Sistema de Recomendación Piloto para Programación de Clases del Departamento de Ingeniería de Sistemas en la Pontificia Universidad Javeriana.

Edward Andrés Cortés Peña
Pablo Miguel Núñez González
Sara Isabela Vergara Aguilar

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERIA
MAESTRÍA EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
BOGOTÁ, D.C.
2021

Sistema de Recomendación Piloto para Programación de Clases del Departamento de Ingeniería de Sistemas en la Pontificia Universidad Javeriana.

Autor:

Edward Andrés Cortez Peña

Pablo Miguel Núñez González

Sara Isabela Vergara Aguilar

MEMORIA DEL TRABAJO DE GRADO REALIZADO PARA CUMPLIR UNO DE LOS REQUISITOS PARA OPTAR AL TITULO DE

MAGÍSTER EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

Director

Juan Erasmo Gómez Morantes

Comité de Evaluación del Trabajo de Grado

Mariela Josefina Curiel Huérfano

Andrés Darío Moreno Barbosa

Página web del Trabajo de Grado

https://livejaverianaedu.sharepoint.com/sites/Ingsis/TGMISC/213005

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERIA
MAESTRÍA EN INGENIERIA DE SISTEMAS Y COMPUTACIÓN
BOGOTÁ, D.C.
Noviembre, 2021

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERIA
MAESTRÍA EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

Rector Magnífico

Jorge Humberto Peláez, S.J.

Decano Facultad de Ingeniería

Ingeniero Lope Hugo Barrero Solano

Director Maestría en Ingeniería de Sistemas y Computación

Ingeniera Angela Carrillo Ramos

Director Departamento de Ingeniería de Sistemas

Ingeniero Efraín Ortiz Pabón

Artículo 23 de la Resolución No. 1 de Junio de 1946

"La Universidad no se hace responsable de los conceptos emitidos por sus alumnos en sus proyectos de grado. Sólo velará porque no se publique nada contrario al dogma y la moral católica y porque no contengan ataques o polémicas puramente personales. Antes bien, que se vean en ellos el anhelo de buscar la verdad y la Justicia"

AGRADECIMIENTOS

Agradecemos a Dios por permitirnos tener esta oportunidad de afrontar el proceso de formación académica en esta disciplina, con la cual esperamos día a día crecer como mejores profesionales al servicio de las personas.

A nuestro director Juan Erasmo Gómez Morantes, por su apoyo y guía incondicional, durante todo este proceso, sin lugar a duda sus aportes y enseñanzas fueron valiosos para lograr con éxito el desarrollo y culminación de este proyecto.

A nuestro asesor Andrés Darío Moreno Barbosa, quien nos compartió todos sus conocimientos de este tema y nos guio durante todo este camino.

Al analista Nelson Jovanny Rodríguez Bohórquez y la directora de la Carrera de Ingeniería de Sistemas Alexandra Pomares Quimbaya, por toda su ayuda, tiempo y conocimientos impartidos para lograr el éxito de este proyecto.

A la ingeniera Angela Cristina Carrillo, directora de la Maestría en Ingeniería de Sistemas y Ciencias de la Computación, por acompañarnos y motivarnos durante en este proceso para culminar nuestra formación profesional.

A cada una de nuestras familias por impulsarnos a afrontar nuevos retos apoyándonos siempre de manera incondicional.

Contenido

LISTA DE TABLAS

LISTA DE FIGURAS

LISTA DE ECUACIONES

LISTA DE DOCUMENTOS ANEXOS

Los documentos anexos del trabajo de grado pueden ser encontrados en el repositorio asignado:

SRDIS _ESPECIFICACIÓN_REQ_02122020

SRDIS_CRISP-DM_02122020

ABSTRACT

The present degree work consists of developing a business intelligence tool for the Systems Engineering department of the Pontificia Universidad Javeriana which facilitates decision-making in the programming of the subjects that must be offered at the beginning of each semester. The proposed solution focuses on reducing the time it takes to carry out this entire process. Using the information provided by the university, data is collected from the last 5 years of this process; These data are subjected to a transformation and optimization process to implement predictive analytical models that make it possible to identify the subjects that must be programmed according to student demand each semester. The results will be presented on a control board displaying the subjects and the approximate number of places required for each one.

RESUMEN

El presente trabajo de grado consiste en desarrollar una herramienta de inteligencia de negocios para el Departamento de Ingeniería de Sistemas de la Pontificia Universidad Javeriana el cual facilite la toma de decisiones en la programación de las asignaturas que se deben ofertar al inicio de cada semestre. La solución planteada se enfoca en reducir el tiempo que conlleva realizar todo este proceso. Haciendo uso de la información proporcionada por la universidad se recopila una data procedente de los últimos 5 años; a estos datos se los somete a un proceso de transformación y optimización para poder implementar los modelos analíticos predictivos que permitan identificar las asignaturas que se deben programar según la demanda de estudiantes cada semestre. Los resultados se presentarán mediante un tablero de control visualizando las materias y el aproximado de cupos necesarios por cada uno.

RESUMEN EJECUTIVO

El área de admisiones y registro de la Pontificia Universidad Javeriana se encarga de brindar soporte técnico y funcional al proceso de programación de clases, catálogo de asignaturas, horarios y asignación de docentes en cada periodo académico. Paralelamente el Departamento de Ingeniería de Sistemas en el grado académico de pregrado se encarga de gestionar y registrar la programación de todos los cursos a ofertar en cada periodo para cumplir con el plan académico establecido. Posteriormente se solicita a la oficina de admisiones y registro la apertura de nuevos grupos o el ajuste en la capacidad de inscripción de las clases programadas. La planeación se maneja bajo un supuesto de estudiantes que se van a matricular, el cual puede variar de un periodo a otro, dependiendo también de si los estudiantes son o no de primer semestre, ya que a partir del segundo semestre en adelante las asignaturas empiezan a tener diferentes requisitos.

El programa de ingeniería de sistemas expresa que el proceso de programación de clases es una tarea compleja y demanda mucho tiempo, debido a variables o factores externos que impiden la planeación de manera adecuada. Uno de los factores evidenciados se debe a que actualmente se cuenta dos planes de estudio en la carrera de pregrado de ingeniería de sistemas; Con el fin de mejorar este proceso se realizaron dos encuestas en momentos diferentes del semestre académico, de las cuales no se obtuvo información suficiente ya que solo una pequeña parte de los estudiantes la respondieron; por lo que se descartó esta opción como una solución al problema definido.

Para dar solución a la problemática planteada anteriormente se propone implementar una herramienta basada en analítica de datos, la cual se debe ejecutar siguiendo la metodología de CRISP-DM. Realizando todo el proceso de limpieza y transformación de los datos obtenidos de las asignaturas programadas en ciclos anteriores, partiendo de estos nuevos datos homogéneos se desarrolla el modelo de predicción de asignaturas.

Se desarrollan varios modelos analíticos usando diferentes algoritmos de predicción como SVM y redes neuronales de los cuales se escogerá el modelo con los mejores parámetros de predicción. Esto con el fin de entregarle a la universidad una herramienta como un tablero de visualización que sea de ayuda en la toma de decisiones del proceso de programación de asignaturas cada inicio de semestre.

1. INTRODUCCIÓN

La oficina de admisiones y registro es el ente encargado de brindar soporte técnico y funcional al proceso de programación de clases y catálogo de asignaturas en cada periodo académico. Por su parte, el Departamento de Ingeniería de Sistemas en el grado académico de pregrado y posgrado se encarga de registrar la programación de clases conforme a las solicitudes del programa de Ingeniería de Sistemas, gestionando la información necesaria para programar los cursos a ofertar en cada periodo. Actualmente cada uno de los departamentos de la universidad tienen la libertad de programar las clases necesarias para cumplir los planes académicos establecidos. Posteriormente se solicita a la oficina de admisiones la apertura de nuevos grupos o el ajuste en la capacidad de inscripción de las clases programadas. Para la planeación se tiene un supuesto de estudiantes que se van a matricular y esto puede cambiar de un periodo a otro, dependiendo de si los estudiantes son o no de primer semestre, ya que a partir del segundo semestre en adelante las asignaturas empiezan a tener prerrequisitos y correquisitos.

El proceso de la programación de asignaturas enmarca el inicio de operaciones en cada semestre, lo que conlleva a validar una cantidad de información importante, generando una carga operativa relevante para los encargados, dado que si realiza una programación errada conllevaría a dejar a varios estudiantes sin cupo o a la no apertura de una materia de gran demanda lo que retrasaría académicamente a los estudiantes. Este proyecto se enfoca en automatizar este proceso a partir de la implementación de una solución de analítica de datos para que los empleados del Departamento de Ingeniería de Sistemas y el área de admisiones y registro cuenten con una herramienta de apoyo que les brinde una visión más completa y acertada de las asignaturas a programas cada inicio de semestre.

1.1. Oportunidad y problemática

El programa de Ingeniería de Sistemas expresa que el proceso de programación de clases es una tarea compleja, ya que hay variables o factores externos que impiden la planeación de manera adecuada; actualmente se cuenta dos (2) planes de estudio; uno antiguo que consta de 10 semestres y uno nuevo que consta de 8 semestres para pregrado. Tratando de eludir los inconvenientes generados, se pusieron en marcha una estrategias para la planeación de las asignaturas a ofertar, esta consta de realizar dos encuestas en momentos diferentes del semestre académico, la primera se realiza al finalizar un semestre con la intensión de conocer que materias planean inscribir los estudiantes en el siguiente semestre y con esto poder tener un estimado de cuantos estudiantes se van a inscribir por asignatura; pero esta primera encuesta no tuvo una buena acogida, ya que solo el 40% de los estudiantes la diligencio. La segunda encuesta se realiza después de la primera fecha de inscripción de materias para el nuevo semestre, esta con el fin de conocer cuántos estudiantes se quedaron sin cupo en las asignaturas abiertas, esta tampoco tuvo una buena respuesta de parte de los estudiantes. Con base en los datos recopilados de la primera encuesta se realiza una solicitud de creación preliminar de las asignaturas que cree necesarias el Departamento de Ingeniería de Sistemas y con los datos de la segunda encuesta se realizan las modificaciones y adiciones necesarias para culminar con la tarea de programación de asignaturas.

Por todo lo anterior, se presenta este trabajo de grado, cuyo propósito es presentar un sistema de recomendaciones piloto para que el Programa de Ingeniería de Sistemas en el grado académico de pregrado pueda planear de forma más asertiva la programación de asignaturas, franjas horarias, capacidad de estudiantes por asignatura en cada nuevo periodo académico.

2. DESCRIPCIÓN GENERAL

2.1. Objetivos

En esta sección se enuncia el objetivo general y sus objetivos específicos, los cuales se desarrollarán durante la ejecución del trabajo de grado y serán evaluados para dar cumplimiento y finalización a este proyecto.

2.1.1. Objetivo General

Desarrollar un sistema de recomendación piloto para optimizar el proceso de programación de asignaturas en el Departamento de Ingeniería de Sistemas en el grado académico de pregrado.

2.1.2. Objetivos Específicos:

- Establecer una arquitectura que permita alinear la tecnología disponible identificando las posibles mejoras en el proceso de programación de asignaturas
- Desarrollar un modelo analítico para un sistema de recomendaciones piloto utilizando la información de datos históricos proporcionados por el Departamento de Ingeniería de Sistemas para obtener una predicción de las asignaturas a inscribir.
- Implementar un tablero de control de visualización de los resultados del sistema de recomendaciones piloto, para que el Departamento de Ingeniería de Sistemas pueda tomar decisiones en la programación de las asignaturas.
- Validar el modelo y su visualización de las recomendaciones obtenidas con fin de presentar los resultados óptimos para el Departamento de Ingeniería de Sistemas.

2.2. Requisitos de manejo de información confidencial

De conformidad con los requisitos legales de manejo de la información confidencial Habeas Data, ley 1266 de 2008, donde se establecen las disposiciones legales para hacer uso, almacenamiento y transformación de la información suministrada por el Departamento de Ingeniería de Sistemas, se lleva a cabo un acuerdo de confidencialidad entre las dos partes interesadas, teniendo como premisa velar por la privacidad y uso adecuado de esta, dando como resultado un beneficio de mejora en uno de los procesos administrativos realizados por el área de admisiones y registro de la Pontifica Universidad Javeriana.

3. MARCO TEÓRICO

Durante el desarrollo de este trabajo de grado se identificaron los conceptos más importantes, los cuales están relacionados con la solución planteada a nuestro problema, a continuación, se explicarán varios de estos conceptos con el fin de contextualizar los términos más usados:

3.1. Inteligencia de negocios

La inteligencia de negocios es un término que se acuño en la década de 1990 cuando el analista de "Gartner Group" Howard Dresner [1], lo utilizo por primera vez para describir los conceptos y métodos que mejorarían la toma de decisiones a nivel empresarial mediante el uso de un sistema de apoyo basado en hechos. La inteligencia de negocios es un conjunto de metodologías, procesos, arquitecturas y tecnologías que limpian y transforman datos sin procesar en información útil e importante, que sirve para generar conocimiento, armar estrategias y tomar decisiones de manera táctica y eficaz [2].

3.2. ETL (Extract Transformation Load)

El proceso de ETL [3] es un método mediante el cual se extraen, transforman y cargan datos de fuentes internas y externas de organizaciones asociadas a un proceso determinado, esto con el din de organizarlas, transformarlas y estandarizarlas de tal forma que se puedan analizar y generar un entendimiento fácil [4]. Al tener esta información procesada se debe cargar en un almacén de datos, el cual debe ser de fácil acceso y control para poder aprovechar dicha información. Este proceso en el contexto del proyecto de grado resalta en la centralización de la información y su tratamiento aportar información relevante que permitirá un mejor entendimiento de los procesos de negocio.

3.2.1. Vista minable

Posterior a las etapas de limpieza y construcción de datos siguiendo los pasos de la metodología CRISP-DM, se tiene como resultado el conjunto de datos para analizar consolidados en una única tabla que contiene los atributos necesarios para aplicar algoritmos de minería de datos. Esta vista minable contiene los datos que según su nivel de relevancia se consideran los más importantes y que brindan la información concisa para su procesamiento. [5]

3.3. Modelos Analíticos

Los modelos analíticos son implementaciones de herramientas, algoritmos y técnicas de estadística y machine learning que permiten caracterizar los comportamientos o la probabilidad futura de la operación de una organización a partir del uso de su información histórica. Dicho proceso se enfoca en tomar como insumo los datos históricos de una compañía para realizar un ejercicio de analítica predictiva. La elección de la herramienta, algoritmo o técnica esa determinada por la forma de la información histórica con la que se cuenta o del comportamiento que se desea predecir [6]. Los datos históricos que se van a manejar en el desarrollo de este proyecto se encuentran como archivos de texto ".xlsx"

3.3.1. Regresión Lineal

Es un método que permite analizar la variabilidad de una variable determinada en función de la información que le proporcionan una o más variables, teniendo como objetivo la descripción clara y concisa de la relación presente entre las variables y los resultados de realizar una predicción de valores sobre la variable de respuesta en función de otras variables [7].

Teniendo en cuenta la posible relación entre dos variables, se establece la siguiente estructura de relación:

$$Y = \propto + \beta X + \varepsilon$$

Ecuación 1. Modelo de Regresión Lineal Simple

Teniendo en cuenta que la variable X es la variable independiente, ésta puede ser fija, cuyos valores se encuentre predeterminados o de manera aleatoria tomando al azar una muestra de los sujetos. La variable Y es la variable de respuesta y Ɛ es el error del modelo

3.3.2. SVM (Support Vector Maching)

Es un método de aprendizaje supervisado que genera funciones de mapeo de entrada-salida a partir de un conjunto de datos de entrenamiento etiquetados [8]. Este algoritmo es muy completo por lo que se puede usar tanto como para clasificación como para la regresión de varias clases; cuando los datos son linealmente separables haciendo uso de la regresión lineal obtenemos un algoritmo que separa los datos mediante un hiperplano. Sin embargo, pueden existir muchos hiperplanos que pueden separar los datos, el modelo de SVM escoge el hiperplano con mayor margen de los puntos de datos, teniendo como ventaja adicional que es un algoritmo que solo requiere el conocimiento de los vectores o puntos de soporte que se encuentran en el margen [9].

