DISEÑO DEL PROCESO OPERACIONAL DE ANALÍTICA DE DATOS BASADO EN EL SISTEMA DE INFORMACIÓN DE ACADEMUSOFT

autor
HUGO ALEXANDER PRADA ORJUELA

Director
AVILIO VILLAMIZAR ESTRADA
Magister en gestión de proyectos informáticos

INGENIERÍA DE SISTEMAS
ELECTRÓNICA, ELÉCTRICA, SISTEMAS Y TELECOMUNICIONES
INGENIERÍAS Y ARQUITECTURA

UNIVERSIDAD DE PAMPLONA
PAMPLONA, 25 de noviembre de 2019

TABLA DE CONTENIDO

LISTA DE ILUSTRACIONES

LISTA DE TABLAS

1. RESUMEN DEL PROYECTO

La analítica de datos consiste en examinar grandes volúmenes de datos con el fin de descubrir patrones ocultos, correlaciones, tendencias y otras ideas útiles en esos grandes almacenes de información para así obtener ventajas competitivas a través de organizaciones rivales y dar beneficios para el negocio como el marketing más efectivo y generar mayores ingresos. Este término usa los grandes volúmenes de información para inspeccionar, limpiar, transformar y modelar datos con el fin de descubrir información útil que se necesita. En este proyecto se hará énfasis en una categoría de la analítica de datos la cual se llama análisis predictivo, esta usa algoritmos avanzados para pronosticar comportamientos y tendencias implicando el uso de modelos estadísticos, consultas analíticas y algoritmos de aprendizaje automático con el fin de crear modelos predictivos que sitúen un valor numérico o puntuación de que ocurra un evento particular.

Este trabajo consiste en analizar y determinar las distintas técnicas de análisis predictiva usadas para diseñar un documento que indique como se debe realizar de manera correcta y eficaz el diseño de estos procesos para determinar tendencias sobre la información registrada en el sistema de Academusoft.

1.1 Palabras claves

Analítica de datos, análisis predictivo, bases de datos No-SQL.

2. PLANTEAMIENTO DEL PROBLEMA

La analítica de datos es uno de los temas actuales que está tomando relevancia en todas las industrias, ya que en muchas empresas se están enfrentando al crecimiento del volumen, velocidad y variedad de datos, las empresas que cuentan con grandes volúmenes de información y que en gran parte no han sido usadas, son una gran oportunidad para describir comportamientos y tendencias de los distintos datos, pero la gran cantidad de volúmenes de información demandan el uso de herramientas informáticas para el correcto almacenamiento y gestión que permita obtener un mecanismo eficiente para el análisis, extracción de información y comportamiento. Teniendo en cuenta lo anterior se tiene que el volumen creciente de los datos genera una oportunidad que requiere de técnicas de análisis de datos que permitan el aprovechamiento de estos con el fin de conseguir conocimientos para tener un soporte en la toma de decisiones de las empresas. El interrogante de esta propuesta es, ¿De qué modo puede beneficiar el desarrollo de un proceso operacional de analítica de datos en una empresa con grandes volúmenes de información?

3. JUSTIFICACION

En los últimos años la cantidad de datos que poseen las empresas u organizaciones han sobrepasado niveles sin precedentes, gracias a los avances tecnológicos se han creado nuevos modelos de base de datos NO-SQL los cuales permiten procesar de mejor manera la información alojada en los grandes volúmenes de esta. Si se analizan estos datos las empresas podrán descubrir patrones ocultos, desconocidos, tendencias del mercado, preferencias de clientes y otra información que puede beneficiar la toma de decisiones. La importancia de desarrollar el proyecto radica en definir los pasos correspondientes para aplicar de manera correcta y eficaz la analítica de datos en el sistema de información de Academusoft, por medio de la analítica predictiva se podrá identificar riesgos y oportunidades que tiene la Universidad de Pamplona para obtener ventajas competitivas frente a otras empresas del mismo sector.

4. MARCO TEORICO

Es el proceso que inspecciona, limpia, transforma y modela datos con el fin de descubrir información útil que se necesita. Al realizar un análisis de datos se obtendrá además de información de utilidad, sugerencias y conclusiones que ayudará a la toma de decisiones (Education, s.f.).

Existen varios tipos de analítica de datos según el nivel de valor de la información los cuales son:

- **Analítica descriptiva**

Analítica de datos que consiste en almacenar y realizar agregaciones de datos históricos, visualizándolos de forma que puedan ayudar a la comprensión del estado actual y pasado del negocio. Esta analítica determina el funcionamiento de la empresa hasta la fecha.

- **Analítica predictiva**

Analítica que se construye sobre la analítica descriptiva y usa modelos estadísticos avanzados para añadir a una base de información de datos que se desconoce. Esto se traduce en técnicas de análisis estadísticos, consultas analíticas y algoritmos de aprendizaje automáticos de datos para crear modelos que sitúen un valor numérico o puntuación en la probabilidad de que ocurra un evento en particular

- **Analítica prescriptiva**

Es la analítica de más alto nivel que junta las anteriores y agrega estrategias de optimización operativa con la finalidad de indicar acciones de negocio para proporcionar los mejores resultados (Routledge, 2013). Mediante esta analítica se puede obtener recomendaciones

automatizadas sobre el momento idóneo para ejecutar pedidos, mantenimientos u otras operaciones cuantificables.

Sistemas de información

Es un sistema encargado de coordinar los flujos y el registro de la información necesaria para llevar a cabo las funciones de una empresa determinada (Andreu R, 1996). Los sistemas de información sirven para dar apoyo a las organizaciones en la planificación, diseño, ejecución y control.

Existen cuatro maneras de ver un sistema de información (Boell S.K, 2015):

1. Visión Tecnológica: sistema de información que incluye el procesamiento de datos, hardware, software, modelos de análisis entre otros. Estos hacen énfasis exclusivamente en la información.
2. Visión social: Sistema de información que está inmerso dentro de un contexto tecnológico y contienen como base los ámbitos sociales teniendo en cuenta la repercusión de estos.
3. Visión Socio-técnica: Sistema de información que estudia las repercusiones técnicas, sociales y a la vez las interacciones que surgen a partir de la combinación de estos dos ámbitos
4. Visión del proceso: Los sistemas de información apoyan y mejoran los procesos dentro de las organizaciones a través de transmitir, manipular, recuperar, capturar y mostrar la información alojada en estos.

Base de datos No-SQL

Este tipo de base de datos están diseñadas para modelos de datos y poseen esquemas flexibles, estas son fáciles de desarrollar permitiendo funcionalidad, rendimiento. Procuran dar solución a las problemáticas, dando la posibilidad de abordar la forma de gestionar la información de una manera diferente a como se viene realizando en las bases de datos tradicionales.

Características

- Estructura distribuida. Se distribuyen los datos mediante mecanismos de tabla de hash distribuidas, donde varias máquinas cooperan en grupos para ofrecer a los clientes datos con el fin de obtener una alta disponibilidad
- Escalabilidad horizontal. Implementación realizada en muchos nodos de capacidad de procesado limitado, por lo tanto, logra tener capacidades de procesamiento general.
- Construcción para grandes volúmenes. Estas bases de datos fueron construidas con el fin de ser capaz de almacenar y procesar enormes cantidades de datos de forma rápida.
- Tolerancia a fallos y redundancias

5. ESTADO DEL ARTE

- **Diseño de un sistema de información, bajo un enfoque de inteligencia de negocios, para el proceso de toma de decisiones, Juan Cadena, 2016.**

La gran cantidad de datos que se han acumulado en las empresas no han sido utilizadas de forma adecuada generando inconsistencias, múltiples versiones y un desperdicio de tiempo y recursos. Manejándose buenas prácticas en la utilización de datos, en la medición de mejoras alcanzadas. La selección de herramientas tecnológicas apropiadas y de la analítica de datos permiten obtener una madurez de las empresas que beneficien en el proceso de toma de decisiones.

