distintas fuentes, formatos y registros.</p> <p>Existe el riesgo que datos erróneos puedan provocar una mala toma de decisiones o pérdidas de oportunidades.</p>
2	Reactivo (Piensa Globalmente, Actúa Localmente)	<p>Este es el estado inicial de la Gobernabilidad de Datos. Existe mucha actividad y esfuerzo relacionados con la reconciliación de inconsistencias, imprecisiones y datos poco fiables.</p> <p>Se tienen avances y más experiencia a nivel departamental —gerencias.</p>
3	Proactivo (Piensa Globalmente, Actúa Colectivamente)	<p>Es muy difícil llegar a este nivel. La empresa debe entender el valor de contar con una visión unificada de la información y del conocimiento. La empresa comienza a pensar en desarrollar la Gestión de Datos Maestros —</p>

		Máster Data Management (MDM). Ejemplos: Datos de Clientes, Proveedores, Productos, Repuestos, etc. La empresa está aprendiendo y preparándose para el nivel siguiente. Está en desarrollo un cambio cultural en la organización.
4	Gobernado (Piensa Globalmente, Actúa Globalmente)	La información esta unificada en todas las áreas de la empresa. Se cuenta con una estrategia y metodología para la gestión de los datos. Se ha producido un cambio cultural en la empresa. Los colaboradores han integrado la idea que la información es un activo clave de la empresa.

Fuente: (Saffirio, 2017)

Muchos niveles se deben de superar para poder alcanzar un nivel de madurez en la administración de la información, pero una de las observaciones más importantes que se citan es el cambio de cultura organizacional, debe de ser incluido dentro de los planes operativos anuales la correcta administración de los datos desde su almacenamiento y compartimiento. Algunas empresas importantes introducen conceptos interesantes de este modelo de Data-Driven, entre más publica la información entre los empleados y bajo términos legales que adecuen a las políticas de las empresas, muchas más oportunidades de poder innovar y crecer podrán existir.

2.2.1 Democratización de la información

La democratización de los datos es una de las ideas más poderosas de la ciencia de datos. Todos en una organización deben tener acceso a la mayor cantidad de datos legalmente posible. (Patil & Mason, 2015)

Si bien el acceso amplio a los datos se ha vuelto más común en las ciencias, Facebook fue una de las primeras empresas en dar acceso a sus empleados a los datos. A escala desde el principio, Facebook se dio cuenta de que dar a todos el acceso a los datos era algo bueno. Los empleados no tuvieron que presentar una solicitud, esperar la priorización y recibir datos que podrían estar desactualizados. Esta idea fue radical porque la creencia predominante era que los empleados no sabrían cómo acceder a los datos, los datos incorrectos se usarían para tomar decisiones empresariales deficientes y los costos técnicos serían prohibitivos. Aunque ciertamente hubo desafíos, Facebook descubrió que los beneficios superaban con creces los costos; se convirtió en una empresa más ágil que podía desarrollar nuevos productos y responder rápidamente a los cambios del mercado. El acceso a los datos se convirtió en una parte fundamental del éxito de Facebook, y sigue siendo algo en lo que invierte agresivamente. (Patil & Mason, 2015)

2.2.2 Como saber si se manejan las mejores prácticas de data-Driven

Para empresas donde uno de sus productos principales es la información, Facebook practica e invierte en estas buenas prácticas por lo que está probada la hipótesis de que una correcta administración de la información y dando acceso de los datos hasta cierto punto a los empleados genera buenas prácticas para innovar en desarrollo de nuevos productos y servicios. A continuación, te compartimos una de las señales de que estas en un buen camino gracias a esa guía de esta útil herramienta según (Análisis de datos: 10 señales de que eres “Data-Driven”, n.d.):

- Siempre hay que tener a un responsable de la calidad de los datos
- En cada proyecto siempre se debe de establecer bien cuales deben de ser los indicadores claves.
- La mayoría de las opiniones están fundamentadas en los datos.

- Los números siempre comunican la verdad.
- Todos tienen acceso a los datos que necesiten trabajar.
- Los objetivos siempre deben de tener buenas métricas definidas.
- Se capacita al equipo en administrar y trabajar con datos.
- Los proyectos de gestión de datos no deben de tener problemas de financiamiento.
- Los datos nunca se utilizan para encausar a alguien de un error.
- El análisis de datos día a día es algo más natural.

Muchas buenas prácticas existen con las que se pueden comenzar el análisis de los datos, Esto es una de las actividades más proactivas de cualquier área en cualquier institución vale la pena mantener estas lecturas y que estén incluidas en investigaciones como la presente ya que esto deja como valor agregado para cualquier lector como se debe de comenzar a construir una correcta gerencia de datos.

Como ultima observación data-Driven debe de ser implementado en cualquier universidad o institución educativa entre más información centralizada y a mano se pueda tener acceso mucho mejor será porque se podrán analizar muchas más correlaciones, y la cantidad de mejoras continuas que se pueden encontrar son más fáciles de justificar y llevar acabo ya que estas siempre van acompañadas de una hipótesis generada o inducida de comportamientos históricos y claramente bien diseñadas.

2.3 Ciencia de datos en instituciones educativas

La ciencia de datos es una disciplina que es aplicada a cualquier área de negocio esta puede ser útil desde funciones biomédicas, industrias de video juegos, fábricas de textiles, educación,

gobierno, etc. Por lo cual es muy importante definir un par de conceptos sobre que abarca esta ciencia y por qué es tan transversal.

2.3.1 Que es ciencia de datos

La ciencia de datos combina diferentes áreas de desarrollo profesional, esta ciencia es la que se encarga de responder cualquier pregunta o problema de negocio mediante el análisis estadístico avanzado de los datos, trata de buscar patrones para perfilar o pronosticar a una variable. Al mismo tiempo se encarga de las buenas prácticas para formular hipótesis, normalmente estas son brechas que mediante la aplicación de algo se tratan ya sea de reducir o aumentar todo depende de la pregunta de negocio. También la ciencia de datos explica mediante historia de datos y visualizaciones como estos evolucionan atreves del tiempo o como describen comportamientos no esperados. Esta área sabe cómo diseñar preguntas para poder indagar y comprender el negocio.

2.3.2 Que hace un científico de datos y por qué son tan necesarios hoy en día

Un científico de datos es aquel que maneja básicamente 4 grandes áreas de dominio para poder responder preguntas de negocio a acoplarse bajo las mejores prácticas de data-Driven, estas áreas son las siguientes:

Tabla 2 funciones de los científicos de datos

Áreas	Descripción
Manejo y transformación de datos	<p>Cualquier Científico de datos debe de saber cómo manipular grandes conjuntos de datos, normalmente se realizan tareas de limpieza, transformación y unión de ellas, estas tareas siempre están ligadas a un objetivo específico, y este es la construcción de métricas. La mayoría de las operaciones entre conjuntos de información es para responder una pregunta de negocio ya que la métrica se encarga de medir la información, dependiendo de la lectura de la métrica se definen distintos tipos de intervenciones.</p>
Manejo de lenguajes de programación	<p>Las operaciones de datos normalmente conllevan mucha de mucha técnica de optimización de procesamiento, aquí se podría señalar que se combina con la ciencia de BIG DATA ya que el procesamiento a grande escala de la información requiere de un conocimiento especializado. Pero para realizar las operaciones cotidianas de limpieza y transformación se requiere de una gran habilidad de conocimientos de programación.</p>
Conocimiento Estadístico y Matemático	<p>El querer encontrar patrones en la información requiere de un conocimiento sobre aplicaciones estadísticas, de aquí se</p>

	desprende el uso de máquinas de aprendizaje supervisadas y no supervisadas. Las cuales se definirán con más detalle. Saber que algoritmo estadístico hace mejor encaje con el problema de negocio que tenemos.
Amplio dominio del negocio y comunicación	Me atrevería a decir que esta es la habilidad que más cuesta desarrollar, ya que en esta se comunica los resultados obtenidos o el nivel de entendimiento que tenemos de la información. Estas se abarcan mediante visualizaciones de datos y contar historias basado en ellos.

Fuente: Propia

2.3.3 El manejo y definición de las métricas

Todos los días trabajamos bajo un propósito u objetivo que queremos lograr, que generalmente se define en un tiempo específico. Estos pueden variar dependiendo de la complejidad de estos. Por ejemplo, imagine que un estudiante quiere completar su carrera universitaria en 4 años. Puede medir el éxito de esta meta por la cantidad de cursos aprobados por año. Medir la métrica ayuda al estudiante a reconocer si su desempeño ha sido bueno o debería ser mejor. Entonces, cuando el estudiante ha terminado el año y ha fallado algunos cursos, puede redefinir la cantidad de cursos que se deben aprobar por año o, en el escenario opuesto, puede establecer el objetivo de finalización en menos tiempo. Podemos encontrar alternativas para mejorar nuestro rendimiento en función de la cantidad de cursos aprobados si el alumno no está teniendo un buen desempeño, como encontrar programas de tutoría. Esto es exactamente lo que hace una buena métrica. Motiva a la persona a buscar la mejora continua.

Las métricas tienen las siguientes propiedades, que siempre tenemos que encontrar. Como hemos mostrado anteriormente, estos son:

- La métrica debe medir algo importante: la lectura continua de las métricas significa que son importantes porque nos proporciona un estado de cómo está progresando la meta que queremos lograr.
- Las métricas podrían mejorarse: una buena métrica es aquella que permite ser modificada a través del tiempo en función del rendimiento. Podemos hacer pequeños ajustes para garantizar el logro de la meta.
- Las métricas permiten intervenciones: una vez que sabemos que las métricas son importantes y podrían mejorarse, significa que pueden proporcionarnos pautas para aplicar las intervenciones. Indirectamente, son una guía de recomendaciones y buenas prácticas futuras. Cómo puede elegir qué intervenciones realizar es haciéndose preguntas sobre cómo mejorar el servicio o el proceso.

Como puede ver, tenemos mucho trabajo por hacer para lograr objetivos críticos y estos están fuertemente relacionados con varias métricas principales. Normalmente los llamamos indicadores clave de rendimiento (KPI). Un KPI podría tener muchas métricas relacionadas, como por ejemplo un estudiante en la plataforma de e-learning:

- Número de cursos inscritos.
- Número de comentarios a la semana.
- Número de logins en la semana.
- Promedio de notas de todos los cursos ya sea semana o acumulado.

Muchas métricas podrían ser monitoreadas para lograr un objetivo principal. Un KPI es un valor medible que demuestra la eficacia con la que una empresa está logrando objetivos de negocio clave. Una buena métrica tiene todas las propiedades que hemos descrito hasta ahora, pero ¿Cómo podemos diseñar nuestros KPI? Existen muchas formas, pero una de las más comunes es utilizar el S.M.A.R.T. metodología. Por ejemplo, la métrica que tiene una empresa para aumentar las ventas en un 20% al finalizar el año en comparación con el año anterior.

Tabla 3 Ejemplo HAVE para definir un KPI.

Especifico	Al finalizar el año, queremos aumentar nuestras ventas en un 20% en comparación con el año anterior.
Medible	Porcentaje de ventas (ventas del año en curso / ventas del año anterior).
Alcanzable	Una vez que hemos identificado todas las métricas con las que queremos trabajar y en base al pronóstico, acordamos que esto es factible, y todos los objetivos del área se alinean con este.
Relevante	Para todas las áreas este es un excelente indicador para mejorar las operaciones. La medición de métricas basadas en este objetivo clave es una buena práctica y una guía para definir su plan anual.
A tiempo	Un año es alcanzable y realista para la meta clave.

Fuente: Propia.

2.3.4 Como está compuesto un proyecto de ciencia de datos

Los proyectos de ciencia de datos están compuestos por tres proyectos macros, el primero de ellos es la recolección de la información, en este proyecto se hacen muchas iteraciones en caso de que la institución no maneje un correcto gobierno de los datos, se trata de buscar que la

información si empate con la pregunta de negocio. El segundo proyecto es todo el análisis exploratorio de datos y modelado estadístico, en esta fase también se valida con la institución de que los análisis realizados encajen con el negocio. Y por último y el que no es muy mencionado en libros es el proceso de automatización en el caso de esta investigación este proyecto se lleva bajo una implementación en la nube.

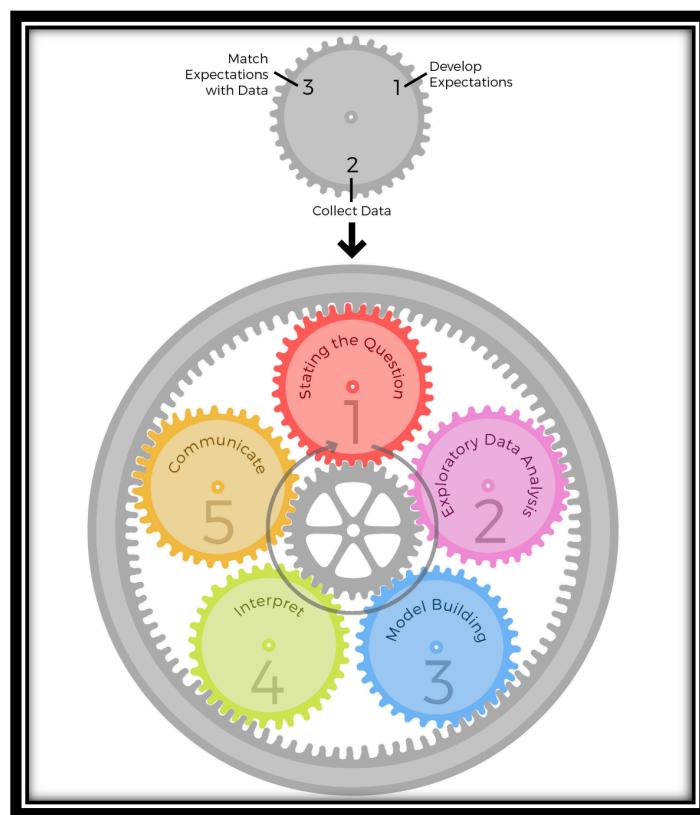


Figura 2 Epicycles of Analysis.

Fuente: (Peng, The Art of Data Science, 2015).

Como se puede ver en la imagen anterior los dos primeros engranajes denotan los dos primeros proyectos a un nivel macro. Estos siempre comienzan con una pregunta de negocio clave, en nuestro caso, ¿Cómo se puede identificar tempranamente a los estudiantes de posgrado que

tienen un alto riesgo de deserción? Esta pregunta siempre será tratada de responder con datos por lo que identificar todos los origines de datos es una clave principal.

Desarrollar expectativas: Aquí nace la pregunta de negocio y se hace el primer filtro de si esta pregunta realmente se puede contestar basado en que es lo que siente de información y si realmente se puede plantear una hipótesis para ella, por ejemplo, imagina que se quiere contestar la pregunta ¿Quiénes son los estudiantes que solo aprueban un 30% de las clases? y se quiere identificar ¿Cuál es el motivo por el cual sucede esto?. Asumiendo que para esta pregunta actualmente se cuenta con toda la información académica tanto de registro como actividades en línea y participaciones en clase, no podemos asegurar con ciencia exacta en una primera instancia la razón, pero la primera mitad de la pregunta si la podemos responder, de igual manera se puede pasar a la fase de recolectar la información. Ahora imagina que se quiere responder la pregunta, ¿Porque los estudiantes no comen tanta verdura? Para esta pregunta no se tiene información sobre su alimentación, por lo que la expectativa de esta pregunta no se puede cumplir. Sin embargo, se puede recomendar una solución para comenzar a capturar esta información.

Recolección de información: Esta fase tiende a ser muy técnica y en ocasiones la que más tiempo toma, todo dependerá del nivel de madurez de la institución u organización con la gobernabilidad de los datos, aquí se construyen los posibles diccionarios de datos que se le solicita al equipo de TI.

Empatar expectativas con los datos: Esta fase es una primera pregunta de qué tan correlacionada puede llegar a estar una información con nuestra pregunta, por ejemplo, la información financiera se puede mencionar que si está relacionada dado que un estudiante que no realiza sus pagos puede tener el riesgo de no regresar al siguiente periodo, otro ejemplo, imagináte

que se tiene la base de datos de que estudiantes fueron buenos jugando boliche hace 10 años, probablemente esta es una información que no empate muy bien con nuestra pregunta.

Una vez se ha terminado esta primera fase de expectativas y recolección de la información se comienza a socializar con el equipo de científicos y analistas de datos los diferentes conjuntos de datos. La primera fase de esta transferencia es formular la correcta pregunta, normalmente la pregunta que surge de la fase o proyecto uno es una duda, problema u oportunidad de negocio, por lo que los científicos formulan una pregunta e hipótesis más formal con la que se podrá empezar a trabajar con los conjuntos de datos.

Siempre hay algo que se debe de mantener en mente y es el riesgo de ejecución de cualquier proyecto, recuerda que fase uno es un proyecto de diseño y expectativas y como todo proyecto de IT siempre hay riesgo que se va evaluando a medida el proyecto va evolucionando estos pueden ir siendo ajustado tanto la pregunta como el alcance. Qué riesgo se puede manejar una vez terminada la fase uno comenzando la fase dos, definitivamente es la calidad de la información y la correlación que esta puede llegar a tener con nuestra variable de respuesta.

Ahora comentemos un poco sobre la segunda fase de este proyecto, esta tiende a ser un poco menos corta en tiempo dependiendo de la pregunta que se desea contestar, pero conlleva una parte más creativa que se debe de explotar y definitivamente esta se va mejorando en base a dos factores, número uno es el entendimiento del negocio y la segunda es la experiencia trabajando en diferentes proyectos.

A un nivel macro la segunda fase está compuesta por tres principales actividades, el análisis exploratorio de datos (EDA), la segunda es el modelado y validación estadística y la tercera es

compartir los hallazgos encontrados. Esta investigación se centra en estas tres fases primarias por lo que se dedicara una sección en este marco teórico por lo importante que es.

2.4 Análisis Exploratorio de datos.

Y entramos a una de las fases más fascinantes que cualquier científico de datos disfruta trabajar pues es aquí donde la creatividad y artes empieza a tomar forma, porque denotamos que se ocupa de mucho arte en esta fase del proyecto, porque aquí es donde se combina el conocimiento del negocio, con la habilidad de manipular conjuntos de datos, transformar variables en el sentido que crea información interesantes para el tomador de decisiones, como modelamos estadísticamente una herramienta que me ayude a tener una posible visión del futuro o ayude a entender más a mis clientes, negocio, empleados, etc. El análisis exploratorio de datos está compuesto por dos fases: la limpieza e ingeniería de variables.

2.4.1 Limpieza de datos.

La limpieza de los datos puede ser atacada de una manera fácil haciendo un análisis descriptivo de la información, este tipo de análisis consiste en ver las características de los datos iniciando por entender el tipo de estos.

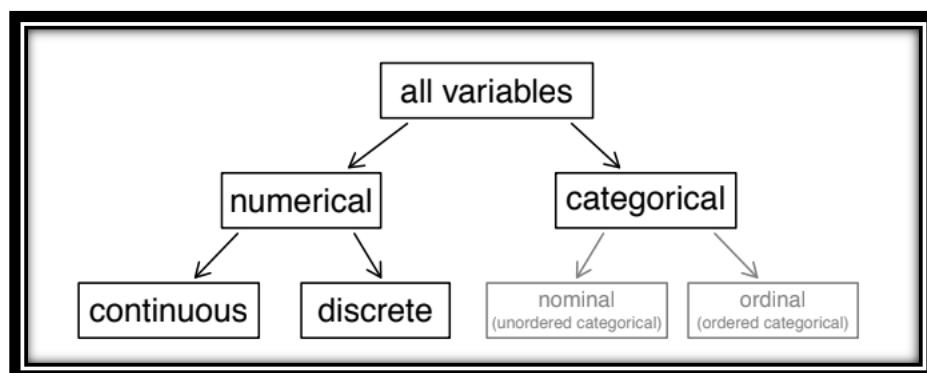


Figura 3 Tipos de variables.

Fuente: (Diez, Barr, & etinkaya-Rundel, 2015).

Partiendo de las variables numéricas continuas podríamos decir que son aquellas a las cuales se puede hacer el análisis de varianza, media, moda, cuartiles (más adelante se explicara con más detalle esto) ya que si les afecta agregar un valor más. (Diez, Barr, & etinkaya-Rundel, 2015). Tomando como ejemplo un conjunto de datos de notas la variable promedio obtenido es continua ya que puedo agregarlas a nivel de promedio entre todos los cursos inscritos por el estudiante. Ahora moviéndonos a las variables discretas que me dirías del conteo de los cursos matriculados esta es considerada discreta ya que pueden partir de cálculos simples como el conteo, analizando específicamente la información de registro estudiantil, el número de cursos después de ciertas semanas cambia con poca frecuencia por lo que esta podría ser otra pista para considerar este tipo de variable, sin embargo hay que tener mucho cuidado ya que de aquí normalmente se hace un tipo de limpieza por ejemplo, estudiantes con un número de cursos matriculados igual a 0 no hace sentido por lo que esta sería una observación a ser excluida, de igual manera el promedio de cursos matriculados es un indicador excelente para automatizar anomalías en la información, por lo que el cálculo de estas variables numéricas discretas sirven en gran medida también al momento de automatizar nuestra solución.

Ahora como podríamos detectar anomalías en variables numéricas continuas, este proceso se hace de una manera casi automática ya que se busca excluir esos valores extremos o anómalos de la información, excelente herramienta para hacer esto es mediante boxplots esta nos da visualmente donde estas esos valores extremos.

Boxplot: las gráficas de caja son una representación visual del resumen de cinco números más un poco más de información. En particular, los diagramas de caja suelen trazar valores atípicos que

van más allá de la mayor parte de los datos. Esto se implementa a través de la función boxplot () con R. (Peng, Exploratory Data Analysis with R, 2015, p. 43)

Tal como mencionar (Peng, Exploratory Data Analysis with R, 2015) estos gráficos muestran cómo se ubican prácticamente los cuartiles la mediana y la moda, muy interesante que la amplitud de cada uno de estos cuartiles nos da gráficamente como están dispersa la varianza en cada una de esas secciones.

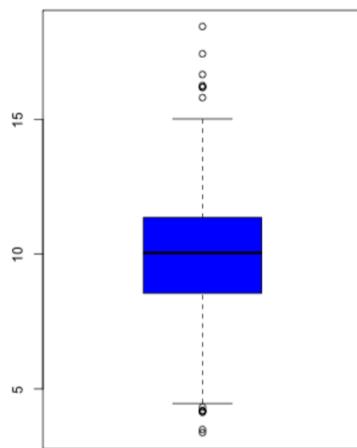


Figura 4 Ejemplo Boxplot.

Fuente: (Peng, Exploratory Data Analysis with R, 2015)

Expliquemos primero las divisiones de cada uno de los cuartiles, comenzando de abajo hacia arriba, los puntos inferiores por debajo de 5 se dice que son valores atípicos inferiores porque son muy pocas las observaciones que tienen ese valor en el conjunto de datos. Q1 denominamos lo que está entre 0 a 25% de la representatividad de la información este como podemos observar esta entre 5 y posiblemente un valor siete u ocho, Q2 es la representatividad de los valores entre un 25% y 50% comparado con Q1 podemos ver que Q2 está un poco más ajustado por lo que

puede decir que tiene una varianza menos dispersa el valor máximo de Q2 es normalmente la mediana, Q2 son los valores que están entre 50% a 75% visualmente se ve que esta igual de ajustado que Q2 y su valor inferior es claramente la mediana. Q4 comprende de 75% a 100% en algunos casos de negocios estos suelen ser excluidos reduciendo hasta un 90%; también suelen ser reducidos los valores atípicos extremos, como se puede observar estos valores son arriba de 15.

Tener en consideración el promedio de las variables numéricas continuas es una importante práctica ya que los modelos estadísticos normalmente discriminan información a veces basadas en las medias, por lo que, en modelos productivo si la variable numérica considerada es importante viene con valores muy desviados del entrenamiento, la predicción no es muy certera.

Variables categóricas como se puede observar tienen dos tipos las ordinarias y nominales, las primeras de ellas son aquellas que representan un orden, por ejemplo, el grado en el que se encuentra un estudiante de primaria; primero, segundo, tercero. Estas sirven para agrupar lógicamente varias observaciones y proveen un sentido de jerarquía. A nivel de limpieza estas son importantes para determinar estudiantes que pertenecen a grupos donde su representatividad es menor del 1%. A nivel de negocio estas pueden ser excluidas del estudio o reagrupadas en una nueva categoría.

Para el segundo caso también son útiles para poder agrupar lógicamente las observaciones solo que en este caso no denota ningún orden jerárquico por ejemplo los estudiantes de primer año (ordinaria) que pertenece ya sea a la escuela (nominal) de ciencias económicas, sociales, informáticas etc. Estas de igual manera se pueden limpiar como las ordinarias aquellas que representen un porcentaje muy bajo de la muestra estas se pueden excluir o reagrupar en otra categoría.

2.4.2 Ingeniería de variables

Esta es una de las partes más creativas del proyecto por que busca eliminar variables que se encuentran apalancadas entre ellas o lo que normalmente en estadística se le denomina multicolinealidad, por ejemplo, una variable es notas y la otra es la misma variable, pero multiplicada por dos. Además, se comienza a estandarizar las variables que conlleva un poco lo mencionado anteriormente como limpieza de la información, eliminar valores extremos o agrupar variables que no representan mucho en la muestra, sin embargo, se debe de tener cuidado de la correlación que tiene esta con la variable de respuesta. Normalmente están son validadas con el negocio.

2.4.2.1 Transformaciones

Una variable transformada puede ser aquella que sea combinada con otra o se aplique alguna regla de asociación para poder aumentar su poder predictivo, de la misma manera se busca combinar en los casos cuando una variable tiene el mismo significado con otra ya sea de una manera positiva o negativa, por ejemplo, si un estudiante ha matriculado 5 clases y al final del periodo académico este aprobó 4 y reprobó una. Ambas suman las mismas 5 por lo que son tres variables que están altamente correlacionadas en lugar de introducir las tres variables se podría considerar colocar solamente una que denote el compromiso académico con registro.

$$\text{Compromiso Academico} = \frac{\text{clases aprobadas}}{\text{clases matriculadas}}$$

Figura 5 Ejemplo de transformación.