3.3.3. Redes neuronales

Es un modelo matemático computacional basado en redes neuronales biológicas, que consiste en interconectar un grupo de neuronas artificiales y procesarlas, este es un sistema adaptativo que puede cambiar su estructura en base de la información interna o externa que fluye por estas durante la fase de entrenamiento [10]. Las redes neuronales se componen de tres partes esenciales: la arquitectura o modelo, el algoritmo de aprendizaje y las funciones de activación. Estas redes están entrenadas para almacenar, reconocer, recuperar patrones, resolver problemas de optimización combinatoria, filtrar ruidos entre otras, este modelo se adapta perfecto cuando se trata de estimar funciones muestreadas que no se conocen. Sin embargo, las redes neuronales también son criticadas por no ser las más adecuadas para la minería de datos, ya que necesitan aprender las reglas de clasificación sobre el conjunto de datos de entrenamiento lo cual aumenta el tiempo de aprendizaje; el proceso de clasificación está enterrado en la estructura como en los pesos asignados a los enlaces entre los nodos lo que dificulta articular las reglas de clasificación. [11]. A pesar de las criticas anteriores, las redes neuronales proveen un error de clasificación más bajo que el modelo de árbol de decisiones.

3.4. Metodología

Para la implementación de la solución basada en analítica de datos, se debe ejecutar y desplegar una secuencia de métodos a través de los cuales se presentará la información obtenida por el Departamento de Ingeniería de Sistemas, esto con el fin de homogenizar todas las variables y proceder con el desarrollo e implementación de los algoritmos de análisis para hallar la diversas correlación entre las variables, escoger las más significativas para cada modelo y dar el valor agregado a la información resultante para mejorar el proceso de predicción de clases. Durante el desarrollo de este trabajo de grado se van a usar como marco de referencia la metodología CRISP-DM [12], adecuando cada una de sus fases a las necesidades se solución de la problemática de programación de asignaturas que presenta el Departamento de Ingeniería de Sistemas; dando como resultado entregable un tablero de visualización que sirva como herramienta en la toma de decisiones del proceso de programación de asignaturas cada inicio de semestre.

3.4.1. CRISP – DM

El modelo de referencia CRISP-DM [12] proporciona una descripción general del ciclo de vida para proyectos de minería de datos, contando en cada fase del proyecto con unas tareas y resultados respectivos. El ciclo de vida de un proyecto de minería de datos se descompone en seis fases (*Figura 1*), la secuencia demarcada por las fechas no es estricta, en esta secuencia el camino de las flechas indica las dependencias más importantes y su frecuencia; en un proyecto específico esta secuencia puede varias dependiendo de la prioridad de cada fase y el resultado obtenido

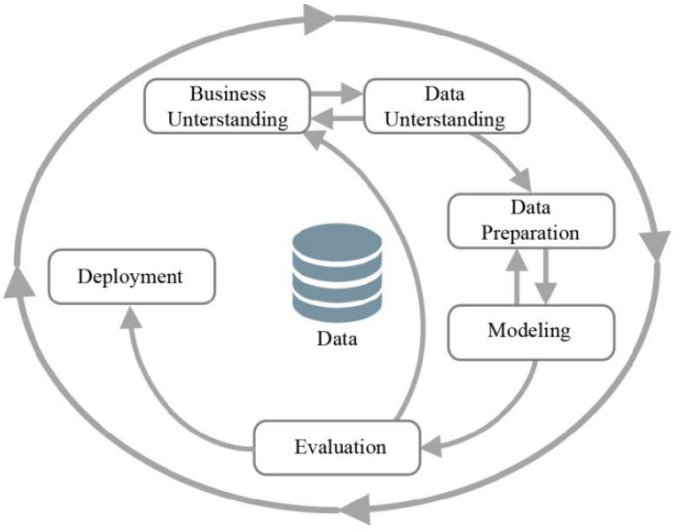


Figura 1. Modelo CRISP-DM

La figura anterior muestra las seis fases del modelo con sus interacciones, a continuación, se explicará brevemente en que consiste cada fase [13]

- Entendimiento del Negocio: La fase inicial se centra en el entendimiento del proyecto, sus objetivos y requerimientos desde una perspectiva de negocio, teniendo como resultado el diseño de un plan preliminar para cumplir con los objetivos
- Entendimiento de los datos: Esta fase inicia con la recolección de datos, su entendimiento e identificación de problemas de calidad de los datos, identificando los primeros los primeros subconjuntos y datos interesante de los datos para plantear posibles hipótesis sobre dicha información.
- Preparación de la información: Esta fase cubre todas las actividades de transformación y procesamiento de los datos inicial para la construcción del conjunto final de datos con el que se va a trabajar. las tareas de preparación de datos que se van a realizar son limpieza de datos, construcción de nuevos atributos y transformación de datos, apoyándonos en herramientas de modelado.
- Modelo: En esta fase se seleccionan y aplican diferentes técnicas de modelado, con las cuales se calibra y se evalúan los mejores parámetros de resultados teniendo en cuenta el problema de minería inicial identificado en los procesos anteriores
- Evaluación del Modelo: En esta etapa del proyecto se deben de tener construidos uno o más modelos de alta calidad para validar los resultados de dicho modelo y confirmar si se cumplió con el alcance de los objetivos planteados en el proceso analítico.
- Implementación: en esta etapa se deben presentar los resultados al usuario final de manera que este pueda hacer uso de este para apoyar la toma de decisiones o en un proceso gerencial.

Con la finalidad de poner en práctica todos estos procesos se obtiene un insumo del ejercicio analítico, con el cual se puede conocer el comportamiento, beneficio y posibles modificaciones que se pueden presentar con la transformación de las bases de datos

4. TRABAJOS RELACIONADOS

Teniendo en cuenta la importancia del buen manejo de la información en los procesos realizados diariamente dentro de las organizaciones y la importancia de los procesos analíticos de los cuales se obtiene una vista general del comportamiento de una organización, por lo que es necesario presentar una recopilación de diferentes modelos analíticos ejecutados en por otras organizaciones con contextos similares, para lograr un mejor entendimiento de dichas implementaciones y su valor aportado en la mejora de los procesos. A continuación, se presentan algunos de ellos.

Proyecto	Características	Descripción
Personalized course sequence recommendations [14]	Objetivo	Desarrollar un sistema automatizado de recomendación de la secuencia de cursos a inscribir que aprenda del desempeño del estudiante por sus cursos anteriores y sea capaz de recomendar de manera adaptativa las secuencias personalizadas a cada estudiante.

	Contexto	La mayoría de los estudiantes universitarios de EE. UU. no logran terminar sus estudios en cuatros que es el tiempo de cada carrera, lo que les implica una inversión económica mayor; esto se debe a varios factores como: créditos perdidos por transferencias entre universidades, desinformación, malos asesoramientos o inscripción tardía de cursos obligatorios los cuales son prerrequisitos
	Solución	Se proponen dos modelos de recomendaciones para la secuencia de inscripción de materias y lograr reducir el tiempo de graduación, el primer algoritmo es fase de inducción hacia atrás; el segundo es un algoritmo de aprendizaje para reducir el arrepentimiento de las selecciones previas. Por lo que la simulación de este método permite acortar el tiempo de graduación de los estudiantes.
Predicting Grades [15]	Objetivo	Desarrollar un algoritmo que predice la nota final de cada estudiante basado en el historial de desempeño individual en un curso y proponer las medidas correctivas en los estudiantes con desempeño deficiente
	Contexto	La educación se ha transformado y el conocimiento es cada vez más accesible para todos, sin embargo, esto ha llevado a que los instructores y asistentes de enseñanzas no den abasto con las clases que tienen un gran volumen de estudiantes presenciales y más con los cursos masivos abiertos en línea (MOOC) limitándoles el seguimiento del desempeño de cada estudiante, lo que conlleva a que los estudiantes fracasen y reprueben.
	Solución	Desarrollo de un algoritmo que predice de manera oportuna y personalidad las calificaciones finales de los estudiantes, mediante la puntuación de evaluaciones de desempeño tempranas, tareas y exámenes. Demostrando que los exámenes dan mejores precisiones que las tareas asignadas al estudiante.
Análisis de rendimiento académico en los estudiantes de Informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos [16]	Objetivo	Aplicar técnicas de minería de datos para estudiar la influencia de parámetros socioeconómicos y datos personales, sobre el rendimiento académico de un estudiante de primer semestre, teniendo solamente la información aportada por el estudiante en el momento de su matricula
	Contexto	Debido a la creciente ola académica, el rendimiento de cada estudiante se encuentra influenciado por factores externos que afectan directamente la estancia en la universidad, prolongando el tiempo de duración de cada carrera
	Solución	Desarrollar un algoritmo predictivo basado en técnicas de minería de datos, por lo que se hará uso de los modelos predictivos de árboles de decisión y regresión multivariable

Construcción e implementación de un modelo para predecir el rendimiento académico de estudiantes universitarios mediante el algoritmo Naïve Bayes [17]	Objetivo	Construir y evaluar la aplicación de un modelo predictivo sobre el rendimiento académico de estudiantes universitarios por medio del algoritmo Naïve Bayes con el fin de utilizar estas predicciones para identificar estudiantes vulnerables a reprobar y diseñar estrategias de preventivas.
	Contexto	Teniendo en cuenta el problema que se presenta en una institución educativa frente al rendimiento insuficiente de los estudiantes, lo que conlleva a la perdida de una o varias materias, limitando su avance en su vida académica. La reprobación de materias puede ocasionar la deserción estudiantil por lo que las instituciones educativas deben crear estrategias para evitar el fracaso de los estudiantes y evitar su deserción.
	Solución	La información recopilada de los estudiantes sobre la aprobación y reprobación de las materias se utiliza para la construcción de un modelo predictivo sobre el rendimiento académico haciendo uso del algoritmo de Naïve Bayes. Los resultados obtenidos muestran como a partir de los datos académicos y algunos datos personales el algoritmo predice el porcentaje de rendimiento académico del estudiante al final del curso
Characterizing Curriculum Prerequisite Networks by a Student Flow Approach [18]	Objetivo	Construir una herramienta de software para plantear el plan de estudios determinado por los requisitos previos haciendo uso de un modelado de eventos discretos, estimando el tiempo presupuestado hasta graduarse y cuales curso son los que impactan este tiempo, esta herramienta es de gran apoyo para las áreas administrativas.
	Contexto	Las instituciones educativas tienen como objetivo aumentar la tasa de graduados en el tiempo establecido o menos, formando buenos especialistas, por lo que el plan de estudios debe estar regulado y cumplir con ciertos requisitos, especificados por un orden secuencial de materias cumpliendo con los objetivos de aprendizaje que se espera para los estudiantes; impactando directamente en las tasas de deserción estudiantil y el tiempo que estos se gastan en graduarse.
	Solución	Desarrollar una aplicación Web que proporcione la información sobre la red de clases basadas en sus prerrequisitos, calculando factores de aplazamiento, bloqueos y destaca los cursos más críticos.

Tabla 1. Comparación Trabajos Relacionados

Teniendo en cuenta las consultas realizadas en el estado del arte anterior, las cuales nos sirvieron de guía para iniciar el desarrollo del trabajo de grado podemos concluir de estos que es de vital importancia para su desarrollo el tener acceso a datos personales de los estudiantes y en algunos casos de sus tutores o padres para conocer variables económicas que en nuestro caso no le aportan valor a este desarrollo, además no se contaba con este tipo de información debido a políticas de tratamiento de datos y privacidad de la información personal de los estudiantes.

De la investigación anterior cabe resaltar el documento titulado "*Characterizing Curriculum Prerequisite Networks by a Student Flow Approach* [18]" el cual introduce el concepto de factor de aplazamiento que hace referencia al momento de que un estudiante pierde una materia y los efectos que conlleva, este análisis lo aplicamos a nuestro trabajo ya que parte de la información suministrada por la carrera de ingeniería de sistemas hace referencia a la deserción o perdida de asignaturas.

La caracterización de los requisitos previos que se necesitan para cursar determinadas asignaturas y el flujo de estudiantes que estas presentan son variables que se usan para realizar el modelado, además que aporta información relevante sobre el peso de las asignaturas que influenciaran de forma directa el tiempo de graduación, finalmente esta solución es presentada como una aplicación web que muestra la red de requisitos previos para cumplir con el plan de estudio.

Esta herramienta sin lugar a duda fue la que más influyo y aporto en el desarrollo de este trabajo de grado por la proximidad del tipo de datos usados, la importancia de priorizar las asignaturas y resaltar el efecto de la pérdida o deserción de las asignaturas por parte de los estudiantes; aunque los objetivos sean diferentes ya que esta predicción se hace para beneficio de los estudiantes y nosotros con el fin de beneficiar a la carrera de Ingeniería de Sistemas.

5. DESARROLLO DEL PROYECTO

5.1. Requisitos

El desarrollo de este trabajo de grado propone una solución que facilite la toma de decisiones para la asignación de materias cada inicio de semestre para el programa de Ingeniería de Sistemas en pregrado, lo que se realizó un levantamiento de requisitos inicial partiendo de la información recopilada a través de entrevistas con el personal encargado de esta tarea; En esta sección se exponen los requisitos de negocio, funcionales y no funcionales. Para un mayor detalle ver el anexo SRDIS_ESPECIFICACIÓN_REQ_02122020

5.1.1. Requisitos de negocio:

En la siguiente *Tabla 2*, se relacionan los requisitos del negocio que permiten identificar las principales necesidades del Departamento de Ingeniería de Sistemas, los cuales se asocian a un modelo o proceso para dar cumplimiento al requisito

Identificador	Nombre del Requisito	Prioridad
RN_001	Realizar la recolección y centralización de las fuentes de datos en un solo lugar para conocer el tamaño y la calidad de datos entregados por el Departamento de Ingeniería de Sistemas	Alta
RN_002	Mecanismo que permita el cargué de datos de periodos históricos extraídos de las fuentes entregadas por los interesados (de las cuales se realizan copias de seguridad con los originales) de los cuales se analizara	Alta
RN_003	Análisis de la información histórica para conocer la cantidad de inscritos en periodos anteriores utilizando algoritmos de analítica con la finalidad de obtener las asignaturas recomendadas para el periodo siguiente y así poder mejorar los tiempos de gestión de solicitud de programación de asignaturas del programa al departamento.	Alto
RN_004	Mecanismo que evalué la cantidad de estudiantes que inscribieron asignaturas durante los periodos anteriores para pregrado, con la finalidad de obtener el mejor pronóstico de estudiantes por asignatura y permitir una mejor gestión y solicitud de aulas para cada asignatura al departamento de registro y admisiones.	Alto
RN_005	Mecanismo de visualización de la información de los resultados de asignaturas recomendadas para programación de estudiantes, y cantidad de estudiantes con la finalidad de que el programa de ingeniería de sistemas pueda tomar decisiones y gestionar con mayor asertividad su periodo académico.	Alto

Tabla 2. Requisitos del Negocio

5.1.2. Requisitos funcionales:

En la *Tabla 3* se describe cada requisito y su nivel prioridad, evaluados según las necesidades que se identificaron durante las entrevistas con el encargado del proceso.

Identificador	Nombre del Requisito	Prioridad
RF_001	El sistema debe permitir cargar los datos históricos de todos los periodos con el formato definido por ambas partes.	Alto
RF_002	El sistema debe ejecutar automáticamente el modelo de predicción sobre los históricos cargados.	Alto
RF_003	El sistema debe permitir predecir la cantidad de estudiantes	Alto

	según la asignatura, teniendo en cuenta que los niveles de confianza en la predicción pueden variar por la cantidad de datos cargados.	
RF_004	El sistema debe visualizar las predicciones por medio de un tablero de control que muestra el detalle de los datos históricos, el retiro de las asignaturas por parte de los estudiantes y las predicciones realizadas.	Medio

Tabla 3. Requisitos Funcionales.

5.1.3. Requisitos No Funcionales

En la *Tabla 4*. Se describen los requisitos no funcionales los cuales describen la visualización de la información final y necesaria para la toma de decisiones.