- **Importancia y decisión de uso de almacenes de datos en la toma de decisiones en empresas corporativos, Benítez Astudillo, Javier Vallejo, Jorge Bermúdez, 2018.**

Las herramientas tecnológicas de analítica de datos permiten mejorar la gestión de la información almacenada en las bases de datos empresariales, resolviendo problemas relacionados como los costes, redundancias de datos y escasa visión corporativa. La implementación de estas herramientas conlleva mejoras del ejercicio empresarial, mejorando el comportamiento de la empresa y promoviendo procesos de cambio en la organización para una excelente toma de decisiones.

- **Análisis comparativo de herramientas de software libre y propietario para la gestión de Big data en empresas de comercialización masiva, cabezas Jácome, John Steven, 2015.**

Las empresas de hoy en día optan por introducir este concepto como una oportunidad para explotar al máximo toda la información que tienen alojada, esta información permitirá ayudar a mejorar la cultura de toma de decisiones dentro de sus diferentes áreas operativas, así como

también el manejo de dicha información en cada una de ellas. Como consecuencia es necesario analizar varios conceptos del Big data, entre ellos se incluyen definición, características, ventajas y desventajas para tener más claro el panorama de funcionamiento dentro de una organización particular.

- **Big Data: Procesamiento y calidad de datos, Sergio García, Julián Luengo, Francisco Herrera, 2014.**

En los últimos años el crecimiento masivo de datos está siendo un factor clave en el escenario de procesamiento de datos. La extracción de conocimiento depende en gran medida de la calidad de los datos, el cual se garantiza por algoritmos de procesamiento. En esta era los algoritmos poseen dificultad para operar con grandes cantidades de información, por lo tanto, se ve la necesidad de crear nuevos modelos que mejoren la capacidad de escalado por medio de las herramientas y técnicas de análisis de datos.

6. OBJETIVOS

6.1 OBJETIVO GENERAL

Diseñar un proceso operacional de la analítica de datos para procesar información de la base de datos del sistema Academusoft de la Universidad de Pamplona.

6.2 OBJETIVOS ESPECIFICOS

- Realizar un estudio del arte para el diseño operacional de aplicación de analítica de datos para sistemas de información.
- Analizar la información de las diferentes estrategias, metodologías y herramientas para la analítica de datos.
- Presentar el diseño operacional de análisis de datos para el sistema de información Academusoft.

6.3 ACOTACIONES

El presente trabajo culmina con el diseño del proceso operacional de analítica de datos, teniendo en cuenta, que se tomará la parte académica de la base de datos del sistema Academusoft. Las herramientas y tecnologías serán analizadas y definidas como parte del trabajo.

7. CRONOGRAMA DE ACTIVIDADES

Tabla 1 Cronograma y descripción de actividades

	QUINCENAS							
ACTIVIDAD	1	2	3	4	5	6	7	8
1 Estudio para el diseño operacional de aplicación de analítica de datos								
2 Analizar las estrategias de analítica de datos.								
3 Analizar las metodologías de analítica de datos.								
4 Analizar las tecnologías de analítica de datos.								
5. Creación del diseño operacional de análisis de datos para el sistema de información Academusoft								
6 Presentar el diseño operacional de analítica de datos.								

8. METODOLOGIA

Tipo de investigación

El tipo de investigación es de corte descriptiva ya que según (Arias, 2006) define: La investigación descriptiva consiste en la caracterización de un hecho, fenómeno, individuo o grupo, con el fin de establecer su estructura o comportamiento. Este tipo de investigación permite el estudio de un fenómeno por medio de variables para poder describir lo que se está investigando permitiendo la posibilidad de realizar predicciones. Dada las características relacionadas con el manejo y tratamiento de información almacenadas en la base de datos de Academusoft y los procesos que contempla la temática de analítica de datos, el resultado implica un proceso operacional que permite el procesamiento de datos en el objeto de estudio.

El desarrollo de este proyecto consta de 3 etapas

1. Se realizó un estudio del arte sobre diseños de procesos operacionales de analítica de datos.
2. Se llevó a cabo un análisis de las diversas estrategias, metodologías y herramientas de la analítica de datos. Mediante al análisis se determinó ventajas y desventajas de cada una de estas, seleccionando la más óptima para la creación de un modelo de analítica de datos.
3. Se realizó el diseño del proceso operacional de analítica de datos en el sistema de información Academusoft con énfasis en lo académico donde se describió el objetivo del proceso, métricas, herramientas, recolección de datos, tratamiento de datos, modelos y evaluación de modelos.

9. RESULTADOS

La creación del proceso operacional de analítica de datos basado en el sistema de información Academusoft, está dividido en tres etapas donde cada una de estas corresponden al debido cumplimiento de los objetivos específicos previamente planteados.

9.1 Etapa 1 Estudio del arte

- **Analítica predictiva de big data en sistemas de base de datos relacionales, Aquino Ximena, 2015**

El avance de las tecnologías y la gran cantidad de información que consiguen y manejan las grandes empresas, representan un nuevo desafío para estas organizaciones debido a que, esta información les permite analizar, revelar y deducir nuevas tendencias que se obtiene al realizar un respectivo procesamiento de datos con metodologías de análisis estadísticos junto con algoritmos de analítica de datos permitiendo dar sentido a todos los datos que poseen las organizaciones logrando mejores decisiones de negocios. Este trabajo realizado permite crear un modelo de análisis predictivo en una base de datos relacional, realizando una respectiva recolección, almacenamiento, tratamiento y visualización de datos permitiendo que la organización afectada por este modelo pueda tomar optimas decisiones empresariales y entender los comportamientos de sus usuarios.

- **Implementación de un sistema de información de análisis predictivo para la toma de decisiones en el proceso de atención médica del hospital Victor lazarte, Horna Carolina, Rodríguez Lorenzo, 2016**

La implementación de los sistemas de información con la aplicación de analítica predictiva otorga beneficios en las organizaciones de cualquier ámbito. Algunas organizaciones cuentan con grandes volúmenes de información que no se encuentran alojadas en bases de datos, es necesario realizar un análisis de como estructurar y modelar las bases de datos para el respectivo almacenamiento de los volúmenes de información. La adecuada implementación de los sistemas de información y la analítica de datos con herramientas open source permite potenciar el proceso de toma de decisiones y en este caso mejorar el proceso de atención medica tales como gestión de citas, manejo de medicamentos entre otros.

- **Modelo predictivo de persistencia universitaria: Alumnado con beca salario, Silvente Vanesa, Gazo María, Fanals Ernert, 2018**

El uso de los modelos predictivos en la educación superior permite buscar factores precisos en la situación de cómo se encuentra los alumnos en las instituciones educativas, la implementación de las becas salario garantizan el acceso y la persistencia de los alumnos en las instituciones educativas que se encuentran en grupos más desfavorecidos económicamente, el uso de estos modelos permiten obtener resultados de la probabilidad de que un alumno becado conserve la persistencia universitaria, para que el modelo predictivo sea válido se necesita de diversas técnicas que se encuentran dentro de este modelo, en este caso se realizó con la técnica de regresión lineal permitiendo seleccionar factores determinantes dentro del sistema de información para aumentar la probabilidad de que la población objeto de estudio permanezca en las instituciones educativas.

- **Modelo predictivo de deserción estudiantil basado en arboles de decisión, Cuji Blanca, Gavilanes Wilma, Sánchez Rina, 2017**

Las instituciones de educación superior en América Latina sufren de un nivel de deserción estudiantil alto, las instituciones generan grandes volúmenes de información de los estudiantes teniendo información como datos personales, económicos, socioeconómicos entre otros. La aplicación de los modelos predictivos permite saber con qué posibilidad puede un estudiante abandonar su programa académico, el modelo predice esto por medio de datos históricos con los que cuentan la institución de educación superior que se encuentran almacenadas en los sistemas de información, teniendo como principales factores el rendimiento académico y socioeconómicos como su estrato. La aplicación de técnicas arboles de decisión junto con el algoritmo de Classification and Regression Tree (CART) permite validar que los principales factores de deserción estudiantil son las notas que presentan los estudiantes en sus programas académicos, permitiendo que las instituciones tomen decisiones al respecto para reducir la gran probabilidad de deserción académica.