Fuente: Propia.

2.4.2.2 Imputaciones

Imputaciones claramente viene de los procesos de limpieza de información, en ocasiones por reglas de negocio se deben de excluir cierta información como por ejemplo, inscripciones que se hicieron y antes de la fecha límite de inscripción el estudiante decidió retirar todo, a lo que denominamos como una deserción precoz, por la naturaleza de las otras variables este estudiante no tendrá actividad alguna a ser modelada, excluirlo es lo más correcto y buscar otro tipo de intervención para este alumno es la operación a seguir.

2.4.2.3 Agregaciones

Agregaciones se pueden dar por dos razones una para eliminar la baja representatividad de un nivel de una variable categórica, como, por ejemplo, se tienen 10 carreras donde están distribuidos los estudiantes y estos están representados de la siguiente manera:

Tabla 4 Distribución de Estudiantes por carrera

Carrera	Frecuencia en porcentaje
Carrera 1	40%
Carrera 2	30%
Carrera 3	20%
Carrera 4	5%
Carrera 5	2%
Carrera 6	1%
Carrera 7	1%
Carrera 8	0.5%
Carrera 9	0.25%
Carrera 10	0.25%

Fuente: Propia

Como podemos observar la mayor representatividad de los estudiantes inscritos están en las 3 primeras carreras por lo que hacer un entrenamiento con 10 niveles de una variable no es recomendado aparte de que no es buena práctica en caso que se utilice por ejemplo una regresión logística ya que esta generaría 10 variables a ser evaluadas en la ecuación lineal que esta arroja, que es lo que se puede hacer siempre y cuando el negocio lo valide es agregar las últimas 7 carreras como una sola para tener mayor representatividad en este último nivel.

Tabla 5 Tabla agregada en base a una regla de negocio.

Carrera	Frecuencia en porcentaje
Carrera 1	40%
Carrera 2	30%
Carrera 3	20%
Otros	10%

Fuente: Propia.

Ahora con esta transformación podemos hacer análisis más fáciles de interpretar y comunicar con el negocio ya que se podría por ejemplo mencionar cuál de esas categorías representa mayor correlación con nuestra variable de respuesta. Por ejemplo:

Tabla 6 Ejemplo de correlación con variable de respuesta.

	Desertor	No Desertor
Carrera 1	10%	90%
Carrera 2	20%	80%

Carrera 3	30%	70%
Otros	90%	10%

Fuente: Propia.

Si compartiéramos con el equipo de retención un hallazgo basada en esta información podríamos mencionar que los estudiantes que están clasificados en la categoría otros tienen un 90% de riesgo más alto que las demás carreras por lo que un plan de intervención en este nivel debería de ser obligatorio de diseñar, sin embargo, se debe de evaluar si el 10% que deserta de carrera 1 genera mayor perdida en costos que el 90% de la clasificación de otros, por lo que se debe de tener mucho cuidado con las recomendaciones de esta.

2.4.2.4 Estandarizaciones

Como último vamos a listar esta práctica de transformación, normalmente esta la hacemos para poder distribuir bien la información o hacerla más normal basado en distancias de medias y medianas, está la podemos realizar mediante análisis de percentiles. siempre los percentiles cero y cien son aquellos que tiene los valores extremos, en ocasiones estos son valores atípicos, como podemos eliminarlos, número uno es mediante una regla de imputación que debe de ser aprobada por el negocio la segunda es mediante un remplazo. Por ejemplo, imaginemos tenemos la siguiente información:

```
> x <- c(sample(1:5,10,TRUE), sample(1000:1200,200,TRUE), sample(1000000:1000009,10,TRUE) )
> x
 [1] 3 2 3 5 1 3 2 5 1 4 1176 1106 1180 1194 1141
[16] 1147 1125 1005 1107 1149 1094 1032 1011 1104 1173 1021 1039 1179 1181 1186
[31] 1012 1008 1004 1188 1090 1167 1012 1101 1110 1130 1042 1017 1115 1052 1143
[46] 1161 1198 1053 1171 1138 1162 1158 1198 1121 1048 1027 1191 1032 1068 1183
[61] 1030 1076 1054 1133 1074 1157 1160 1057 1000 1000 1197 1152 1169 1033 1199
[76] 1164 1100 1142 1117 1170 1073 1084 1083 1017 1164 1175 1066 1157 1089 1035
[91] 1150 1043 1026 1082 1173 1062 1177 1146 1130 1131 1056 1047 1093 1106 1027
[106] 1120 1062 1105 1073 1057 1058 1018 1058 1186 1026 1147 1061 1177 1012 1058
[121] 1128 1147 1142 1015 1127 1152 1161 1089 1077 1156 1086 1046 1112 1141 1157
[136] 1059 1164 1156 1020 1115 1139 1197 1018 1019 1190 1001 1157 1106 1199 1051
[151] 1140 1060 1148 1170 1081 1061 1097 1048 1125 1120 1032 1075 1121 1104 1088
[166] 1036 1110 1159 1073 1156 1096 1123 1058 1120 1189 1109 1097 1123 1163 1070
[181] 1131 1169 1099 1082 1184 1019 1181 1135 1168 1196 1078 1153 1139 1168 1047
[196] 1032 1084 1151 1056 1195 1094 1110 1193 1159 1037 1076 1121 1110 1103 1190
[211] 1000009 1000000 1000008 1000006 1000000 1000005 1000009 1000005 1000003 1000009
```

Figura 6 ejemplo de lista de números.

Fuente: Propia.

Este es un ejemplo sencillo de cómo generar mediante R una colección de elementos enteros aleatorios ahora miremos un resumen de esta información:

```
> summary(x)
   Min. 1st Qu. Median     Mean 3rd Qu.    Max.
1      1056    1110    46461    1159 1000009
```

Figura 7 Resumen de una variable numérica.

Fuente: Propia.

Como podemos observar tenemos como mínimo 1 y máximo 1000009 son valores totalmente extremos comparados tanto en la media como mediana, aparte de que tenemos estos dos muy lejanos. Como podemos ver visualmente que estos datos efectivamente son valores atípicos, mediante un boxplot.

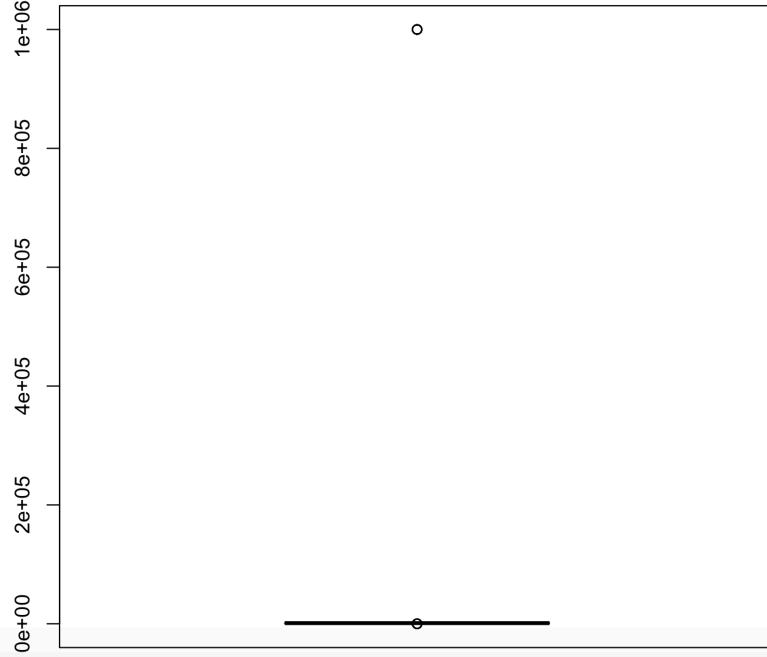


Figura 8 Boxplot variable numérica sin estandarizar.

Fuente: Propia.

Tratemos ahora de encontrar una regla para poder estandarizar esos valores atípicos por ejemplo utilizando la información de los percentiles en 5,10,90,95.

> quantile(x, prob=seq(0, 1, length = 101))												
0%	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%	11%	
1.00	2.00	3.00	3.57	5.00	1000.00	1004.14	1008.99	1012.00	1014.13	1017.00	1018.09	
12%	13%	14%	15%	16%	17%	18%	19%	20%	21%	22%	23%	
1019.28	1023.35	1026.66	1029.55	1032.00	1032.23	1035.42	1038.22	1042.80	1046.99	1048.00	1051.37	
24%	25%	26%	27%	28%	29%	30%	31%	32%	33%	34%	35%	
1053.56	1056.00	1057.00	1058.00	1058.32	1060.51	1061.70	1065.56	1070.24	1073.00	1074.46	1076.00	
36%	37%	38%	39%	40%	41%	42%	43%	44%	45%	46%	47%	
1077.84	1082.00	1083.22	1084.82	1088.60	1089.79	1093.98	1096.17	1097.72	1100.55	1103.74	1104.93	
48%	49%	50%	51%	52%	53%	54%	55%	56%	57%	58%	59%	
1106.00	1107.62	1110.00	1110.00	1114.64	1117.21	1120.00	1121.00	1122.28	1124.66	1127.02	1130.00	
60%	61%	62%	63%	64%	65%	66%	67%	68%	69%	70%	71%	
1131.00	1134.18	1138.78	1139.97	1141.16	1142.35	1146.54	1147.00	1148.92	1151.11	1152.30	1156.00	
72%	73%	74%	75%	76%	77%	78%	79%	80%	81%	82%	83%	
1156.68	1157.00	1158.06	1159.25	1161.00	1162.63	1164.00	1167.01	1168.20	1169.39	1170.58	1173.00	
84%	85%	86%	87%	88%	89%	90%	91%	92%	93%	94%	95%	
1175.96	1177.30	1180.34	1182.06	1185.44	1187.82	1190.00	1191.58	1194.48	1196.67	1197.86	1199.00	
96%	97%	98%	99%	100%								
1000000.00	1000003.86	1000005.62	1000008.81	1000009.00								

Figura 9 Información percentiles.

Fuente: Propia.

Como podemos observar entre percentil 4 y 5 hay una distancia bastante grande entre 5 y 10 no es tanta la diferencia por lo que podemos tomar 5 como valor a remplazar todo aquello que este debajo de él, y en el caso de los valores extremos entre percentil 95 y 96 una distancia bastante grande entre 95 y 90 no es tanto por lo que se puede tomar como regla de que todos aquellos valores mayores a percentil 95 serán remplazados por él.

Haciendo este cambio tendríamos la siguiente distribución y boxplot:

```
> summary(x)
  Min. 1st Qu. Median     Mean 3rd Qu.    Max.
 1000    1056    1110    1106    1159    1199
```

Figura 10 Resumen de variable estandarizada.

Fuente: Propia.

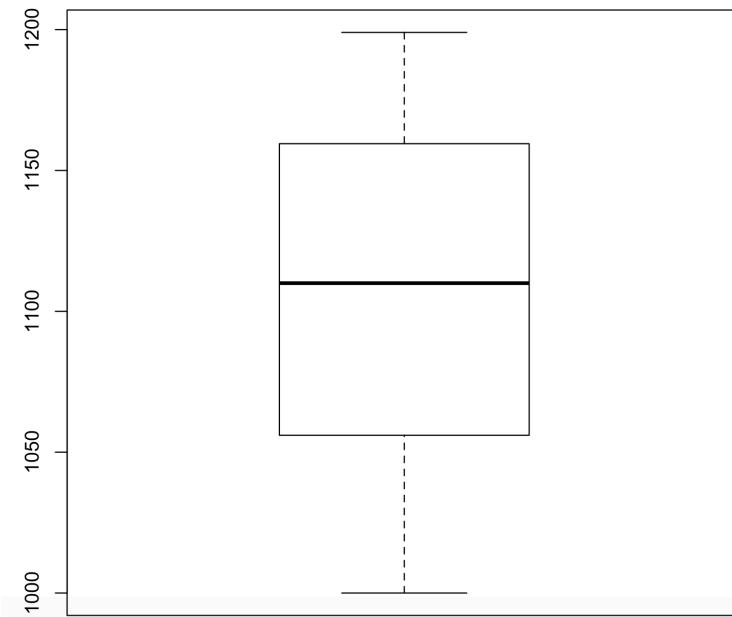


Figura 11 Boxplot variable estandarizada.

Fuente: Propia.

Como podemos observar tenemos ahora una distribución más normal, la distancia entre la media y la mediana están más cerca y la información mostrada en el boxplot está más distribuida mostrando cada uno de los cuartiles casi del mismo tamaño.

Muchas transformaciones pueden formar parte de este proceso y dependerá del negocio en caso de que sean aprobadas, pero normalmente estas son las buenas prácticas que se siguen en esa fase de análisis exploratorio, en algunos casos después de ese tratamiento también suelen restar a todos los valores la media, donde se tendrían más valores decimales y se pueden tener predicciones más suaves.

2.5 Modelado estadístico

Una vez terminada la ingeniería de variables se debe de comenzar el proceso de seleccionar que variables deberían de entrar a mi modelo, que técnica estadística puedo usar y que método de validación cruzada debería de usar. Son tres procesos los que se deben de considerar, lastimosamente no hay una formula en la que se diga en base a tal información te recomendamos utilizar este modelo, estas variables o estas validaciones. Esto es dado por que depende de la información de las variables no siempre es la misma información y varían de conjuntos a conjuntos.

2.5.1 Selección de variables

Selección de variables consta con dos técnicas principales, opción uno podría ser “La mejor selección de subconjuntos”, y como segunda opción “procedimientos de selección de modelos paso a paso”.

2.5.1.1 La mejor selección de subconjuntos

Este algoritmo va eligiendo variables aleatorias y va evaluando el modelo seleccionado a este se le puede almacenar la suma del residuo de error evaluado y se puede comparar basado en eso cuál de todos esos modelos es el mejor.

Algorithm 6.1 Best subset selection

1. Let \mathcal{M}_0 denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
 2. For $k = 1, 2, \dots, p$:
 - (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
 - (b) Pick the best among these $\binom{p}{k}$ models, and call it \mathcal{M}_k . Here *best* is defined as having the smallest RSS, or equivalently largest R^2 .
 3. Select a single best model from among $\mathcal{M}_0, \dots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2 .
-

Figura 12 Algorithm Best subset selection.

Fuente: (James, Witten, Trevor , & Tibshirani, 2015).

2.5.1.2 Procedimientos de selección de modelos paso a paso

Procedimientos de selección de modelos paso a paso tiene dos formas, una es tomando todas e ir agregando una a una esta a su vez puede ir combinando con la anterior de ir agregando una a una para tener N modelos baso a p-1 como primera iteración, y después de eso se puede hacer lo mismo para p-1, donde tendremos una combinación N para p-2, esto como se puede observar generaría una gran cantidad de modelos por lo que se podría concluir que esto genera mucho poder computacional por lo que se ocuparía mucho costo económico y en tiempo.

Algorithm 6.2 *Forward stepwise selection*

1. Let \mathcal{M}_0 denote the *null* model, which contains no predictors.
 2. For $k = 0, \dots, p - 1$:
 - (a) Consider all $p - k$ models that augment the predictors in \mathcal{M}_k with one additional predictor.
 - (b) Choose the *best* among these $p - k$ models, and call it \mathcal{M}_{k+1} . Here *best* is defined as having smallest RSS or highest R^2 .
 3. Select a single best model from among $\mathcal{M}_0, \dots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2 .
-

Figura 13 Forward stepwise (Adelante paso a paso).

Fuente: (James, Witten, Trevor , & Tibshirani, 2015).

Como se puede observar este se evalúa también mediante la suma del residuo del error, y así como existe un modelo donde se va agregando también existe uno donde se comienza todo y se van quitando.

Algorithm 6.3 *Backward stepwise selection*

1. Let \mathcal{M}_p denote the *full* model, which contains all p predictors.
 2. For $k = p, p - 1, \dots, 1$:
 - (a) Consider all k models that contain all but one of the predictors in \mathcal{M}_k , for a total of $k - 1$ predictors.
 - (b) Choose the *best* among these k models, and call it \mathcal{M}_{k-1} . Here *best* is defined as having smallest RSS or highest R^2 .
 3. Select a single best model from among $\mathcal{M}_0, \dots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2 .
-

Figura 14 Backward stepwise (Hacia atrás paso a paso).

Fuente: (James, Witten, Trevor , & Tibshirani, 2015).

2.5.2 Selección de modelos estadísticos

La selección de modelos depende del problema de negocio que se desea abordar, se comienza número uno si es una predicción lo que se desea hacer, normalmente estas predicciones están basadas en una variable de respuesta, por lo que se requiere de información histórica para tener finalmente esa ecuación lineal o sistema de ecuaciones hacer evaluadas en nueva información. La segunda consideración pueda ser para clasificar o perfilar algo, estas se construyen normalmente con técnicas de agrupación (clúster). Estas lo que hacen es explicar mediante la media distintos grupos basado en P variables, en ocasiones estas convienen reducirlas mediante técnicas vectoriales. Y por último dependiendo de la técnica que se deseé utilizar también se debe de considerar el nivel de interpretación que se desea ya que se puede obtener alta precisión, pero baja interpretación o baja precisión, pero alta interpretación.

Estas técnicas pueden ser divididas en dos grandes temas de estudios, una de ellas es máquinas de aprendizaje supervisadas, las que llevan una leyenda a la que se le quiere predecir y la segunda son las no supervisadas, las que basadas en la interpretación del negocio se le puede colocar esa leyenda.

En esta investigación nos centramos en las técnicas supervisadas ya que lo que se desea calcular son los estudiantes con alto riesgo de deserción y los motivos del por qué esto puede llegar a suceder.

2.5.2.1 Máquinas de aprendizaje supervisadas

Para explicar los distintos tipos de modelos que se pueden aplicar vamos a resumir estos con una tabla que explique los principales a ser tomados y en qué caso de negocio se pueden aplicar.

Técnica	Explicación
Naïve Bayes	<p>Permite, dado un ejemplo, encontrar la hipótesis que mejor lo describe. Para poder llegar a una conclusión, necesita que el sistema se nutra de datos suficientes, aunque, cuando el volumen de información es muy grande, debe recurrirse a la hipótesis de la independencia condicional, que permite simplificar la expresión del Teorema de Bayes, factorizando la probabilidad.</p> <p>En las organizaciones, este tipo de algoritmos de machine learning se utilizan para softwares de reconocimiento facial o para determinar si la emoción contenida en un texto es positiva o negativa. Sin embargo, su uso también alcanza a realidades más cotidianas, como el marcar un correo electrónico como spam o no spam. (Logicalis, 2017)</p>
Árboles de decisión	<p>En base a un gráfico, esta herramienta logra servir como apoyo a una toma de decisiones informada, al exponer las distintas opciones y sus posibles consecuencias, incluidos los resultados de eventos fortuitos, los costos de recursos y la utilidad. A la hora de trabajar con este algoritmo es necesario tener en cuenta que hay que conocer el número mínimo de preguntas simples (es decir, las que puedan responderse con un sí o un no) que es preciso lanzar para evaluar la probabilidad de tomar una decisión</p>

	<p>correcta. La ventaja de los árboles de decisión es que permiten abordar el problema de una manera estructurada y sistemática para llegar a una conclusión lógica. Pueden emplearse para predecir la respuesta del público ante el lanzamiento de un nuevo producto o para averiguar la idoneidad de una campaña de marketing. (Logicalis, 2017)</p>
Modelos de regresión lineal	<p>El método de los mínimos cuadrados ordinarios permite es un método para realizar la regresión lineal que puede aplicarse al análisis de relaciones entre variables financieras. El análisis de regresión permite desde desarrollar previsiones de futuro, hasta identificar los factores que mayor incidencia tienen en la generación de beneficios de una corporación o determinar cuánto afectará un cambio en las tasas de interés a una cartera de bonos. (Logicalis, 2017)</p>
regresión logística	<p>Una vez más, la estadística hace posible modelar un resultado binomial con una o más variables explicativas. En este caso, se encarga de medir la relación entre la variable dependiente categórica y una o más variables independientes. Así, aplicando una función logística se pueden estimar las probabilidades de ocurrencia de un suceso. Esta aplicación de learning machine</p>

	sirve a las empresas para elaborar, por ejemplo, sus pronósticos acerca de los ingresos que obtendrán con la venta de un determinado producto o las condiciones climatológicas en un área determinada y una fecha en concreto que pudieran afectar al transporte de mercancías. (Logicalis, 2017)
Máquinas de vectores de soporte	Las denominadas SVM (support vector machines, en inglés) basan su funcionamiento en un algoritmo de clasificación binario que facilita a los negocios identificar el género de los usuarios de sus webs o escoger el tipo de publicidad que debe aparecer en la pantalla, entre otras aplicaciones. (Logicalis, 2017)

Como se puede observar hay muchas técnicas que utilizar dependiendo de la interoperabilidad se puede seleccionar cualquier de estas.

2.6 Como llevar un modelo estadístico a la nube

2.6.1 Software como un servicio (SaaS)

El software como un servicio es un concepto que lleva vigente en la industria de la tecnología desde hace mucho tiempo. Desde que los ingenieros se dieron cuenta que era más barato correr aplicaciones en un centro de datos y distribuirlas a las terminales clientes, el software se convirtió en un servicio. Sin embargo, fue a comienzo del siglo XXI cuando SaaS realmente se convirtió en un término importante para las empresas un poco más pequeñas.

¿Qué es exactamente el SaaS? La explicación más sencilla se puede ver en el modelo de distribución. El software tradicional simplemente se vendía y quedaba alojado en los equipos del comprador. El cambio de modelo enfocado en un servicio implica que el software queda guardado en un servidor remoto y es administrado por la empresa que lo ofrece. Gartner, una de las empresas de investigación de mercados más importante del mundo, define SaaS como: “software que es propiedad, es entregado y es administrado por algún proveedor”. (Santos, 2015)

2.6.2 Solución basada en SaaS

La solución está diseñada como una arquitectura SaaS que permitirá a Laureate cambiar a una vista más estratégica y facilitar implementaciones más rápidas y estandarizadas. Otros beneficios de las soluciones entregadas por software como servicio, Implementaciones rápidas de productos analíticos ofrecidos a las universidades en el nivel de procesamiento de datos y un conjunto estándar de visualizaciones. Disponibilidad automática de información que se adapta a los ciclos académicos de cada institución. Notificaciones automatizadas y alertas tempranas de los estudiantes y profesores para apoyar la toma de decisiones. Base de datos de éxito de estudiantes de reutilización donde la información histórica y centralizada nos permite desarrollar e implementar diferentes soluciones.

Para el software como servicio, MS Azure satisface nuestras necesidades, ya que es un conjunto integral de servicios en la nube que los desarrolladores y profesionales de TI utilizan para crear, implementar y administrar aplicaciones a través de su red global de centros de datos. Otras ventajas se resumen en la siguiente lista: Aprovecharemos Azure Machine Learning para optimizar la forma en que las instituciones obtienen acceso oportuno a información práctica. Azure Data Factory Versión 2 es un servicio totalmente administrado para componer los servicios de almacenamiento, procesamiento y movimiento en servicios simplificados. Canalizaciones de producción confiables MS Azure facilita la integración de las herramientas de BI y la plataforma de CRM Todos los componentes del desarrollador se encuentran en el mismo entorno. Siga el estándar de Office 365 Laureate MS Azure nos brinda diferentes servicios analíticos con documentación sólida y soporte sólido comunidad. MS Azure admite una gama de opciones de implementación, marcos e idiomas populares y un conjunto completo de motores de datos a través de datos relacionales, NoSQL y big data. Utilice esta flexibilidad más el rendimiento, la escala y la seguridad que brindan las tecnologías de Microsoft. El API de búsqueda de MS Azure facilita la búsqueda de sus estudiantes en el Portal.

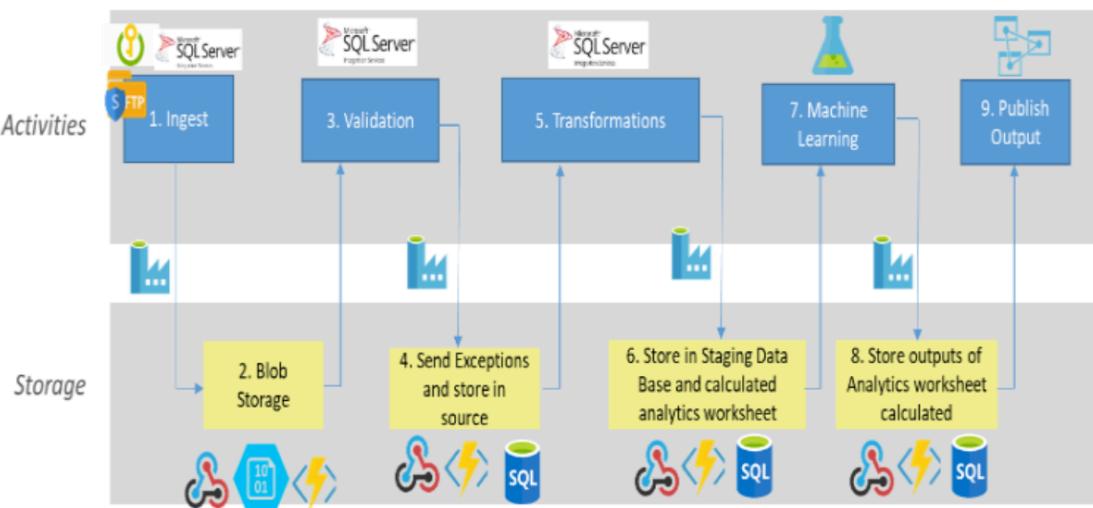


Figura 15 Arquitectura usando componentes en la nube

2.6.3 Microsoft Azure Machine Learning Services

Simplifica y acelera la creación, capacitación y despliegue de sus modelos de aprendizaje automático. Utilice el aprendizaje automático para identificar algoritmos adecuados y ajustar los hiperparámetros más rápido. Mejore la productividad y reduzca los costos con el cálculo automático de escala y DevOps para el aprendizaje automático. Implementar sin problemas en la nube y el borde con un solo clic. Acceda a todas estas capacidades desde su entorno Python favorito utilizando los últimos marcos de código abierto, como PyTorch, TensorFlow y scikit-learn. (Microsoft, 2019)

2.6.4 Azure Data Factory Versión 2

La versión 2 (V2) de Azure Data Factory le permite crear y programar flujos de trabajo controlados por datos (denominados canalizaciones) que pueden ingerir datos de distintos almacenes de datos, procesar / transformar los datos mediante el uso de servicios informáticos como Azure HDInsight Hadoop, Spark, Azure Data Lake Analytics y Azure Machine Learning, y publican datos de salida en almacenes de datos como Azure SQL Data Warehouse para aplicaciones de inteligencia empresarial (BI) para consumir. En última instancia, a través de Azure Data Factory, los datos sin procesar se pueden organizar en almacenes de datos significativos y lagos de datos para tomar mejores decisiones comerciales. (Microsoft, Azure Data Factory version 2 (V2), 2016)

2.6.5 Azure Databricks

Acelera las soluciones de análisis de bigdata y de inteligencia artificial (AI) con Azure Databricks, un servicio de análisis rápido, fácil y colaborativo basado en Apache Spark.