Identificador	Nombre del Requisito	Prioridad
RNF_001	Se define el formato de datos cargados el cual debe mantenerse con el tiempo, de requerir ajustes en la plantilla principal de cargue datos es necesario realizar el cambio en el sistema de recomendaciones para obtener el resultado adecuado.	Alta
RNF_002	La cantidad de datos mínima debe ser igual o superior a 1500 datos debido a que puede afectar el resultado de sistema de recomendaciones	Alto
RNF_003	Los ajustes después de la entrega de plantilla de cargue, modelo y visualización no corresponden no se asumirían como parte del proyecto sino a mejoras a futuro o desarrollos para otro proyecto de grado.	Bajo
RNF_004	Documentación que describa la información y fuentes involucradas en el proceso	Medio

Tabla 4. Requisitos No Funcionales

5.2. Conocimiento del Negocio

La Pontificia Universidad Javeriana, con el fin de siempre brindar la mejor calidad en la educación que ofrece a sus alumnos prepara, organiza y crea un pensum educativo para cara carrera, el cual está enfocado en que el estudiante construya sus bases de conocimiento y secuencialmente estas aumenten durante el tiempo estipulado de duración de cada carrera. Para este trabajo de grado nos enfocamos específicamente en el proceso de programación de clases del Departamento de Ingeniería de Sistemas el cual actualmente cuenta con dos planes de estudio debido a reestructuraciones. Teniendo en cuenta la situación actual del departamento la programación de clases actualmente es un proceso complejo que depende de muchos factores y requiere de un gasto de tiempo importante.

El área admisiones y registro espera por los datos suministrados por el Departamento de Ingeniería de Sistemas para cargar las materias que se van a ofertar y posteriormente realizar la creación de cupos por asignatura y la respectiva selección y asignación de docentes. Esta es la manera actual que se usa para realizar este proceso.

Para entender más a fondo esta problemática se entrevistó a la directora de la carrera de Ingeniería de Sistemas y su analista los cuales nos ayudaron a mejorar el entendimiento del problema y proponer una mejor solución, inicialmente se tuvo una reunión con la directora, la cual nos ayudó a tener una visión más amplia de este problema y semanalmente nos reunimos con el analista para presentarle los avances del modelo y validar el uso de las variables dependientes; esto con el fin de que el modelo final tuviera el menor índice de error posible; durante cada sesión se acordaban nuevos cambios y de ser necesario el analista nos proporcionaba nuevas fuentes de datos, manteniendo siempre la confidencialidad de datos personales de los estudiantes. Las últimas reuniones con el analista se acordaron para validar la forma de presentar los datos resultantes a través de tableros interactivos y explicarle su funcionamiento para que ellos puedan tomar las mejores decisiones respecto a la programación de las clases; después de esta presentación el analista nos da su retroalimentación y nos expone varios puntos en la presentación de las asignaturas canceladas que es importante para ellos en el momento de la toma de decisiones. Teniendo en cuenta esta retroalimentación se realizan las modificaciones solicitadas y se presenta el tablero de visualización con las respectivas modificaciones a la directora de carrera para su aprobación.

5.3. Arquitectura

Para dar inicio con el planteamiento de la solución, se realizó un diseño general de la arquitectura, en donde se evidencian las funcionalidades y responsabilidades de cada área involucrada en el proceso.

5.3.1. BPM (Business Process Model)

De acuerdo con las necesidades encontradas en el Departamento de Ingeniería de Sistemas se plantea el mapa de procesos por el cual se debe atravesar para cumplir con el objetivo inicial de facilitar la elección de las asignaturas más relevantes en cada semestre, esto con el fin de automatizar el proceso y priorizar las materias más críticas para culminar el programa.

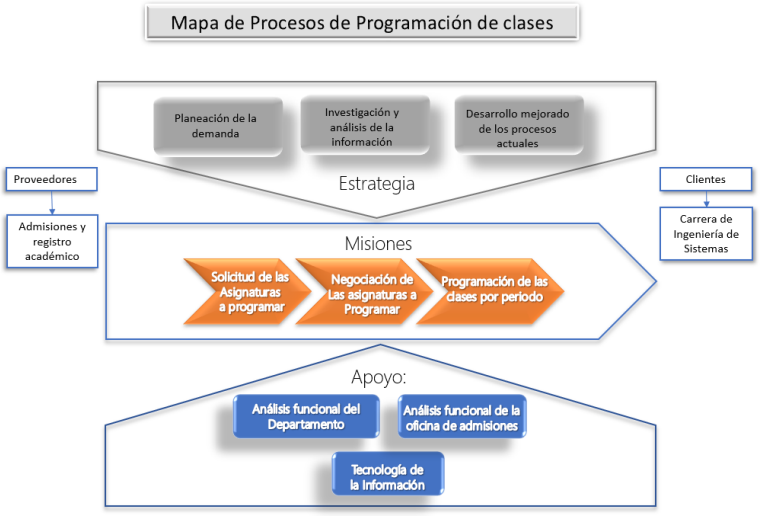


Figura 2. Mapa de procesos BPM – Elaboración propia

5.3.2. Arquitectura As Is

El modelo de arquitectura AS-IS está basado en el levantamiento de datos y recopilación de información que se realizó con los actores involucrados en el proceso los cuales describieron sus procesos diarios y el tiempo que tardan en cada una de las actividades que se deben realizar; todas las actividades descritas se utilizaron para plantear el modelo del proceso actual que se trabajando para realizar la programación de las asignaturas al inicio de cada semestre, a continuación se presenta el modelo AS-IS el cual describe secuencialmente los procesos que deben las tres áreas involucradas, las cuales son: la dirección de carrera, los departamentos y el área de admisiones y registro.

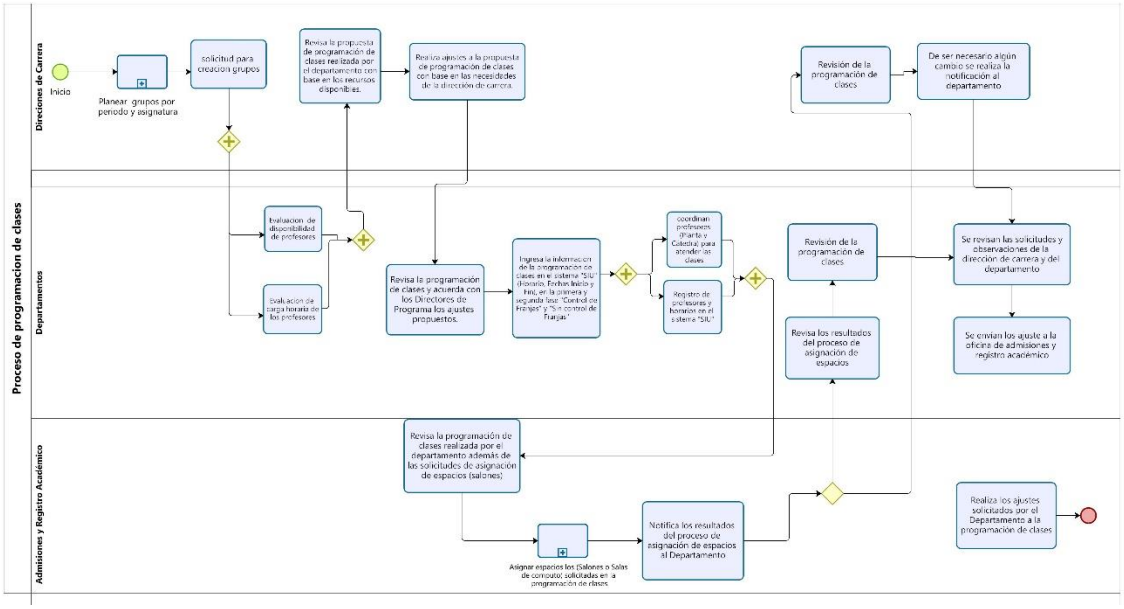


Figura 3. Modelo AS-IS, elaboración propia

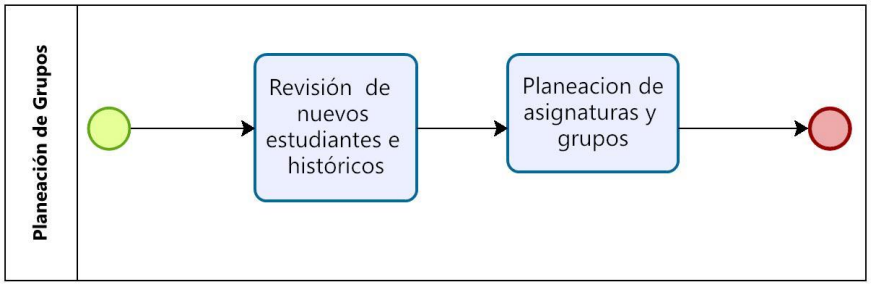


Figura 4. Descripción del módulo de Planeación de Grupos

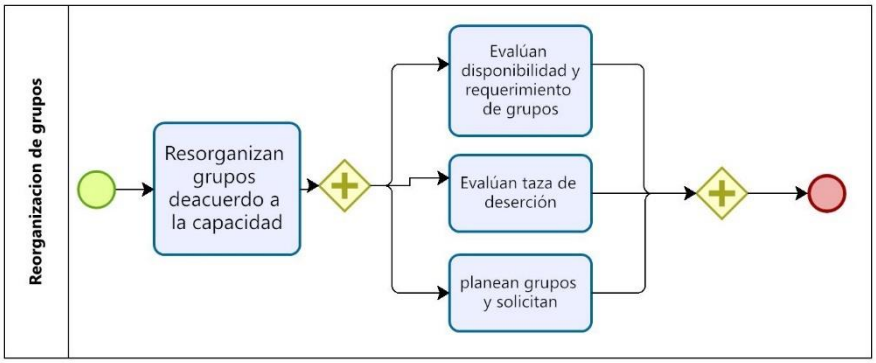


Figura 5. Descripción del módulo de Reorganización de grupos

5.3.3. Arquitectura To Be

El modelo de arquitectura TO-BE, parte del diagnóstico anterior del cual se detectan diferentes falencias, por lo que se plantea una nueva solución que mejore el sistema y añada valor al proceso realizado logrando mayor eficiencia en este proceso, estos cambios y mejoras planteadas, se validan con el cliente final, el cual ratifica que el nuevo sistema se encuentra acorde con su necesidad y da una correcta solución a su problema

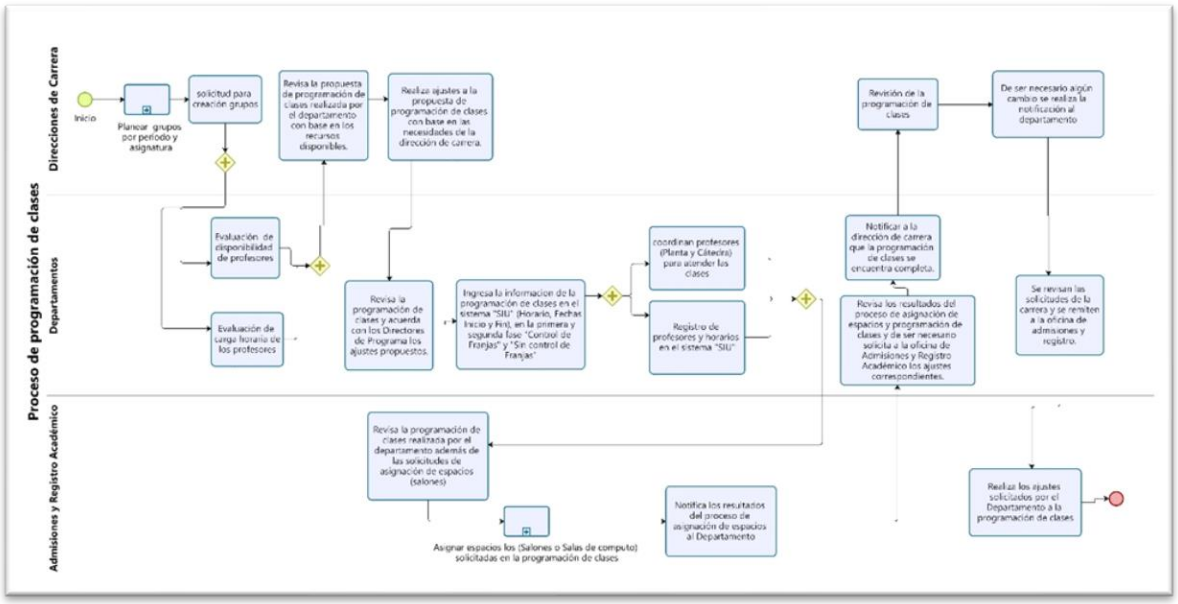


Figura 6. Modelo TO-BE, elaboración propia

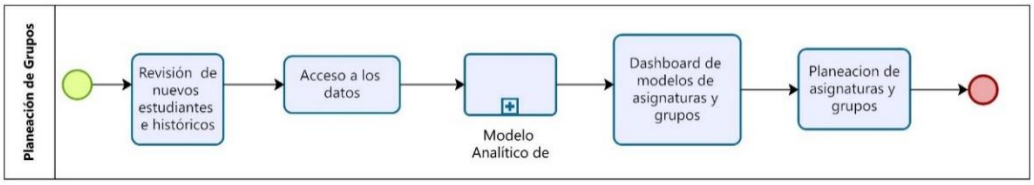


Figura 7. Descripción del módulo de Planeación de Grupos – Modificado

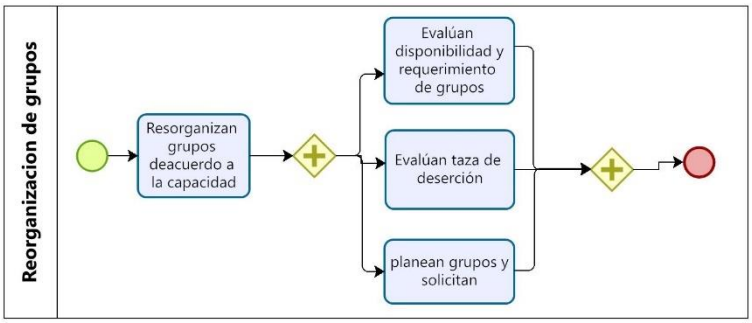


Figura 8. Descripción del módulo de Reorganización de grupos – Modificado

5.4. Origen de datos

A partir de la información brindada por el Departamento de Ingeniería de Sistemas se extrae la siguiente información de los datos, la cual se consolida en la siguiente tabla:

Datos	Donde se Encuentra	Formato	Confiabilidad	Frecuencia de Actualización	Veracidad	Fuente	Exactitud
Solicitud para Creación de Grupos	Archivo de la Dirección de programa envía a la dirección de Departamento	Excel	Media	Semestral	Media		Baja
Encuestas Planeación Asignaturas	En un repositorio de SharePoint	Excel	Medio	Semestral	Media	Se envía por correo Electrónico	Medio
Planeación Profesores	*Departamento*						
Base Matriculados	En un repositorio de SharePoint	Excel	Alta	Semestral	Alta	Sistema de Información	Alta
Solicitudes Servicios Externos (Carrera)	En un repositorio de SharePoint	Excel	Alta	Semestral	Alta	Sistema de Información	Alta
Solicitudes de servicios Externos (Departamento)	*Departamento*						
Asignaturas Por Programar por la Carrera	En un repositorio de SharePoint	Excel	Alta	Semestral	Alta	Documentos Carrera	Alta
Encuesta Previa a la cita de inscripción	Encuesta realizada en LimeSurvey	LimeSurvey	Media	N/A	Media/Baja	N/A	Media/Baja
Plan de estudios Vigente	Página de la universidad	Web	Alta	N/A	Alta	Carrera de Ing. de Sistemas	Alta
Plan de Estudios Antiguo	Página de la universidad	Web	Alta	N/A	Alta	Carrera de Ing. de Sistemas	Alta
Horario Programación de clases	Plataforma PeopleSoft (Sistema de información Universitario)	Excel	Alta	Semestral	Alta	Sistema de información Universitario SIU	Alta

Tabla 5. Origen de Datos

5.5. Modelos Iniciales

En el desarrollo del proyecto se consideraron en primer lugar las necesidades del negocio y como la tecnología se podía alinear con ellas, es decir se buscó integrar en el desarrollo un modelo que respondiera de forma efectiva a cada uno de los objetivos planteados, debido a que este proyecto involucra como componente principal la supervisión de una variable Y (Número total de inscritos para una asignatura en un periodo específico) con base en características X, se consideró realizar un tratamiento previo de las variables y una selección posterior de las que mejor explican el comportamiento de la variable objetivo, para revisar los detalles de este procedimiento ver el anexo SRDIS _CRISP-DM_02122020, es necesario señalar que para el primer conjunto de ensayos que se revisara a continuación se usaron todas las variables disponibles, posteriormente se eliminan algunas de ellas y se crean otras.