- **Analítica de datos de aprendizaje ADA y gestión educativa**

En la actualidad diversas instituciones educativas cuentan con una gran cantidad de datos e información tales como la asistencia, nivel socioeconómico, desempeño académico entre otros, esta información que albergan estas instituciones son importantes para mejorar estas organizaciones en cuestión de desempeño, alertas tempranas y toma de decisiones. Por lo tanto, es necesario indagar acerca de la temática de análisis de datos para generar un proceso que pueda brindar una mejora a la gestión educativa.

9.2 Etapa 2. Análisis de estrategias, metodologías y herramientas de analítica de datos

9.2.1 Estrategias

Dentro de la analítica de datos existen diversas estrategias según la información con la que se cuente y que se requiera dentro de estas se encuentra la descriptiva, predictiva y prescriptiva. En este trabajo se hará uso de la estrategia predictiva debido a que se busca predecir un comportamiento de los individuos respecto a las actividades académicas de la universidad de Pamplona. Este tipo de modelo da uso de diversas técnicas de análisis estadísticos, utilizando la información alojada en una base de datos extrayendo las características de los individuos objeto de estudio como entrada y asignando un valor numérico o calificación como salida entre más alta sea la puntuación es más grande la probabilidad de que ocurra un evento en particular.

Dentro de la analítica de datos existen diversas estrategias que permiten tratar la información que se encuentra alojada en los sistemas de información según la necesidad que requieren las organizaciones, dentro de estas se encuentran las siguientes estrategias:

- **Prescriptivo**

La estrategia prescriptiva usa los datos obtenidos del sistema de información para establecer acciones o decisiones que permiten obtener mejores resultados, este tipo de estrategia necesita de un modelo predictivo previo que se le agregan dos módulos adicionales los cuales son datos procesables y un sistema de retroalimentación que monitoriza los efectos de las acciones sugeridas. Permitiendo a una organización obtener un conjunto de reglas que permitan producir o pronosticar un resultado deseado para lo que se requiera. Este tipo de estrategia se usa cuando una organización requiere manejar la información y actuar de manera inmediata, esta no solo

pronostica un futuro posible, sino varios dependiendo de las operaciones que se tomen permitiendo mejorar la toma de decisiones a tiempo real dentro de una organización.

- **Predictiva**

La estrategia predictiva utiliza diversos modelos estadísticos y técnicas de aprendizaje automático para analizar datos recientes e históricos calculando la probabilidad de que un elemento que posea características similares al conjunto de datos demuestre un comportamiento específico. Este tipo de estrategia utiliza datos existentes alojadas en una base de datos para pronosticar datos o comportamientos de los cuales no se dispone, para crear un modelo predictivo se necesitan de dos conjuntos de datos los cuales se dividen en datos de entrenamiento y prueba. Los datos de entrenamiento se usan para alimentar el modelo recogiendo todas las características necesarias para obtener una predicción correcta de lo que se desea y los datos de prueba se usan para garantizar y validar el rendimiento del modelo permitiendo mejorar las operaciones, tendencias y tomas de decisiones de una organización.

Las calificaciones que generan los modelos predictivos son de especial cuidado y puede necesitar el uso de varias técnicas de análisis estadísticos para garantizar la efectividad de estas. Estas calificaciones muestran tendencias en un grupo adecuadamente amplio con el fin de certificar que la respectiva predicción se cumpla en cada caso donde se ponga a prueba. Este tipo de análisis valora la correspondencia que existe entre una gran cantidad de elementos para aislar los datos que informan sobre un hecho, permitiendo tener una mejor toma de decisiones.

- **Descriptivo**

La estrategia descriptiva usa conjuntos de datos históricos que se encuentran en los sistemas de información identificando las relaciones entre los datos para poder clasificar individuos en

grupos, esta estrategia genera resúmenes de lo que ha ocurrido en el pasado y como se encuentra actualmente la organización proporcionando información relevante permitiendo que esta pueda descubrir, investigar, calcular e identificar diversos indicadores para adquirir una visión de lo que está pasando, proporcionando un reducción de costes y gestión inteligente en la organización.

Validación de modelos

Luego de la creación de un modelo predictivo es necesario comprobar la validez de este, para una validación correcta se debe contar con dos tipos de datos los cuales se dividen en entrenamiento y prueba. Estos tipos deben tener una cierta cantidad de datos para que el modelo pueda predecir con un alto índice de acierto, por lo general los datos de entrenamiento deben poseer 2 terceras partes de la muestra total y los datos de prueba debe contener las muestras sobrantes. Para determinar si el modelo está clasificando correctamente se debe realizar una matriz de confusión la cual mostrara como está clasificando las muestras mostrando el nivel de precisión del modelo.

9.2.2 Metodologías

- **Metodología de regresión**

Las metodologías de regresión son la columna principal de la analítica predictiva, se basan en técnicas estadísticas que calculan la relación entre variables, estas metodologías son planteados para datos continuos siguiendo una distribución normal (SAS, 2017). Con la finalidad de representar las relaciones entre las diferentes variables a usar. Dentro de las metodologías de regresión se pueden encontrar las siguientes técnicas:

- **Regresión lineal**

La regresión lineal es una técnica de aprendizaje supervisado que se utiliza en Machine Learning y en estadística, la cual estudia la correlación entre variables además de utilizarse para la predicción de varios fenómenos, esta compara la relación que existe entre la variable dependiente aquella que contiene la característica de respuesta de un grupo en el modelo y un conjunto de variables independientes que contiene características únicas de los grupos a clasificar, por medio de una ecuación que predice la variable dependiente como una función lineal de los parámetros.

Para que la regresión lineal funcione de manera óptima es necesario ajustar los parámetros según el conjunto de datos que se use. Esta técnica reduce el coste de una función de error cuadrático y los coeficientes dados pertenecerá a la recta óptima (Roman, s.f.) .

Características

- Eficaz con datos lineales.
- Sensible a valores extremos de los datos.

Condiciones para asegurar y garantizar la validez de un modelo de regresión lineal

- Linealidad

La variable dependiente debe ser la suma de un conjunto de elementos como lo son: origen de la recta, combinación de variables independientes y los residuos. La falta de uno de estos elementos se denomina error de especificación.

- Independencia

Los residuos son autónomos entre ellos ya que estos establecen una variable aleatoria.

- Homocedasticidad

La varianza de los residuos es constante en cada valor de las variables independientes.

- Normalidad

Los residuos se distribuyen con media cero en cada valor de la variable independiente.

- **Análisis de supervivencia o duración**

El análisis de supervivencia es una metodología que se utiliza principalmente en campos de las ciencias médicas y biológicas, esta técnica es de carácter inferencial teniendo como objetivo principal modelar el tiempo en que ocurra un determinado fenómeno o suceso, los individuos objeto de estudio entran en tiempos distintos y se va realizando un seguimiento desde su entrada hasta que se produzca el fenómeno o suceso. Al ser un análisis con técnicas no paramétricas ha triunfado en ámbitos de salud y economía provocando que se generalice más este tipo de técnicas desarrollando una mayor eficacia en sus modelos (Perez, 2013).

La técnica no paramétrica más usada es el estimador de Kaplan-Meier conocido como el estimador del límite del producto este se puede usar en conjunto de datos que poseen datos con censuras es necesario poseer dos características de los individuos los cuales son el tiempo de la última observación que son los días que han pasado hasta que suceda el fenómeno y el estado del individuo que indica el tiempo en que ha ocurrido el fenómeno.

Características

- **Análisis de sobrevida:** conjuntos de métodos de carácter estadísticos que permiten analizar los datos de supervivencia.

- **Datos de supervivencia:** Los datos se crean si generan beneficio en estudiar el tiempo en que ocurre entre un evento inicial el cual determina la inclusión en un individuo y un evento final el cual se considera como falla. El tiempo entre los dos eventos se denomina como tiempo de falla o tiempo de muerte.
- Este tipo de análisis pertenecen a estudios longitudinales en donde se monitorea a un individuo a través del tiempo, este monitoreo inicia desde el ingreso al estudio de objeto hasta la ocurrencia de su falla (Sociedad colombiana de cardiologia y cirugía, 2017).
- **Observaciones censuradas:** en este tipo de análisis si algún individuo objeto de estudio abandona su monitoreo antes de su falla generará información parcial o censurada, estos datos faltantes es el mayor hincapié que genera esta técnica requiriendo mayor cantidad de procesos para lograr el funcionamiento correcto.