Configura su entorno Spark en minutos y realice una auto escala rápida y fácilmente. Los científicos de datos, ingenieros de datos y analistas de negocios pueden colaborar en proyectos compartidos en un espacio de trabajo interactivo. Aplica sus habilidades existentes con soporte para Python, Scala, R y SQL, así como marcos de aprendizaje profundos y bibliotecas como Tensor Flow, Pytorch y Scikit-learn. La integración nativa con Azure Active Directory (Azure AD) y otros servicios de Azure le permite crear su moderno almacén de datos y sus soluciones de aprendizaje automático y de análisis en tiempo real.

2.6.6 Docker administración de contenedores

Un contenedor es una unidad estándar de software que empaqueta el código y todas sus dependencias para que la aplicación se ejecute de forma rápida y confiable de un entorno informático a otro. Una imagen de contenedor de Docker es un paquete de software liviano, independiente y ejecutable que incluye todo lo necesario para ejecutar una aplicación: código, tiempo de ejecución, herramientas del sistema, bibliotecas del sistema y configuraciones.

Las imágenes de los contenedores se convierten en contenedores en el tiempo de ejecución y, en el caso de los contenedores Docker, las imágenes se convierten en contenedores cuando se ejecutan en un motor de Docker. Disponible para aplicaciones basadas tanto en Linux como en Windows, el software en contenedores siempre se ejecutará de la misma manera, independientemente de la infraestructura. Los contenedores aíslan el software de su entorno y

aseguran que funcione de manera uniforme a pesar de las diferencias, por ejemplo, entre el desarrollo y la puesta en escena. (Inc D. , s.f.)

CAPÍTULO III. METODOLOGÍA

En este capítulo se explica la metodología utilizada para responder nuestras preguntas de investigación. Al mismo tiempo se explica la razón sobre el desarrollo de nuestra hipótesis para abordar la solución de nuestro problema y como se combina ese componente de tecnologías emergentes para poder llevar el desarrollo del producto propuesto a una arquitectura en la nube usando componentes de bigdata.

3.1 Planteamiento de la investigación

Los métodos de investigación aplicados en los proyectos de ciencia de datos normalmente combinan varias metodologías. Como el núcleo de esta investigación consta con todas las fases del ciclo de vida de un proyecto de ciencia de datos se irá describiendo en qué fase se encuentran los métodos científicos con mayor correlación al estudio aplicado.

Comenzando con el nivel Exploratorio donde se analizan todos los eventos fenomenológicos relacionados al problema principal, tomo como experiencia vivida de que yo fui un desertor de la maestría de posgrado y pude observar otros estudiantes que también desertaron por distintos tipos de motivos, por lo que esto me da la apertura de formularme la pregunta de investigación ya descrita anteriormente.

Como una segunda fase a nivel descriptivo se hace el análisis de todas las variables de los distintos modelos de datos recolectados, dando como resultado la frecuencia de variables categóricas y el resumen de las variables numéricas, por lo que se menciona que este análisis en su fase de construcción es un estudio retrospectivo. Se describirá la relación que tienen las categorías con la deserción estudiantil mediante tablas de proporciones.

Posterior a la fase descriptiva se aborda el nivel de investigación relacional donde se demuestra que los grupos involucrados en el desarrollo comparativo son independientes y si son aplicables para la construcción de los modelos correlacionales. En esta fase hacemos la división de nuestro conjunto de datos históricos.

Explicativo, este nivel de nuestro método científico explica las razones del por qué los estudiantes tienden a desertar de la universidad, para alcanzar este nivel de explicación de factores, se aplica previamente un modelo estadístico que normalmente involucra la regresión entre las variables una vez estas están expresadas en un modelo lineal, como por ejemplo la linealización de una regresión logística, los coeficientes nos ayudan a explicar esos factores de riesgo.

Predictivo, este utiliza los pasos utilizados antes, normalmente cuando se construye o se aplica un modelo estadístico este puede ser con fines predictivos como por ejemplo la técnica de potenciación de árboles de decisiones, da como resultado un modelo lineal que se puede emplear para predecir el comportamiento o el estado de una nueva observación. En este proceso de modelado se compara contra el conjunto de validación (validación cruzada) para medir la precisión de la técnica aplicada.

Y por último el nivel aplicativo, el desarrollo de los modelos estadísticos que predicen y explican por qué los estudiantes pueden desertar son desarrollados utilizando tecnologías en la nube y la puesta en producción de estos modelos se realizan a través de computación distribuida configurada en Databricks y Spark.

3.2 Operatividad de variables

Tabla 7 Operación de variables- Estudiantes nuevos.

Problema Específico	Pregunta de investigación	Variable Independientes	Variable dependiente	Hipótesis que comprobar
¿Es posible la construcción de un motor estadístico en la nube para la detección temprana de estudiantes en riesgo de deserción?	¿Quiénes son los estudiantes de primer año de posgrado en riesgo de abandonar sus estudios en el siguiente ciclo?	- Información académica. - Información Financiera. - Información demográfica. - Información del LMS. - Información de ticktes	- Desertor.	Si correlacionamos las variables de información de estudiantes y aplicamos un modelo estadístico podremos predecir que estudiantes tienen mayor riesgo de deserción dado qué no se está analizando la información histórica y operativa.

Fuente: Propia.

Tabla 8 Operación de variables - factores de riesgo.

Problema Específico	Pregunta de investigación	Variable Independientes	Variable dependiente	Hipótesis para comprobar
¿Es posible la construcción de un motor estadístico en la nube para la detección temprana de estudiantes en riesgo de deserción?	¿Cuáles son los principales factores o variables significativas que influyen en la deserción estudiantil?	<ul style="list-style-type: none"> - Transformación de información académica. - Transformación de Información Financiera. - Transformación de Información demográfica. - Transformación de Información del LMS. - Transformación de Información de ticktes 	<ul style="list-style-type: none"> - Factores de riesgo. 	Si aplicamos un modelo estadístico lineal podremos identificar los factores de riesgo más importantes sustentados mediante estadística descriptiva dado que los modelos lineales son los más fáciles de interpretar.

Fuente: Propia.

Tabla 9 Operación de variables - Calculo de deciles.

Problema Específico	Pregunta de investigación	Variable Independientes	Variable dependiente	Hipótesis para comprobar
¿Es posible la construcción de un motor estadístico en la nube para la detección temprana de estudiantes en riesgo de deserción?	¿Quiénes son los estudiantes con mayor probabilidad de abandonar estudios en el siguiente ciclo académico?	- Información académica. - Información Financiera. - Información demográfica. - Información del LMS. - Información de ticktes	- Calificación de riesgo. - Decil de riesgo	Si calculamos una calificación de riesgo y los agrupamos en deciles se podrá intervenir más efectivamente los estudiantes con mayor riesgo de deserción dado que en modelos aplicados en pregrado ha dado buenos resultados y se observa que la detección de estos estudiantes es más efectiva.

Fuente: Propia

3.3 Enfoque de la investigación

Esta investigación tiene un efecto de metodologías mixtas que combina el marco de trabajo tradicional aplicado por científicos de datos en resolución de problemas. Ya que esta lleva una serie sistemática de pasos de cómo a través de los datos se pueden construir productos estadísticos y gracias a los marcos de trabajo tecnológicos que tenemos hoy en día podemos implementar estos en arquitecturas basadas en la nube.

Gracias a Laureate mediante el departamento de “Marketing and business Insights” se podrá implementar este modelo de postgrado de la misma manera que se elaboró con pregrado en UNITEC Honduras, dando como valor agregado y por la complejidad de la información de los estudiantes de posgrado, acceso a más modelos de datos como el detalle de la información por curso de la misma manera se está evaluando incluir información de Blackboard e información de tickets.

3.4 Alcance de la investigación

Esta investigación tiene como alcance dejar conceptualizado el marco de trabajo principal que debe de seguir cualquier científico de datos y deja como entregable modelos estadísticos entrenados que pueden ser implementados en tecnologías basadas en la nube para su fácil consumo e interpretación.

Como toda investigación también se deja como alcance: métricas bien definidas en el sentido que siempre miden algo importante que está relacionado con dimensiones de riesgo y permiten claramente una intervención, por ejemplo, mediante un percentil determinar quiénes son esos estudiantes que son buenos o malos pagadores.

Un proyecto de ciencia de datos siempre presenta muchas hipótesis ya que al momento de diseñar el modelo estadístico se tuvo que haber pasado por un proceso de potenciación de variables relacionado con nuestra variable principal de estudio.

3.5 Diseño de la investigación

La primera fase de diseño de investigación consta de lo que es la limpieza de las variables de los modelos de datos entregados, esto conlleva lo que es estandarización de variables numéricas basado en análisis de valores extremos e imputaciones validadas por la institución y con respecto a las categóricas se realizan técnicas de agregación para eliminar sesgos en niveles de baja representatividad y baja correlación con nuestra variable de respuesta.

La segunda fase de esta investigación consiste en la aplicación de modelos exploratorios de datos, en esta explicamos las variables más importantes según análisis de correlaciones y opiniones de expertos comparadas con nuestra variable de respuesta, también, se trata de eliminar la multicolinealidad combinando variables por lo cual se vuelve a justificar que esta investigación utiliza ese método “Exploratorio”.

En la tercera fase comienza todo el modelado estadístico, aquí se debe de elegir tres cosas; número uno es la selección de variables, este como se mencionó en el marco teórico se puede hacer basado en dos métodos o usando la opinión de un experto. La segunda es selección de técnica estadística, en esta se pueden elegir métodos de regresión y técnicas vectoriales. La tercera es selección de validaciones cruzadas para validar que el modelo estadístico tiene buena presión.

La cuarta fase consiste en hacer el resumen de las variables importantes y el análisis de las variables de salida, como por ejemplo los deciles de riesgo y el análisis de las variables importantes mediante estadística descriptiva.

La quinta y última fase es la implementación de los modelos en la nube, básicamente aquí se programan todas las transformaciones hechas en el modelo exploratorio utilizando R, Python y SPARK. Utilizamos servicios web para llevar el modelo estadístico, serverless para ejecutar el modelo correcto en determinada semana y fábricas de datos para orquestar el llamado de todas las actividades.

3.6 Herramientas e Instrumentos

La recolección de información fue mediante reuniones con el equipo de TI compartiendo los modelos de datos necesarios para iniciar el proceso de limpieza y modelado exploratorio de datos.

Información semanal:

Columna	Definición	% de nulos
CUENTA	Número de cuenta del estudiante.	0%
ANIO	Año de matrícula.	0%
SEMESTRE	Semestre de matrícula (1 y 2).	0%
MODULO	Trimestre o ciclo de matrícula (1,2,4 y 5).	0%
SEMANA	Semana del año y ciclo correspondiente.	0%
FACULTAD	Escuela o facultad a la que pertenece la carrera.	0%

GRADO	Grado académico a la que pertenece la carrera (Licenciatura, maestría, doctorado, etc).	0%
CARRERA	Carrera del estudiante para el año y ciclo correspondiente.	0%
SEDE	Campus donde cursa el estudiante en el año y ciclo correspondiente.	0%
JORNADA	Jornada de clases que cursa el estudiante en el año y ciclo correspondiente.	0%
TIPO ESTUDIANTE	Clasificación del estudiante para el año y ciclo correspondiente.	0%
NOTA PROMEDIO	Nota promedio en la semana del año y ciclo correspondiente.	0%
PORCENTAJE ASISTENCIA	Porcentaje de asistencia en la semana del año y ciclo correspondiente.	0%
DESERTOR	Indica si el estudiante es un desertor temprano en la semana correspondiente.	0%
TEMPRANO		

Información de Registro:

Columna	Definición	% de nulos
CUENTA	Número de cuenta del estudiante.	0%

ANIO	Año de matrícula.	0%
SEMESTRE	Semestre de matrícula (1 y 2).	0%
MODULO	Trimestre o ciclo de matrícula (1,2,4 y 5).	0%
SEXO	Sexo del estudiante.	0%
EDAD	Edad del estudiante.	0%
ZIPCODE	Código postal (No disponible).	100%
DEPTO INSCRIPCION	Departamento de inscripción.	0%
MUNICIPIO	Municipio de inscripción.	0%
INSCRIPCION		
NIVEL	Nivel a la que pertenece la carrera (Bachillerato, tecvoc, pregrado y postgrado).	0%
FACULTAD	Escuela o facultad a la que pertenece la carrera.	0%
GRADO	Grado académico a la que pertenece la carrera (Licenciatura, maestría, doctorado, etc).	0%
CARRERA	Carrera del estudiante para el año y ciclo correspondiente.	0%
MARCA	Marca a la que pertenece el estudiante (UNITEC, CEUTEC).	0%
CIUDAD MARCA	Ciudad de la marca donde el estudiante cursa su carrera (TEGUCIGALPA, SAN PEDRO SULA, LA CEIBA, DUV).	0%

JORNADA	Jornada de clases que cursa el estudiante en el año y ciclo correspondiente.	0%
TIPO ESTUDIANTE	Clasificación del estudiante para el año y ciclo correspondiente.	0%
ANIO INGRESO	Año de ingreso a la universidad.	0%
SEMESTRE INGRESO	Semestre de ingreso a la universidad (1 y 2).	0%
MODULO INGRESO	Trimestre o ciclo de ingreso a la universidad (1,2,4 y 5).	0%
PERIODOS HISTORICOS	Cantidad de ciclos matriculados desde el ingreso a la universidad hasta el año y ciclo correspondiente (este último se incluye).	0%
CLASES HISTORICAS CURSADAS	Cantidad de clases cursadas desde el ingreso a la universidad hasta el año y ciclo correspondiente (este último se incluye).	0%
CLASES HISTORICAS APB	Cantidad de clases aprobadas desde el ingreso a la universidad hasta el año y ciclo correspondiente (este último se incluye).	0%
CLASES HISTORICAS REP	Cantidad de clases reprobadas desde el ingreso a la universidad hasta el año y ciclo correspondiente (este último se incluye).	0%

CLASES HISTORICAS RET	Cantidad de clases retiradas desde el ingreso a la universidad hasta el año y ciclo correspondiente (este último se incluye).	0%
CLASES HISTORICAS SD	Cantidad de clases sin derecho desde el ingreso a la universidad hasta el año y ciclo correspondiente (este último se incluye).	0%
UVs HISTORICAS CURSADAS	Cantidad de créditos cursados desde el ingreso a la universidad hasta el año y ciclo correspondiente (este último se incluye).	0%
UVs HISTORICAS APB	Cantidad de créditos aprobados desde el ingreso a la universidad hasta el año y ciclo correspondiente (este último se incluye).	0%
UVs HISTORICAS REP	Cantidad de créditos reprobados desde el ingreso a la universidad hasta el año y ciclo correspondiente (este último se incluye).	0%
UVs HISTORICAS RET	Cantidad de créditos retirados desde el ingreso a la universidad hasta el año y ciclo correspondiente (este último se incluye).	0%
UVs HISTORICAS SD	Cantidad de créditos sin derecho desde el ingreso a la universidad hasta el año y ciclo correspondiente (este último se incluye).	0%

PROMEDIO HISTORICO	Promedio simple histórico calculado con la fórmula (suma (nota clase histórica) / cuenta (clase histórica)).	0%
FECHA INICIO PERIODO	Fecha de inicio del año y periodo correspondiente.	0%
FECHA MATRICULA	Fecha de matrícula de clases del estudiante en el año y periodo correspondiente..	0%
DIFERENCIA FECHAS	Resultado en días de la diferencia FECHA MATRICULA - FECHA INICIO PERIODO.	0%
TIPO MATRICULA	Modalidad de matrícula (WEB, PRESENCIAL).	0%
UVs MATRICULADAS	Créditos matriculados en el año y ciclo correspondiente.	0%
CLASES MATRICULADAS	Clases matriculadas en el año y ciclo correspondiente.	0%
CLASES APROBADAS	Clases aprobadas en el año y ciclo correspondiente.	0%
CLASES REPROBADAS	Clases reprobadas en el año y ciclo correspondiente.	0%
CLASES RETIRADAS	Clases retiradas en el año y ciclo correspondiente.	0%
CLASES SIN DERECHO	Clases sin derecho en el año y ciclo correspondiente.	0%
INDICE PERIODO	Índice del estudiante para el año y ciclo correspondiente calculado con la fórmula (suma (nota clase * créditos) / suma(créditos))	0%

PORCENTAJE ASISTENCIA	Porcentaje de asistencia en el año y ciclo correspondiente.	0%
DESERTOR TEMPRANO	Indica si el estudiante retiró y/o quedó sin derecho en todas sus clases matriculadas en el año-ciclo correspondiente (SI, NO).	0%
DESERTOR SIG PERIODO	Indica si el estudiante estuvo matriculado en el siguiente año-ciclo, considerando que no haya egresado. (SI, NO)	0%
EGRESADO_BANDERA	Indica si el estudiante lleva la última clase del programa de estudio en el año y ciclo correspondiente.	

Información de pagos:

Columna	Definición	% de nulos
CUENTA	Número de cuenta del estudiante.	0%
ANIO	Año de matrícula.	0.12%
SEMESTRE	Semestre de matrícula (1 y 2).	0.12%
MODULO	Trimestre o ciclo de matrícula (1,2,4 y 5).	0%
TIPO BENEFICIO	Beneficio brindado al estudiante para el ANIO-SEMESTRE-MODULO en cuestión.	92.28%

SALDO	Saldo pendiente de los documentos generados en el ANIO-SEMESTRE-MODULO en cuestión.	0.12%
SALDO ACUMULADO	Saldo pendiente de los documentos generados históricamente (incluyendo los generados en el ANIO-SEMESTRE-MODULO en cuestión).	0.12%
PAGADO	Monto pagado entre las fechas de inicio y final de clases.	0.12%
FECHA PAGO	Fecha del pago realizado.	0%
FORMA PAGO	Forma en que pagó el estudiante para la cuota correspondiente.	0%
FECHA A PAGAR	Fecha en que debió pagarse la cuota del documento.	0%
MULTA	Recargo generado a la cuota pagada.	0.12%

Información de estudiantes de primer ingreso:

Columna	Definición	% de nulos
CUENTA	Número de cuenta del estudiante	0%
ANIO	Año de matrícula	0%
SEMESTRE	Semestre de matrícula (1 y 2)	0%
MODULO	Trimestre o ciclo de matrícula (1,2,4 y 5)	0%
SEXO	Sexo del estudiante	8.94%
EDAD	Edad del estudiante	0.85%

ZIPCODE	Código postal (No disponible)	100%
DEPTO INSCRIPCION	Departamento de inscripción	0%
MUNICIPIO	Municipio de inscripción	0%
INSCRIPCION		
MARCA	Marca a la que pertenece el estudiante (UNITEC, CEUTEC)	0%
CIUDAD MARCA	Ciudad de la marca donde el estudiante cursa su carrera (TEGUCIGALPA, SAN PEDRO SULA, LA CEIBA, DUV)	0%
NIVEL	Nivel a la que pertenece la carrera (Bachillerato, tecvoc, pregrado y postgrado).	0%
FACULTAD	Escuela o facultad a la que pertenece la carrera.	0%
GRADO	Grado académico a la que pertenece la carrera (Licenciatura, maestría, doctorado, etc).	0%
CARRERA	Carrera del estudiante para el año y ciclo correspondiente.	0%
JORNADA	Jornada de clases que cursa el estudiante en el año y ciclo correspondiente.	0%
COLEGIO	Nombre del colegio de procedencia.	32.08%
TIPO COLEGIO 1	Tipo del colegio de procedencia (Privado, Público).	13.18%
TIPO COLEGIO 2	Tipo del colegio de procedencia (Bilingüe, Monolingüe).	13.18%

ANIO GRADUCION	Año de graduación del colegio (NA).	100%
COLEGIO		
NIVELACION	Resultado de programa de nivelación o curso	100%
MATEMATICAS	propedéutico en matemáticas (NA).	
NIVELACION	Resultado de programa de nivelación o curso	100%
LECTURA-ESCRITURA	propedéutico en lectura y escritura (NA).	
FORMA INGRESO	Forma de ingreso a la institución (NA).	100%
NACIONALIDAD	Nacionalidad extraída a través del número de identidad (Extranjera, hondureña).	0%
NACIONALIDAD2	Nacionalidad Identidad ingresada por registro (hondureña, panameña, salvadoreña, etc....).	0%

CAPÍTULO IV. RESULTADOS

En este capítulo se comienza analizar todas las fuentes primarias de datos proporcionadas por la dirección tecnológica de UNITEC, el contenido de este será desarrollado en cuatro fases mostrando el resultado de cada una de ellas, comenzando con el análisis exploratorio de nuestra variable de estudio, subsecuente a este se muestra una propuesta de diseño de modelos basado en las variables más significativas, como tercera fase mostraremos la selección de variables y el método de validación cruzada utilizada para demostrar nuestra hipótesis y como última fase se valida como esta investigación puesta en entorno productivo puede tener éxito al momento de hacer las intervenciones.

4.1 Análisis exploratorio de datos de registro

Como se definió el enfoque de esta investigación utiliza métodos retrospectivos por lo que se cuenta con un conjunto de datos históricos para elaborar este estudio. Comencemos analizando cuales son las observaciones con las que contamos.

Contamos con un total de 20,373 observaciones, donde 2,427 casos son estudiantes que desertaron de un periodo a otro, representando un 12%, esto comprendido en 7 periodos académicos desde 20171 hasta 20184.

La distribución de estudiantes por carrera está comprendida de la siguiente manera:

	CAREER	n	percentage
1	MAESTRÍA EN DESARROLLO LOCAL Y TURISMO	1	0.00
2	PROGRAMA DE POSTGRADO NEGOCIOS INTERNACIONAL...	1	0.00
3	MAESTRÍA EN CONTADURÍA PÚBLICA	2	0.00
4	MAESTRÍA EN DIRECCIÓN DE MERCADOTECNIA	7	0.00
5	MAESTRÍA EN GESTIÓN DE OPERACIONES Y LOGÍSTICA	101	0.00
6	POSTGRADO EN ECONOMÍA Y EMPRESA	147	0.01
7	SISTEMAS DE GESTIÓN DE CALIDAD INTEGRADOS EN EL ...	362	0.02
8	MAESTRÍA EN INGENIERÍA DE ESTRUCTURAS	454	0.02
9	MASTER EN GESTIÓN DE TECNOLOGÍAS DE LA INFORMA...	578	0.03
10	GESTIÓN DE ENERGÍAS RENOVABLES	579	0.03
11	DIRECCIÓN DE LA COMUNICACIÓN CORPORATIVA	719	0.04
12	PROGRAMA DE POSTGRADO EN DERECHO EMPRESARIAL	1005	0.05
13	PROGRAMA DE POSTGRADO DE DIRECCIÓN DE RECURS...	1162	0.06
14	FINANZAS	2684	0.13
15	PROGRAMA DE POSTGRADOS DE ADMINISTRACIÓN DE P...	3974	0.20
16	DIRECCIÓN EMPRESARIAL	8597	0.42

Figura 16 Distribución Estudiantes por carrera.

Fuente: Propia.

Como se puede observar contamos con 16 carreras de las cuales tenemos algunas que representan casi 0% de la población ya que contamos con carreras que tienen menos de 10 estudiantes, como primera regla de limpieza de datos hemos decidido excluir esta información ya que puede ser considerada como valores atípicos. Por lo tanto, las siguientes carreras a ser eliminadas son:

	CAREER	n
	<chr>	<int>
1	MAESTRÍA EN CONTADURÍA PÚBLICA	2
2	MAESTRÍA EN DESARROLLO LOCAL Y TURISMO	1
3	MAESTRÍA EN DIRECCIÓN DE MERCADOTECNIA	7
4	PROGRAMA DE POSTGRADO NEGOCIOS INTERNACIONALES - GLOBAL MBA-	1

Figura 17 Carreras excluida.

Fuente: Propia.

Como observación para la lectura de las siguientes tablas hay que tomar en cuenta de que “0” representa los estudiantes que no desertaron el periodo y siguiente y “1” representa los estudiantes que si desertaron.

Ahora si analizamos nuestra distribución por periodo académico podremos observar la siguiente distribución, donde se puede señalar que al inicio de cada semestre (periodos 1 y 4) hay un mayor porcentaje de deserción:

	0	1
20171	0.87442472	0.12557528
20172	0.88479101	0.11520899
20174	0.86246922	0.13753078
20175	0.90037594	0.09962406
20181	0.87265799	0.12734201
20182	0.90345772	0.09654228
20184	0.87068966	0.12931034

Figura 18 Porcentaje de deserción por periodo académico.

Fuente: Propia.