Conjunto Inicial de Variables:

- Grado: Identificador para postgrado y pregrado (PREG/GRAD)
- Catálogo: Descriptor del número de Catálogo
- Sesión: Calendario usado para programar la clase (18 semanas, 20 semanas, 22 semanas)
- ID Curso: Identificador de la asignatura
- Sección: Descriptor del consecutivo del número de clases por asignatura
- Nº Clase: Número de la clase
- ID Instalación: Salón asignado a la clase
- F Inicial: Fecha inicial de la clase
- Fecha Final: Fecha final de la clase
- Ciclo: Periodo académico al cual corresponde la clase
- Estado Clase: Estado de la clase (Sección cancelada, Activa)
- Tipo Clase: Hace referencia a la clasificación de una clase (Sección Inscripción o Sección sin Inscripción)
- Componente: Define si una clase es de alguno de los siguientes tipos (Teórico, Práctico o Teórico-Práctico)
- Gp Acad: Facultad o Instituto a la que pertenece la asignatura
- Modalidad: Modalidad de asociación del profesor a la clase
- Nro Horas: Número de horas de la clase a la semana
- Número Semanas: Número de semanas de clase
- Horas Semanales: Número de horas de clase al semestre
- Lunes: “Y” sí la clase se toma el lunes de lo contrario “N”
- Martes: “Y” sí la clase se toma el martes de lo contrario “N”
- Miércoles: “Y” sí la clase se toma el miércoles de lo contrario “N”
- Jueves: “Y” sí la clase se toma el jueves de lo contrario “N”
- Viernes: “Y” sí la clase se toma el viernes de lo contrario “N”
- Sábado: “Y” sí la clase se toma el sábado de lo contrario “N”
- Domingo: “Y” sí la clase se toma el domingo de lo contrario “N”

- Acceso: El nivel de acceso de la asociación del profesor
- Fecha Novedad: Fecha de asociación de un profesor en la plataforma PeopleSoft
- Estado de Asociación: Estado de la asociación del profesor (Activo o Inactivo)
- ***Total de Inscripciones: Número total de estudiantes inscritos para una clase específica, es importante señalar que esta es la variable dependiente.***
- Nº Modelo Clase: Consecutivo del número de modelos de reunión de una clase
- Rol Profesor: Tipo de asociación (Gestor o Profesor)

El conjunto de datos se dividió en dos: entrenamiento y pruebas; inicialmente se dispuso del 80% de los datos para entrenar el modelo y el 20% restante para pruebas; pero se decidió caracterizar los datos en función de su uso, es decir agrupar los datos por periodo académico, desde el 1510 hasta el 2030 para entrenar el modelo, y 2110 para realizar pruebas. Es necesario señalar que esta es solamente una de las pruebas realizadas para encontrar el mejor modelo, más delante se realizan nuevos experimentos y se dispone del uso de validación cruzada para mejorar la precisión.

- Métrica de evaluación de los modelos (RMSE):

Con el fin de establecer cuál era el modelo que tenía mejor desempeño, se decidió usar como métrica la raíz cuadrada del error cuadrático medio ***RMSE***, este indicador compara los valores predichos vs. los valores reales. Representa la variación de las predicciones obtenidas y que tan cerca están los datos reales de los valores predichos, a continuación, describimos la fórmula implementada al evaluar los modelos.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

Ecuación 2. Raíz cuadrada del error cuadrático medio

Los métodos implementados son:

Support Vector Machine: Es una técnica clásica de aprendizaje de máquina principalmente diseñada para resolver problemas de clasificación, pero adaptada también para responder de forma efectiva en problemas de regresión, sin embargo, es matemáticamente compleja y de alto costo computacional.

Existen varios métodos para implementar este tipo de modelos, aparte de ajustar los hiperparametros es necesario revisar cuál de los núcleos disponibles presenta mejor desempeño. Durante el desarrollo del modelo se implementaron 3 tipos de Núcleo (Linear, Radial y Polinomial), que básicamente consiste en transformar las observaciones originales a un espacio de diferente dimensionalidad en el que pueda dibujarse un hiperplano.

Redes Neuronales: Es un modelo matemático computacional basado en redes neuronales biológicas, que consiste en interconectar un grupo de neuronas artificiales y procesarlas [5]. Para este caso de estudio específico su diseño se centra en la resolución de un problema de regresión.

Los modelos fueron desarrollados en notebooks de Python, allí se realizó la limpieza selección e integración de los datos necesarios para enriquecer las variables, desde la validación de datos faltantes hasta la selección e implementación de los métodos de imputación, para ampliar los detalles al respecto ver el anexo SRDIS _CRISP-DM_02122020.

Modelo 1: Se diseña un modelo SVM con núcleo radical sobre el conjunto de datos de entrenamiento.

Posteriormente se evalúa el desempeño del modelo en el conjunto de datos de prueba, como se indicó previamente el conjunto de datos usado para pruebas es el periodo 2110.

Creamos una tabla de validación de los datos reales vs. los datos predichos, esto lo hacemos con el ánimo de revisar el desempeño del modelo propuesto.

```
In [474]: df = pd.DataFrame({'RealValues':sc_y.inverse_transform(y_test.reshape(-1)), 'PredictedValues':y_pred})
          df
```

Out[474]:

	RealValues	PredictedValues
0	21.0	25.610085
1	18.0	22.569954
2	18.0	22.853291
3	19.0	23.614969
4	19.0	21.304669
5	19.0	23.532167
6	18.0	21.122395

Figura 9. Predicciones vs. valor real SVM Kernel Radial - 2110

Con estos datos evaluamos el rendimiento del modelo bajo la métrica RMSE (Raíz cuadrada del error cuadrático medio), es necesario señalar que entre más alto sea el resultado de esta medida el desempeño del modelo será más bajo, el resultado de la evaluación es: "*7.92"*.

```
In [475]: #importar las bibliotecas necesarias
          from sklearn.metrics import mean_squared_error
          from math import sqrt

          #calcular RMSE
          sqrt (mean_squared_error (df['RealValues'],df['PredictedValues']))
```

```
Out[475]: 7.9280243009964
```

Figura 10. RMSE del modelo SVM Kernel Radial – 2110

Modelo 2: Con base en el resultado del modelo planteado con un Kernel radial, diseñamos uno diferente con un Kernel "linear"

Revisamos por medio del uso de una tabla de comparación los valores predichos vs. los valores reales en el conjunto de datos de prueba (2110), el proceso de ajuste del modelo se hace de forma iterativa comparando los resultados obtenidos y el valor de las métricas de una prueba a otra, para observar más detalles al respecto por favor consultar el documento anexo SRDIS _CRISP-DM_02122020 capítulo 5 "Evaluación".

De igual forma que en la instancia previa evaluamos el resultado del modelo en el conjunto de datos de pruebas, estas son las predicciones realizadas.

```
In [382]: df = pd.DataFrame({'RealValues':sc_y.inverse_transform(y_test.reshape(-1)), 'PredictedValues':y_pred})
          df
```

Out[382]:

	RealValues	PredictedValues
0	21.0	33.732895
1	18.0	22.799567
2	18.0	22.996913
3	19.0	23.465716
4	19.0	23.312385
5	19.0	23.586892
6	18.0	21.581023
7	18.0	21.085366
8	18.0	21.644339
9	18.0	21.148681

Figura 11. Predicciones vs. valor real SVM Kernel Linear - 2110

Con estos datos evaluamos el rendimiento del modelo bajo la métrica RMSE (error cuadrático medio), es necesario señalar que entre más alto sea el resultado de esta medida el desempeño del modelo será más bajo, el resultado de la evaluación de este segundo modelo usando un Kernel linear fue de "*6.92"* un poco más bajo que al usar un Kernel Radial.

```
In [383]: #calcular RMSE
          sqrt (mean_squared_error (df['RealValues'],df['PredictedValues']))
Out[383]: 6.9215916433743585
```

Figura 12. RMSE del modelo SVM Kernel Linear – 2110

Modelo 3: Dentro del proceso de experimentación se dispone del desarrollo de un SVM usando un núcleo Polinomial que considere los siguientes hiperparámetros, (C=5, gamma=0.05, degree=2):

Se usa el modelo desarrollado previamente para predecir la variable objetivo sobre el conjunto de datos de prueba (2110).

Comparamos los valores predichos vs los valores reales del conjunto de datos de prueba (2110), el resultado obtenido es el siguiente:

38	17.0	22.875697

Figura 13. Predicciones vs. valor real SVM Kernel Polinomial - 2110

Con estos datos evaluamos el rendimiento del modelo bajo la métrica RMSE (Raíz cuadrada del error cuadrático medio), el resultado de la evaluación de este tercer modelo usando un Kernel polinomial fue de 8.06 el más alto de los 3 modelos desarrollados, por ende, él menos eficiente de esta ronda de pruebas.

```
8.06833706047319
```

Figura 14. RMSE del modelo SVM Kernel Polinomial- 2110

Continuando con el esquema de experimentación se realiza validación cruzada para dividir el conjunto de datos en:

Tres periodos para entrenamiento y uno para pruebas, así:

- 1510, 1530, 1610 (Entrenamiento)
- 1630 (Pruebas)

Se validan los modelos anteriormente desarrollados con los nuevos conjuntos de datos de entrenamiento y pruebas.

Modelo 4: Modelo SVM Radial

Es válido destacar los resultados hallados en esta prueba, al comparar los modelos desarrollados usando un núcleo radial con el primer conjunto de datos de entrenamiento vs. el segundo conjunto de datos de entrenamiento (validación cruzada), se observa que el indicador RMSE mejora en la segunda prueba *3.79* respecto del resultado obtenido en la primera instancia *7.92*.

Modelo 5: Modelo SVM Linear

Nuevamente se revisan los resultados y se comparan los indicadores de los dos modelos con núcleo linear, el primero tiene un valor RMSE de *6.92 vs.* La segunda instancia que tiene un valor inferior de *3.63*, por lo tanto, el mejor desempeño hasta el momento lo tiene el modelo que se entrenó usando los periodos (1510 al 2030).

Modelo 6: Modelo SVM Polinomial

De la comparación entre los modelos polinomiales desarrollados previamente con los diferentes conjuntos de entrenamiento, este presenta un mejor rendimiento del indicador RMSE, en donde se observa "*4.06"* respecto de los "*8.06"* del modelo entrenado inicialmente.

Continuando con el esquema de experimentación se realizan cambios en la selección de las variables del conjunto de entrenamiento, *se eliminan los modelos de reunión, por tanto, los días en que se dictan las clases y se agrupan algunos elementos para dejar únicamente el número de inscritos por asignatura y periodo,* para revisar los detalles de este cambio ver el anexo SRDIS _CRISP-DM_02122020 "selección y limpieza de datos".

Es importante señalar que los datos usados provienen de diferentes fuentes de información y algunas de ellas tienen una estructura diferente, como por ejemplo la información relacionada con los retiros y las notas de los estudiantes. Para solucionar esta situación se emplearon llaves, es decir campos en común que pudieran servir para relacionar e integrar los diferentes documentos, los campos utilizados fueron los siguientes (Id asignatura, Periodo, Número de clase). Estos datos se cargaron a un DataFrame individual y posteriormente se concatenaron para formar dos conjuntos de datos llamados “Retiros” y “Notas”. Es fundamental señalar que los documentos de Excel que se usaron para crear el dataframe “Retiros” se diligencian de forma manual por parte del departamento de Ingeniería de Sistemas, con base en las solicitudes de retiro realizadas por los estudiantes en fechas específicas, por tanto, algunos de los documentos tienen variables diferentes o la definición del nombre de las columnas cambia de un archivo a otro. La técnica de concatenación descrita previamente nos permitió centrar en dos Data Frames la información requerida para enriquecer el conjunto de variables.

Cabe resaltar que para este punto del desarrollo se involucran variables nuevas con base en el diseño de un *Grafo dirigido*, se dispone del uso de la librería *NetworkX de Python* para modelar el nuevo plan de estudios del programa de Ingeniería de Sistemas, se adicionan nodos, los cuales corresponden a los Id de las asignaturas que lo componen y enlaces (Edge) los cuales corresponden a las dependencias entre asignaturas, es importante destacar que estas dependencias se encuentran estructuradas con base en el nuevo plan de estudios del programa de Ingeniería de Sistemas.

La variable Id corresponde al número de Id de asignaturas precedentes de otras asignaturas denominadas IdEdge.

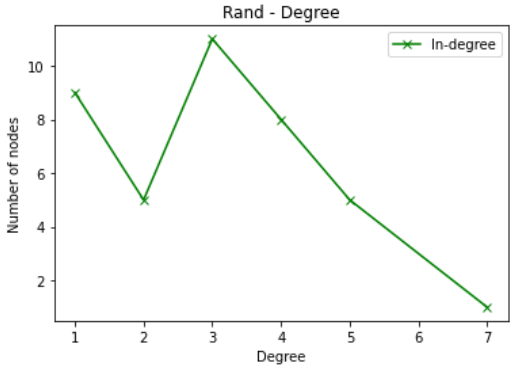


Figura 15. Gráfico del Grafo NetworkX - Plan de Estudios

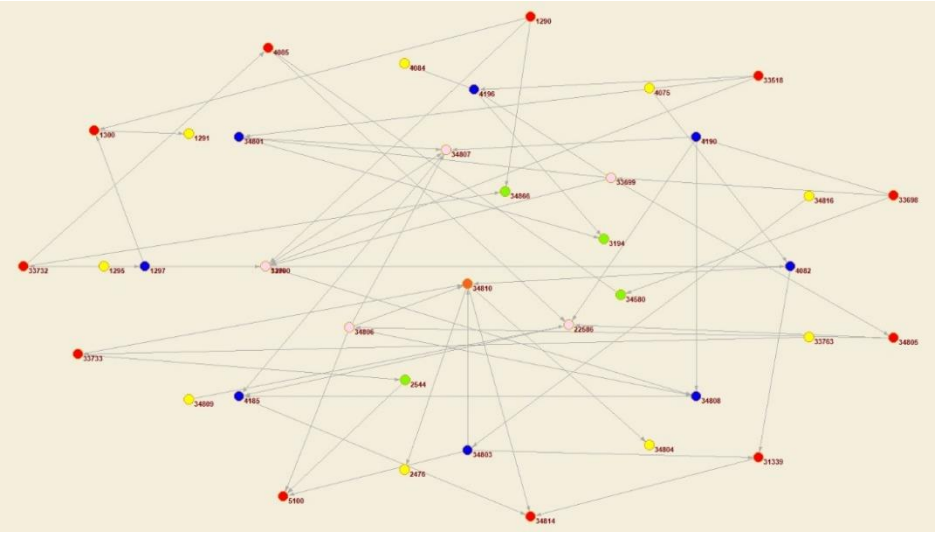


Figura 16. Grafo NetworkX - Plan de Estudios - All Degree

Al ser un grafo dirigido es posible modelar el comportamiento de las asignaturas que preceden a otras signaturas, por tanto, es posible revisar características propias de esta estructura, como, por ejemplo:

- Grado de un nodo
- Centralidad de un nodo
- Suma de los estudiantes que se encuentran inscritos en asignaturas precedentes de un nodo específico.

En la anterior imagen (*Figura*) es posible observar la descripción del grado (All degree) de cada uno de los nodos que componen el plan de estudios, entendiendo que un nodo corresponde al Id de una asignatura, en donde hallamos que la media de los grados del *Grafo* es 3 y hay pocas asignaturas que tienen un grado de 7, sería válido entender que estas últimas formarían un cuello de botella por lo tanto se considerarían nodos más importantes que los otros.