- **Arboles de clasificación y regresión**

Los arboles de clasificación y regresión (CART) es una técnica alternativa al análisis de discriminación, este tipo de técnica posee aprendizaje de árboles de decisión no paramétricas generando arboles de clasificación o regresión dependiendo de la variable dependiente, si esta es de carácter continua son arboles de regresión y si es de carácter cuantitativo son arboles de clasificación. La finalidad de esta técnica es generar un esquema de múltiples divisiones, anidadas en forma de árbol, con la finalidad de recorrer cada una de las ramas para alcanzar el resultado esperado.

Reglas para la generación del modelo

- Los valores de las variables objeto de estudio se deben seleccionar y tratar para obtener la mejor división con la finalidad de diferencias las características de cada grupo.
- Seleccionado los valores para el modelo se divide un nodo en dos, aplicando el mismo proceso a cada nodo subsiguiente.
- Las divisiones se interrumpen cuando el modelo detecta que no se cumplen algunas reglas determinadas por los valores.

Al final de cada rama creada termina en un nodo final, donde estos nodos poseen un conjunto de reglas únicas categorizando los grupos de clasificación.

Características

- Fácil interpretación del modelo.
- Poco tratamiento de datos.
- No requiere de normalización.
- Capacidad de manejo de datos de tipo numérico y categorizado.
- Validación por medio de pruebas estadísticas.
- Capacidad de predecir eficazmente con grandes volúmenes de información.

- **Análisis discriminante lineal**

El análisis discriminante lineal (LDA) es una técnica estadística multivariante que usa variables cualitativas cuya finalidad es determinar si existen diferencias entre los grupos objeto de estudio respecto a un conjunto de variables medidas sobre los mismos clasificando en función de las características de cada individuo, esta técnica estima la probabilidad de que una

observación, dado una determinada característica de los valores pertenezca a cada una de las clases de las variables objeto de estudio.

Reglas para la creación de un modelo LDA (Rodrigo, 2016)

- Poseer un conjunto de datos de entrenamiento en donde se conoce a que grupo pertenece.
- Calcular las probabilidades previas o proporción esperada de observaciones de cada grupo.
- Comprobar si la matriz de covarianzas es homogénea en todos los grupos.
- Calcular el resultado de la función discriminante para evaluar si clasifica correctamente los individuos en cada grupo.
- Utilizar validación cruzada para determinar las probabilidades de clasificaciones falsas.

Características

- Sencillez del modelo predictivo.
- Las probabilidades previas son fáciles de generar.
- Simplicidad de generar observaciones y clasificar individuos en grupos.

- **Metodología de aprendizaje computacional**

Estas metodologías permiten diseñar y desarrollar algoritmos capaces de aprender a identificar patrones complejos y tomar decisiones inteligentes basados en datos empíricos, estas se encuentran en el campo de machine learning. El aprendizaje computacional posee una variedad de métodos estadísticos avanzados para la clasificación predictiva, permitiendo que estas técnicas tengan un amplio campo de acción usándose en diagnósticos médicos, fraudes

bancarios, reconocimiento facial, análisis de mercados entre otros. Los algoritmos generados por esta metodología pretenden simular el pensamiento humano y aprender de datos históricos para predecir eventos fututos.

Unas de las técnicas más importantes de aprendizaje computacional dentro de la analítica de datos son:

- **Redes neuronales**

Este tipo de metodología de analítica de datos son técnicas de modelado no lineal idóneas para formar funciones complejas, las redes neuronales simulan la estructura y comportamiento del cerebro, usando los procesos de machine learning para buscar la solución o clasificación a un problema, debido a su naturaleza son predilectas para tareas como lo es el reconocimiento de patrones, modelos de clasificación y ayuda en la toma de decisiones usándose en áreas como finanzas, psicología, medicina, ingeniería y física.

En las redes neuronales existen cuatro aspectos que determinan las características de estos los cuales son:

- **Topología**

Es la organización y disposición de las neuronas en la red formando capas o agrupaciones de neuronas. Dentro de las redes neuronales existen parámetros fundamentales los cuales son: el número de capas, neuronas por capa, nivel de conectividad y tipo de conexión entre neuronas.

Las clasificaciones de las redes neuronales por su topología son:

- Redes monocapa

Este tipo de redes construyen conexiones laterales entre neuronas que corresponden a la única capa que constituye la red. Este tipo de red neuronal se usa en campos relacionados como la auto asociación la cual reconstruye los datos de entrada que se encuentran como incompletas.

- Redes multicapa

Las redes multicapas cuentan con neuronas agrupadas en diferentes niveles, contando con una cantidad amplia de capas permite que las neuronas sean de conexión hacia adelante FeedForward o de conexión hacia atrás FeedBack.

- **Mecanismo de aprendizaje**

Mecanismo que permite alterar los pesos de un modelo en respuesta a la información de entrada. Estos cambios se dan durante la etapa de aprendizaje realizando la reducción, modificación y creaciones de conexión entre neuronas.

Dentro de este mecanismo existen criterios para asignar reglas de aprendizaje las cuales determinan como modificar sus pesos y clasificar los individuos en grupos. Existen dos tipos de aprendizaje según la necesidad del modelo los cuales son:

- Aprendizaje supervisado

Modelo de aprendizaje que requiere de un conjunto de pares, entradas y salidas para poder encontrar un modelo que logre aprender la relación entre estos y predecir sucesos no alojados en los volúmenes de información, para lograr esta predicción necesita contar con la salida de los datos de entrenamiento.

- Aprendizaje no supervisado

Modelo de aprendizaje que requiere de un conjunto de entradas el cual busca la relación entre estos, generando grupos para encontrar un patrón que los una por medio de las características de cada individuo.

La diferencia entre estos dos tipos de aprendizajes es que el supervisado necesita de un agente externo que permita obtener la respuesta deseada en la red neuronal mientras que el no supervisado solo requiere del conjunto de datos para generar una respuesta.

- **Tipo de asociación entre información de entrada y salida**

Todo modelo de red neuronal requiere de información para entrenar, estos datos se les asigna un determinado peso para realizar las conexiones entre las neuronas. Se necesita constituir una relación entre la información de entrenamiento y la salida ofrecida por el modelo a esto se le conoce como memoria asociativa.

Los Tipos de asociación según la información de entrada y salida son los siguientes:

- Redes heteroasociativas

Asociación de correspondencia entre datos de entrada y salida, esta correspondencia entre la diferente información requiere de al menos dos capas una que permite captar y retener la información de entrada y la otra mantener la salida con la información asociada. Este tipo de asociación se usa en el mecanismo de aprendizaje supervisado.

- Redes autoasociativas

Asociación de autocorrelacion donde la red aprende de los datos de entrada, este tipo de asociación por lo general se implementa en modelos de una capa, utilizando el mecanismo de

aprendizaje no supervisado. Estas redes se usan para filtrar información, reconstruir, eliminar datos que generen distorsión permitiendo resolver problemas de optimización.

- **Representación de información de entrada y salida**

La representación de información de entrada y salida se clasifican en las siguientes:

- Redes continuas

En este tipo de red los datos de entrada y salida poseen valores reales continuos y normalizados por lo tanto su valor absoluto será menor que la unidad original, estos tipos de datos son de naturaleza analógica. Las funciones de activación de las neuronas serán de carácter lineal o sigmoidal.

- Redes discretas

En las redes discretas los datos de entrada y salida deben ser de tipo discreto por lo tanto los valores deben oscilar entre 0 y 1, generando una respuesta de tipo binario. Las funciones de activación para este tipo de red son de tipo escalón.

- Redes hibridas

En las redes hibridas los datos de entrada son de carácter continua y la salida de los datos en el modelo deben ser de carácter binaria. Las funciones de activación pueden ser de tipo escalón o lineal.

- **Naive Bayes**

Naive bayes es una técnica que usa algoritmos de aprendizaje automático, donde las variables de predicción son independientes entre sí por lo tanto la representación de una característica en

un conjunto de datos no está relacionada con la presencia de otra característica permitiendo la construcción de un modelo de manera fácil e intuitiva.