Antes de seguir analizando cada una de las variables categóricas y como estas se correlacionan con nuestra variable de respuesta primero analicemos, descartemos variables que solo tienen un nivel dentro de sus categorías o valores demasiados extremos de los cuales no hace sentido tomarlos crudos, estos normalmente requieren de un tipo de tratamiento.

	column	different_levels
5	NIVEL	1
6	FACULTAD	1
9	MARCA	1
15	TIPO.MATRICULA	2
16	DESERTOR.SIG.PERIODO	2
17	BANDERA.EGRESADO	2
2	SEXO	3
7	GRADO	3
12	TIPO.ESTUDIANTE	3
10	CIUDAD.MARCA	4
11	JORNADA	5
13	FECHA.INICIO.PERIODO	7
20	ANIO.MODULO	7
19	CAREER	12
18	Translation	13
8	CARRERA	22
3	DEPTO.INSCRIPCION	35
4	MUNICIPIO.INSCRIPCION	298
1	CUENTA	6003
14	FECHA.MATRICULA	19188

Figura 19 Niveles por variable categóricas.

Fuente: Propia.

Como podemos observar inicialmente contamos con tres variables que solo tienen un nivel, las cuales podemos descartar ya que no aportan información en nuestro análisis. CARRERA es una columna que viene con caracteres especiales por lo que se limpió y quedo en una nueva llamada CAREER, translación sirvió en el proceso de la limpieza por lo que se puede eliminar también.

Es interesante notar que variables como CUENTA lo que nos trasmite es el número único de estudiantes que ha habido en estos últimos periodos, 6003 estudiantes. Por lo que podemos concluir que este estudio se tomó en cuenta esa cantidad de estudiantes. CAREER después de su limpieza queda con un total de 12 carreras.

Otra variable interesante y con valor extraño que podemos notar es DEPTO.INSCRIPCION ya que este cuenta con 35 valores, si hacemos un conteo rápido de los distintos niveles de este, notaremos lo siguiente:

> table(TrimestralPosgrado\$DEPTO_INSCRIPCION)								
ATLANTIDA	ATLÁNTIDA	CHOLUTECA	COLON	COLÓN	COMAYAGUA	COPAN	COPÁN	
203	125	765	38	21	686	142	66	
CORTES	EL PARAISO	EL PARAÍSO	FRANCISCO MORAZÁN	FRANCISCO MORAZÁN	GRACIAS A DIOS	INTIBUCA	INTIBUCÁ	
3774	252	118	2701	1464	18	52	34	
ISLAS DE LA BAHIA	ISLAS DE LA BAHIA	LA PAZ	LEMPIRA	NINGUNO	OCOTEPEQUE	OLANCHO	SANTA BARBARA	
4	3	230	166	302	140	702	133	
SANTA BÁRBARA	VALLE	YORO	ATLANTIDA	COLAN	COPUN	EL PARAOSO	FRANCISCO MORAZÁN	
74	397	844	348	88	214	394	5446	
INTIBUCÁ	ISLAS DE LA BAHIA	SANTA BJRBARA						
111	9	298						

Figura 20 Distribución Departamentos de Honduras sin limpiar.

Fuente: Propia.

Este igual que Carrera requiere de una limpieza, por esos valores que están codificados de manera errónea. Después de hacer la limpieza de esta notamos la siguiente distribución:

	0	1
ISLAS DE LA BAHIA	0.7500000	0.2500000
COLON	0.7959184	0.20408163
LEMPIRA	0.8253012	0.17469880
CORTES	0.8330684	0.16693164
OCOTEPEQUE	0.8357143	0.16428571
INTIBUCA	0.8373206	0.16267943
COPAN	0.8412322	0.15876777
YORO	0.8483412	0.15165877
SANTA BARBARA	0.8534653	0.14653465
ATLANTIDA	0.8535503	0.14644970
COMAYAGUA	0.8717201	0.12827988
OLANCHO	0.8831909	0.11680912
NINGUNO	0.8907285	0.10927152
FRANCISCO MORAZAN	0.9038602	0.09613984
VALLE	0.9042821	0.09571788
EL PARAISO	0.9096859	0.09031414
CHOLUTECA	0.9241830	0.07581699
LA PAZ	0.9304348	0.06956522
GRACIAS A DIOS	0.9444444	0.05555556

Figura 21 Deserción por departamento (Valores limpios).

Fuente: Propia.

Como podemos observar tenemos varios departamentos con altos porcentajes de deserción sin embargo son departamentos con muestras pequeñas, aparte de que esta categoría se podría considerar que tiene muchos niveles por lo que transformar esta columna es buena práctica, como por ejemplo las 6 zonas geográficas del país.

	0	1
occidental	0.8365385	0.16346154
noroccidental	0.8375952	0.16240484
nororiental	0.8449704	0.15502959
centro occidental	0.8773333	0.12266667
NINGUNA	0.8907285	0.10927152
centro oriental	0.9029521	0.09704794
sur	0.9173838	0.08261618

Figura 22 Deserción por zona geográfica del país.

Fuente: Propia.

Ahora podemos hacer un par de conclusiones que pueden generar efectivamente intervenciones estudiantes que provienen de las zonas del occidente y el norte tiene mucho mayor riesgo de deserción. “NINGUNA” es una categoría para estudiantes que no se pudo capturar su información.

```
> table(TrimestralPosgrado$GEO.ZONE)
centro occidental    centro oriental          NINGUNA      noroccidental      nororiental      occidental      sur
           1125            11077             302            5123              845             728            1162
```

Figura 23 Número de estudiantes sin clasificación de zona geográfica.

Fuente: Propia.

Si observamos estos estudiantes podemos observar que son muy pocos, representando menos del uno por ciento, y la zona que tiene una mayor cantidad de inscripciones es centro oriental. Por lo que podemos agregar a esos estudiantes bajo esa categoría.

	0	1
occidental	0.8365385	0.16346154
noroccidental	0.8375952	0.16240484
nororiental	0.8449704	0.15502959
centro occidental	0.8773333	0.12266667
centro oriental	0.9026276	0.09737235
sur	0.9173838	0.08261618

Figura 24 Deserción por zona geográfica - datos limpios.

Fuente: Propia.

Como podemos observar los porcentajes de centro occidental no se movieron mucho por cual se puede concluir que esta agregación es correcta.

Otra variable importante que podemos analizar es la variable tipo de estudiante, en primer lugar, podemos ver que tan representativos son cada una de las categorías con respecto al total.

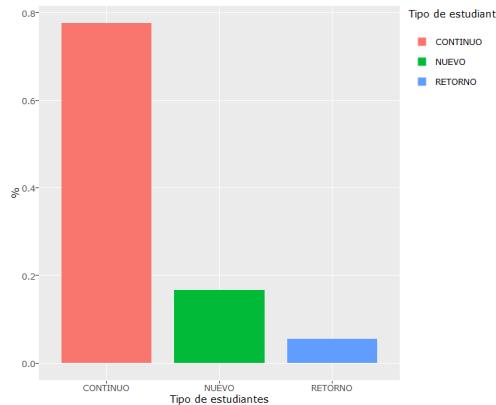


Figura 25 Porcentaje por tipo de estudiantes.

Fuente: Propia.

Como podemos observar la mayor categoría es la de estudiantes continuos en segundo lugar los estudiantes nuevos y por último tenemos los estudiantes de retorno, ahora si analizamos cada una de las categorías con respecto a su deserción podemos observar lo siguiente:

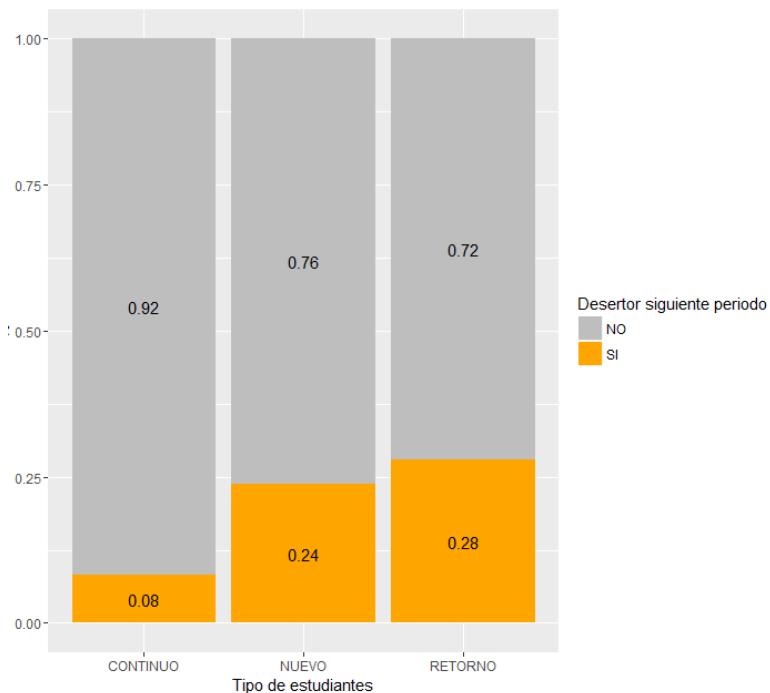


Figura 26 Porcentaje de deserción por tipo de estudiantes.

Fuente: Propia.

A pesar de que los continuos tienen mayor población su deserción es más baja comparada con los nuevos y retorno, ahora si revisamos individualmente estas categorías podemos observar que solamente hay cuatro puntos porcentuales de diferencia por lo que podemos agregar estas dos en una sola categoría y más aun sabiendo que retorno realmente tiene una representación bastante baja.

Haciendo este cambio rápidamente podemos observar la siguiente distribución:

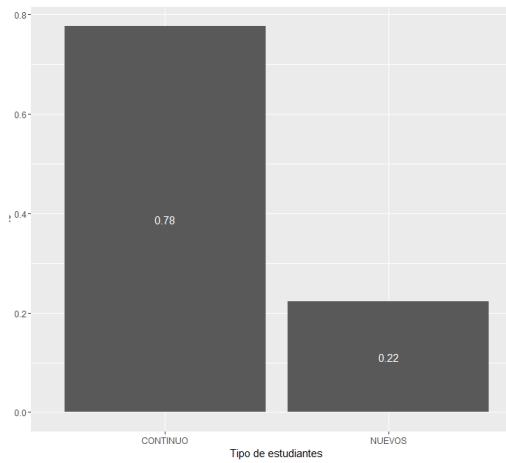


Figura 27 Distribución de estudiantes agrupados.

Fuente: Propia.

La nueva categoría de NUEVOS representa un 22% que es la suma de las categorías anteriores, ahora si volvemos a analizar la distribución con respecto a nuestra variable de respuesta tendremos el siguiente comportamiento:

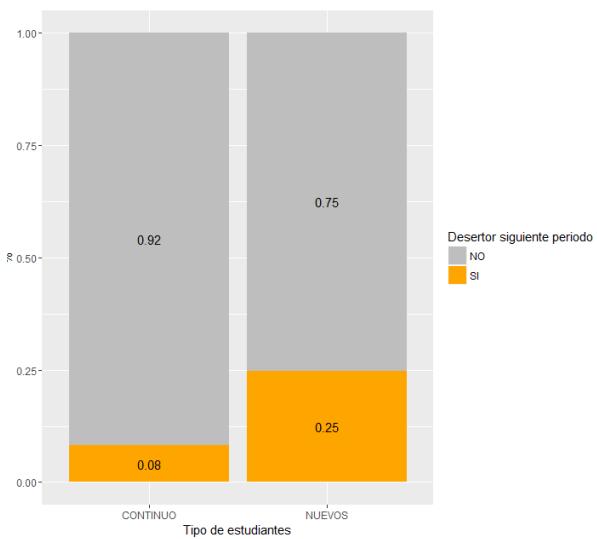


Figura 28 Porcentaje de deserción por tipo de estudiantes agrupados.

Fuente: Propia.

NUEVOS ahora tiene un punto porcentual de mayor representatividad comparado a como estaba antes, esto es bueno ya que gracias a esta agregación podemos hacer modelos dedicados en base a esta categoría ya que tenemos una fuerte diferencia.

Esto nos da una buena primera apertura para construir los dos primeros modelos para nuestro Inicio de semestre, un modelo será para estudiantes nuevos y el otro será para estudiantes de modalidad continua.

Para cada uno de estos modelos se considerarán el compromiso con registro, este será dado como CLASES.RETIRADAS/CLASES.CURSADAS en un inicio esta es la única forma que podemos medir esto, vale la pena aclarar que para los estudiantes continuos tendrás esta variable tanto para lo matriculado como para lo histórico. Al hacer este tipo de transformaciones tratamos de eliminar la multicolinealidad. Las unidades valorativas serán excluidas de este análisis ya que estas son consideradas en el cálculo del índice.

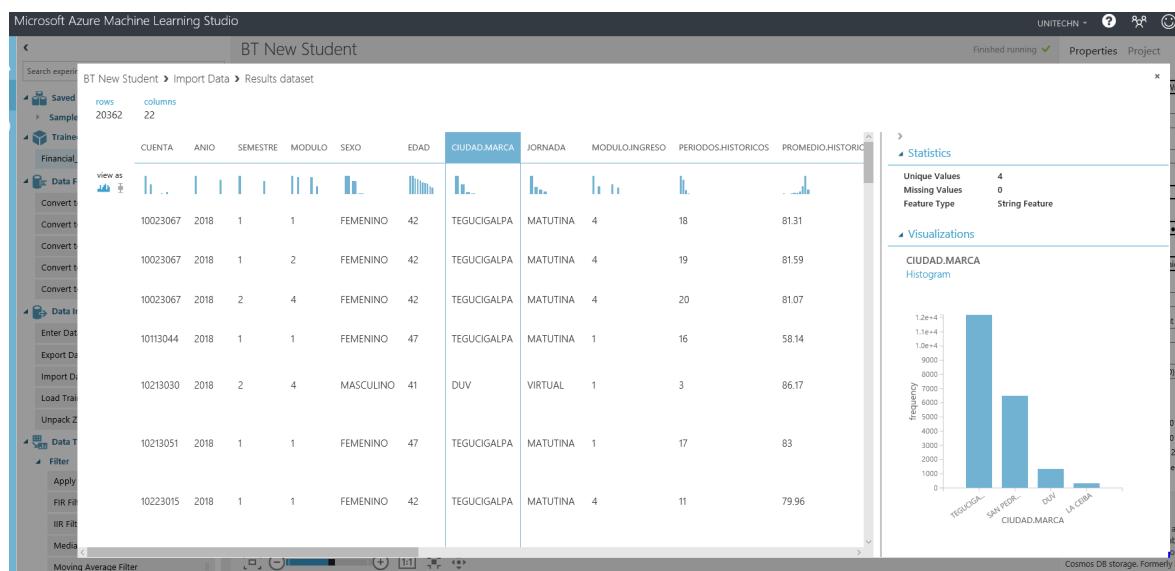


Figura 29 Distribución mostrada desde Microsoft Azure.

Fuente: Propia.

Como podemos observar obtuvimos una correcta carga de la información ahora lo que se procede es filtrar la información a nivel de registros y columnas, recordemos que el primer modelo que entrenaremos es para estudiantes nuevos por lo que toda la información relacionada con estudiantes continuos se debe de excluir, de la misma manera se hará un proceso de validación cruzada para revisar la precisión de nuestro entrenamiento.

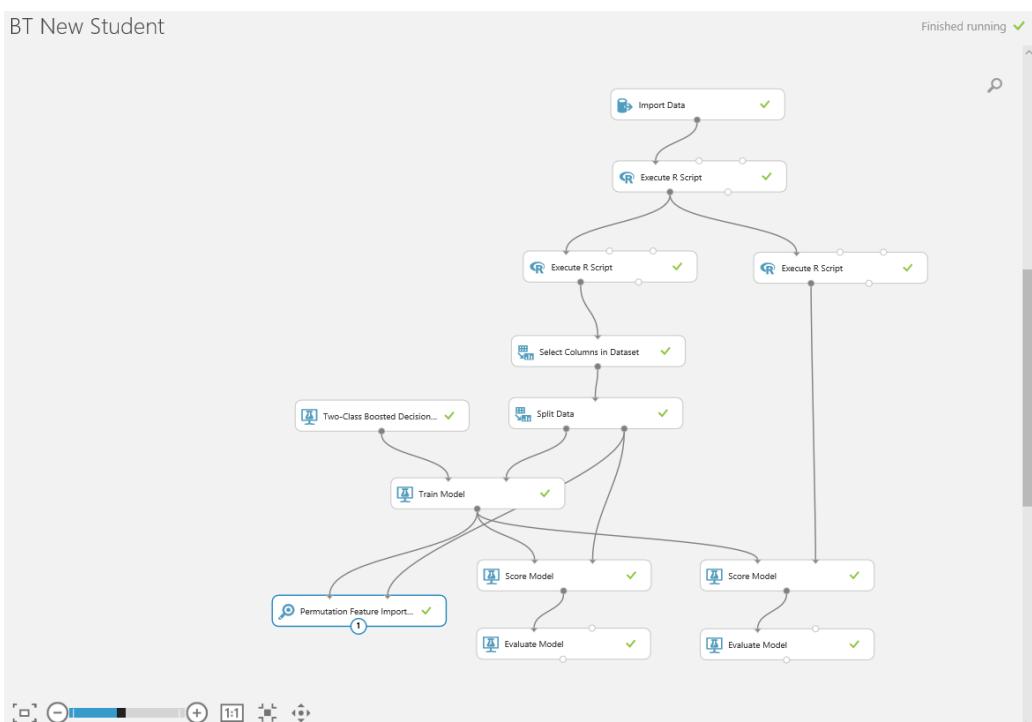


Figura 30 ETL Entrenamiento para nuevos estudiantes Inicio del Periodo.

Fuente: Propia.

Este modelo de inicio de semestre para estudiantes nuevos como se puede observar consta de dos ramas principales, la rama de la izquierda se hace la primera validación cruzada para entrenar y hacer prueba y la rama de la izquierda es la que se utiliza para validar. Los resultados de este modelo se consideran no malos ya que en este modelo no contamos con mucha información.

Como información extra este modelo se entrenó con información de 2017 utilizando todos los módulos encontrados para este año, para 2018 se utilizaron modulo uno y dos. Como se mencionó en la rama de la izquierda se utiliza el módulo de 4 de 2018 para hacer toda la validación de este mismo.

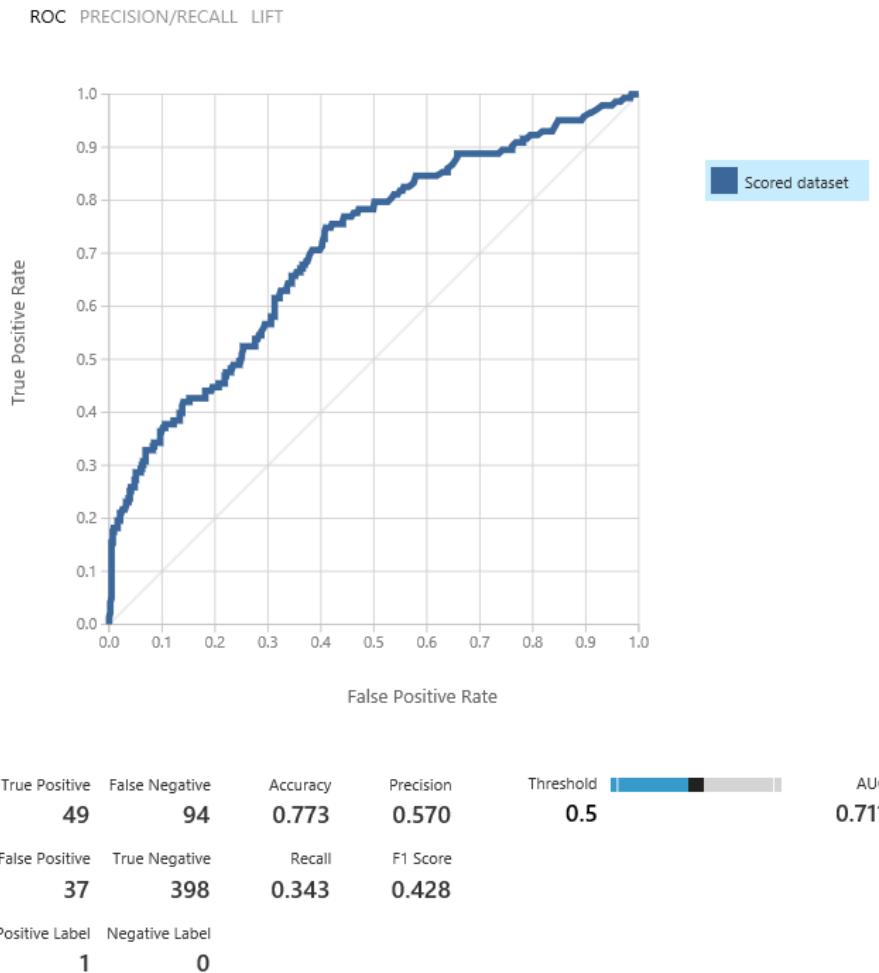


Figura 31 Validación del modelo de inicio de ciclo.

Fuente: Propia.

Como se puede observar este modelo tiene varias formas de validarla podríamos decir que tiene una buena área bajo la curva, dándonos nuestra primera referencia con una calificación de

0.7 ahora en referencia a la precisión tenemos un valor tanto bajo con 0.57 sin embargo para ser un modelo de inicio de semestre está muy bien valorado.

Otra información importante para leer son sus variables importantes ya que este nos ayudara hacer todo el trabajo de intervención, para este modelo podemos encontrar los siguientes valores.

rows	columns	
11	2	
view as	Feature	Score
	START.ENDGAGEMENT	0.068884
	DIFERENCIA.FECHAS	0.017815
	JORNADA	0.017221
	MODULO.INGRESO	0.016033
	PERIODOS.HISTORICOS	0.010095
	CIUDAD.MARCA	0.007126
	GEO.ZONE	0.006532
	EDAD	0.005344
	CAREER	0.003563
	SEXO	0.001188
	TIPO.MATRICULA	-0.001188

Figura 32 Variables más importantes modelo inicio de semestre.

Fuente: Propia.

Este modelo nos indica que aquellos estudiantes que comienzan a retirar clases tienen una mayor probabilidad de deserción. Seguido de la jornada y así sucesivamente muy importante revisar mediante deciles cómo está la interpretación del modelo. Antes de ver los deciles podemos ver cómo está el histograma de esta predicción.



Figura 33 Distribución predicción modelo inicio de semestre.

Fuente: Propia.

Esta distribución se ve bastante normalizada ya que la mayoría de los estudiantes están distribuidos de lado izquierdo tal cual como lo muestra la mediana. Ahora recordando un poco los conceptos básicos de los deciles estos agrupan la distribución en diez grupos iguales y para cada barra mide el desempeño del modelo, lo que se espera es ver en los primero deciles ordenando de mayor a menor los estudiantes con mayor riesgo y que a medida este se comience a mover este pierda precisión por que claramente baja el riesgo.

► Graphics

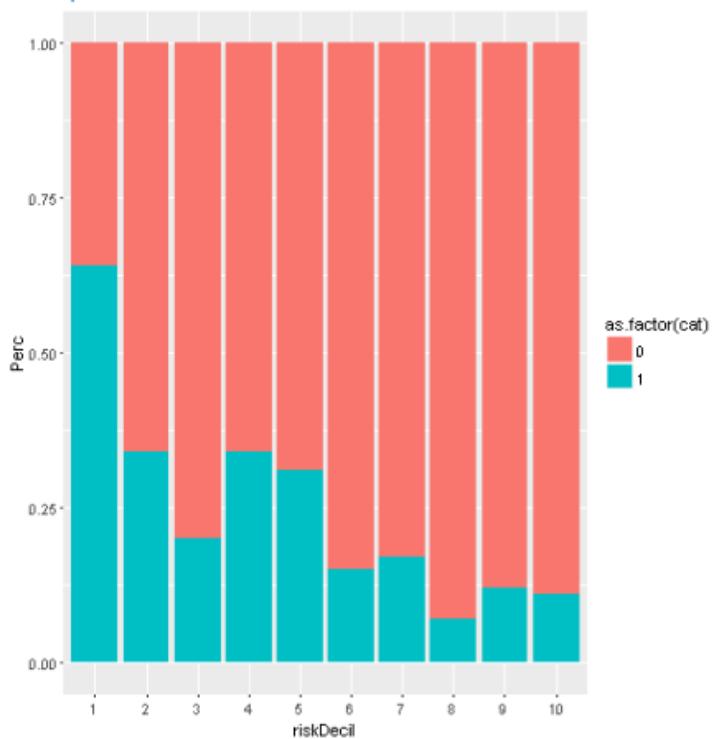


Figura 34 Validación por deciles modelo inicio de semestre.

Fuente: Propia.

Ahora si queremos hacer un modelo para estudiantes de retorno siempre analizando el inicio del periodo, solamente agregamos dos variables de las ya calibradas y modificamos los filtros de que registros se usaran para entrenar, sin embargo, si se ha creado un nuevo experimento para este modelo. A diferencia con el modelo de estudiantes nuevos la variable de respuesta es solamente de un 10% a calcular, por lo que la precisión de esta tiende a ser mucho más baja que un modelo donde la variable dependiente tiene una representación de casi un 20%.