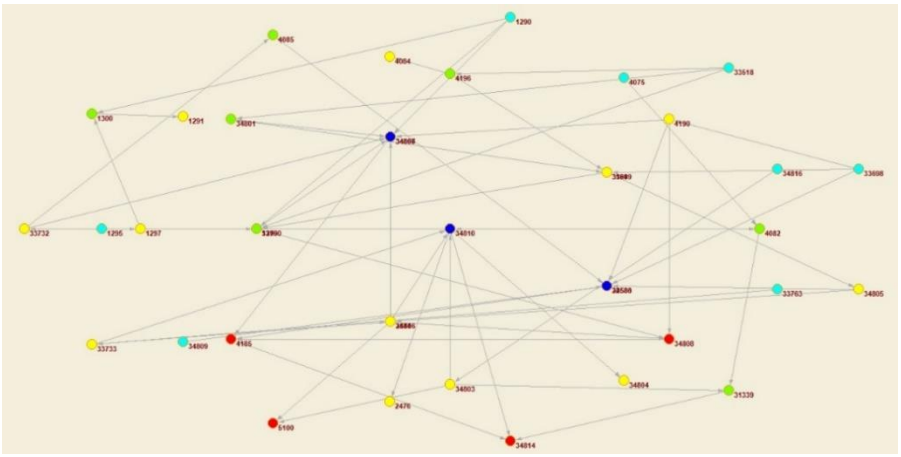


Figura 17. Grafo NetworkX - Plan de Estudios – Input Degree

La estructura Input Degree (*Figura* **¡Error! No se encuentra el origen de la referencia.**) se divide en varios grupos, los nodos que se encuentran marcados de color azul tienen un grado de entrada de 4, los marcados en color amarillo tienen un grado de entrada de 1, los marcados en color azul celeste tienen un grado de entrada de 0, los marcados en color verde tienen un grado de entrada de 2, los marcados en color rojo tienen un grado de entrada de 3. Mayoritariamente los nodos predominantes en esta estructura son los que tiene un grado de entrada de 1, los cuales corresponden al grupo amarillo de los que se cuentan 14 Id, seguidos de los de color Azul Celeste de los cuales hay 11 Id. De la anterior observación se entiende que hay varias asignaturas que no tiene prerrequisitos, pero mayoritariamente tiene por lo menos uno.

Con base en las anteriores observaciones y con el ánimo de enriquecer el conjunto de datos de entrenamiento se dispuso de la implementación de algunas líneas de Python que tiene por fin adicionar 3 columnas nuevas al conjunto de datos:

- Sum_antecestors: Este campo corresponde a la suma de los estudiantes inscritos en las asignaturas precedentes de una asignatura que tenga por lo menos una asignatura como requisito de inscripción.
- Grados: Representa el Grado (All Degree) de un nodo especifico.
- Centralidad: Representa la importancia de un nodo específico respecto de los demás nodos que componen el grafo.
- Descripción del segundo conjunto de variables:

Para ejecutar el conjunto de pruebas disponemos de un nuevo set de variables, las cuales se describen a continuación:

- Sesión: Corresponde al calendario usado para programar una asignatura, es decir el número de semanas que la componen.
- IDCurso: Corresponde al Id de la asignatura.
- Ciclo: Corresponde al periodo académico en el cual se programó la asignatura.
- TipoClase: Sección de inscripción o Sección sin inscripción.
- Días: Días transcurridos desde la fecha inicial de la clase hasta la fecha final de la asignatura.
- NroHoras: Número de horas por cada modelo de reunión, es decir el número de horas que la asignatura se dicta a la semana.
- HorasSemanales: Corresponde al número total de horas de clase de una asignatura al semestre.
- TotalInscripciones: Corresponde al número de estudiantes inscritos para una asignatura en un periodo especifico.
- Retiros: Corresponde al número de estudiantes que se retiran de la asignatura para un periodo específico.
- Perdidas: Corresponde al número de estudiantes que pierden la asignatura para un periodo específico.
- Aprobados: Corresponde al número de estudiantes que aprueban la asignatura para un periodo específico.
- ProbAprobar: Corresponde a la probabilidad de que un estudiante aprueba la asignatura para un periodo específico.

- ***InscritosSiguientePeriodo: (Variable Objetivo), Corresponde al número de estudiantes que se inscribieron en la asignatura para el siguiente periodo académico.***
- Grados: (*All Degree),* número total de grados de un nodo.
- sum_antecestors: Este campo corresponde a la suma de los estudiantes inscritos en las asignaturas precedentes de una asignatura que tenga por lo menos una asignatura como requisito de inscripción.
- Centralidad: Representa la importancia de un nodo específico respecto de los demás nodos que componen el grafo.

Con base en este nuevo conjunto de datos, se fija un esquema de experimentación para todas las pruebas de aquí en adelante, en donde se toman 3 periodos académicos para entrenar los diferentes modelos y un periodo académico para probarlos.

A continuación, se relacionan los resultados de las pruebas sobre cada uno de los datasets de entrenamiento y validación, es importante indicar que estos conjuntos de datos se dividieron así:

- Primer conjunto de datos de entrenamiento (1510, 15030 y 1610)
- Primer conjunto de Pruebas (1630)
  - Segundo conjunto de datos de entrenamiento (1630, 1710 y 1730)
  - Segundo conjunto de datos de Pruebas: (1810)
    - ♦ Tercer conjunto de datos de entrenamiento (1810, 1830 y 1910)
    - ♦ Tercer conjunto de datos de Pruebas (1930)
      - ★ Cuarto conjunto de datos de entrenamiento (1930, 2010, 2030)
      - ★ Cuarto conjunto de datos de pruebas (2130)

Primer conjunto de datos de entrenamiento (1510, 15030 y 1610) y Primer conjunto de datos de Pruebas (1630):

Modelo 7: Modelo SVM radial

Dado que esta prueba se realizó con base en un nuevo conjunto de datos, es necesario revisar los resultados sobre cada uno de los núcleos del SVM, como se observa en la imagen previa se entrena un nuevo modelo radial, a continuación, se presentan los resultados obtenidos:

```
[91] df = pd.DataFrame({'RealValues':sc_y.inverse_transform(y_test.reshape(-1)), 'PredictedValues':y_pred})
     df
```

	RealValues	PredictedValues
0	49.0	28.582588
1	0.0	59.676039
2	106.0	93.733860
3	26.0	36.161009
4	218.0	176.950471
5	20.0	47.462565
6	98.0	45.831484
7	98.0	41.942097
8	98.0	43.512764

Figura 18. Predicciones vs. datos reales SVM Kernel radial - 1630

```
#importar las bibliotecas necesarias
from sklearn.metrics import mean_squared_error
from math import sqrt

#calcular RMSE
sqrt (mean_squared_error (df['RealValues'],df['PredictedValues']))
```

```
35.99715104775627
```

Figura 19. RMSE del modelo SVM Kernel radial - 1630

Los resultados sobre el nuevo conjunto de entrenamiento y pruebas para un Kernel radial difieren en gran medida con los anteriores modelos, debido a la agrupación de los datos, esta vez *el número de inscritos se calculó por asignatura y por periodo*, anteriormente esta cifra se calculaba por modelo de reunión, es decir, por día para cada clase. Es necesario revisar el resultado con los demás kernels.

Modelo 8: SVM linear

```
[95] df = pd.DataFrame({'RealValues':sc_y.inverse_transform(y_test.reshape(-1)), 'PredictedValues':y_pred})
     df
```

	RealValues	PredictedValues
0	49.0	12.170527
1	0.0	70.524685
2	106.0	76.297210
3	26.0	20.486433
4	218.0	178.156618

Figura 20. Predicciones SVM Kernel linear - 1630

```
[96] #calcular RMSE
     sqrt (mean_squared_error (df['RealValues'],df['PredictedValues']))

     29.165800868148082
```

Figura 21. RMSE del modelo SVM Kernel linear - 1630

Para este otro modelo observamos que la medida RMSE es similar a la calculada en la modelo radial *29.16*, es necesario continuar con la experimentación para encontrar la mejor selección de características.

Modelo 9: SVM Polinomial

```
[101] #calcular RMSE
      sqrt (mean_squared_error (df['RealValues'],df['PredictedValues']))

      44.72395712018452
```

Figura 22. RMSE del modelo SVM Kernel Polinomial – Modelo 9

Con base en el mismo conjunto de datos para entrenamiento y pruebas se implementa el uso de un modelo basado en Redes Neuronales, a continuación, se presenta el desarrollo y los resultados:

Modelo 10: Disponemos del uso del Python y de la librería *Keras* para diseñar las diferentes estructuras de las redes neuronales que se implementarán, en primer lugar, normalizamos los conjuntos de entrenamiento y pruebas usando la medida de Z-Score y posteriormente definimos la función de la red neuronal como “regresión”.

El diseño del primer modelo de RN considera el uso de una sola capa intermedia con 50 neuronas y una capa de salida, la capa de entrada debe tener el mismo número de neuronas que la cantidad de variables de los datos de entrenamiento después de usar el algoritmo de One Hot Encoding para numérizar las variables categóricas.

La estructura del primer modelo diseñado usando redes neuronales tiene una capa de entrada con 113 neuronas que se activan usando una función Sigmoidal, una capa intermedia con 50 neuronas y una capa de salida de una sola neurona, a continuación, se describe gráficamente esta distribución:

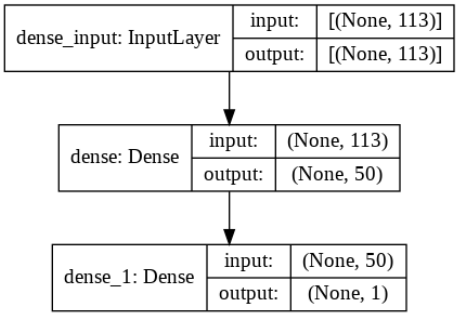


Figura 23. Estructura modelo 10 RN

EL modelo 10 se compila usando 300 épocas con un Batch Size de 32 registros para calcular la gradiente y una tasa de aprendizaje de 0.015, lo cual modifica el algoritmo original de optimización, para este caso puntual se seleccionó el algoritmo "Adam". Como se indicó previamente la métrica de evaluación seleccionada fue el RMSE, por lo tanto, se espera que esta cifra disminuya conforme avanzan las épocas, lo cual al final podría resultar en un modelo sobre ajustado, para determinar esto es necesario revisar el resultado del modelo en el dataset de entrenamiento como en el dataset de pruebas.

Las métricas de desempeño del modelo 10 tanto en el conjunto de datos de entrenamiento con en el de pruebas fueron las siguientes:

- Desempeño en Entrenamiento:
  - Coeficiente de determinación: 0.804601204338104
  - Coeficiente de Correlación: 0.8969956545815059
  - RMSE: 48.29430108272599

- Desempeño en Pruebas:
  - Coeficiente de determinación: 0.5254213175208764
  - Coeficiente de Correlación: 0.7248595157138219
  - RMSE: 60.89433334482907

El desempeño del modelo 10 con el dataset de pruebas no es eficiente si lo comparamos con los resultados obtenidos con el mismo conjunto de datos sobre el modelo 8 SVM Linear, el cual tuvo un RRMSE de *29.16* vs los *60.89* del modelo 10. A continuación se describe el desempeño del RMSE por época junto con la función de perdida. Cabe señalar que el rendimiento óptimo se da sobre las 140 épocas, después de esto la eficiencia disminuye.

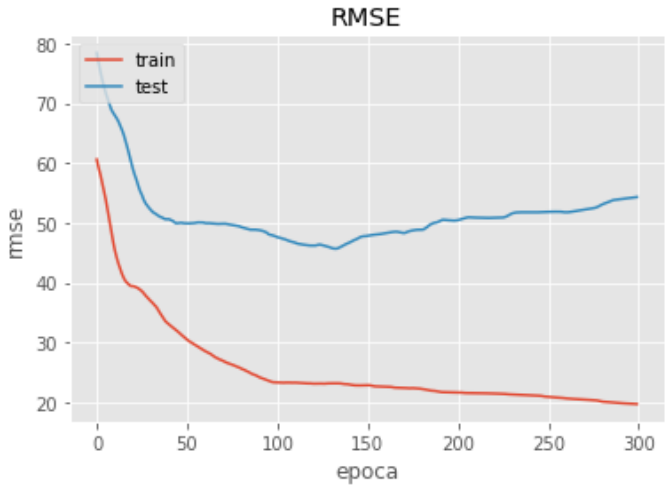


Figura 24. RMSE por época Modelo 10

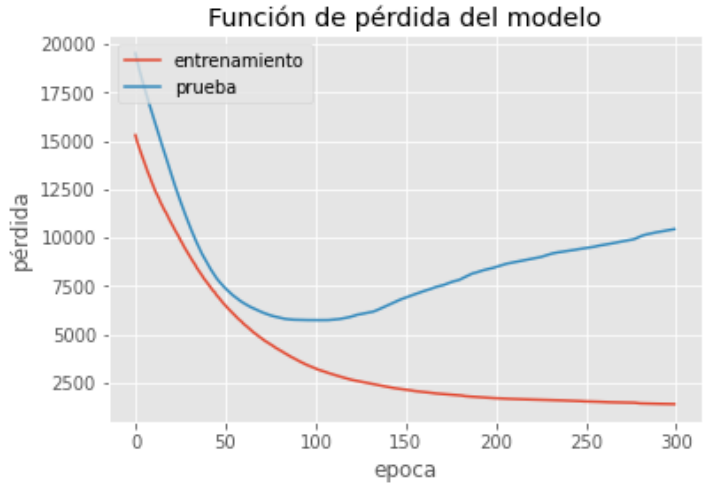


Figura 25. Función de perdida por época Modelo 10

Atendiendo a las anteriores observaciones para los siguientes modelos se reduce el número de épocas, es necesario revisar el rendimiento y ajustar los parámetros en los esquemas de experimentación.

Modelo 11: La estructura de este modelo se ajusta con base en los aprendizajes obtenidos del modelo anterior, se configura un modelo con una capa inicial de 113 neuronas, con dos capas intermedias de 50 neuronas cada una, se adiciona una neurona de sesgo “siempre activa” en cada una de las capas intermedias, nuevamente se deja una sola neurona en la capa de salida con una función de activación “linear”. Esta vez se establecen solamente 150 épocas se usa un Bacht Size de 1 para calcular la gradiente y una taza de aprendizaje de 0.015, los resultados son los siguientes:

- Desempeño en Entrenamiento:
- Coeficiente de determinación: 0.9896291054675725
- Coeficiente de correlación: 0.9948010381315314
- RMSE: 11.126102404302738

- Desempeño en Pruebas:
- Coeficiente de determinación: 0.5836066731493661
- Coeficiente de correlación: 0.7639415377824184
- RMSE: 57.03936047485954

El indicador RMSE mejoro respecto del modelo 10, pero aún su eficiencia es inferior a las de los modelos SMV, por lo cual es conveniente continuar con el esquema de experimentación.

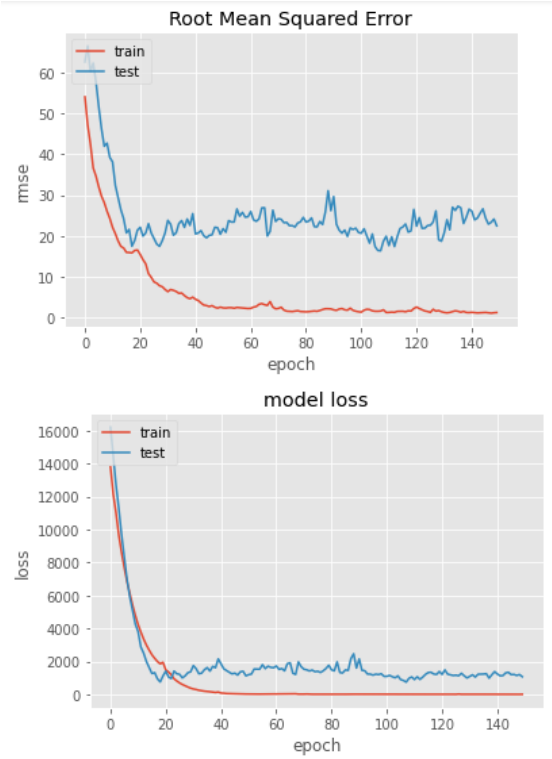


Figura 26. RMSE por época Modelo 11

Modelo 12: Este modelo presenta una variación respecto de los dos anteriores, se usan 3 capas ocultas cada una con 50 neuronas y con activación Sigmoidal y una capa de salida con una neurona con una función de activación linear, nuevamente se compila con 300 épocas y con un Bacht Size de 32, los resultados conseguidos son mejores que los de los modelos previos. Es importante destacar en primer lugar que los coeficientes de correlación aumentaron significativamente, en segundo lugar, que no se presenta sobre ajuste debido a que el indicador RMSE es similar tanto en entrenamiento como en pruebas y en tercer lugar que después de las 250 épocas el modelo se estabiliza debido a que el indicar RMSE se mantiene hasta el final. La eficiencia es similar en comparación con los modelos SMV diseñados previamente.