Las técnicas bayesianas en tareas de aprendizaje poseen las siguientes características:

- Por cada objeto de estudio que se encuentre en el modelo va a alterar la probabilidad de que la predicción sea correcta, por lo tanto, esta predicción de un evento será más grande o pequeña según el conjunto de datos mas no será desechada.
- Estos modelos son susceptibles al ruido o anomalía de datos presentes en los datos de entrenamiento y la probabilidad de tener predicciones erróneas al analizar datos incompletos (Luque, 2003).
- Los modelos bayesianos reconocen y tienen en cuenta la predicción del conocimiento a priori por lo tanto si no se cuenta con una cantidad moderada de datos de entrada fallara en las predicciones de eventos.

- **K-vecinos**

La técnica de los K-vecinos KNN (Nearest Neighbor) pertenece al aprendizaje automático y a métodos estadísticos de reconocimiento de patrones, este algoritmo ordena y clasifica los datos de entrada en un grupo que posea similares características, esto es posible calculando la distancia de un elemento nuevo a cada uno de los grupos existentes y establece las distancias de menor a mayor para lograr clasificar el nuevo elemento en un grupo. Este algoritmo es de aprendizaje supervisado por lo tanto requiere de un agente externo por lo tanto necesita que los datos de entrenamiento puedan permitir al modelo clasificar correctamente los datos en grupos (ANALYTICS, 2017).

Para que los modelos realizados con esta técnica posean un rendimiento aceptable deben tener en cuenta los siguientes aspectos:

- La distancia utilizada para localizar a los vecinos más cercanos.
- La regla de decisión para realizar una predicción correcta.
- La cantidad de vecinos para clasificar los datos nuevos.

Las características de esta técnica son:

- Es un algoritmo de carácter lazy lo que significa que durante el proceso de entrenamiento solo guarda instancias y clasifica cuando se está realizando las pruebas.
- No es paramétrico por lo tanto no hace suposiciones sobre cómo se encuentran los datos.
- No requiere de adaptación si se usan más de dos clases en el modelo.

9.2.3 Herramientas

Hoy en día la analítica de datos cuenta con una gran cantidad de herramientas disponibles que ayudan a la ejecución de modelos de análisis predictivo. Estas pueden necesitar de conocimientos avanzados por parte de los usuarios o usuarios con conocimientos específicos para realizar estas tareas. Estas herramientas permiten un gran nivel de personalización y según la cantidad de datos que se manejen afectara el rendimiento en los modelos.

Todos los modelos que se presentarán a continuación serán de software libre ya que permiten una independencia tecnológica, migración a diferentes entornos y posibilidad de mejorar las herramientas, además al ser de esta naturaleza permite agregar un valor añadido a las organizaciones sin necesidad de generar un gasto.

En este análisis se hablará sobre las diferentes herramientas en el ámbito de la analítica de datos dentro de las cuales se encuentran:

- **Python**

Python es un lenguaje de alto nivel creado por Guido Van Rossum en los países bajos siendo sucesor del lenguaje de programación ABC. Python hace hincapié en tener una sintaxis limpia favoreciendo un código legible favoreciendo la productividad y permitiendo una curva de aprendizaje suave, se trata de un lenguaje de carácter multiparadigma por lo tanto soporta programación orientada a objetos, imperativa y en mejor medida funcional. Este lenguaje posee un punto fuerte para el desarrollo de aplicaciones web ya que cuenta con frameworks como Diango y Flask, además permite y facilita la extracción de información de páginas web debido a técnicas de scraping y crawling por medio de la herramienta scrapy (Alvarez, 2018).

Características

- Lenguaje de propósito general

Al ser de propósito general significa que no está orientado a un fin en concreto permitiendo la creación de páginas webs, scripts o software para un sistema operativo.

- Lenguaje interpretado

Este tipo de lenguaje no requiere que el código sea preprocesado mediante un compilador debido a que, posee un intérprete el cual ejecuta el programa basándose en el código directamente.

- Multiplataforma

Al ser de carácter multiplataforma permite que este lenguaje sea usado en diferentes sistemas operativos y dispositivos.

- Tipado dinámico

Permite que las variables puedan tomar valores de diferentes tipos, por lo tanto, las variables definidas se adaptan a los que se escribe cuando se ejecuta el programa. Permitiendo la creación de software sin lidiar con particularidades propias del lenguaje.

- Versatilidad

Python como es un lenguaje de propósito general permitiendo integrar sus códigos con varias aplicaciones ampliando las funcionalidades de estos como lo son la lectura de archivos planos, conexión a bases de datos entre otras. Esto se debe a la gran cantidad de librerías que poseen.

Este lenguaje cuenta con un amplio soporte para el desarrollo rápido e interactivo de la analítica de datos debido a que, cuenta con un amplio abanico de librerías y entornos de desarrollo para cada una de las fases de esta temática. Teniendo esto en cuenta se puede resaltar las siguientes librerías para la creación de un modelo de analítica de datos.

- Scikits-Learn

Librería de código abierto enfocada para el análisis de datos y machine learning proporcionando una gran cantidad de algoritmos de aprendizaje supervisado y no supervisado. Scikit-Learn al estar construida sobre Scientific Python (SciPy) posee otras librerías como lo son Numpy, Pandas, Matplotlib y SymPy las cuales permiten el manejo de matrices, estructura y análisis de datos y trazado de datos (Moreno, s.f.).

Características

- Herramienta simple y eficaz para el uso de minería de datos y analítica de datos.
- Flexible y reutilizable en numerosos contextos.
- Robustez para un proyecto de aprendizaje automático de inicio a fin.
- Gran comunidad.

Esta librería al estar centrada en el aprendizaje automático cuenta con algoritmos de regresión, agrupación, arboles de decisión, redes neuronales, máquinas de soporte vectorial y naive bayes.

- TensorFlow

Librería de código abierto desarrollada por Google enfocada en el desarrollo de algoritmos de aprendizaje automática mediante la construcción y entrenamiento de redes neuronales mediante la computación numérica permitiendo descubrir y descifrar correlaciones y patrones en datos sujetos de estudio. Esta librería permite aprovechar el hardware de mejor manera ya que cuenta con diferentes versiones donde implementa el cálculo en una CPU o GPU según la necesidad del usuario (Buhigas, s.f.).

Características

- Manejo eficiente con expresiones matemáticas donde se incluyan matrices multidimensionales.
- Buen manejo y soporte en la construcción e implementación de redes neuronales.
- Alta escalabilidad de computación entre máquinas y grandes volúmenes de información.

- Keras

Librería de código abierto desarrollada por Francois Chollet diseñada para la construcción y manejo de redes neuronales y modelos de machine learning y Deep learning utilizando como backend otras librerías como lo son Theano, Toolkit y TensorFlow. Esta librería permite añadir más funcionalidades a una red neuronal implementando funciones objetivos, funciones de activación y optimizadores que mejoren el rendimiento de estas.

Características

- Modularidad al permitir usar librerías externas para ampliar la funcionalidad de un modelo.
- Entrenamiento de redes neuronales de manera rápida y eficaz.
- Compatibilidad para el desarrollo de aplicaciones móviles.

- Pandas

Librería de código abierto diseñada para la gestión y análisis de datos la cual posee diversas operaciones que permiten un fácil tratamiento de datos por medio de estructuras flexibles, esta librería permite la generación de arrays unidimensionales, dataframe las cuales son estructuras similares a una tabla de base de datos y estructuras con más de dos dimensiones para manejo de datos.

Características

- Facilidad de manejo de grandes volúmenes de información.
- Eficiencia en la creación e implementación de estructuras para el almacenamiento de información.

- Indexado multinivel permitiendo tener control sobre tablas y grandes volúmenes de información.
- Soporte de ficheros CSV.

Ventajas y Desventajas

Tabla 2. Ventajas y desventajas de Python

Ventajas	Desventajas
• Curva de aprendizaje corta. • Conexión y soporte amplio en bases de datos. • Buena gestión de errores. • Enfocado a la productividad. • Gran comunidad de desarrollo. • Diversidad de librerías para análisis de datos. • Excelente integración con aplicaciones web.	• Bajo rendimiento en multihilo. • Poca documentación. • Complejidad alto para desarrollo web.