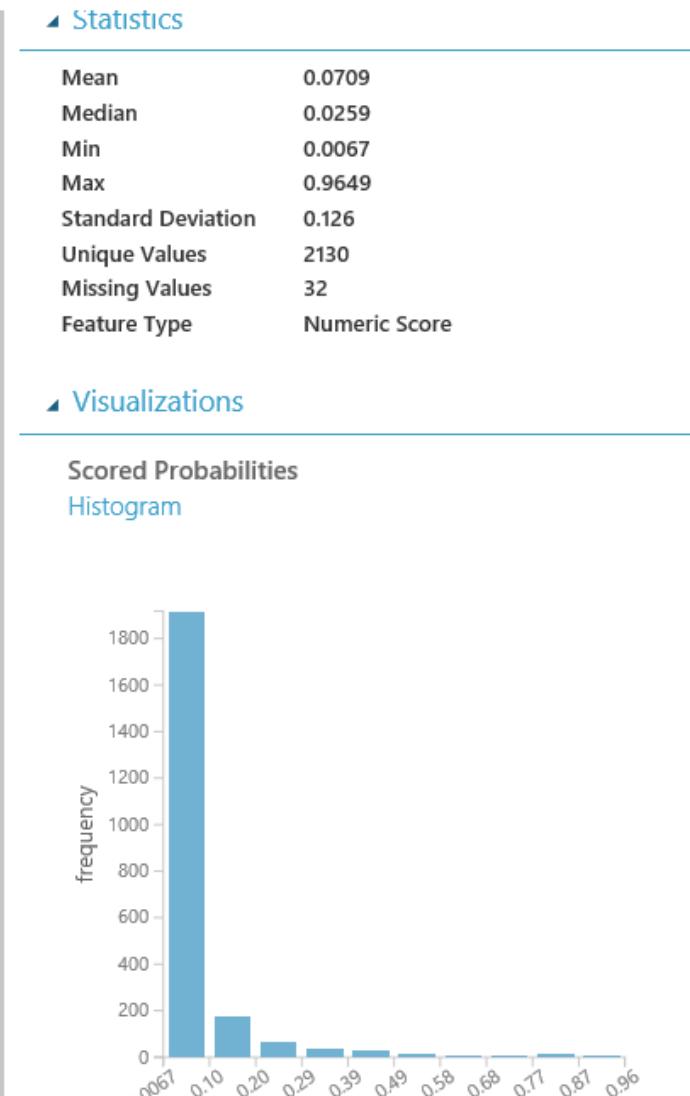


Figura 35 Distribución modelo para estudiantes continuos inicio de semestre.

Fuente: Propia.

Notemos en primer lugar de que la distribución de este modelo es menos concentrada en la izquierda y que la mayoría está en la derecha. Analizando rápidamente Curva ROC, matriz de confusión y deciles tendremos menos precisión, sin embargo, para ser un modelo de inicio de

semestre para una muestra donde la variable de respuesta no tiene mucha representación podemos decir que está bastante bien.

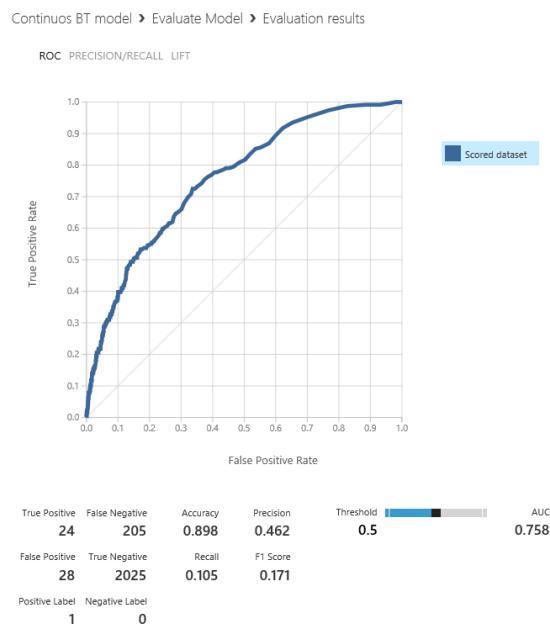


Figura 36 Validaciones modelo inicio de semestre, estudiantes continuos.

Fuente: Propia.

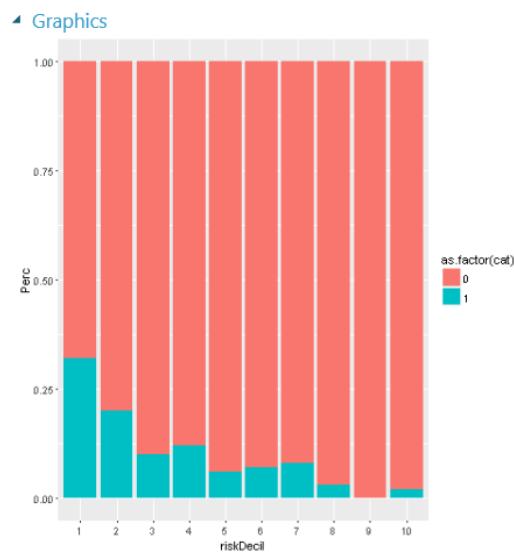


Figura 37 Validación por decil modelos inicio de semestre estudiantes continuos

Fuente: Propia.

4.2 Análisis de información financiera

La información financiera paso por un proceso de transformación ya que el modelo original estaba presentado de una manera donde la correlación con la variable de respuesta era muy baja y el nivel de agregación estaba no muy optimizada para leer mediante rutinas de predicción y correlación después de la limpieza la información se muestra de la siguiente manera:

	CUENTA	SEMANA	ANIO	SEMESTRE	MODULO	NUMERO.PAGOS	SUM.PAGOS	DIAS.SIN.ACTIVIDAD	SUM.MULTA	NUMERO.PAGOS.ATRASADOS	
1	10023067	1	2018	1	1	0	0	7	0	1	
2	10113044	1	2018	1	1	0	0	7	0	1	
3	10213051	1	2018	1	1	0	0	7	0	1	
4	10223015	1	2018	1	1	0	0	7	0	1	
5	10313130	1	2018	1	1	0	0	7	0	1	
6	10323032	1	2018	1	1	0	0	7	0	1	
						NUMERO.DIAS.PAGOS.ANTES.FECHA	NUMERO.DIAS.PAGOS.DESPUES.FECHA	DESERTOR.SIG.PERIODO			
1						0	0	0	0		
2						0	0	0	0		
3						0	0	0	0		
4						0	0	0	0		
5						0	0	0	0		
6						0	0	0	0		

Figura 38 Modelo de datos origina de información financiera.

Fuente: Propia.

Ahora con este nuevo formato podemos hacer análisis más puntuales, uno de nuestros enfoques es saber el acumulado de días que el estudiante pago antes y después de las fechas, aparte nos interesa saber a la semana cuantos pagos ha hecho y el acumulado de los pagos. Con esta información lo que pretendemos es poder crear un nuevo modelo que nos prediga cual es el riesgo de que el estudiante no pague su siguiente cuota esta predicción la vamos a utilizar después como una variable predictora en nuestro modelo de retención. Antes de comenzar a construir estos modelos lo primero que tenemos que observar es en que semanas hace sentido modelar un nuevo comportamiento:



Figura 39 Comportamiento de pago semana 1

Fuente: Propia.

Antes de explicar la interpretación de estas variables notemos que la primera grafica lo que nos quiere comunicar es que en la semana uno más del 70% de los estudiantes no han realizado ni un pago, por lo tanto, en la segunda grafica lo que queremos comunicar es aquellos estudiantes que han realizado más de dos pagos y la tercera más de tres pagos.

Muy interesante de que en la semana uno la actividad financiera por parte de los estudiantes es casi nula. Por lo que esta no sería una buena semana para construir un modelo, si aplicamos la misma lógica y nos movemos a una semana 4 podremos notar el siguiente comportamiento de cuotas:



Figura 40 Comportamiento de pago semana 4

Fuente: Propia.

Perfecto, lo que podemos observar ahora en la primera grafica es que más del 70% de los estudiantes ya realizaron su primer pago por lo que aquí podremos realizar un modelo estadístico de que estudiantes basado en sus características tienen alto riesgo de no realizar ese primer pago, la segunda grafica en una semana cuatro no es de mucho interés ya que sigue predominando la cantidad de estudiantes que no han hecho más de dos pagos:



Figura 41 Comportamiento de pago semana 8

Fuente: Propia.

Ahora notemos que para una semana ocho predomina más los estudiantes con más de dos pagos hecho, vale la pena observar también que son muy pocos los estudiantes en esta semana que no han realizado ni un pago, de la misma manera para pagos mayor a tres en esa semana aún sigue siendo inferior los estudiantes que si lo hayan completado:

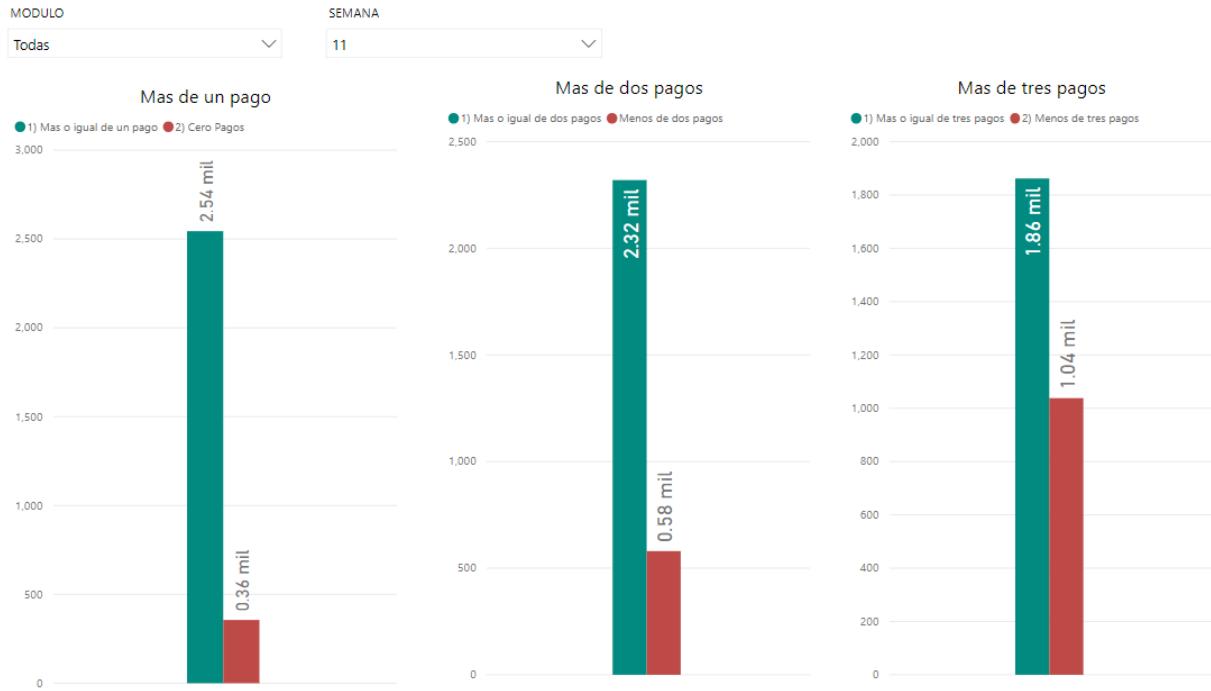


Figura 42 Comportamiento de pago semana 11.

Fuente: Propia.

Ahora para esta semana, considerada como semana de fin de ciclo, la mayoría de los estudiantes han realizado más de tres pagos. Por lo que es clave identificar que la construcción de los modelos será basada en semana cuatro, ocho y fin de ciclo.

De la misma manera es buena práctica revisar como están las variables antes de hacer una estandarización, por lo que revisaremos rápidamente los resúmenes de todas las variables:

```

> summary(final_financial)
      CUENTA      SEMANA      ANIO      SEMESTRE      MODULO      NUMERO.PAGOS      SUM.PAGOS
Length:98749   Min.   : 1   Min.   :2018   Min.   :1.000   Min.   :1.000   Min.   : 0.000   Min.   : 0
Class :character 1st Qu.: 3   1st Qu.:2018   1st Qu.:1.000   1st Qu.:1.000   1st Qu.: 0.000   1st Qu.: 0
Mode  :character Median : 6   Median :2018   Median :1.000   Median :2.000   Median : 2.000   Median : 5318
                           Mean   : 6   Mean   :2018   Mean   :1.323   Mean   :2.295   Mean   : 1.857   Mean   : 6145
                           3rd Qu.: 9   3rd Qu.:2018   3rd Qu.:2.000   3rd Qu.:4.000   3rd Qu.: 3.000   3rd Qu.:10072
                           Max.  :11  Max.  :2018   Max.  :2.000   Max.  :4.000   Max.  :17.000   Max.  :65000
DIAS.SIN.ACTIVIDAD  SUM.MULTA      NUMERO.PAGOS.ATRASADOS  NUMERO.DIAS.PAGOS.ANTES.FECHA  NUMERO.DIAS.PAGOS.DESPUES.FECHA
Min.   : 1.00   Min.   : 0.00   Min.   : 0.000   Min.   : 0.00   Min.   : 0.00
1st Qu.: 7.00   1st Qu.: 0.00   1st Qu.: 0.000   1st Qu.: 0.00   1st Qu.: 0.00
Median :17.00   Median : 0.00   Median : 1.000   Median : 0.00   Median : 0.00
Mean   :22.08   Mean   :33.13   Mean   : 1.343   Mean   :12.31   Mean   : 40.07
3rd Qu.:30.00   3rd Qu.: 0.00   3rd Qu.: 2.000   3rd Qu.: 9.00   3rd Qu.: 10.00
Max.  :77.00   Max.  :4500.00  Max.  :16.000   Max.  :388.00  Max.  :4622.00
DESERTOR.SIG.PERIODO
Length:98749
Class :character
Mode  :character

```

Figura 43 Modelo de datos de información financiera transformada.

Fuente: Propia.

Aquí podemos observar un par de datos interesantes como que número de pagos tiene valores bastante extremos en la parte superior esto lo podemos observar rápidamente mediante un análisis de boxplot.

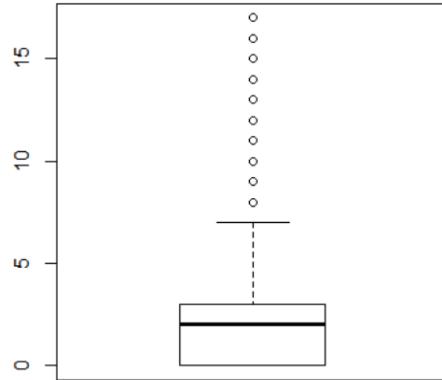


Figura 44 Boxplot número de pagos

Fuente: Propia.

Como se puede observar el boxplot tiene varios valores atípicos después de Q4 por lo que podríamos hacer una estandarización rápida de que después del valor máximo de Q4 todo será

llevado a ese mismo número. Después de hacer esa transformación observamos nuestro grafico de la siguiente manera:

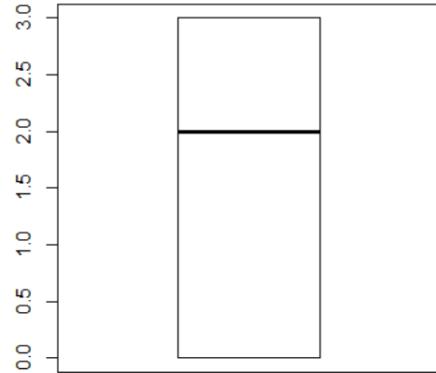


Figura 45 Boxplot número de pagos tratado

Fuente: Propia.

Ahora con esto podríamos decir que tenemos una variable normalizada, siguiendo esta misma lógica y dependiendo de la lógica de negocio se continuara haciendo la limpieza de estas variables.

```
> summary(final_financial)
   CUENTA      SEMANA     ANIO     SEMESTRE    MODULO    NUMERO.PAGOS    SUM.PAGOS
Length:98749  Min.   : 1   Min.   :2018   Min.   :1.000  Min.   :1.000  Min.   :0.000  Min.   : 0
Class :character  1st Qu.: 3   1st Qu.:2018   1st Qu.:1.000  1st Qu.:1.000  1st Qu.:0.000  1st Qu.: 0
Mode  :character  Median : 6   Median :2018   Median :1.000  Median :2.000  Median :2.000  Median : 5318
                           Mean   : 6   Mean   :2018   Mean   :1.323  Mean   :2.295  Mean   :11.581  Mean   : 5764
                           3rd Qu.: 9   3rd Qu.:2018   3rd Qu.:2.000  3rd Qu.:4.000  3rd Qu.:3.000  3rd Qu.:10072
                           Max.  :11   Max.  :2018   Max.  :2.000  Max.  :4.000  Max.  :3.000  Max.  :14216
  DIAS.SIN.ACTIVIDAD  SUM.MULTA  NUMERO.PAGOS.ATRASADOS  NUMERO.DIAS.PAGOS.ANTES.FECHA  NUMERO.DIAS.PAGOS.DESPUES.FECHA
Min.   : 1.00  Min.   : 0.00  Min.   :0.000          Min.   : 0.000          Min.   : 0.00
1st Qu.: 7.00  1st Qu.: 0.00  1st Qu.:0.000          1st Qu.: 0.000          1st Qu.: 0.00
Median :17.00  Median : 0.00  Median :1.000          Median : 0.000          Median : 0.00
Mean   :20.74  Mean   : 33.13  Mean   :1.259          Mean   : 6.126          Mean   :10.41
3rd Qu.:30.00  3rd Qu.: 0.00  3rd Qu.:2.000          3rd Qu.: 9.000          3rd Qu.:10.00
Max.  :49.00  Max.  :4500.00  Max.  :3.000          Max.  :29.000          Max.  :58.00
  DESERTOR.SIG.PERIODO
Length:98749
Class :character
Mode  :character
```

Figura 46 Nuevo resumen de variables financieras tratadas.

Fuente: Propia.

Ahora podemos ver variables con mayor sentido, vale la pena señalar que la variable SUM.MULTA (Sumatoria acumulada de todas las multas de un estudiante) es una variable que no fue tratada ya que realmente hace sentido ver un valor máximo de 4500, ahora algo muy interesante que se puede observar es que las variables de DIAS.PAGOS tienen valores extremos en comparación en la fotografía anterior, pero aún se puede ver que la mediana y media están algo lejos, lo cual hace mucho sentido ya que estas son las variables a través de todas las semanas y en las primeras semanas predominan los valores bajos.

Como paso adicional se debe de calcular la nueva variable de respuesta para las semanas en las que tenemos interés hacer nuestro entrenamiento, esto significa que podemos excluir la mayoría de las semanas y enfocarnos en esa nueva dependiente que la denominaremos PAGO.ESPERADO, será una variable binaria por lo que nuestra predicción será una clasificación. Las semanas para utilizar serán (SEMANA.ESPERADA {4,8,11} – 1) esto es porque lo que nos interesa predecir es el comportamiento de SEMANA.ESPERADA.

En temas de entrenamiento recordemos que para información financiera contamos con tres períodos académicos de 2018 por lo que entrenaremos con modulo uno y dos, como validación cruzada usaremos el 75% para entrenar, el otro 25% para test y el módulo 4 para validar. A nivel ETL tendremos el siguiente proceso:

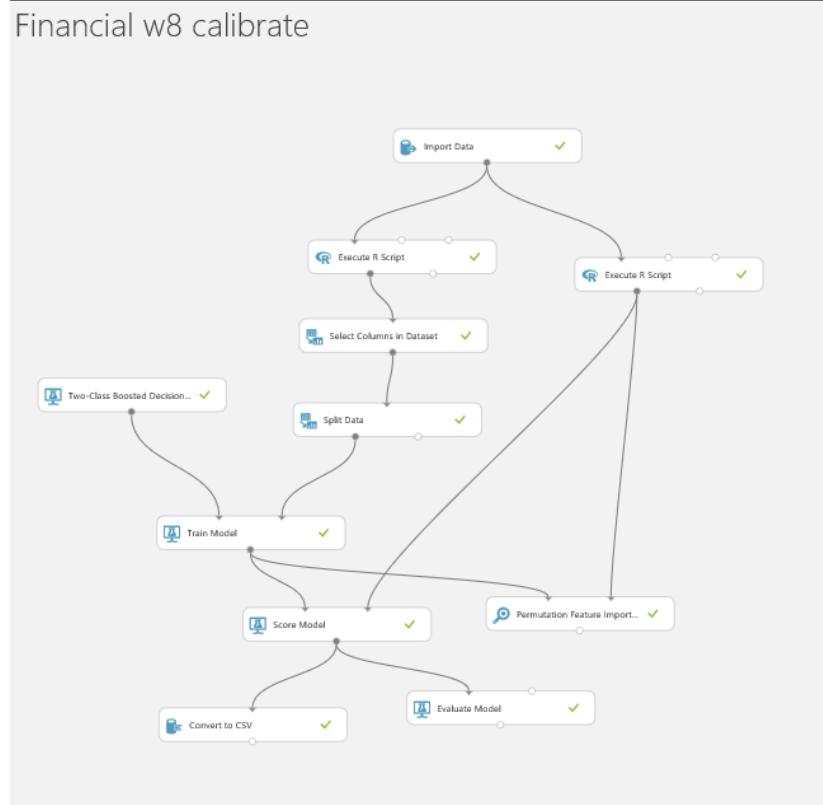


Figura 47 ETL entrenamiento de modelo financiero.

Fuente: Propia.

Este ETL como se muestra en las ramas de la derecha tenemos la división y de lado derecho tenemos el conjunto de datos para ser validado, algo muy importante que notar en este ETL es que estamos utilizando como técnica “Clasificación binaria mediante potenciación de árboles de decisión” los parámetros con los que fueron calibrados son los siguientes:

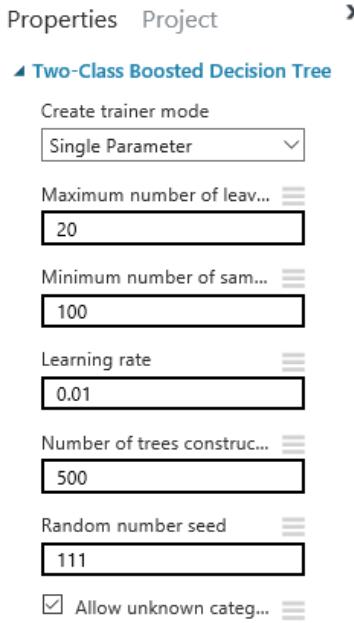


Figura 48 Ajuste de potenciación mediante arboles de decisiones.

Fuente: Propia.

Estas configuraciones nos indican el nivel de profundidad que tendrá esta técnica como por ejemplo “Calificación de aprendizaje” de 0.01 busca mayor precisión esto es bueno cuando contamos con un conjunto de datos grandes como el que utilizamos. Este como es una técnica de potenciación mediante arboles de decisiones debemos de especificar cuantos arboles serán construidos para hacer el ajuste de entrenamiento, “Semilla aleatoria” es una propiedad muy importante ya que este es el que nos permite que nuestro experimento sea auto reproducible.

Comencemos analizando el modelo entrenado de semana cuatro, este modelo es interesante ya que es muy complicado predecir quien no hará un primer pago y lo que nosotros buscamos mediante esta investigación es una alta interpretabilidad. Como deducimos que este es un modelo que si lo entrenamos estará sobre ajustado, mediante el siguiente análisis:

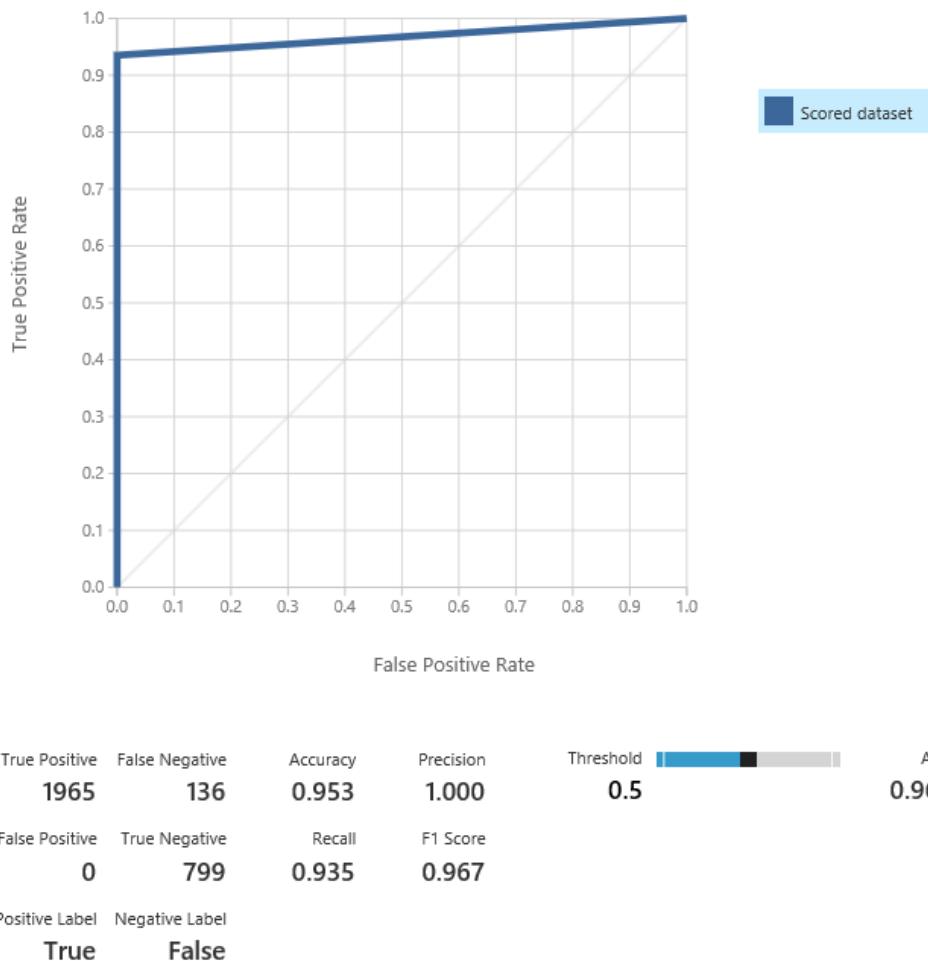


Figura 49 Validación modelo de información financiera.

Fuente: Propia.

Notemos que los verdaderos positivos y falsos positivos en el eje de las Y se encuentra totalmente lineal, cuando la función dibuja en el eje de la Y una recta totalmente vertical como el caso que se muestra en la figura 50 esto nos da como señal que no es una función, aparte de que en la matriz de confusión False positive es 0 dando como resultado 1.0 de precisión.

Dado este comportamiento **queda descalificado este modelo** en un 100%, esto es porque si la curva ROC se muestra ajustada en ambos ejes normalmente se considera como un modelo

sobre ajustado, matemáticamente se le puede llamar también sesgado, ahora analicemos el modelo de la semana ocho que nos predice que estudiantes podrían a llegar a tener una mora.