- Desempeño en Entrenamiento:
- Coeficiente de determinación: 0.9354612825429054
- Coeficiente de correlación: 0.9671924744035726+0j
- RMSE: 27.755247816104795

- Desempeño en Pruebas:
- Coeficiente de determinación: 0.9066112375260632
- Coeficiente de correlación: 0.9521613505735587+0j
- RMSE: 27.01282949713059

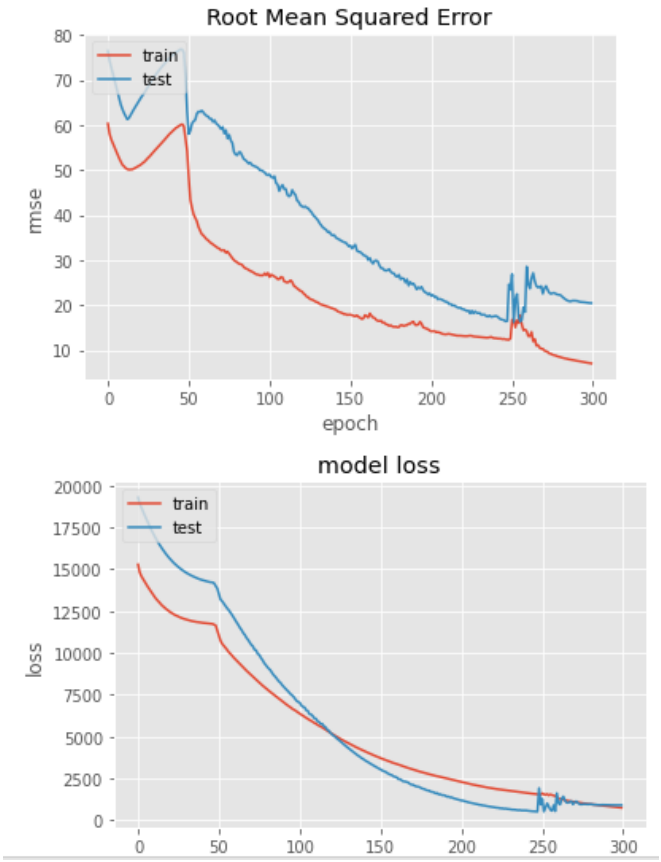


Figura 27. RMSE por época Modelo 12

Segundo conjunto de datos de entrenamiento (1630, 1710 y 1730) y Segundo conjunto de datos de Pruebas (1810):

Debido a que contamos con un nuevo conjunto de datos para entrenamiento y pruebas, revisamos el comportamiento de los modelos SVM y RN para validar su desempeño.

Modelo 13: Este modelo es del tipo SVM y usa un núcleo linear, probamos su desempeño en el dataset de pruebas, encontramos que el indicador RMSE es de 29.71.

En la siguiente tabla se relacionan la predicción realizada por este modelo ***para los primeros cinco datos de prueba***:

Asignatura	Valores Reales	Valores Predichos
Administración Bases de Datos	43.0	41.13
Administración Básica Linux	0.0	38.38
Admón. Sistemas de Información	103.0	105.71
Análisis de Algoritmos	35.0	33.36
Análisis y Diseño O.O.	204.0	209.54

Tabla 6. Predicciones Modelo 13

Modelo 14: Este modelo es del tipo SVM y usa un núcleo linear, probamos su desempeño en el dataset de pruebas, encontramos que el indicador RRMSE es de 27.31.

En la siguiente tabla se relacionan la predicción realizada por este modelo ***para los primeros cinco datos de prueba***:

Asignatura	Valores Reales	Valores Predichos
Administración Bases de Datos	43.0	38.10
Administración Básica Linux	0.0	44.38
Admón. Sistemas de Información	103.0	107.03
Análisis de Algoritmos	35.0	31.41
Análisis y Diseño O.O.	204.0	234.72

Tabla 7. Predicciones Modelo 14

Modelo 15: Este modelo es del tipo SVM y usa un núcleo Polinomial, probamos su desempeño en el dataset de pruebas, encontramos que el indicador RMSE es de 26.86.

En la siguiente tabla se relacionan la predicción realizada por este ***modelo para los primeros cinco datos de prueba***:

Asignatura	Valores Reales	Valores Predichos
Administración Bases de Datos	43.0	48.03
Administración Básica Linux	0.0	48.68
Admón. Sistemas de Información	103.0	98.77
Análisis de Algoritmos	35.0	35.39
Análisis y Diseño O.O.	204.0	198.56

Tabla 8. Predicciones Modelo 15

Modelo 16: Este modelo es del tipo RN (Redes Neuronales) y se diseñó usando los siguientes parámetros:

- Capa de entrada: 111 Neuronas, Función de activación (Sigmoidal), Neurona de Sesgo (Verdadero).
- Capa oculta: 50 Neuronas, Función de activación (Sigmoidal), Neurona de Sesgo (Verdadero).
- Capa de Salida: 1 Neurona, Función de activación (Linear), Neurona de Sesgo (Verdadero).
- Batch Size: 32.
- Learning_rate=0.015
- Epocas: 300

El desempeño de este modelo es el siguiente:

Desempeño en entrenamiento

- Coeficiente de determinación 0.7958150345108024
- Coeficiente de correlación: 0.8920846565829963
- RMSE: 39.03321493442109

Desempeño en Pruebas

- Coeficiente de determinación: 0.7122181464368749
- Coeficiente de correlación: 0.8439301786503874
- RMSE: 45.6325349893067

Modelo 17: Este modelo es del tipo RN (Redes Neuronales) y se diseñó usando los siguientes parámetros:

- Capa de entrada: 111 Neuronas, Función de activación (Sigmoidal), Neurona de Sesgo (Verdadero).
- 2 capas ocultas: 50 Neuronas, Función de activación (Sigmoidal), Neurona de Sesgo (Verdadero).
- Capa de Salida: 1 Neurona, Función de activación (Linear), Neurona de Sesgo (Verdadero).
- Batch Size: 1.
- Learning_rate=0.015
- Epocas: 150

El desempeño de este modelo es el siguiente:

Desempeño en entrenamiento:

- Coeficiente de determinación: 0.8972499097465817

- Coeficiente de correlación: 0.9472327642911121
- RMSE: 27.689400214465387

Desempeño en Pruebas

- Coeficiente de determinación: 0.9237448571917611
- Coeficiente de correlación: 0.9611164639063058
- RMSE: 23.489699561930912

Modelo 18: Este modelo es del tipo RN (Redes Neuronales) y se diseñó usando los siguientes parámetros:

- Capa de entrada: 111 Neuronas, Función de activación (Sigmoidal), Neurona de Sesgo (Verdadero).
- 3 capas ocultas: 50 Neuronas, Función de activación (Sigmoidal), Neurona de Sesgo (Verdadero).
- Capa de Salida: 1 Neurona, Función de activación (Linear), Neurona de Sesgo (Verdadero).
- Batch Size: 32.
- Learning_rate=0.015
- Epocas: 300

El desempeño de este modelo es el siguiente:

Desempeño en entrenamiento

- Coeficiente de determinación: 0.9272507629913733
- Coeficiente de correlación: 0.9629386081113236+0j
- RMSE: 23.29896687395755

Desempeño en Pruebas

- Coeficiente de determinación: 0.89218388114565
- Coeficiente de correlación: 0.9445548587274589+0j
- RMSE: 27.930887771260227

Tercer conjunto de datos de entrenamiento (1810, 1830, 1910) y Tercer conjunto de datos de Pruebas (1930):

Modelo 19: Este modelo es del tipo SVM y usa un núcleo radial, probamos su desempeño en el dataset de pruebas, encontramos que el indicador RMSE es de 47.18.
En la siguiente tabla se relacionan la predicción realizada por este modelo ***para los primeros cinco datos de prueba***:

Asignatura	Valores Reales	Valores Predichos
Administración Bases de Datos	41.0	46.47
Administración Básica Linux	119.0	79.44
Admón. Sistemas de Información	22.0	29.57
Análisis de Algoritmos	27.0	30.38
Análisis y Diseño O.O.	0.0	28.84

Tabla 9. Predicciones Modelo 19

Modelo 20: Este modelo es del tipo SVM y usa un núcleo linear, probamos su desempeño en el dataset de pruebas, encontramos que el indicador RMSE es de 23.86.

En la siguiente tabla se relacionan la predicción realizada por este modelo ***para los primeros cinco datos de prueba***:

Asignatura	Valores Reales	Valores Predichos
Administración Bases de Datos	41.0	44.87
Administración Básica Linux	119.0	91.05
Admón. Sistemas de Información	22.0	27.80
Análisis de Algoritmos	27.0	32.52
Análisis y Diseño O.O.	0.0	27.32

Tabla 10. Predicciones Modelo 20

Modelo 21: Este modelo es del tipo SVM y usa un núcleo polinomial, probamos su desempeño en el dataset de pruebas, encontramos que el indicador RMSE es de 28.66.

En la siguiente tabla se relacionan la predicción realizada por este modelo ***para los primeros cinco datos de prueba***:

Asignatura	Valores Reales	Valores Predichos
Administración Bases de Datos	41.0	34.90
Administración Básica Linux	119.0	90.87
Admón. Sistemas de Información	22.0	39.25
Análisis de Algoritmos	27.0	21.50
Análisis y Diseño O.O.	0.0	37.62

Tabla 11. Predicciones Modelo 21

Modelo 22: Este modelo es del tipo RN (Redes Neuronales) y se diseñó usando los siguientes parámetros:

- Capa de entrada: 111 Neuronas, Función de activación (Sigmoidal), Neurona de Sesgo (Verdadero).
- 1 capa oculta: 50 Neuronas, Función de activación (Sigmoidal), Neurona de Sesgo (Verdadero).
- Capa de Salida: 1 Neurona, Función de activación (Linear), Neurona de Sesgo (Verdadero).
- Batch Size: 32.
- Learning_rate=0.015
- Epocas: 300

El desempeño de este modelo es el siguiente:

Desempeño en entrenamiento:

- Coeficiente de determinación: 0.6908801713449766
- Coeficiente de correlación: 0.8311920183357012
- RMSE: 48.618505938602574

Desempeño en Prueba

- Coeficiente de determinación: 0.6080958007387118
- Coeficiente de correlación: 0.779804976092556
- RMSE: 51.81522084595333

Modelo 23: Este modelo es del tipo RN (Redes Neuronales) y se diseñó usando los siguientes parámetros:

- Capa de entrada: 111 Neuronas, Función de activación (Sigmoidal), Neurona de Sesgo (Verdadero).
- 2 capas ocultas: 50 Neuronas, Función de activación (Sigmoidal), Neurona de Sesgo (Verdadero).
- Capa de Salida: 1 Neurona, Función de activación (Linear), Neurona de Sesgo (Verdadero).
- Batch Size: 1.
- Learning_rate=0.015
- Epocas: 150

El desempeño de este modelo es el siguiente:

Desempeño en entrenamiento:

- Coeficiente de determinación: 0.5211432165281777
- Coeficiente de correlación: 0.7219024979373445
- RMSE: 60.511929154962225

Desempeño en prueba:

- Coeficiente de determinación: 0.7535962731450305
- Coeficiente de correlación: 0.8680992300106195
- RMSE: 41.085732210564046

Modelo 24: Este modelo es del tipo RN (Redes Neuronales) y se diseñó usando los siguientes parámetros:

- Capa de entrada: 111 Neuronas, Función de activación (Sigmoidal), Neurona de Sesgo (Verdadero).
- 3capas ocultas: 50 Neuronas, Función de activación (Sigmoidal), Neurona de Sesgo (Verdadero).
- Capa de Salida: 1 Neurona, Función de activación (Linear), Neurona de Sesgo (Verdadero).
- Batch Size: 32.
- Learning_rate=0.015
- Épocas: 300

El desempeño de este modelo es el siguiente:

Desempeño en entrenamiento:

- Coeficiente de determinación: 0.3871839245458599
- Coeficiente de correlación: 0.62224105019346+0j
- RMSE: 68.4546968166241

Desempeño en prueba:

- Coeficiente de determinación: 0.6390787861921035
- Coeficiente de correlación: 0.7994240340345689+0j
- RMSE: 49.724863507753625

Cuarto conjunto de datos de entrenamiento (1930, 2010, 2030) y Cuarto conjunto de datos de Pruebas (2110), es importante señalar que este conjunto de datos es atípico debido a que se da en periodo de pandemia, a continuación, se relacionan los resultados obtenidos:

Modelo 25: Este modelo es del tipo SVM y usa un núcleo radial, probamos su desempeño en el dataset de pruebas, encontramos que el indicador RMSE es de 94.73.

En la siguiente tabla se relacionan la predicción realizada por este modelo ***para los primeros cinco datos de prueba***:

Asignatura	Valores Reales	Valores Predichos
Administración Básica Linux	68.0	56.73
Admón. Sistemas de Información	0.0	45.20
Análisis de Algoritmos	87.0	94.80
Análisis de Algoritmos	87.0	125.49
Análisis y Diseño O.O.	0.0	47.62

Tabla 12. Predicciones Modelo 25

Modelo 26: Este modelo es del tipo SVM y usa un núcleo linear, probamos su desempeño en el dataset de pruebas, encontramos que el indicador RMSE es de 111.66.

En la siguiente tabla se relacionan la predicción realizada por este modelo ***para los primeros cinco datos de prueba***:

Asignatura	Valores Reales	Valores Predichos
Administración Básica Linux	68.0	63.20
Admón. Sistemas de Información	0.0	66.85
Análisis de Algoritmos	87.0	162.34
Análisis de Algoritmos	87.0	135.10
Análisis y Diseño O.O.	0.0	31.80

Tabla 13. Predicciones Modelo 26

Modelo 27: Este modelo es del tipo SVM y usa un núcleo polinomial, probamos su desempeño en el dataset de pruebas, encontramos que el indicador RMSE es de 97.56.

En la siguiente tabla se relacionan la predicción realizada por este modelo ***para los primeros cinco datos de prueba***:

Asignatura	Valores Reales	Valores Predichos
Administración Básica Linux	68.0	57.85
Admón. Sistemas de Información	0.0	52.90
Análisis de Algoritmos	87.0	114.93
Análisis de Algoritmos	87.0	115.76
Análisis y Diseño O.O.	0.0	39.03

Tabla 14. Predicciones Modelo 27

Modelo 28: Este modelo es del tipo RN (Redes Neuronales) y se diseñó usando los siguientes parámetros:

- Capa de entrada: 111 Neuronas, Función de activación (Sigmoidal), Neurona de Sesgo (Verdadero).
- 3capas ocultas: 50 Neuronas, Función de activación (Sigmoidal), Neurona de Sesgo (Verdadero).
- Capa de Salida: 1 Neurona, Función de activación (Linear), Neurona de Sesgo (Verdadero).
- Batch Size: 32.
- Learning_rate=0.015
- Epocas: 300

El desempeño de este modelo es el siguiente:

Desempeño en entrenamiento:

- Coeficiente de determinación: 0.5677917657995519
- Coeficiente de correlación: 0.7535195855447634
- RMSE: 51.24965158320232

Desempeño en Pruebas:

- Coeficiente de determinación: 0.5700773159796821
- Coeficiente de correlación: 0.7550346455492504
- RMSE: 80.21064858167023

Modelo 29: Este modelo es del tipo RN (Redes Neuronales) y se diseñó usando los siguientes parámetros:

- Capa de entrada: 111 Neuronas, Función de activación (Sigmoidal), Neurona de Sesgo (Verdadero).
- 2 capas ocultas: 50 Neuronas, Función de activación (Sigmoidal), Neurona de Sesgo (Verdadero).
- Capa de Salida: 1 Neurona, Función de activación (Linear), Neurona de Sesgo (Verdadero).
- Batch Size: 1.
- Learning_rate=0.015
- Epocas: 150

El desempeño de este modelo es el siguiente:

Desempeño en entrenamiento:

- Coeficiente de determinación: 0.9640652332633629
- Coeficiente de correlación: 0.9818682362024769
- RMSE: 14.77752878457291

Desempeño en Pruebas:

- Coeficiente de determinación: 0.39746833637264656
- Coeficiente de correlación: 0.6304508992559583
- RMSE: 94.95695620022266

Modelo 30: Este modelo es del tipo RN (Redes Neuronales) y se diseñó usando los siguientes parámetros:

- Capa de entrada: 111 Neuronas, Función de activación (Sigmoidal), Neurona de Sesgo (Verdadero).
- 3capas ocultas: 50 Neuronas, Función de activación (Sigmoidal), Neurona de Sesgo (Verdadero).
- Capa de Salida: 1 Neurona, Función de activación (Linear), Neurona de Sesgo (Verdadero).
- Batch Size: 32.
- Learning_rate=0.015
- Épocas: 300

El desempeño de este modelo es el siguiente:

Desempeño en entrenamiento:

- Coeficiente de determinación: 0.858262678505969

- Coeficiente de correlación: 0.9264246750308246+0j
- RMSE: 29.348540561370168

Desempeño en prueba:

- Coeficiente de determinación: 0.1525489076319536
- Coeficiente de correlación: 0.39057509858150663+0j
- RMSE:112.61446287434437

6. RESULTADOS DEL PROYECTO

6.1. Resultados de los diferentes modelos implementados

A continuación, se relacionan los resultados bajo el indicador RMSE obtenidos al evaluar cada uno de los modelos diseñados:

#Modelo	*Tipo de Modelo*	*RMSE*	*Entrenamiento*	*Test*	*Predicción Modelo*
1	SVM Radial	7.92	1510, 1530, 1610, 1630, 1710, 1730, 1810, 1830 1910, 1930, 2010, 2030	2110	Número de inscritos por Clase
2	SVM Linear	6.92	1510, 1530, 1610, 1630, 1710, 1730, 1810, 1830 1910, 1930, 2010, 2030	2110	Número de inscritos por Clase
3	SVM Polinomial	8.06	1510, 1530, 1610, 1630, 1710, 1730, 1810, 1830 1910, 1930, 2010, 2030	2110	Número de inscritos por Clase
4	SVM Radial	3.79	1510, 1530, 1610	1630	Número de inscritos por Clase
5	SVM Linear	3.63	1510, 1530, 1610	1630	Número de inscritos por Clase
6	SVM Polinomial	4.06	1510, 1530, 1610	1630	Número de inscritos por Clase
7	SVM Radial	35.99	1510, 1530, 1610	1630	Número de inscritos por Asignatura
8	SMV Linear	29.16	1510, 1530, 1610	1630	Número de inscritos por Asignatura
9	SVM Polinomial	44.72	1510, 1530, 1610	1630	Número de inscritos por Asignatura
10	RN	60.89	1510, 1530, 1610	1630	Número de inscritos por Asignatura
11	RN	57.03	1510, 1530, 1610	1630	Número de inscritos por Asignatura
12	RN	27.01	1510, 1530, 1610	1630	Número de inscritos por Asignatura
13	SVM Radial	29.71	1630, 1710, 1730	1810	Número de inscritos por Asignatura
14	SVM Linear	27.31	1630, 1710, 1730	1810	Número de inscritos por Asignatura
15	SVM Polinomial	26.86	1630, 1710, 1730	1810	Número de inscritos por Asignatura
16	RN	45.86	1630, 1710, 1730	1810	Número de inscritos por Asignatura
17	RN	23.48	1630, 1710, 1730	1810	Número de inscritos por Asignatura
18	RN	27.84	1630, 1710, 1730	1810	Número de inscritos

					por Asignatura
19	SVM Radial	47.18	1810, 1830, 1910	1830	Número de inscritos por Asignatura
20	SVM Linear	23.86	1810, 1830, 1910	1830	Número de inscritos por Asignatura
21	SVM Polinomial	28.66	1810, 1830, 1910	1830	Número de inscritos por Asignatura
22	RN	51.81	1810, 1830, 1910	1830	Número de inscritos por Asignatura
23	RN	41.08	1810, 1830, 1910	1930	Número de inscritos por Asignatura
24	RN	49.72	1810, 1830, 1910	1930	Número de inscritos por Asignatura
25	SVM Radial	94.73	1930, 2010, 2030	2110	Número de inscritos por Asignatura
26	RN Linear	111.66	1930, 2010, 2030	2110	Número de inscritos por Asignatura
27	RN Polinomial	97.56	1930, 2010, 2030	2110	Número de inscritos por Asignatura
28	RN	80.21	1930, 2010, 2030	2110	Número de inscritos por Asignatura
29	RN	94.95	1930, 2010, 2030	2110	Número de inscritos por Asignatura
30	RN	112.65	1930, 2010, 2030	2110	Número de inscritos por Asignatura

Tabla 15. Resultados de los Modelos diseñados

En la siguiente *Tabla 16* se relaciona el mejor modelo de acuerdo con el indicador RMSE por grupo de datos de entrenamiento y pruebas para SVM y RN.

#Modelo	*Tipo de Modelo*	*RMSE*	*Entrenamiento*	*Test*	*Predicción Modelo*
8	SVM Linear	29.16	1510, 1530, 1610	1630	Número de inscritos por Asignatura
12	RN	27.01	1510, 1530, 1610	1630	Número de inscritos por Asignatura
15	SVM Polinomial	26.86	1630, 1710, 1730	1810	Número de inscritos por Asignatura
17	RN	23.48	1630, 1710, 1730	1810	Número de inscritos por Asignatura
20	SVM Linear	23.86	1810, 1830, 1910	1930	Número de inscritos por Asignatura
23	RN	41.08	1810, 1830, 1910	1930	Número de inscritos por Asignatura

Tabla 16. Mejores resultados por indicador RMSE

Es importante señalar que se usaron parámetros diferentes en el diseño de los 6 modelos base (3 SVM y 3 RN), posteriormente estos modelos se probaron en cada uno de los conjuntos de datos de entrenamiento y pruebas. A continuación, se describen los parámetros utilizados en el diseño base y el resultado promedio por conjunto de datos de entrenamiento:

- Modelos SVM:

Kernel	*C*	*Degree*	*Epsilon*	*Gamma*	*Max_iter*	*Promedio RMSE*
(Radial)	1	3	0.1	Scale	-1	37,66
(Linear)	1	3	0.1	Scale	-1	26,66
(Polinomial)	5	2	0.1	0.05	-1	33,66

Tabla 17. Resultados promedio de los Modelos diseñados SVM

- Modelos RN:

Capa de Entrada	*Capas intermedias*	*Épocas*	*Batch Size*	*Activación*	*Promedio RMSE*
111 neuronas	1 (50 neuronas)	300	32	Sigmoid	53
111 neuronas	2(50 neuronas)	150	1	Sigmoid	40,33
111 neuronas	3(50 neuronas)	300	32	Sigmoid	35

Tabla 18. Resultados promedios de los Modelos diseñados RN

Para revisar más detalles acerca de los parámetros usados en la configuración de los modelos por favor revisar el documento anexo SRDIS _CRISP-DM_02122020.

Como se puede observar en las tablas previas el mejor rendimiento lo presentan los modelos SVM (Support Vector Machine), en particular el diseñado con un núcleo linear, su promedio RMSE para los conjuntos de entrenamiento y pruebas es de *26.66.*

6.2. Tablero de Control

A partir de los procesos realizados sobre las fuentes de información, se generó unas visualizaciones a través de tableros de control, esta sección describe cada uno de los tableros y la validación de los resultados.

El tablero de control se implementó sobre una herramienta tecnológica llamada Power BI, esta herramienta actualmente se encuentra en uso a nivel administrativo por la parte interesada del proyecto.

La interfaz principal permite la visualización de la solución entre los diferentes componentes como se muestra en el menú con las opciones de informes de Inscripciones General, Inscripciones Detallado, Notas, Retiros, Predicción. esta sección describe el objetivo de cada tablero de control y resultado de cada interfaz

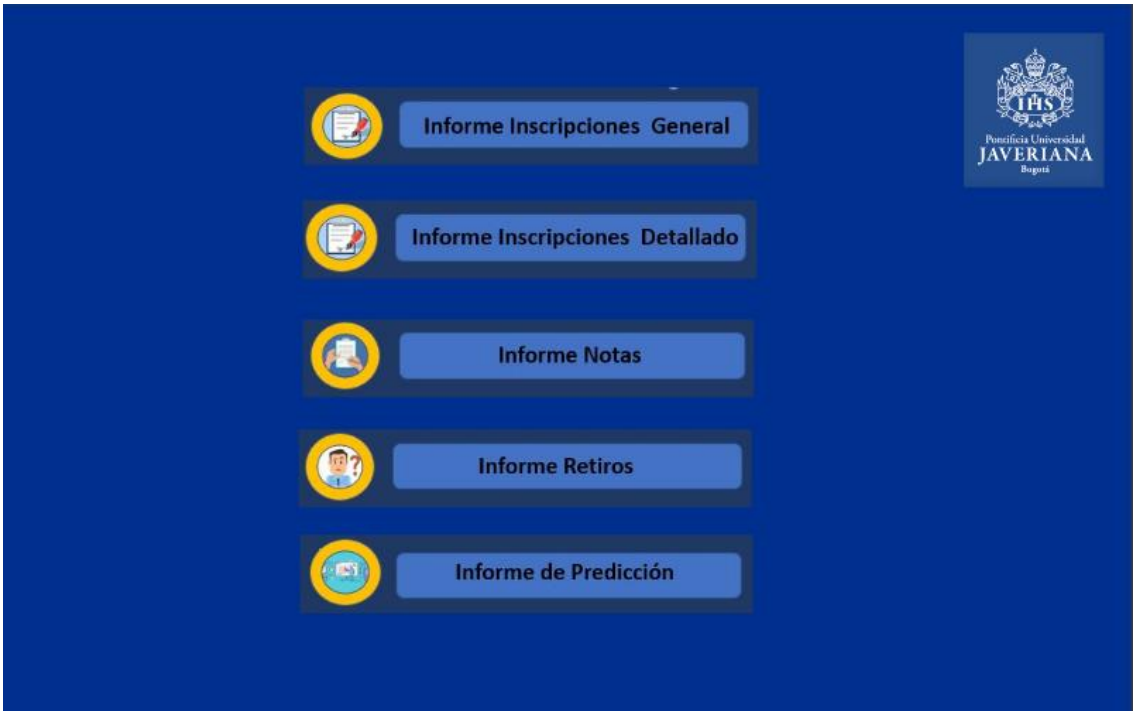


Figura 28. Menú Inicial

6.2.1. Informe Inscripciones General

El tablero de Inscripciones General tiene como objetivo la visualización de la información del compartimiento de inscripciones de las asignaturas a través de cada ciclo histórico registrado. Para desarrollar este tablero se realizó la implementación de filtro de esquema jerárquicos donde se visualiza la descripción de cada asignatura (selección única de un criterio) y ciclo (selección múltiple de criterio). A partir de estos se realizó la inclusión de una matriz de valores con filtros de Grupo en total inscripciones (selección única de un criterio) y grado (selección única de un criterio) para que entregue la información del número de inscripciones de pregrado. se incluyó una gráfica de distribución de total de inscripciones y ciclo para discriminar estos dos criterios de forma visual y promediar la inscripción por asignatura entre todos los ciclos al seleccionar una asignatura en el filtros de segmentación, la inclusión de un gráfico de esquema jerárquico que inicia con el total de inscripciones por asignatura y luego se desglosa por ciclo, tipos de clase, componente y finalizando en la variable estado de clase, se incluye una tabla con la información de los registros totales de cada una o en todas las asignaturas de acuerdo a la selección del filtro de esquema jerárquico.

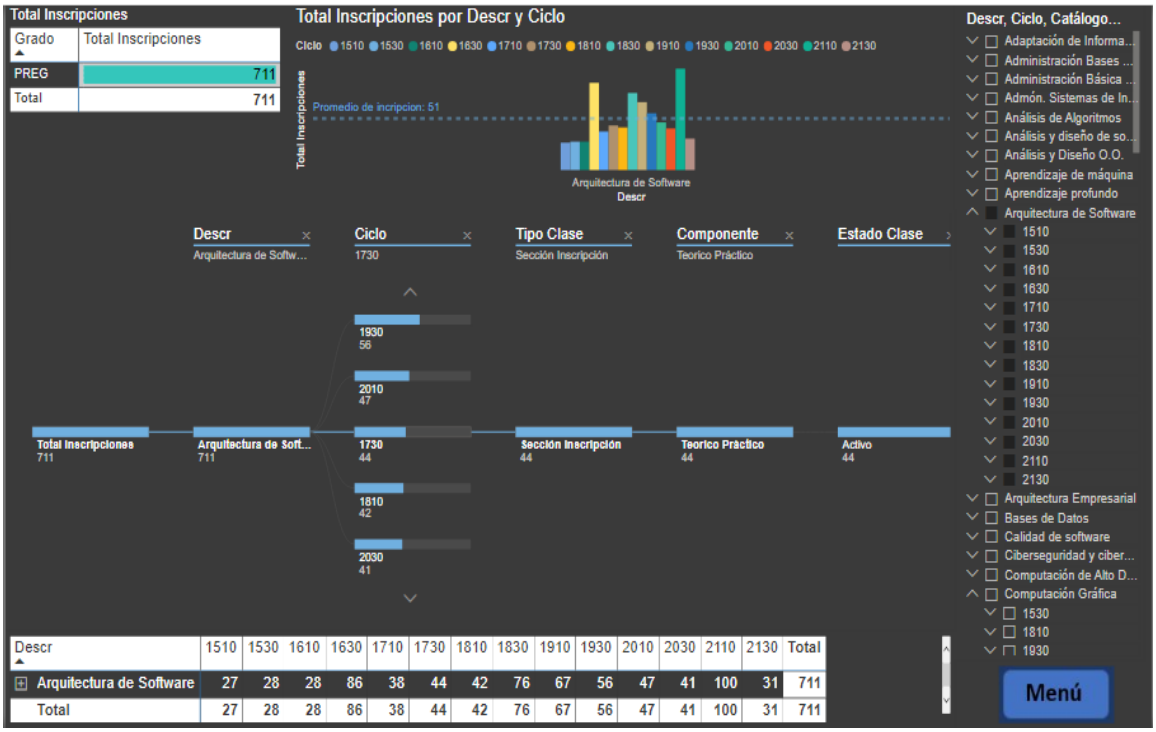


Figura 29. Informe de inscripciones generales.

6.2.2. Informe de Inscripciones Detallado

El tablero de Inscripciones Detallado tiene como objetivo la visualización de la información del compartimiento de inscripciones de forma detallado de las asignaturas a través de cada ciclo histórico. Para desarrollar este tablero se realizó la implementación del filtro de esquema jerárquico donde se visualiza la descripción de cada asignatura (selección única de un criterio) y ciclo (selección múltiple de criterio). A partir de esto se realizó la implementación de una tarjeta visual con las horas semanales por asignatura. Implementación de una tarjeta visual con total de inscripciones por asignatura, se realizó implementación de un gráfico embudo con el total de inscripciones por ciclo de acuerdo a la asignatura, se realizó la implementación de una tarjeta descriptiva con la siguiente información del ciclo, números de clase y total de inscripciones por número de clase y se realiza la implementación de un gráfico de embudo por la cantidad de inscripciones por asignatura y la demanda de inscripción por profesor esto representaría la capacidad de los profesores por ciclo de acuerdo a la asignatura.

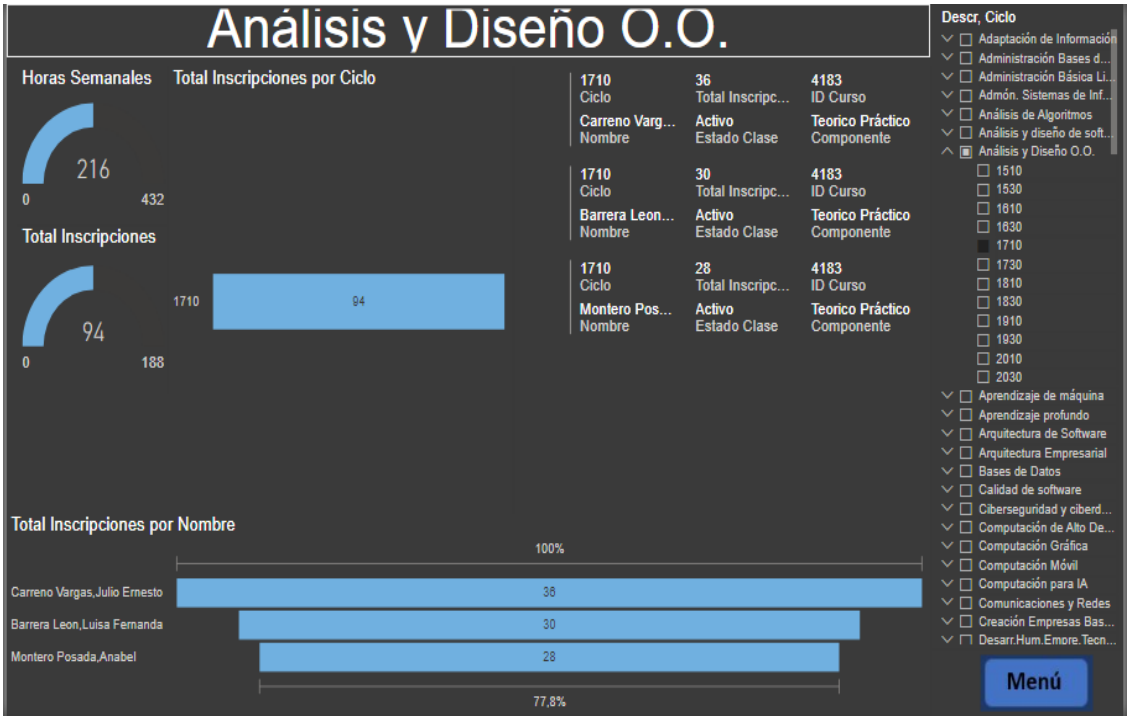


Figura 30. Informe de inscripciones detallado.