- **R**

R es un lenguaje de programación multiplataforma creado por Ross Ihaka y Robert Gentleman siendo sucesor de dos lenguajes como lo son S y Scheme, este tiene un enfoque en el paradigma de orientada a objetos y hacia el análisis estadístico teniendo como prioridad la exploración, limpieza y análisis de grandes volúmenes de información permitiendo crear modelos predictivos.

Para la creación de estos modelos el lenguaje de programación cuenta con un entorno grafico denominado R Studio el cual facilita la codificación y ejecución de scripts además facilita la

visualización y tratamiento de datos que se encuentren cargados o que se generen en la ejecución del código.

Características

- Gran cantidad de herramientas estadísticas

Dentro del lenguaje de programación permite crear algoritmos de carácter lineal y no lineal, realizar test estadísticos y generar modelos para el tratamiento y análisis de datos para clasificar y agrupar individuos por características.

- Extensible

La gran cantidad de librerías pre construidas en el lenguaje de programación permiten un amplio estudio analítico y estadístico, además, permite a los desarrolladores crear sus propias funciones aumentando las funcionalidades de las librerías que posee.

- Funcionalidad

R permite manejar las funciones como vectores, por lo tanto, permite establecer las funciones a variables con el propósito de almacenar y devolver como resultado de otras funciones. Estas funciones se pueden manejar.

- Basado en memoria

En este lenguaje todos los objetos que se definan quedan programados en memoria por lo tanto es importante tener conocimientos sobre cómo gestionar la memoria para mejorar el código evitando ralentizar y consumir más recursos de los necesarios en el hardware (Datahack, 2018).

Este lenguaje es uno de los principales para la implementación de analítica de datos y minería de datos ya que permite utilizar los modelos estadísticos de mejor manera que en otros lenguajes, estos modelos se apoyan en librerías para crear e implementar modelos de analítica de datos predictivos dentro de estas se pueden destacar las siguientes.

- Caret

Librería de machine learning y análisis predictivo desarrollada en el 2016 que incluye una serie de funciones que permiten facilitar el uso de algoritmos complejos de clasificación de datos. Caret contiene herramientas que permiten el preprocesado de datos, selección de variables, separación y estimación de variables y ajustes a modelos a través del re muestreo.

Características

- Permite el uso de código unificado con la finalidad de usar reglas de clasificación distintas.
- Posee herramientas vitales para realizar problemas de clasificación y predicción.

- Tidyverse

Librería que permite al lenguaje de programación R obtener características para la analítica de datos, Tidyverse consta de seis librerías para lograr este objetivo las cuales son ggplot2 la cual permite la visualización y exploración de los datos por medio de gráficos, dplyr permite la transformación de datos siendo un equivalente a un lenguaje sql, readr permite la lectura e integración de datos, purr permite la vectorización de la información para la realización de algoritmos de regresión lineal, tidyr transforma cualquier dato a un formato legible para el

lenguaje de programación, tibble permite almacenar los datos en un dataset, stringr manipula, extrae y sustituye las cadenas de caracteres para el tratamiento de datos y forcast la cual permite manejar variables de tipo categórico.

- Prophet

Librería desarrollada por el equipo Core Data Science que permite la implementación de pronósticos de series temporales fundamentados en modelos aditivos por lo tanto poseen tendencias no lineales que se ajustan según un periodo de tiempo, esta librería es robusta ya que tolera los datos carentes y los cambios en las tendencias.

Características

- Pronostico Ajustable

Al ser una librería enfocada a la realización de pronósticos permite que los usuarios puedan modificar y ajustar los modelos de pronósticos permitiendo ampliar las posibilidades de estos, permitiendo usar parámetros interpretables para los usuarios con el fin de mejorar los pronósticos agregando conocimientos del machine learning.

- Automático

La librería permite generar modelos y reportes de pronósticos sin necesidad de tratar los datos generando menos esfuerzos ya que prophet es robusto a los valores incompletos o atípicos y al cambio brusco de los tiempos.

Ventajas y Desventajas

Tabla 3. Ventajas y desventajas de R

Ventajas	Desventajas
• Variedad de algoritmos estadísticos. • Calidad y variabilidad de gráficos. • Buena documentación. • Facilidad de algoritmos multihilos. • Cantidad amplia de librerías para visualización de datos.	• Mala gestión de errores. • Insuficientes librerías para los proyectos de machine learning. • Curva de aprendizaje alta. • Poca conexión y soporte a base de datos.

9.3 Etapa 3. Diseño del proceso operacional de analítica de datos.

Para la creación del diseño operacional de analítica de datos que procese información de la base de datos de Academusoft se requiere de cumplir tres fases para la correcta creación e implementación de un modelo analítico en el sistema de información Academusoft.

9.3.1 Fase 1

Definición del objetivo

El objetivo del modelo es realizar pronósticos de manera temprana sobre la aprobación o reprobación de materias que este cursando un estudiante desde el primer y segundo corte del periodo académico

Identificación de datos

Los datos que se requieren para el cumplimiento del objetivo son las notas académicas de los estudiantes que cursaron las materias que ofrece la Universidad de Pamplona donde se agrupen por materias para la respectiva predicción.

Selección de herramientas

El modelo planteado pertenece a la estrategia de analítica predictiva, por lo tanto, se dará uso de las diferentes metodologías estudiadas y como herramientas se usarán las siguientes:

- Python (Lenguaje de la programación).
- TensorFlow, keras, pandas y Scikits-Learn (Librerías de analítica de datos).
- Cx_Oracle (Conexión a la base de datos).
- Visual code (Editor de texto).
- Flask (Framework para creación de aplicación web).

Estas herramientas permiten la creación de un modelo de analítica de datos y facilita la conexión e integración con el sistema de información Academusoft.

Definición de métricas

Para la definición de las métricas es necesario tener en cuenta los siguientes criterios en el modelo predictivo.

- El modelo clasifica a los estudiantes en dos grupos los que aprobaron o los que no aprobaron una materia.

- Los datos de los estudiantes recolectados se dividen en dos conjuntos los cuales son de entrenamiento y prueba, estos conjuntos tendrán el 80% y el 20% de los datos respectivamente.
- Mediante los dos conjuntos de datos y la librería Scikits-Learn se pudo generar un reporte y una matriz de confusión de cada modelo que permite medir el porcentaje de precisión y clasificación respectivamente.

Para alcanzar el propósito del modelo se debe cumplir lo siguiente:

- El modelo debe contar una precisión superior al 94%
- El modelo debe tener un puntaje de clasificación superior al 92%

9.3.2 Fase 2

Recolección de datos

Los datos para el uso del modelo predictivo se encuentran en la base de datos de Academusoft donde se extraerán la siguiente información:

- Número de identificación de los estudiantes.
- Notas de trabajos y quices.
- Notas de parciales.
- Cantidad de fallas.
- Periodo cuando cursaron la materia.

Esta información extraída de la base de datos se transformará en un archivo CSV el cual se transformará en un dataframe por medio de la librería pandas con la finalidad de alimentar el modelo predictivo.

Tratamiento de datos

En el tratamiento se debe limpiar las imperfecciones que posean los datos recolectados previamente ya que esto puede generar un mal desempeño y rendimiento a un modelo predictivo.

Se determinará si el dataframe posee valores nulos, categóricos por medio de Python y la librería pandas para su respectivo tratamiento.

Luego del tratamiento de datos atípicos dentro del conjunto de datos, se observa que los datos son de carácter continúo permitiendo un mejor manejo de la información, para que el modelo obtenga un mayor desempeño se debe agregar más información al que se obtiene al combinar la información ya alojada en el dataframe. Se agregará la nota acumulada de trabajos y quices y parciales de cada corte y la nota final de cada estudiante en la materia.

Procesamiento de datos

Al obtener un dataframe ajustado para el modelo predictivo se deben normalizar los datos para permitir que los diversos algoritmos de Machine Learning puedan obtener una clasificación precisa para determinar si los estudiantes que cursan una materia en específico puedan saber si aprueban o no una materia de manera temprana.

9.3.3 Fase 3

Generación del modelo

En esta fase se creará el modelo predictivo donde se seleccionará la metodología según el desempeño y el cumplimiento de las métricas anteriormente establecidas. Estos modelos se crearán mediante el uso de las librerías Scikits-learn y TensorFlow.