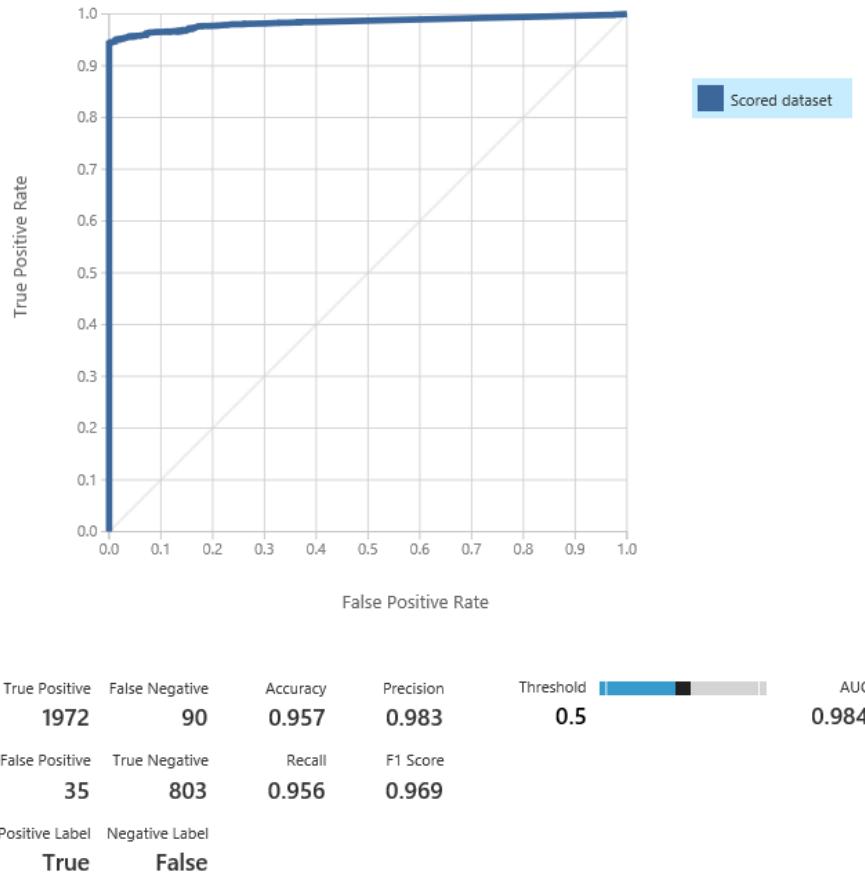


Figura 50 Modelo financiero utilizado en otras semanas.

Fuente: Propia.

Este modelo se mira con un comportamiento menos sobre ajustado gracias a esos 35 false positive y 90 falsos negativos, sin embargo, este modelo se debe de tener en observación cuando mande a producción. Un experimento interesante es utilizar este modelo como un modelo standard para “n” semanas.

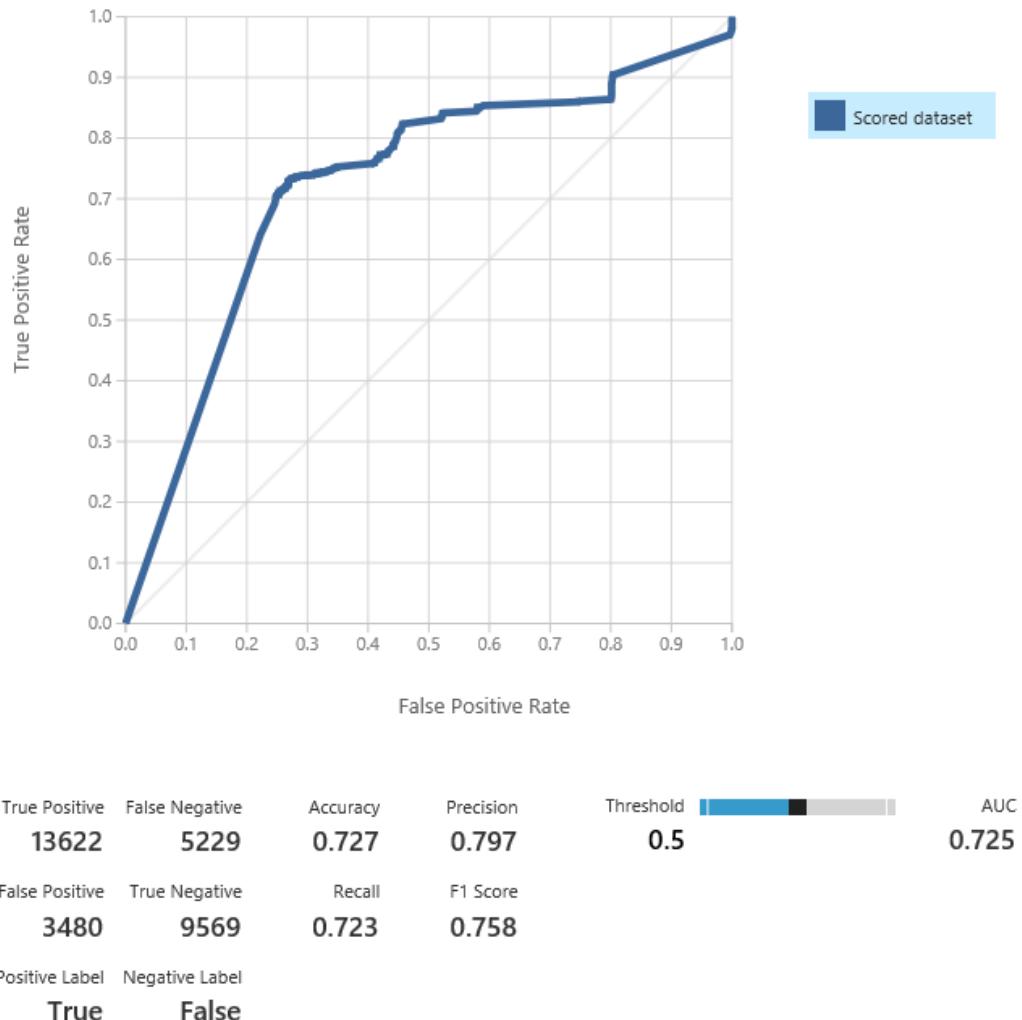


Figura 51 Modelo financiero utilizado en otras semanas.

Fuente: Propia.

Lo que pudimos observar con este modelo financiero es que, si existe una correlación con nuestra principal variable de respuesta, desertor del siguiente periodo académico, mostramos a continuación como es la correlación de esta:

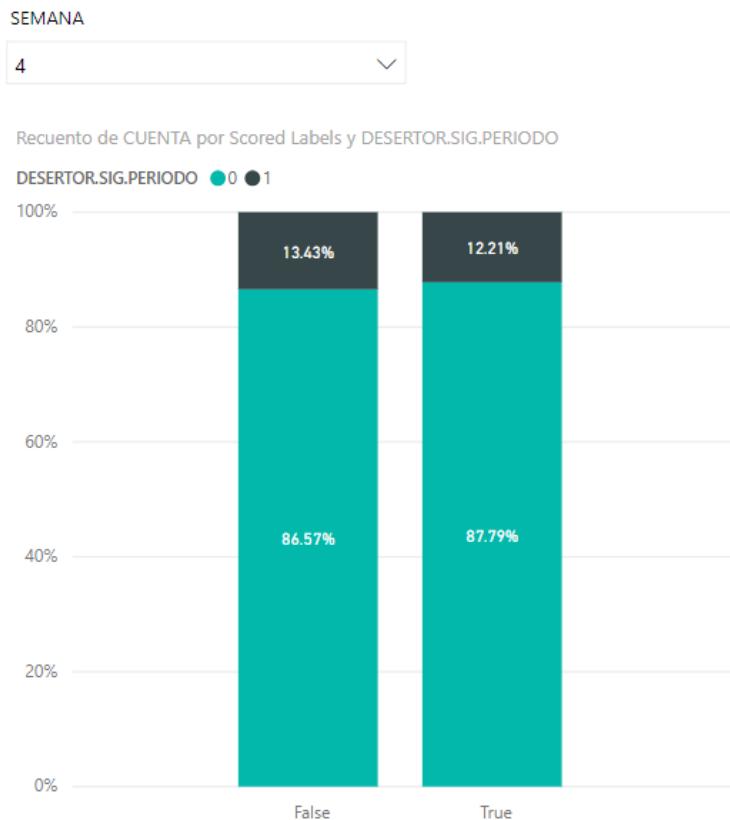


Figura 52 Precisión modelo financiero con variable de respuesta semana 4.

Fuente: Propia.

Comenzando con nuestra predicción menos efectiva dado la poca información con la que contamos, podemos notar que estudiantes con la predicción “false” son los estudiantes que tienen un riesgo de no realizar un pago la siguiente semana, en esta semana 4 lo que podemos notar es que tenemos una diferencia de un uno por ciento. Ahora si analizamos por ejemplo la semana ocho podremos notar mayor precisión con respecto a nuestra variable de respuesta, este comportamiento lo podemos ver a continuación:

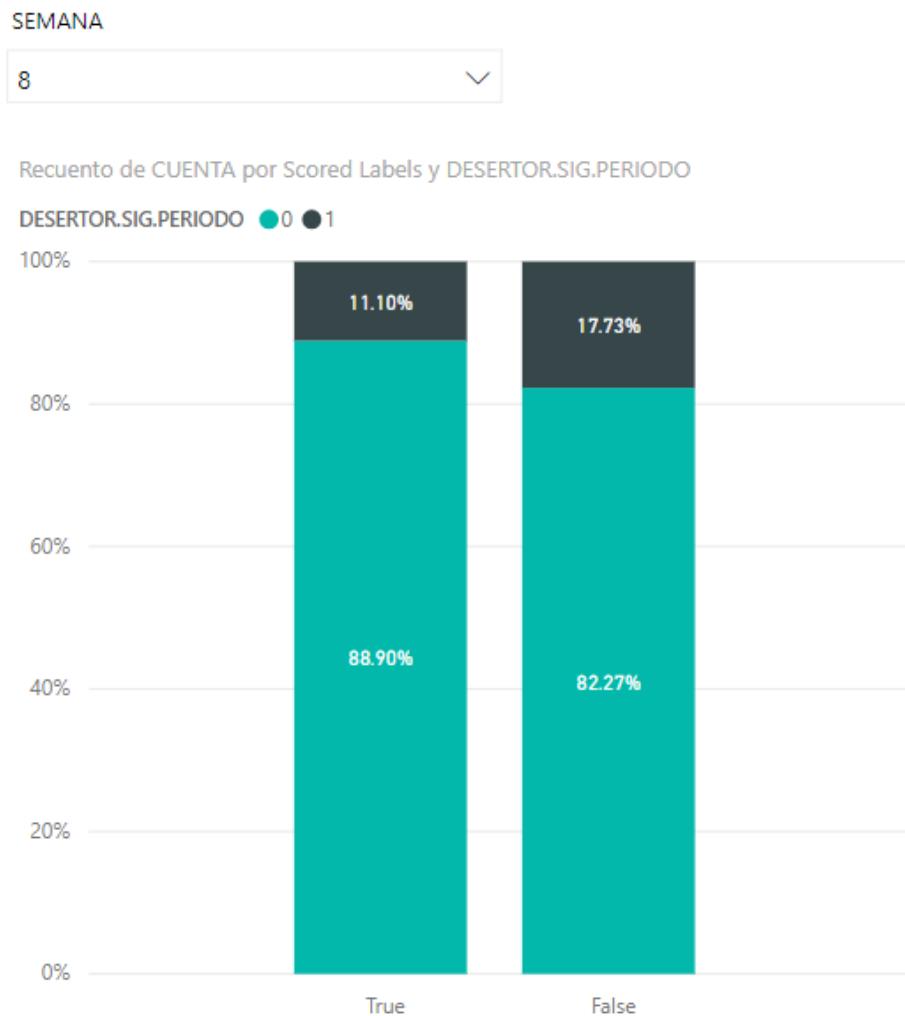


Figura 53 Precisión modelo financiero con variable de respuesta semana 8.

Fuente: Propia.

Claramente podemos observar que los estudiantes clasificados con esta nota para la semana 8 hay casi un seis por ciento de mayor precisión con respecto a nuestra variable de respuesta, ahora si nos vamos a las últimas semanas podremos notar mayor precisión:

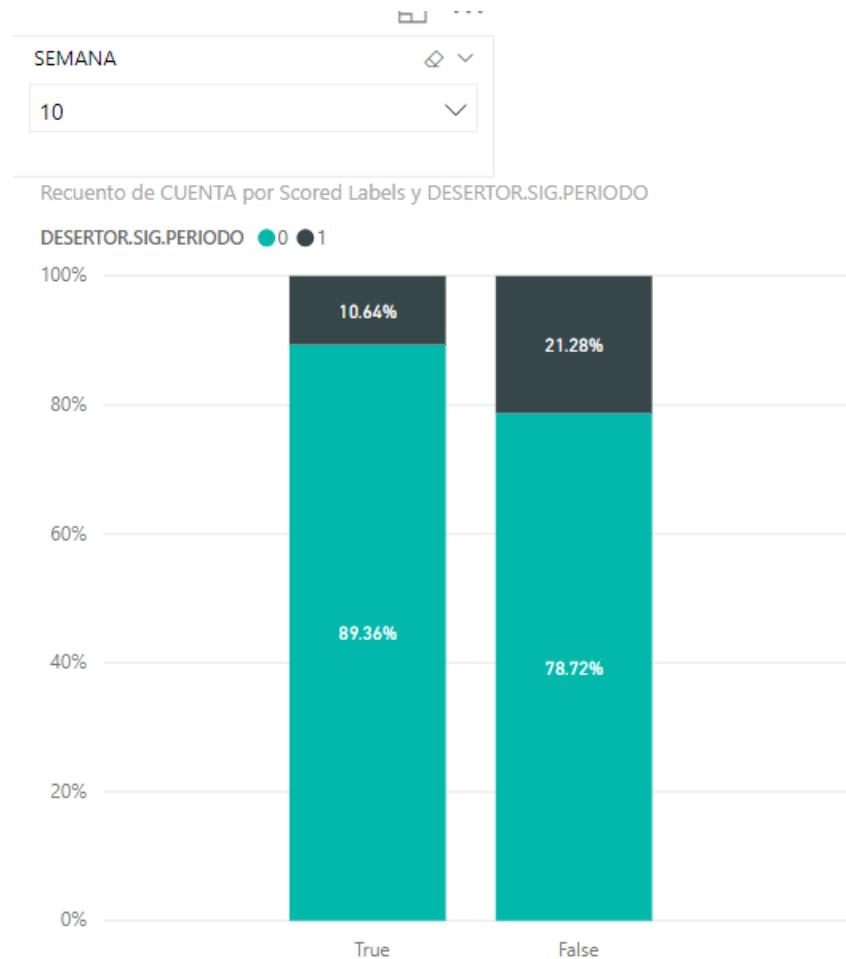


Figura 54 Precisión modelo financiero con variable de respuesta semana 10.

Fuente: Propia.

4.3 Análisis exploratorio información de tickets

La información de los tickets es una fuente valiosa de información ya que por medio de esta podemos ver qué tipo de quejas levantan los estudiantes y el nivel de correlación que esta tiene con la deserción estudiantil, rápidamente podemos ver un análisis del comportamiento que esta tiene a nivel de campus, categoría y días de respuesta.

Para empezar, podemos observar el comportamiento que tiene a nivel de campus y es notorio señalar de que los estudiantes de TGU son los que menos riesgo tienen de deserción

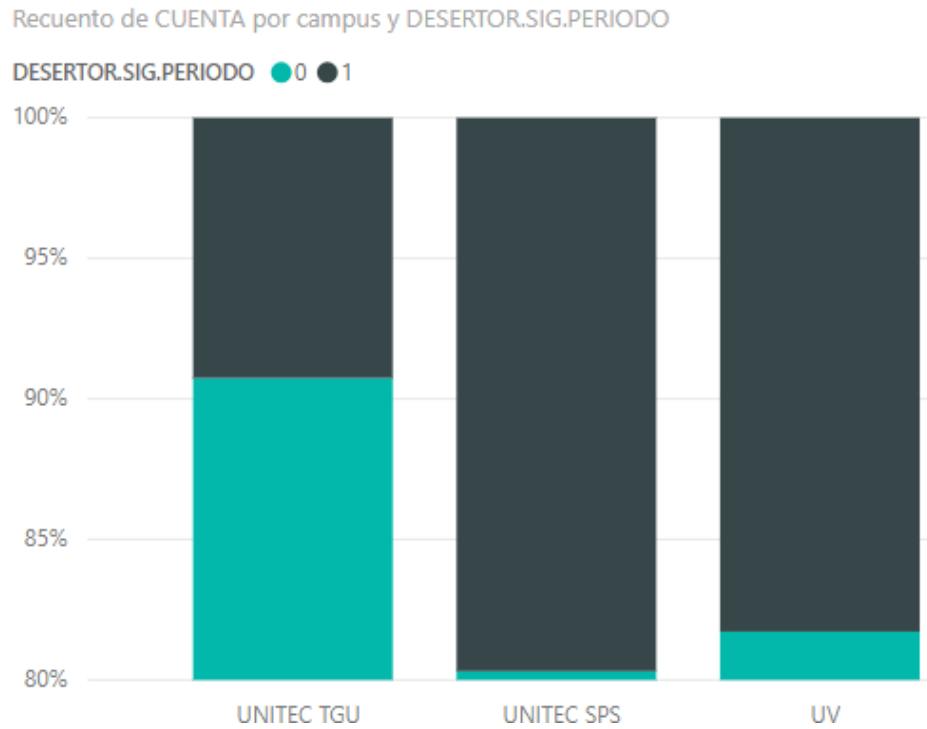


Figura 55 Deserción por campus de estudiantes con tickets.

Fuente: Propia.

Se puede notar como SPS y UV son lo que manejan un mayor riesgo de deserción casi con una diferencia de diez puntos porcentuales.

Ahora podemos analizar la información a través de los días de sin contestar los tickets.

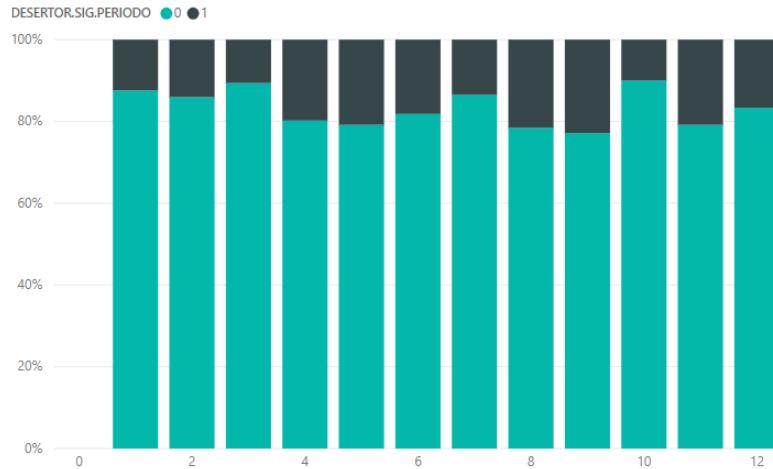


Figura 56 Tickets no contestados por semana y deserción.

Fuente: Propia.

Como primera observación se puede notar que en los tres primeros días se mantiene casi constante por lo que podemos crear una clasificación que sea binaria para una mayor interpretación.

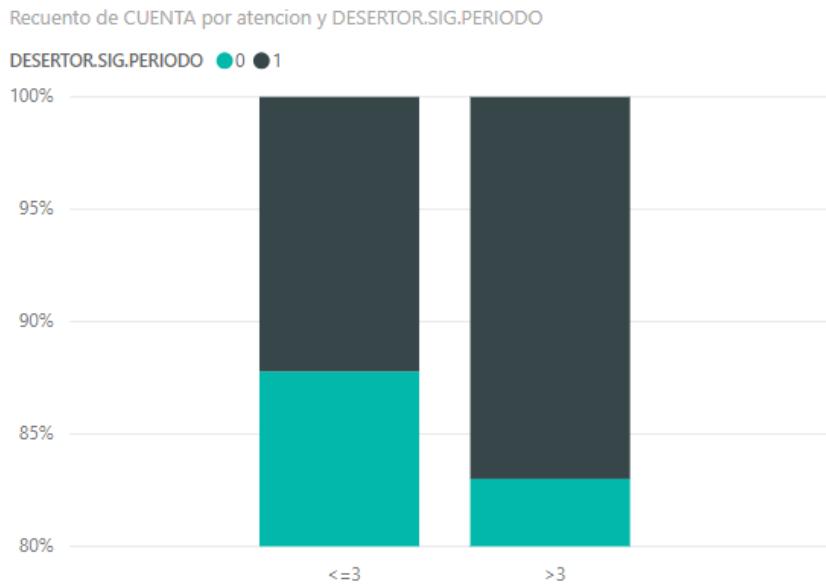


Figura 57 Comportamiento de deserción estudiantes sin respuesta a tickets.

Fuente: Propia.

Con el tipo notamos algo muy curioso para el primer periodo académico del año es cuando más tickets de tipo queja se levantan y este a su vez hace que haya una mayor deserción.

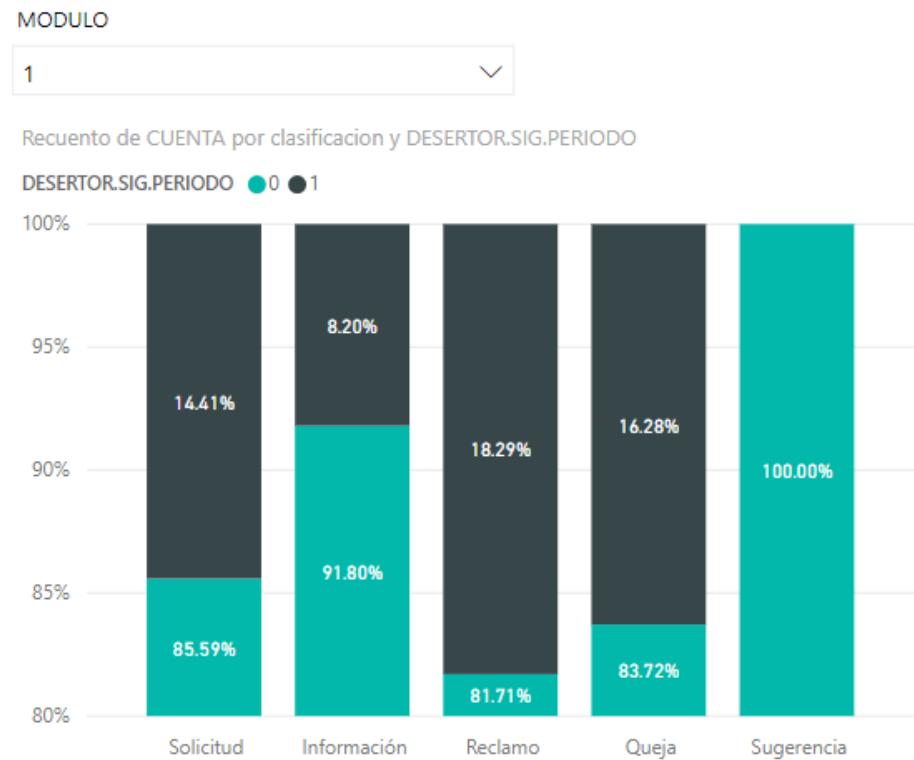


Figura 58 Deserción por clasificación de tickets modulo uno.

Fuente: Propia.

Se puede notar claramente que reclamo y queja tienden a tener esa fuerte tendencia de deserción ahora si observamos para un segundo periodo



Figura 59 Deserción por clasificación de tickets modulo dos.

Fuente: Propia.

El porcentaje de deserción por categoría disminuye y se puede apreciar de que las personas que levantan sugerencias no tienden a desertar.

Por último, si analizamos el número de tickets que envían los estudiantes por periodo, notamos la siguiente tendencia:

```
> prop.table(table(sti$NUMERO.TICKETS,sti$DESERTOR.SIG.PERIODO),1)
      1          0
1 0.09144543 0.90855457
2 0.12116788 0.87883212
3 0.16183575 0.83816425
4 0.13733906 0.86266094
5 0.17730496 0.82269504
6 0.15730337 0.84269663
7 0.15254237 0.84745763
8 0.20437956 0.79562044
```

Figura 60 Deserción por número de tickets.

Fuente: Propia.

Como se puede observar entre más número de tickets mayor riesgo de deserción, claramente entre más números de tickets menor frecuencia se tiene. Pero definitivamente este es un excelente predictor con respecto a nuestra variable de respuesta.

4.4 Análisis Exploratorio de la plataforma de Blackboard

Otra herramienta importante que se debe de analizar es la información de Blackboard, esta plataforma es la que registra todas las actividades que el estudiante realiza en línea, normalmente en esta plataforma se analizan, los logins y las interacciones ya que se puede considerar esto como una asistencia. En análisis más profundos esto se puede combinar con índices de esfuerzo y compromiso.

Primero analizaremos el número de logins que los estudiantes realizan por semana, para analizar cuál es el comportamiento que este tiene:



Figura 61 Suma de logins por semana.

Fuente: Propia.

De esto podemos sacar interesantes conclusiones, en las primeras semanas el número de logins es bastante alta a medida pasan las semanas estas van disminuyendo y después de la semana 5 estas bajan considerablemente, sin embargo, hay que recordar que después de semana 5 se comienza un nuevo curso.

Ahora si analizamos esta misma gráfica, pero la mediada notamos un comportamiento interesante:



Figura 62 Mediana de logins por semana.

Fuente: Propia.

En la semana 7 los estudiantes tienden a conectarse muy poco. Esto claramente nos da una intervención directa hacia los docentes de incentivar mucho más a los estudiantes de usar la plataforma. Claramente si hacemos este análisis separado por estudiantes desertores después de la semana 7 tendremos el mismo comportamiento.