Evaluación de los modelos

Para evaluar la correcta clasificación de los modelos se alimentará mediante dos dataframe uno de entrenamiento que serán los registros de los estudiantes que cursaron una materia desde el año 2010 hasta el 2018 y el conjunto de pruebas será los estudiantes que cursaron la materia en el año 2019.

Para medir su rendimiento se mostrará un reporte de su rendimiento y una matriz de confusión para determinar si clasifica correctamente a los estudiantes, estos indicaran la precisión del modelo, la clasificación, la puntuación global del modelo que indica su rendimiento.

- **Análisis discriminante lineal (LDA)**

Tabla 4. Reporte del modelo LDA

Clases	Precisión	Clasificación	F1-Score	Datos
Aprobado	96%	96%	96%	26
No Aprobado	93%	93%	93%	15
GENERAL	95%	95%	95%	41

Tabla 5. Matriz de confusión LDA

Clases	Clasificación Correcta	Clasificación Incorrecta
Aprobado	25	1
No Aprobado	14	1

- **K-vecinos**

Tabla 6. Reporte del modelo K-Vecinos

Clases	Precisión	Clasificación	F1-Score	Datos
Aprobado	80%	83%	81%	26
No Aprobado	55%	50%	52%	15
GENERAL	67%	66%	67%	41

Tabla 7. Matriz de confusión K-Vecinos

Clases	Clasificación Correcta	Clasificación Incorrecta
Aprobado	21	5
No Aprobado	9	6

- **Naive Bayes**

Tabla 8. Reporte del modelo Naive Bayes

Clases	Precisión	Clasificación	F1-Score	Datos
Aprobado	89%	96%	93%	26
No Aprobado	92%	80%	86%	15
GENERAL	91%	88%	89%	41

Tabla 9. Matriz de confusión Naive Bayes

Clases	Clasificación Correcta	Clasificación Incorrecta
Aprobado	25	1
No Aprobado	12	3

- **Arboles de clasificación y regresión (CART)**

Tabla 10. Reporte del modelo CART

Clases	Precisión	Clasificación	F1-Score	Datos
Aprobado	100%	82%	90%	26
No Aprobado	57%	100%	73%	15
GENERAL	79%	91%	81%	41

Tabla 11. Matriz de confusión CART

Clases	Clasificación Correcta	Clasificación Incorrecta
Aprobado	19	6
No Aprobado	15	0

- **Red neuronal**

Tabla 12. Reporte del modelo Red Neuronal

Clases	Precisión	Clasificación	F1-Score	Datos
Aprobado	93%	93%	93%	26
No Aprobado	85%	85%	85%	15
GENERAL	89%	89%	89%	41

Tabla 13. Matriz de confusión Red Neuronal

Clases	Clasificación Correcta	Clasificación Incorrecta
Aprobado	24	2
No Aprobado	13	2

Selección del modelo

Mediante las pruebas realizadas anteriormente y el desempeño de cada una de estas, el mejor modelo que clasifico a los estudiantes fue el de análisis discriminante lineal, Por lo tanto, este modelo está cumpliendo con la finalidad del objetivo y cumpliendo con las métricas requeridas.

Implementación del modelo

Luego de seleccionar el mejor modelo que cumple con las métricas las cuales son el porcentaje requerido de precisión y el puntaje de clasificación se podrá integrar este modelo al sistema de información Academusoft por medio de un entorno web usando Flask el cual es un

microframework que comunica HTML y Python el cual hará uso de las librerías scikits-learn, pandas y tensorflow para la generación del modelo y predicciones.

10. CONCLUSIONES

Mediante el estudio del arte acerca de la aplicación de analítica de datos en sistemas de información se puede establecer que hoy en día muchas organizaciones no están teniendo en cuenta la importancia de la información, por lo tanto, existen diversos estudios donde se analiza como diseñar e implementar la analítica de datos permitiendo mejorar el desempeño y toma de decisiones de estas además de ayudar a la población que está siendo objeto de estudio.

Tras el análisis de diversas estrategias, metodologías y herramientas dentro de la temática de analítica de datos, permite estar al tanto y saber cuál de todas estas son las más óptimas para la creación de un proceso operacional según la necesidad requerida, además de saber las características que poseen y como realizar un tratamiento correcto de la información.

La creación de un diseño de proceso operacional de analítica de datos permite y facilita la implementación de este, además permite explorar la mejor manera de usar la información alojada en la base de datos para lograr obtener beneficios como tener una visión general del sitio para mejorar la toma de decisiones.

11. RECOMENDACIONES

La implementación del proceso operacional de analítica de datos, con la finalidad de mejorar el rendimiento académico de los estudiantes para reducir la cantidad de materias perdidas por parte de estos en la Universidad de Pamplona.

Ampliar el modelo predictivo previamente diseñado para aumentar la utilidad de este en diversas áreas dentro de la Universidad de Pamplona con el fin de mejorar la toma de decisiones y comprensión de la situación actual que la rodea.

Crear un nuevo modelo prescriptivo basándose en el diseño del modelo predictivo anterior con el fin de que este realice predicciones según la información histórica y la toma de decisiones que realice el usuario para obtener una clasificación más exacta y personalizada.

12. ANEXOS

Ilustración 1. Diseño del modelo predictivo "Menú"

Ilustración 2. Resultado de precisión del modelo.

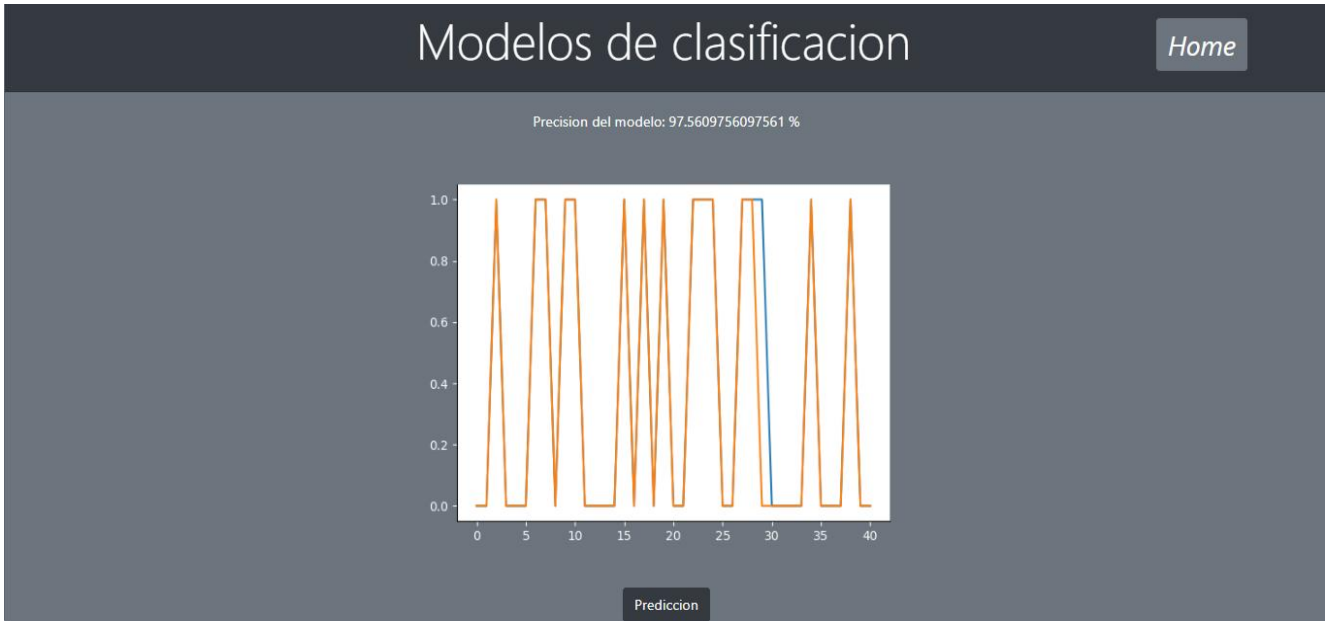


Ilustración 3. Búsqueda del estudiante para predicción.