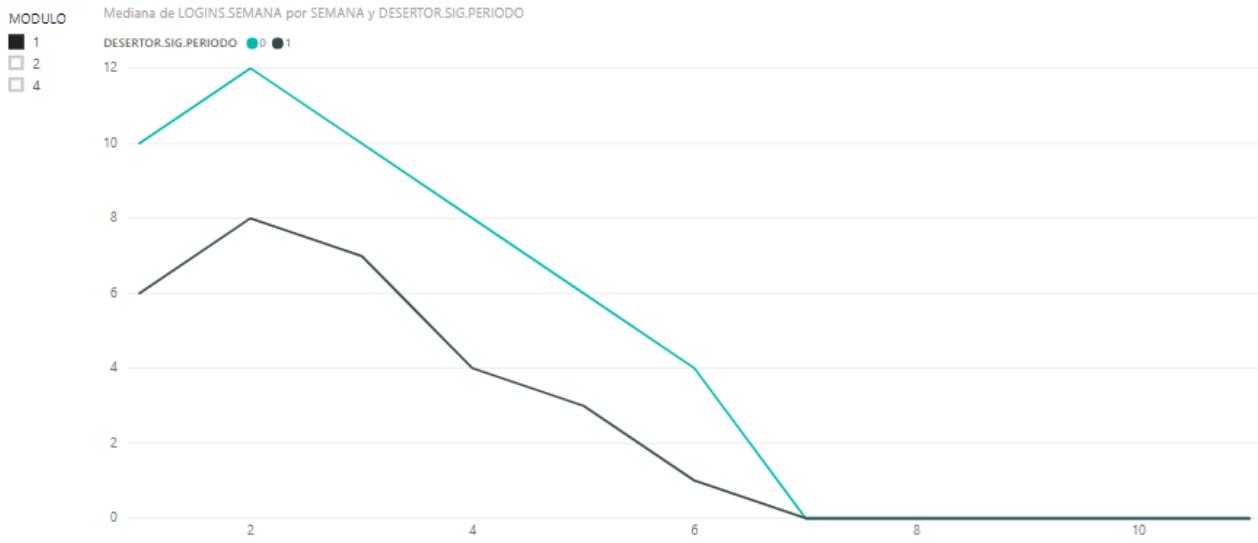


Figura 63 Mediana de logins por semana y desertores.

Fuente: Propia.

Como se puede observar hay una clara diferencia entre los que desertan y no antes de semana 6, lastimosamente la tendencia de que los estudiantes disminuyen su participación semana a semana siempre se conserva. Ahora una pregunta clave que se puede hacer es como hacer una intervención después de semana 6 con un número accionable, esto se puede lograr mediante un análisis de información acumulada.

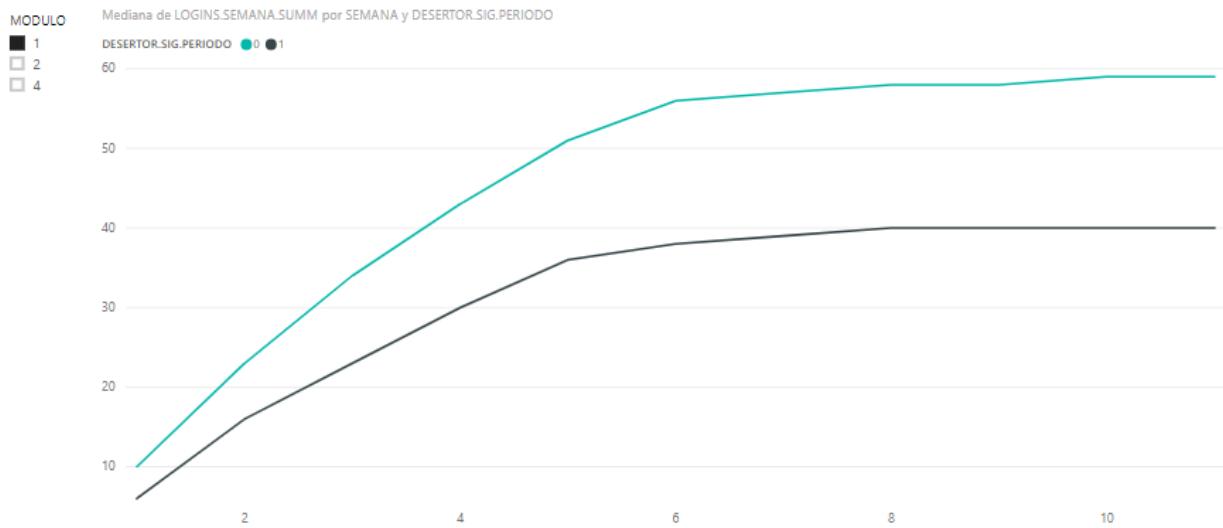


Figura 64 Mediana de logins por semana y desertores acumulado.

Fuente: Propia.

Como Se puede observar la información acumulada de las personas que si tienen interacciones versus los que desertan se encuentra por debajo por lo que eso da una buena regla sobre como intervenir esos estudiantes que no están entrando mucho a Blackboard.

Rápidamente si hacemos este mismo análisis sobre la información de interacciones podremos ver un comportamiento similar donde los estudiantes que abandonan se encuentran por debajo tanto en información semanal actual como la acumulada.

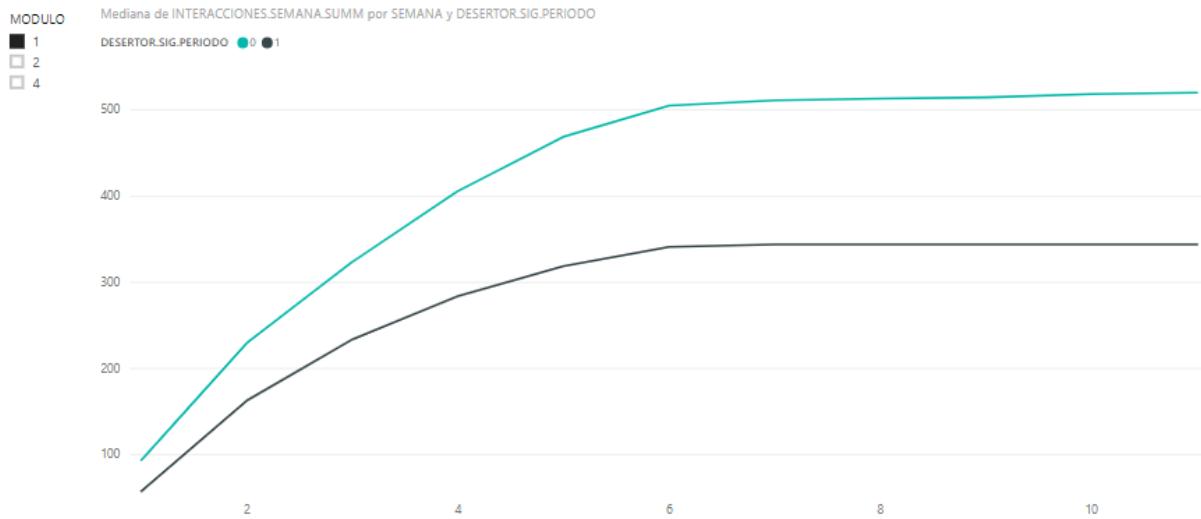


Figura 65 Mediana de accesos por semana y desertores acumulado, modulo uno.

Fuente: Propia.

Nuevamente como recomendación, como los estudiantes que no desertan son la mayoría la conclusión es directa, estudiantes que se encuentren por debajo de la mediana se deben de intervenir sobre su poco acceso e interacción a la plataforma de Blackboard.

Como última observación del análisis de esta información en la data entregada los estudiantes que desertan en el módulo cuatro tienden a tener un comportamiento casi de cero, como se puede observar a continuación:

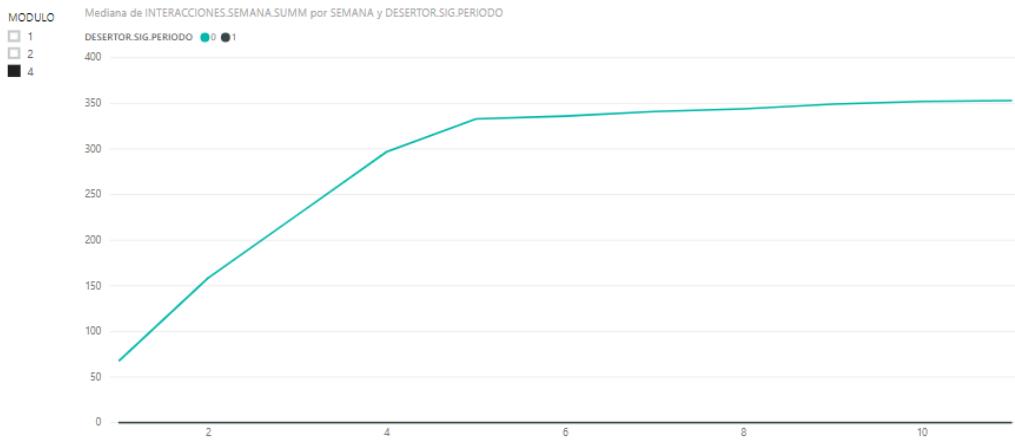


Figura 66 Mediana de accesos por semana y desertores acumulado, modulo cuatro.

Fuente: Propia.

Esto es una clara llamada de atención de que estudiantes que estén teniendo esa participación casi inexistente se debe de contactar rápidamente, ahora si hacemos un acercamiento de estos estudiantes podremos ver mejor cual es el comportamiento acumulado de ellos.

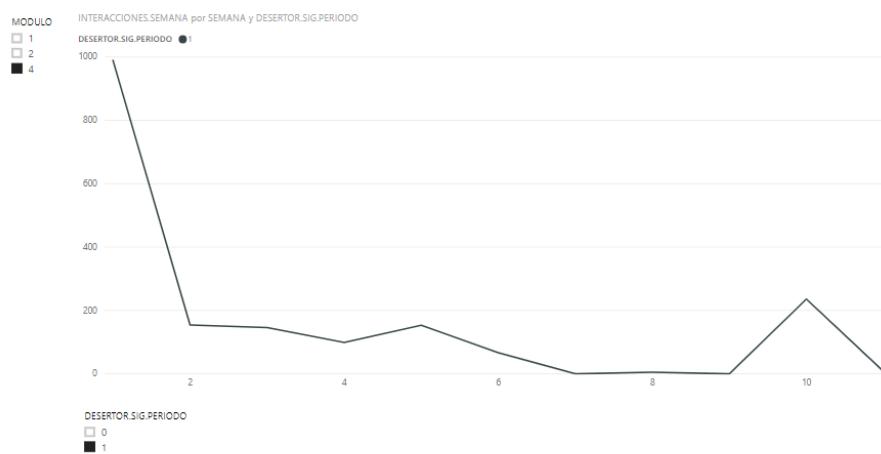


Figura 67 Comportamiento de estudiantes que desertan con sus interacciones.

Fuente: Propia.

4.5 Elaboración de modelos estadísticos

Una vez ya analizada la información que nos ayuda a describir el comportamiento de los estudiantes tanto al inicio del periodo, en medio y al final se puede proceder a la construcción de los modelos finales, en el caso de nuestra investigación pudimos notar tanto financiera como académicamente que hay distintos comportamientos a mitad del ciclo y al final.

Como recordatorio rápido los modelos de inicio de ciclo ya habían sido construidos tanto para estudiantes nuevos como recurrentes, a esta alarma la denotaremos como una alerta temprana, donde se explicó que esta normalmente tiene una precisión muy baja, después de la aplicación de este modelo ahora se construirá de la siguiente manera:



Propuesta de modelos estadísticos para la detección de estudiantes en riesgo de deserción

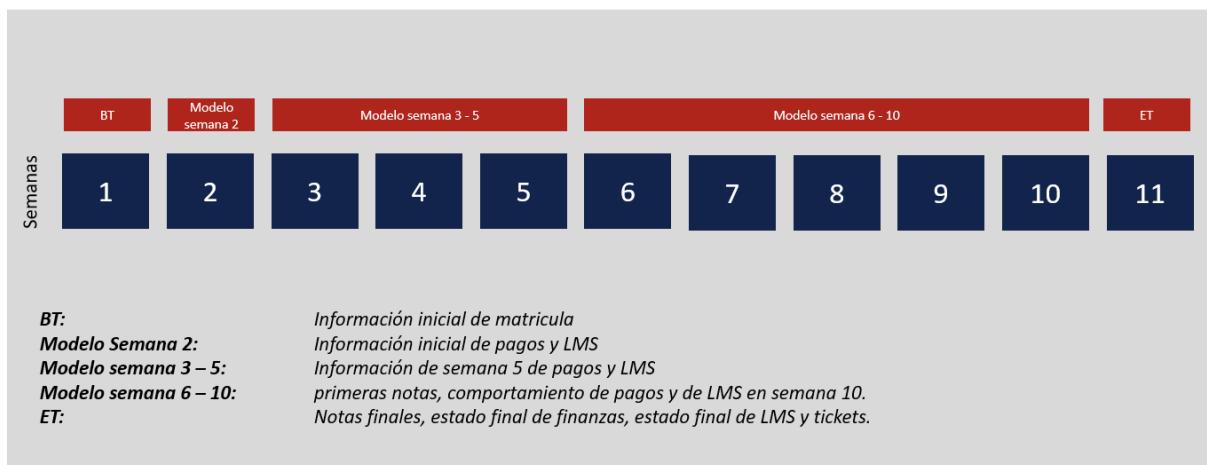


Figura 68 Propuesta final de modelos a construir.

Fuente: Propia.

Como se puede observar atreves de las semanas se construirán 5 modelos ahora, esto se debe de considerar por tipo de estudiante ya que a nivel de deserción los estudiantes nuevos tienen un riesgo más alto que los recurrentes. Por lo que en total se construirán 10 modelos.

4.5.1 Modelo para nuevos estudiantes en semana dos

Para este modelo se considera siempre la información inicial de registro, una de ellas como ya se había explicado es la variable que determina le nivel de compromiso con registro, en una semana dos lo que mide son los cursos cancelados contra los que se hayan matriculado. De igual manera en esta semana se comienza a considerar la información de registro y las actividades que tienen en Blackboard que como bien se pudo comprobar en los análisis de arriba en las primeras semanas es cuando mayor actividad se tiene.

Para este modelo se usará una regresión logística por temas de interpretabilidad de los coeficientes resultantes de la ecuación característica de la regresión logística. La precisión de este modelo luce de la siguiente manera:

ROC PRECISION/RECALL LIFT

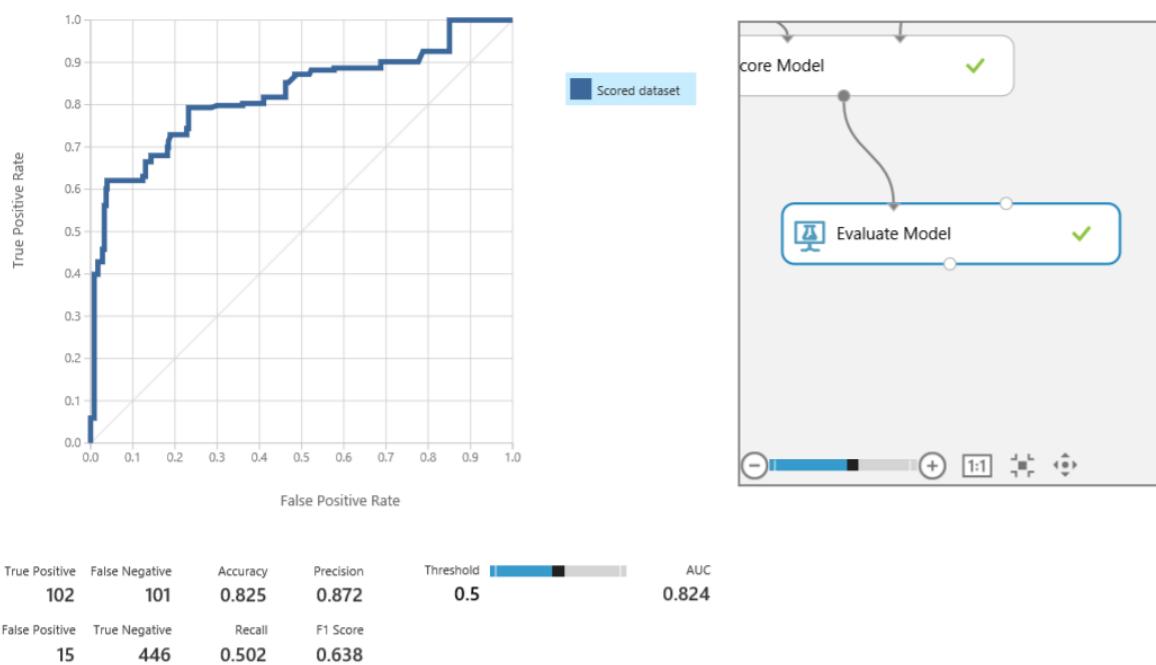


Figura 69 Curva ROC Nuevos estudiantes semana dos.

Fuente: Propia.

Otro método efectivo para medir la precisión es a través de los deciles aplicado a la calificación del modelo evaluado:

Feature Weights

Feature	Weight
START.ENDGAGEMENT	-3.29537
EDAD_NA_36	-2.27935
EDAD_17_4	1.93238
LOGINS.SEMANA	-1.69069
EDAD_26_11	1.68426
Bias	1.53457
CAREER_SISTEMAS DE GESTIÓN DE CALIDAD INTEGRADOS EN EL GRADO DE MAESTRÍA_10	0.942215
EDAD_28_13	0.844032
EDAD_33_18	0.761165
EDAD_34_19	-0.729312
NUMERO.PAGOS.ATRASAD OS	-0.645404
CAREER_MASTER EN GESTIÓN DE TECNOLOGÍAS DE LA INFORMACIÓN_5	-0.630317
NUMERO.DIAS.PAGOS.DESP UES.FECHA	0.570964

Figura 70 Pesos, modelo semana dos estudiantes nuevos.

Fuente: Propia.

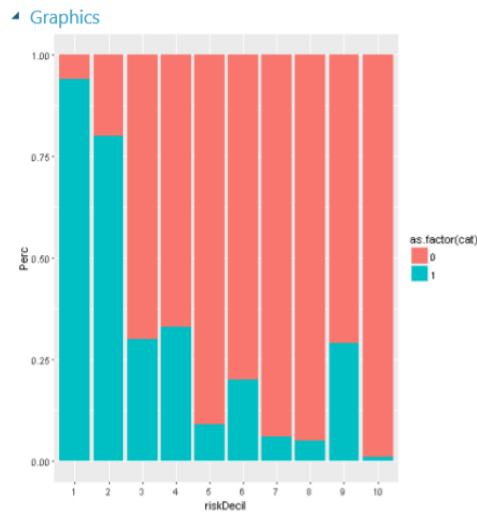


Figura 71 Precisión mediante deciles, modelo semana dos estudiantes nuevos.

Fuente: Propia.

Por último, analicemos las variables más importantes que resultan de este modelo:

rows	columns	
19	2	
view as	Feature	Score
	START.ENDGAGEMENT	0.514652
	EDAD	0.014652
	DIAS.SIN.ACTIVIDAD	0.00693
	LOGINS.SEMANA	0.00693
	CIUDAD.MARCA	0.004538
	NUMERO.PAGOS.ATRASA DOS	0.00223
	SUM.PAGOS	0.001105

Figura 72 variables importantes modelo semana dos estudiantes nuevos.

Fuente: Propia.

Este modelo también predomina el compromiso académico inicial, la edad sin embargo se maneja una mejor precisión debido a la correlación de la información académica y la información del comportamiento financiero.

4.5.2 Modelo para nuevos estudiantes semana tres a cinco

Este modelo sigue el mismo proceso de validación cruzada, selección de variables y elección de técnica. Algo muy importante que señalar con este modelo es que se excluye el comportamiento propio de la semana y se incluye la información acumulado con respecto a la información académica, donde como resultado el siguiente comportamiento a nivel de deciles:

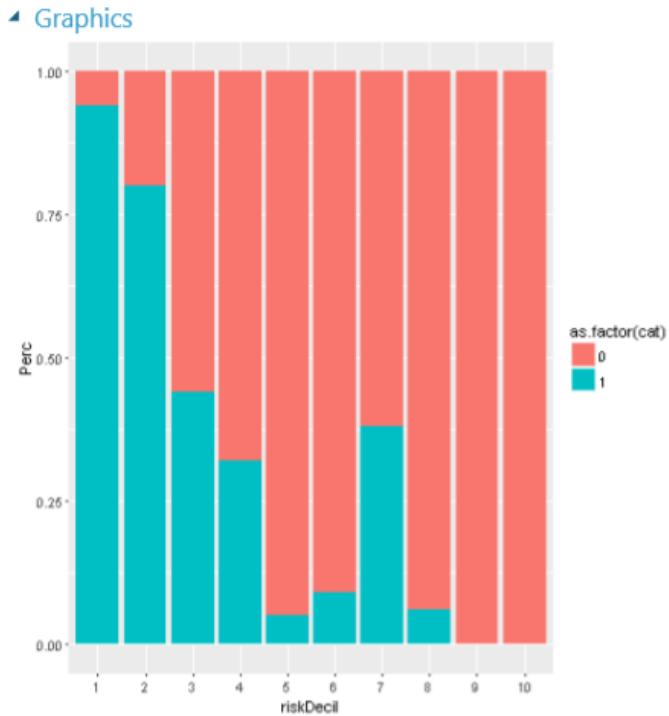


Figura 73 Precisión mediante deciles, modelo semana tres a cinco estudiantes nuevos.

Fuente: Propia.

Observemos como el decil uno mantiene la misma precisión, sin embargo, el decil dos gana mayor precisión y los últimos deciles son más exactos, las variables que comienzan a puntear también son más interpretables y se puede realizar una intervención más efectiva.

Para la construcción de este grafico se construyen los deciles con la variable que se predijo, quedando en el decil uno los estudiantes con mayor riesgo de deserción o calificación más alta, como este conjunto de estudiantes es un conjunto de validación claramente sabemos si el estudiante desertó o no, por lo que esperamos ver que los estudiantes que desertan más se encuentran ubicados en los primeros deciles como lo muestran las imágenes.

Feature	Score
	0.493728
START.ENDGAGEMENT	0.493728
INTERACCIONES.SEMANA.SUMM	0.024303
SUM.PAGOS	0.022956
CALIFICADO.SUMM	0.020486
LOGINS.SEMANA.SUMM	0.017863
JORNADA	0.010628
NUMERO.PAGOS.ATRASADOS	0.002359
REALIZO.POSTS.SEMANA.SUMM	0.000906

Figura 74 variables importantes modelo semana tres a cinco estudiantes nuevos.

Fuente: Propia.

Como se puede observar pierde sesgo el primer factor y se distribuye más en los otros aparte de que ahora se encuentra en segundo lugar una de las variables acumuladas de la información.

4.5.3 Modelo para nuevos estudiantes semana tres a cinco

A este modelo se le considera el modelo con mejor presión ya que en este se encuentra el estado final de la mayoría de las variables, como resumen mostramos a continuación la validación común a leer mediante la curva ROC:

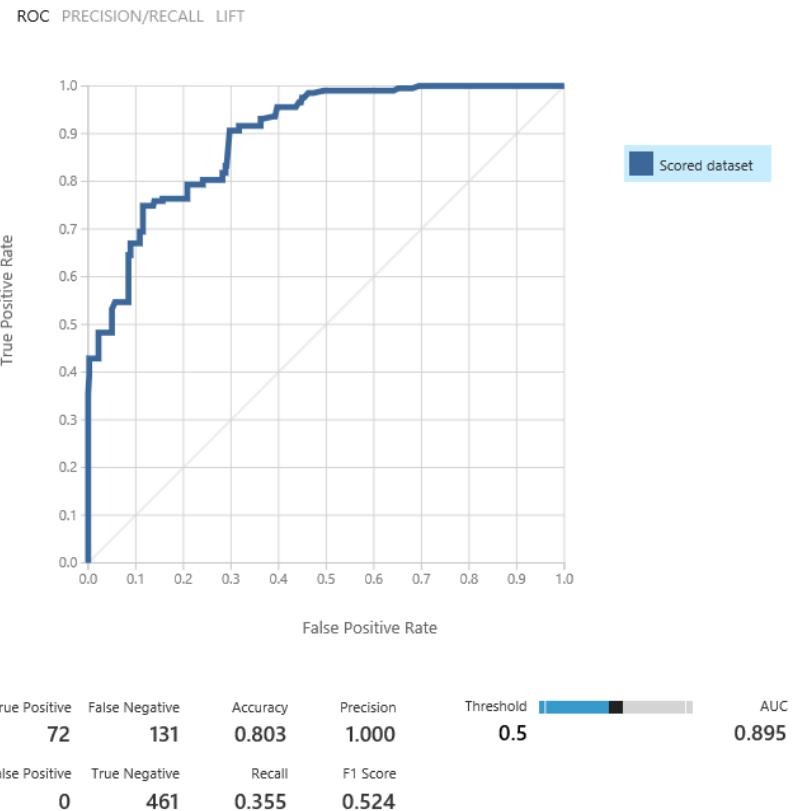


Figura 75 Curva ROC modelo de fin de ciclo estudiantes nuevos.

Fuente: Propia.