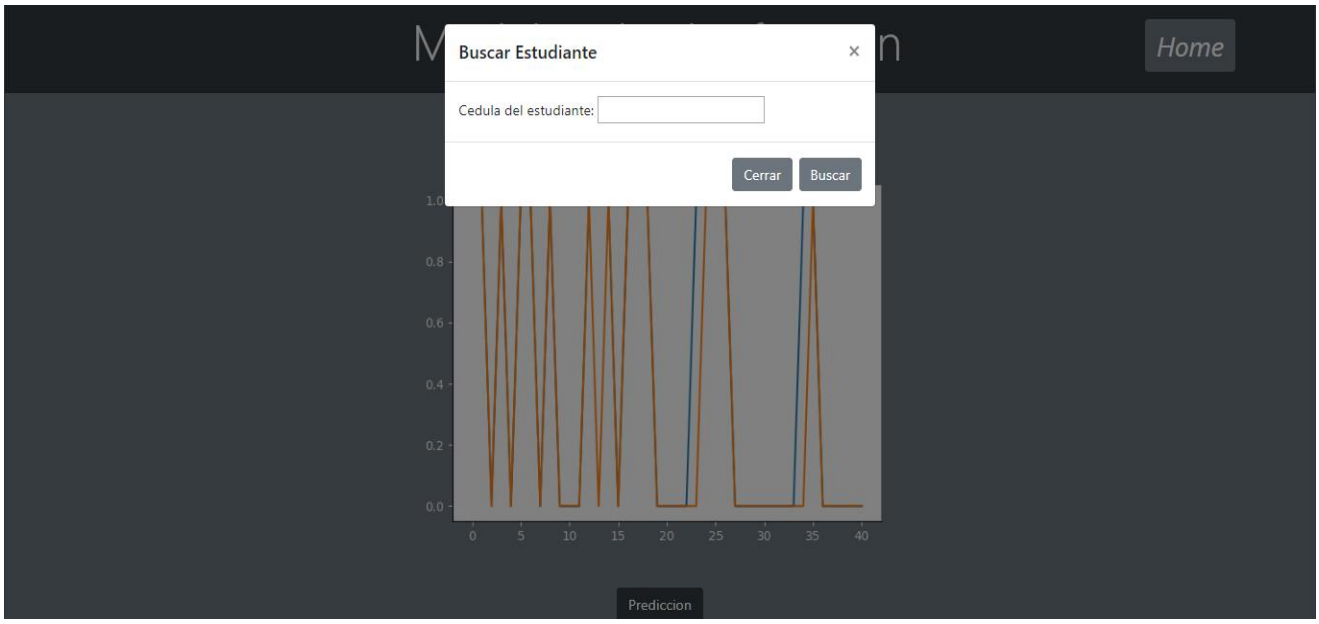
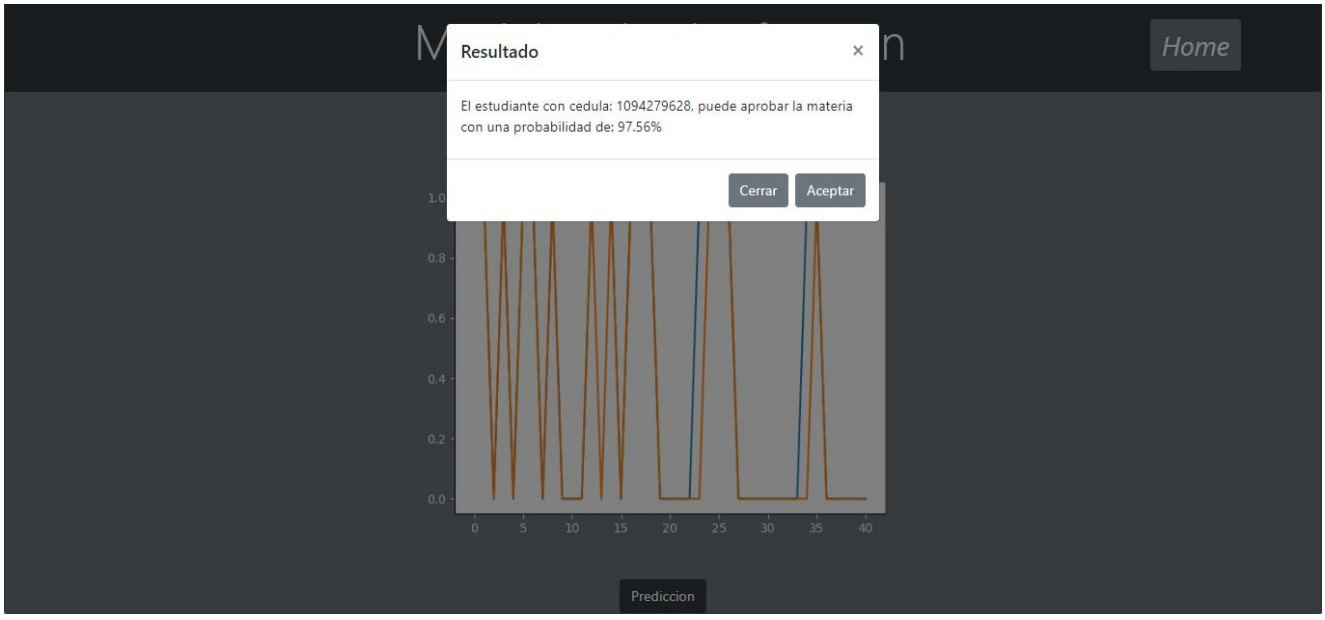


Ilustración 4. Resultado de la predicción.



13. REFERENCIAS BIBLIOGRÁFICAS

Adell F, G. A. (2013). *Big Data y los nuevos metodos de visualizacion de la informacion.*

Alvarez, M. (2018). *Lenguaje de Programacion*. Obtenido de https://lenguajesdeprogramacion.net/python/

ANALYTICS, A. T. (20 de Julio de 2017). *Analitica Web* . Obtenido de https://www.analiticaweb.es/algoritmo-knn-modelado-datos/

Andreu R, R. J. (1996). *Estrategia y Sistemas de Informacion* . McGraw Hill.

Arias, F. G. (2006). El Proyecto de Investigacion . Episteme.

Boell S.K, K. D. (2015). What is an Information System? *Hawaii International Conference on System Sciences*.

Buhigas, J. (s.f.). *Puentes Digitales* . Obtenido de https://puentesdigitales.com/2018/02/14/todo-lo-que-necesitas-saber-sobre-tensorflow-la-plataforma-para-inteligencia-artificial-de-google/

Datahack. (13 de Agosto de 2018). Obtenido de https://www.datahack.es/introduccion-lenguaje-programacion-r/

Education, A. (s.f.). *Fundamentos de Big Data*. Obtenido de www.arcitura.com

Luque, C. M. (14 de Mayo de 2003). *Clasificadores Bayesianos El Algoritmo Naive Bayes.*

Moreno, F. Y. (s.f.). *Ciencia&Datos*. Obtenido de https://fyaromo.com.co/2019/03/20/scikit-learn-el-estandar-de-oro-en-python-para-machine-learning/

Perez, J. L. (07 de Enero de 2013). *La Estadistica: Una Orquesta Hecha Instrumento*. Obtenido de https://estadisticaorquestainstrumento.wordpress.com/2013/01/07/tema-21-analisis-de-supervivencia/

Rodrigo, J. A. (Septiembre de 2016). *Rpubs*. Obtenido de https://rpubs.com/Joaquin_AR/233932

Roman, V. (s.f.). *Medium*. Obtenido de https://medium.com/datos-y-ciencia/machine-learning-supervisado-fundamentos-de-la-regresi%C3%B3n-lineal-bbcb07fe7fd

Routledge. (2013). *Critical questions for Big dara.* Cambridge: Danah Boyd & Kate Crawford. Obtenido de https://www.tandfonline.com/doi/pdf/10.1080/1369118X.2012.678878

SAS. (2017). *Software y Soluciones de Analitica* . Obtenido de https://www.sas.com/es_mx/insights/analytics/predictive-analytics.html

Sociedad colombiana de cardiologia y cirugía. (10 de 2017). Obtenido de http://scc.org.co/wp-content/uploads/2017/10/Supervivencia.pdf