Como se puede observar con este modelo si se alcanza una precisión de 1.0 aunque este modelo presente un recall bajo. Ahora analicemos si la información de los deciles está más suavizada:

◀ Graphics

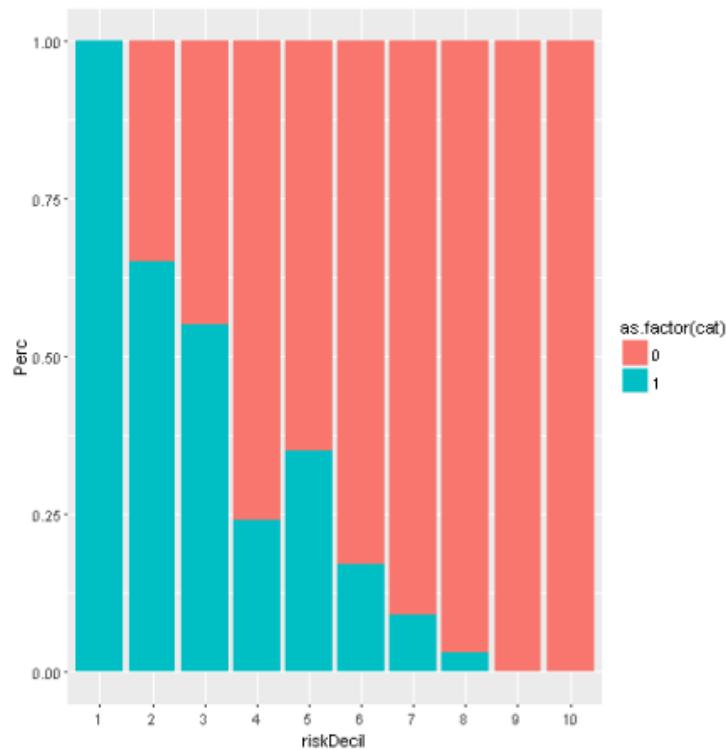


Figura 76 validación deciles modelo fin de semestre estudiantes nuevos.

Fuente: Propia.

Y sucedió exactamente lo que esperábamos un modelo más exacto en el decil uno y mantiene la precisión en los dos últimos deciles, aparte de que el escalonado en cada decil se ve de una manera más normal.

4.5.4 Resumen de todos los modelos.

De la misma manera como se explicó la mayoría de los modelos de inicio de semestre se hace la misma dinámica con los modelos de estudiantes de retorno por lo que se colocara como resumen los resultados de todos los modelos ya entrenados y validados:

Tipo de estudiante	Modelo	AUC	Precisión	Recall	Decil uno	Feature 1	Feature 2
Nuevo	BT	0.7	0.58	0.37	0.64	START.ENDGAGMENT	DIFERENCIA.FECHAS
Nuevo	Semana 2	0.82	0.87	0.5	0.94	START.ENDGAGMENT	EDAD
Nuevo	Semana 3 - 5	0.87	0.85	0.58	0.94	START.ENDGAGMENT	INTERACCIONES.SEMANA.SUMM
Nuevo	Semana 6 - 10	0.84	0.9	0.6	0.92	START.ENDGAGMENT	SUM.PAGOS
Nuevo	ET	0.89	1	0.35	1	END.ENDGAGMENT	NUMERO.TICKETS
Retorno	BT	0.75	0.46	0.11	0.32	PROMEDIO.HISTORICO	PERIODOS.HISTORICOS
Retorno	Semana 2	0.66	0.39	0.08	0.42	HIST.ENDGAGMENT	START.ENDGAGMENT
Retorno	Semana 3 - 5	0.68	0.55	0.15	0.4	PROMEDIO.HISTORICO	HIST.ENDGAGMENT
Retorno	Semana 6 - 10	0.7	0.58	0.18	0.42	PROMEDIO.HISTORICO	START.ENDGAGMENT
Retorno	ET	0.71	0.47	0.15	0.43	NUMERO.DIAS.PAGOS.DESPUES.FECHA	INTERACCIONES.SEMANA.SUMM

La finalización de esta aplicación se puede resaltar que los modelos de estudiantes de retorno no tienen la misma precisión que la tiene la de estudiantes nuevos, sin embargo, se puede ver como aumenta la precisión del percentil uno para la calibración de todos los modelos, de la misma manera se puede observar claramente como aumenta de manera significativa la precisión de los modelos de estudiantes nuevos terminando con una precisión de 1.0 en el decil uno para el último modelo.

4.5.5 Análisis de estudiantes de primer año

De la manera como hemos llegado a este punto solo es de concluir basado en los factores de riesgo quienes son esos estudiantes de primer año que tienen esas características más marcadas de desertar.

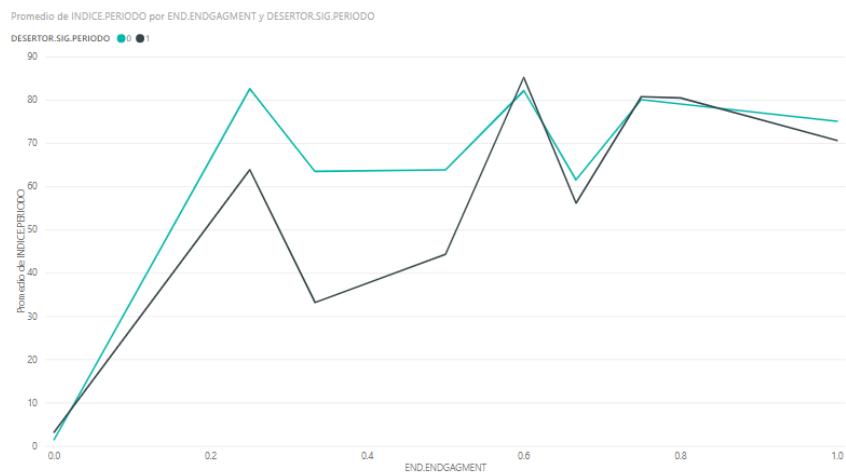


Figura 77 Explicación de factores de riesgo académicos para estudiantes nuevos

Fuente: Propia.

Podemos notar claramente en esta grafica que los estudiantes que desertan son aquellos que caen por debajo de 0.6 en su compromiso académico aparte de que son estudiantes que sacan notas casi por debajo de 60%.

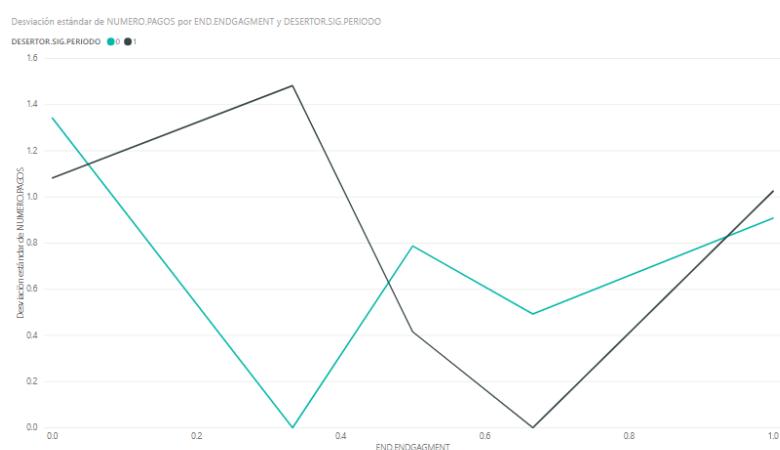


Figura 78 Comportamiento financiero de estudiantes que tienen bajo compromiso.

Fuente: Propia.

De la misma manera se puede observar que estudiantes que tienen bajo compromiso con registro tienen una desviación estándar más alta con respecto a los números de pagos por lo tanto estudiantes arriba 0.6 en desviación tienen más riesgo de deserción.

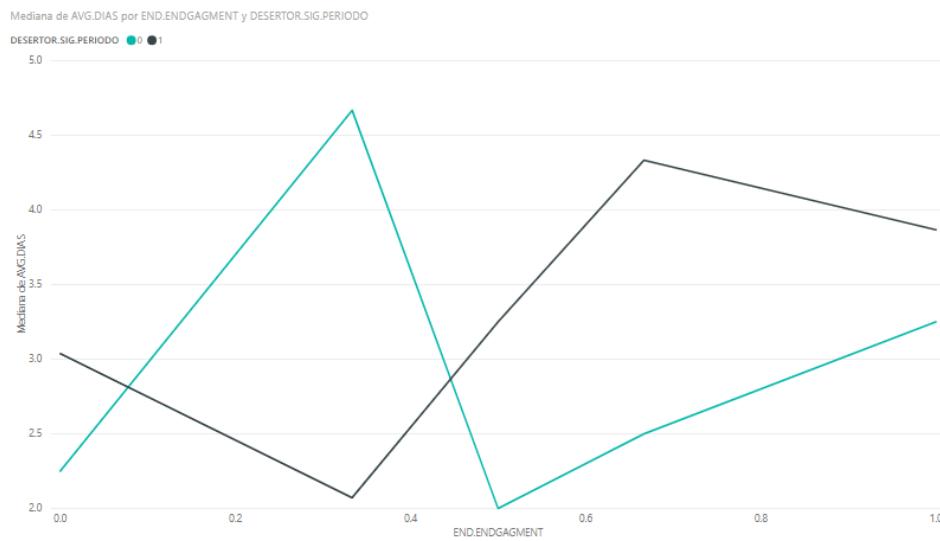


Figura 79 Estudiantes con buen compromiso y alto promedio de respuesta de tickets.

Fuente: Propia.

También en este ejemplo se puede observar que los estudiantes que son responsables con el compromiso de registro y tienen un promedio alto de respuesta de sus tickets tienen mayor riesgo de deserción.

4.6 Migración de modelo a la nube

La migración de estos modelos se puede realizar de manera sencilla en primer lugar se debe de exportar los modelos entrenados para posteriormente ser usados en otros ETL predictivos.

NAME	SUBMITTED BY	DESCRIPTION	DATA TYPE
20180205_UNITEC_W2_NEW_without_sumcalsem4	avilla		ILearnerDotNet
UNITEC_OLD_WEEK_A_7	nquecan		ILearnerDotNet
UNITEC_NEW_FIN	nquecan		ILearnerDotNet
UNITEC_OLD_FIN	nquecan		ILearnerDotNet
UNITEC_OLD_WEEK_B_10	nquecan		ILearnerDotNet
UNITEC_NEW_WEEK_B_10	nquecan		ILearnerDotNet
UNITEC_NEW_WEEK_A_7	nquecan		ILearnerDotNet
UNITEC_OLD_WEEK_1_3	uayeb.caballero		ILearnerDotNet
UNITEC_NEW_WEEK_1_3	uayeb.caballero		ILearnerDotNet

Figura 80 Modelos estadísticos exportados en Microsoft Azure

Fuente: Propia.

Estos modelos son exportados en un tipo de formato específico, este es un buen comienzo para los que desean hacer empaquetados de proyectos, sin embargo, se pueden crear modelos utilizando Microsoft Azure Machine Learning Services, este opera de una manera muy dinámica utilizando Docker al momento que usted quiere exportar el modelo utilizando Python se puede crear una imagen del entrenamiento y especificar la máquina en la que se ejecutarán los llamados.

Microsoft Azure genera un web service este puede ser llamado de distintas maneras, la convencional es a través de un “llamado” y la segunda a través de un llamado BATCH este último genera una estadística de la forma en que se manda a llamar el Web-Service.

unitec: nuevos semana 1-3

The screenshot shows the configuration interface for a web service. At the top, there are tabs for DASHBOARD and CONFIGURATION, with DASHBOARD selected. Below the tabs, there's a section for 'Published experiment' with links to 'View snapshot' and 'View latest'. A 'Description' field contains the note 'No description provided for this web service.' An 'API key' field is present with a redacted value. Under 'Default Endpoint', there are sections for 'API HELP PAGE', 'TEST' (with 'Test' and 'Test Preview' buttons), and 'APPS' (with two Excel 2013 or later workbook icons). In the 'TEST' section, there are also 'REQUEST/RESPONSE' and 'BATCH EXECUTION' buttons. At the bottom, there's a note about additional endpoints and a link to 'Manage endpoints'.

Figura 81 Configuración de web service.

Fuente: Propia.

Cuando se han generado todos los web services se para por el proceso de automatización, para esto se utilizan fábricas de datos, en esta herramienta se agregan todos los pipelines y actividades necesarias para la calendarización.



Figura 82 Fábrica de datos.

Fuente: Propia

Gracias a este tipo de implementación se puede incluir lógica de distribución de correos, donde en ese envío se puede adjuntar un reporte por la jefatura de la carrera con todos los estudiantes matriculados y su calificación de riesgo calculada, gracias a los modelos ya entrenados.

Este servicio de reporte por medio de correo es un gran beneficio ya que el coordinador podrá saber rápidamente cuál es el porcentaje de estudiantes en riesgo de abandono. Al momento en que éste se descargue, tendrá el detalle por estudiante de quiénes son los que están en mayor riesgo, ya sea por la calificación o la clasificación por deciles.

CAPITULO V CONCLUSIONES DE INVESTIGACIÓN

5.1 Conclusiones

La aplicación de modelos estadísticos en cualquier entorno de negocio es muy amplia y esta debe de ser una las buenas prácticas que todas las áreas de las instituciones deben de considerar aplicar, gracias a la ciencia de datos ahora se pueden contar historias más interesantes que aportan mucho valor y ayudan a la generación de productos basados en datos.

También se debe de considerar todo el conjunto de herramientas tecnológicas a ser aplicadas, gracias al poder de la nube es mucho más fácil migrar productos que antes era complicado llevar a producción. En nuestro caso analizamos Azure como herramienta principal de gestión de componentes en la nube y nuestra experiencia fue bastante buena. Ahora muchas maneras de trabajar local y rápidamente migrar y servir mediante el consumo de web services.

Otro de los beneficios de utilizar Azure en nuestro caso fue el fácil acceso gracias a la integración que hay con Office 365, este permite hace SSO con todos los componentes y se pueden crear diversas maquinas tanto para correr de manera distribuida transformaciones como maquinas que alojan servicios webs.

Esta tecnología de la misma manera nos ayudó a la fácil implementación de los reportes que serán distribuidos a los jefes de carrera, siempre se conservan los temas de seguridad de la información ya que los reportes a ser distribuidos son almacenados en el blob storage container mediante encriptaciones especiales.

Incluir información de distintas áreas de negocio es muy importante como ser el caso del CAP y la información de Blackboard gracias a ello pudimos crear transformaciones con gran poder predictivo y una alta interpretabilidad que ayuda a fáciles intervenciones.

Es interesante ver el comportamiento de las precisiones de los modelos estadísticos entrenados durante la fase predictiva de nuestra investigación, se pudo comprobar usando la tabla de resumen que los estudiantes nuevos entre más cerca del fin de ciclo mayor probabilidad de tener un acierto positivo de esos estudiantes que pueden desertar.

Por último, la aplicación de toda esta investigación a un nivel operativo mediante la nube ayudara y facilitara la toma de decisiones del equipo de retención de UNITEC HN teniendo un acceso claro de cuáles son esos factores de riesgo que deben de ser rápidamente intervenidos.

5.2 Recomendaciones

El campo de analítica avanzada para los estudiantes y garantizar el éxito de su vida estudiantil es muy amplio por lo que muy fácilmente varios de los temas tocados en esta investigación pueden desprender otras investigaciones más profundas. De la misma manera por todo el nivel de estadística aplicado y unión con tecnología emergente es un correcto inicio para una investigación a nivel de doctorado, por lo que cualquier persona hace lectura de esta investigación puede hacer referencia de ella como punto inicial y profundizar un poco más en la matemática aplicada mejorando sistema de ecuaciones líneas o transformándolas a clasificación de probabilidades bayesianas para una mejor interpretación de los factores de riesgos individuales.

Es muy importante tomar en cuenta el tiempo de desarrollo de esta investigación, ya se contaba con la experiencia del campo de ciencia de datos y en su mayoría los datos ya estaban

listos para comenzar a ser analizados. Por lo que si otra persona pretende hacer una investigación de este tipo se recomienda considerar altamente los dos puntos anteriores.

Adicionalmente se le recomienda a UNITEC hacer una capacitación a los jefes de carrera, de la correcta lectura de ese reporte, para que sepan cómo intervenir a los estudiantes o escalar con otras áreas; cómo se puede trabajar la calificación de riesgo en función de que esta comience a disminuir.

Es muy importante que los equipos de retención estudiantil y los jefes de carrera puedan contar en un futuro con equipo (tabletas) donde puedan ver información o comportamientos importantes como los descubiertos en el análisis exploratorio de esta investigación, sobre las tendencias de bajo desempeño correlacionados con la deserción o probabilidad de abandono.

BIBLIOGRAFÍA

- Análisis de datos: 10 señales de que eres “Data-Driven”.* (s.f.). Obtenido de NUBLIO: <https://nublio.com/analisis-de-datos-la-empresa-data-driven/>
- Apache. (2019). *Apache Hadoop*. Obtenido de Apache Hadoop: <https://hadoop.apache.org/>
- Camargo-Vega, J., Camargo-Ortega, J., & Joyanes-Aguilar, L. (2015). Conociendo Big Data. *Revista Facultad de Ingeniería, Tomo 24, N.º 38*.
- Diez, D. M., Barr, C., & etinkaya-Rundel, M. (2015). *OpenIntro Statistics*.
- Ferrer-Rodriguez, J. (2015). La retención estudiantil en las instituciones educativas postsecundarias universitarias que ofrecen como máximo el grado asociado. *ProQuest Central*, <https://search.proquest.com/docview/1688680777?accountid=35325>.
- Google. (27 de Marzo de 2018). *Aprendizaje Automatico*. Obtenido de Google Developer: <https://developers.google.com/machine-learning/crash-course/classification/true-false-positive-negative?hl=es-419>
- Guta, M. (2017). *What is Gamification and How Can It Help My Business?* Obtenido de smallbiztrends: <https://smallbiztrends.com/2017/07/what-is-gamification.html>
- Inc, D. (s.f.). *What is a container*. Obtenido de docker: <https://www.docker.com/resources/what-container>
- Inc, L. E. (2019). *Laureate Education Inc*. Obtenido de <https://www.laureate.net/>.
- James, G., Witten, D., Trevor , H., & Tibshirani, R. (2015). *An Introduction to Statistical Learning*. know, T. p. (s.f.). Obtenido de https://www.sas.com/es_ar/insights/big-data/internet-of-things.html
- Logicalis. (2017). *Learning machine, los usos del aprendizaje supervisado*. Obtenido de blog.es.logicalis.com: <https://blog.es.logicalis.com/analytics/learning-machine-los-usos-del-aprendizaje-supervisado>
- Microsoft. (15 de 11 de 2016). *Azure Data Factory version 2 (V2)*. Obtenido de Microsoft Azure: <https://docs.microsoft.com/en-us/rest/api/datafactory/v2>
- Microsoft. (2019). *Azure Machine Learning service*. Obtenido de Microsoft Azure: <https://azure.microsoft.com/en-us/services/machine-learning-service/>
- Nespereira, C. G., Fernández Vilas, A., & Díaz Redondo, R. (2015). Am I failing this course? Risk prediction using e-learning data.
- Patil, D., & Mason, H. (2015). *Data Driven Creating a data culture*. O'Reilly Media.
- Patriarca, M. (2013). La deserción en el inicio de la vida universitaria. http://www.revistaraes.net/revistas/raes6_art3.pdf.
- Peng, R. D. (2015). *Exploratory Data Analysis with R*. Leanpub.
- Peng, R. D. (2015). *The Art of Data Science*.
- Pineda Báez, C., & Pedraza Ortiz, A. (2009). Programas exitosos de retención estudiantil universitaria: las vivencias de los estudiantes. *Revista Virtual Universidad Católica del Norte*.
- Saffirio, M. (2017). *Gobernabilidad de Datos*. Obtenido de Tecnologías de la Información y Procesos de Negocios (BPM): <https://msaffirio.wordpress.com/2017/01/11/gobernabilidad-de-datos/>

- Santos, M. (77 de 01 de 2015). *Enter*. Obtenido de Enter.co: <https://www.enter.co/guias/tecnoguias-para-empresas/saas-iaas-y-paas-que-son-como-usarlos-y-para-que/>
- Vergara Morales, J. R., Eva Boj, d., Barriga, O., & Díaz Larenas, C. (2017). Factores explicativos de la deserción de estudiantes de pedagogía. *Revista Complutense De Educación*, 28(2), 69-630. .
- Vries, W. (2011). ¿Desertores o decepcionados? Distintas causas para abandonar los estudios universitarios. <http://www.scielo.org.mx/pdf/resu/v40n160/v40n160a2.pdf>.

GLOSARIO DE TÉRMINOS

Termino	Concepto
Azure Active Directory	Directorio centralizado donde se encuentran todos los usuarios y las aplicaciones que se pueden gestionar para dar permiso a otros tipos de librerías.
Azure Data Lake Analytics	Tecnología de almacenamiento masivo de diferentes tipos de fuente de datos y tipos.
Azure HDInsight Hadoop	Tecnología es de Azure específica para correr Hadoop.
Azure Machine Learning	Tecnología de Microsoft para realizar aprendizaje automático de los datos.
BATCH	Proceso que corre una gran cantidad de datos en un solo llamado.
Bigdata	Conjunto gigante de datos que requiere de infraestructura especial para procesar datos.
Blackboard	Es una tecnología de LMS.
Blob storage container	Estructura de almacenamiento en la nube que persiste la información encriptada en SHA256.
Data Management Book of Knowledge	Uno de los libros principales donde se describe las mejores prácticas de manejo de datos a nivel organizacional y gobierno.
E-learning	Plataforma en línea de aprendizaje continuo donde los docentes pueden planificar múltiples actividades académicas y los estudiantes pueden utilizarlas de manera que complementa la experiencia estudiantil.
ETL	Extracción, transformación y carga de datos.
False Negatives	Es un resultado en el que el modelo predice incorrectamente la clase negativa.
False Positives	Es un resultado en el que el modelo predice incorrectamente la clase positiva.
Hadoop	La biblioteca de software Apache Hadoop es un marco que permite el procesamiento distribuido de grandes conjuntos de datos en grupos de computadoras utilizando modelos de programación simples. Está diseñado para escalar desde servidores individuales a miles de máquinas, cada una ofrece computación y almacenamiento locales. En lugar de confiar en el hardware para ofrecer alta disponibilidad, la biblioteca está diseñada para detectar y manejar fallas en la capa de la aplicación, por lo que ofrece un servicio de alta disponibilidad sobre un grupo de computadoras, cada una de las cuales puede ser propensa a fallas. (Apache, 2019)
KPI	Indicadores principales de rendimiento.
Linux	Sistema Operativo Libre.
LMS	Sistema administrador de aprendizaje.

Logins	Accesos a la plataforma de e-learning.
Machine Learning	Máquinas de aprendizaje estadísticas para hacer regresiones o clasificaciones.
Marketing and business Insights	MBI, Ideas de marketing y negocios.
Microsoft Azure	Solución de la nube de Microsoft.
Office 365	Producto de Microsoft que centraliza todas sus soluciones usando una sola plataforma y método de autenticación.
Python	Lenguaje de programación multiplataforma y multipropósito.
PyTorch	Librería de Python para inteligencia artificial de Facebook.
R	Lenguaje de programación estadístico.
Recall	Porcentaje de aciertos versus todo el universo real a predecir.
Resiliencia	Perseverancia de no abandonar por más fracaso que se encuentre.
ROC	Curva que mide precisión basado en Positives and Negatives Trues.
Scikit-learn	Librería de Python para máquinas de aprendizaje.
Serverless	Capacidad de configurar y correr rutinas de código teniendo como resultado la exposición rápida de hacer una integración hacia otros servicios sin necesidad de configurar servidores.
Spark	Capa dentro de la arquitectura de un sistema de bigdata para correr procesos de manera distribuida.
SSO	Método de comunicar dos o más servicios de manera segura mediante la nube.
TensorFlow	Librería de Python para inteligencia artificial de Google.
Tickets	Solicitud enviadas por estudiantes y personal de UNITEC para hacer algún llamado de atención.
True Negatives	Es un resultado en el que el modelo predice correctamente la clase negativa.
True Positive	Es un resultado en el que el modelo predice correctamente la clase positiva. (Google, 2018)
UVs	Unidades Valorativas.
Web-Service	Herramienta web para comunicar servicios o iniciar uno nuevo, en nuestro caso se utiliza para ejecutar el ETL de nuevas predicciones.
Multicolinealidad	Afirmamos que hay colinealidad aproximada, cuando una o más variables, no son exactamente una combinación lineal de la otra, pero existe un coeficiente de determinación entre estas variables muy cercano al uno.
Cuantiles	Los cuantiles son medidas de posición que se determinan mediante un método que determina la ubicación de los valores que dividen un conjunto de observaciones en partes iguales.
Cuartiles	Los cuartiles son los tres valores que dividen al conjunto de datos ordenados en cuatro partes porcentualmente iguales.
Deciles	Los deciles son ciertos números que dividen la sucesión de datos ordenados en diez partes porcentualmente iguales. Son los nueve valores que dividen al conjunto de datos ordenados en diez partes iguales, son

	también un caso particular de los percentiles. Los deciles se denotan D1, D2,..., D9, que se leen primer decil, segundo decil, etc.
Probabilidad bayesiana	La probabilidad bayesiana es una de las diferentes interpretaciones del concepto de probabilidad. La interpretación bayesiana de la probabilidad puede ser vista como una extensión de la lógica proposicional que permite razonar con hipótesis, es decir, las proposiciones cuya verdad o falsedad son inciertas.
Curva ROC	En la Teoría de detección de señales, una curva ROC (acrónimo de Receiver Operating Characteristic, o Característica Operativa del Receptor) es una representación gráfica de la sensibilidad frente a la especificidad para un sistema clasificador binario según se varía el umbral de discriminación. Otra interpretación de este gráfico es la representación de la razón o ratio de verdaderos positivos (VPR = Razón de Verdaderos Positivos) frente a la razón o ratio de falsos positivos (FPR = Razón de Falsos Positivos) también según se varía el umbral de discriminación (valor a partir del cual decidimos que un caso es un positivo). ROC también puede significar Relative Operating Characteristic (Característica Operativa Relativa) porque es una comparación de dos características operativas (VPR y FPR) según cambiamos el umbral para la decisión. En español es preferible mantener el acrónimo inglés, aunque es posible encontrar el equivalente español COR.
Sensibilidad	Nos indica la capacidad de nuestro estimador para dar como casos positivos los casos realmente enfermos; proporción de enfermos correctamente identificados. Es decir, la sensibilidad caracteriza la capacidad de la prueba para detectar la enfermedad en sujetos enfermos.
Especificidad	Nos indica la capacidad de nuestro estimador para dar como casos negativos los casos realmente sanos; proporción de sanos correctamente identificados. Es decir, la especificidad caracteriza la capacidad de la prueba para detectar la ausencia de la enfermedad en sujetos sanos.