6.2.3. Informe Notas

El tablero de Notas tiene como objetivo la visualización de la información del compartimiento de las asignaturas por notas de forma detallado en el cual se puede identificar el estado de los estudiantes al cursar las asignaturas en cada ciclo histórico. Para desarrollar este tablero se realizó la implementación de filtro de esquema jerárquico donde se visualiza la descripción de cada asignatura (selección única de un criterio) y ciclo (selección múltiple de criterio). A partir de esto se realizó la implementación de un gráfico de esquema jerárquico que inicia por una descripción, ciclo, estado, motivo. Se implemento un gráfico de anillo el cual representa el número de oferta para la asignatura por su estado. Se implemento un gráfico de columnas para cuantificar el número de oferta por descripción de la asignatura y se implementó una tarjeta visual donde se expone el total de la oferta, se implementó un gráfico de anillos el ciclo con el estado de la oferta por asignatura, se implementó una tarjeta de varias filas para complementar la información de notas donde se exponen los siguiente variables descripción, el número de oferta, el id de curso y el estado del curso.

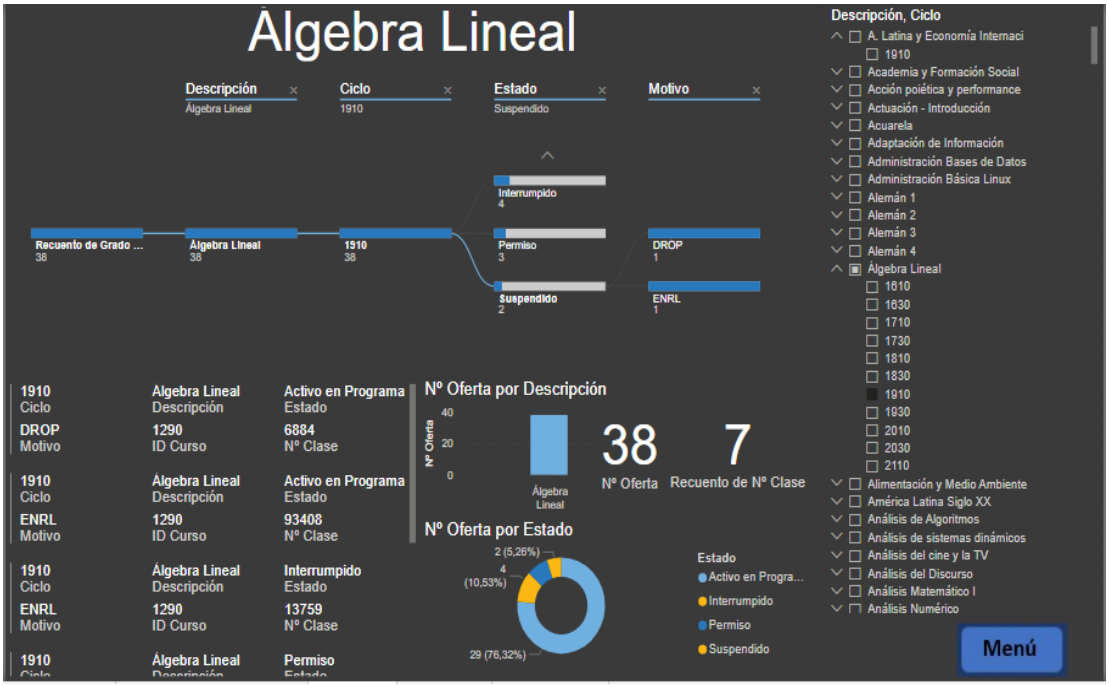
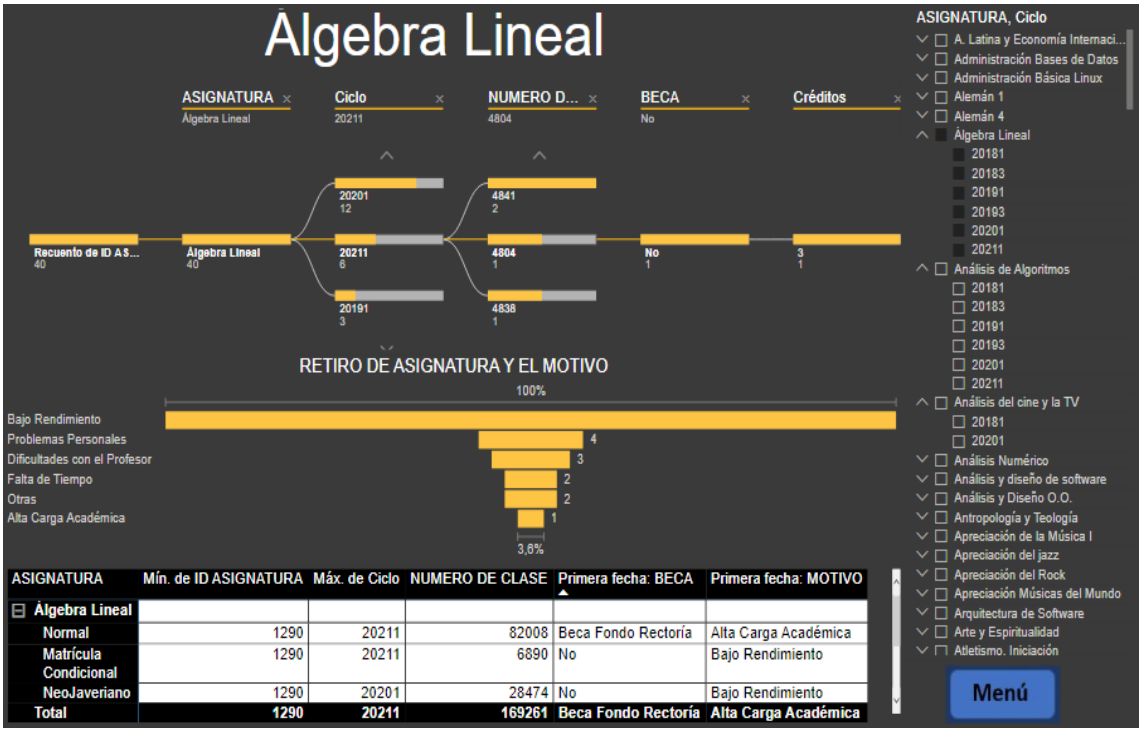


Figura 31. Informe de Notas general.

6.2.4. Informe de Retiros

El tablero de retiros tiene como objetivo la visualización de la información del compartimiento de los estudiantes de forma detallado el motivo de retiro de las asignaturas a través de cada ciclo histórico. Para desarrollar este tablero se realizó la implementación de filtro de esquema jerárquico donde se visualiza la descripción de cada asignatura (selección única de un criterio) y ciclo (selección múltiple de criterio). A partir de esto se realizó la implementación de un gráfico de esquema jerárquico que inicia por el id de clase de la asignatura, descripción de la asignatura, estado, motivo con la finalidad desglosar los retiros de la asignatura de una forma más detallada. Se implemento un gráfico de embudo el cual representa el número de retiros por asignatura y la causal de retiro. Se implemento una matriz con las siguientes variables la descripción de la asignatura y el estado académico, el id de la asignatura, el número de clase, beca, consecutivo de ciclo, cantidad de créditos con la finalidad de complementar la información del retiro de la asignatura.

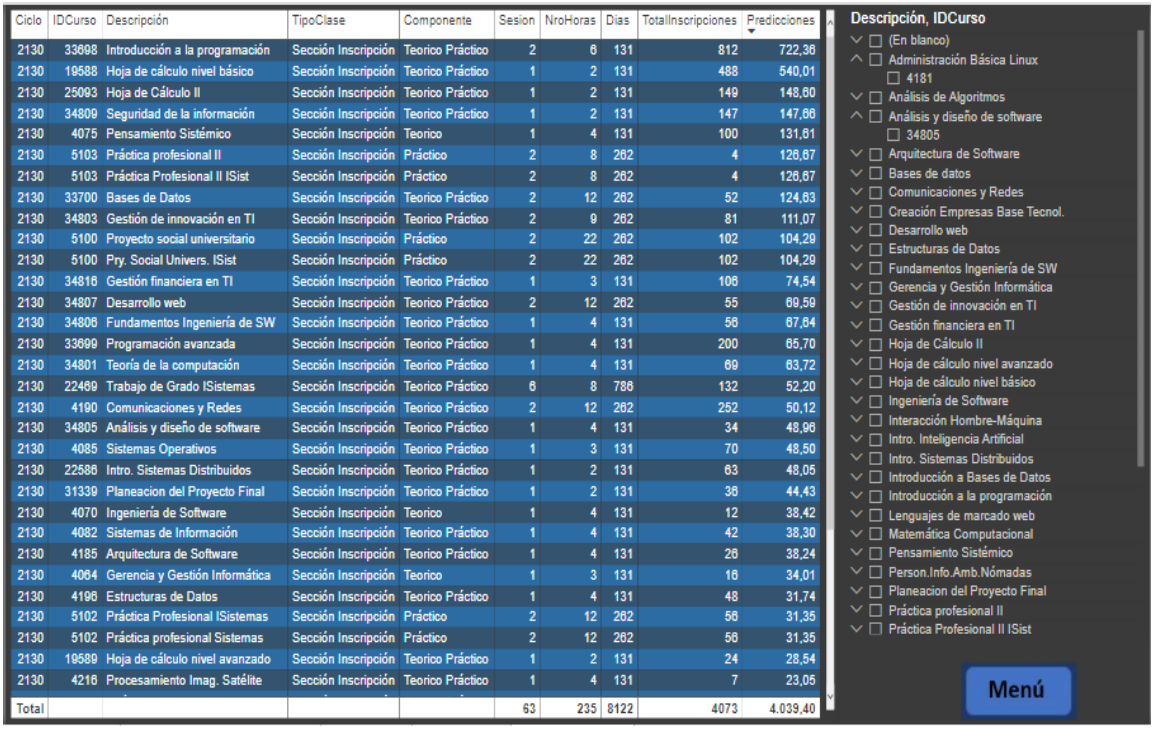


ASIGNATURA	Mín. de ID ASIGNATURA	Máx. de Ciclo	NUMERO DE CLASE	Primera fecha: BECA	Primera fecha: MOTIVO
⊟ Álgebra Lineal					
Normal	1290	20211	82008	Beca Fondo Rectoría	Alta Carga Académica
Matrícula Condicional	1290	20211	6890	No	Bajo Rendimiento
NeoJaveriano	1290	20201	28474	No	Bajo Rendimiento
Total	**1290**	**20211**	**169261**	**Beca Fondo Rectoría**	**Alta Carga Académica**

Figura 32. Informe de Retiros

6.2.5. Informe de Predicción

El tablero de predicción tiene como objetivo mostrar la información de la predicción obtenida por modelos de analítica para brindar una recomendación. Para desarrollar este tablero se realizó la implementación de filtro de esquema jerárquico donde se visualiza la descripción de cada asignatura (selección única de un criterio) e id de curso (selección múltiple de criterio). A partir de esto se realizó la implementación de una tabla con las variables de ciclo, Id curso, descripción, tipo de clase, componente, sesión, número de horas, días, total de inscripciones y predicciones esto con la finalidad de entregar las recomendaciones de inscripción de asignaturas.



Ciclo	IDCurso	Descripción	TipoClase	Componente	Sesion	NroHoras	Dias	TotalInscripciones	Predicciones
2130	33698	Introducción a la programación	Sección Inscripción	Teorico Práctico	2	6	131	812	722,36
2130	19588	Hoja de cálculo nivel básico	Sección Inscripción	Teorico Práctico	1	2	131	488	540,01
2130	25093	Hoja de Cálculo II	Sección Inscripción	Teorico Práctico	1	2	131	149	148,80
2130	34809	Seguridad de la información	Sección Inscripción	Teorico Práctico	1	2	131	147	147,66
2130	4075	Pensamiento Sistémico	Sección Inscripción	Teorico	1	4	131	100	131,61
2130	5103	Práctica profesional II	Sección Inscripción	Práctico	2	8	262	4	126,67
2130	5103	Práctica Profesional II ISist	Sección Inscripción	Práctico	2	8	262	4	126,67
2130	33700	Bases de Datos	Sección Inscripción	Teorico Práctico	2	12	262	52	124,63
2130	34803	Gestión de innovación en TI	Sección Inscripción	Teorico Práctico	2	9	262	81	111,07
2130	5100	Proyecto social universitario	Sección Inscripción	Práctico	2	22	262	102	104,29
2130	5100	Pry. Social Univers. ISist	Sección Inscripción	Práctico	2	22	262	102	104,29
2130	34816	Gestión financiera en TI	Sección Inscripción	Teorico Práctico	1	3	131	106	74,54
2130	34807	Desarrollo web	Sección Inscripción	Teorico Práctico	2	12	262	55	69,59
2130	34806	Fundamentos Ingeniería de SW	Sección Inscripción	Teorico Práctico	1	4	131	56	67,64
2130	33699	Programación avanzada	Sección Inscripción	Teorico Práctico	1	4	131	200	65,70
2130	34801	Teoría de la computación	Sección Inscripción	Teorico Práctico	1	4	131	69	63,72
2130	22469	Trabajo de Grado ISistemas	Sección Inscripción	Teorico Práctico	6	8	786	132	52,20
2130	4190	Comunicaciones y Redes	Sección Inscripción	Teorico Práctico	2	12	262	252	50,12
2130	34805	Análisis y diseño de software	Sección Inscripción	Teorico Práctico	1	4	131	34	48,96
2130	4085	Sistemas Operativos	Sección Inscripción	Teorico Práctico	1	3	131	70	48,50
2130	22586	Intro. Sistemas Distribuidos	Sección Inscripción	Teorico Práctico	1	2	131	63	48,05
2130	31339	Planeacion del Proyecto Final	Sección Inscripción	Teorico Práctico	1	2	131	38	44,43
2130	4070	Ingeniería de Software	Sección Inscripción	Teorico	1	4	131	12	38,42
2130	4082	Sistemas de Información	Sección Inscripción	Teorico Práctico	1	4	131	42	38,30
2130	4185	Arquitectura de Software	Sección Inscripción	Teorico Práctico	1	4	131	26	38,24
2130	4064	Gerencia y Gestión Informática	Sección Inscripción	Teorico	1	3	131	16	34,01
2130	4196	Estructuras de Datos	Sección Inscripción	Teorico Práctico	1	4	131	48	31,74
2130	5102	Práctica Profesional ISistemas	Sección Inscripción	Práctico	2	12	262	56	31,35
2130	5102	Práctica profesional Sistemas	Sección Inscripción	Práctico	2	12	262	56	31,35
2130	19589	Hoja de cálculo nivel avanzado	Sección Inscripción	Teorico Práctico	1	2	131	24	28,54
2130	4216	Procesamiento Imag. Satélite	Sección Inscripción	Teorico Práctico	1	4	131	7	23,05
Total					63	235	8122	4073	4.039,40

Figura 33. Informe de predicción

6.3. Validación del negocio y tableros de control

Para la validación de los tableros de control y la información de los resultados obtenidos a través de los modelos predictivos, con los cuales se busca brindar las mejores recomendaciones de las asignaturas a inscribir, se contó con el acompañamiento permanente del analista asignado por parte de la dirección de carrera de la facultad de ingeniería, el cual basado en su experiencia y conocimientos del tema, validaba que el modelo predictivo se aproximara a los datos reales.

El proceso del desarrollo de este trabajo inicia con la validación y verificación de la información entregada por la dirección de carrera, realizando una exploración previa para determinar la calidad e importancia de los datos que se encuentran relacionados directamente con el objetivo del negocio, construyendo la primera vista minable de los datos limpios, la que posteriormente se presenta como el informe de inscripciones generales y detalladas y el informe de notas y retiros. Dando por cumplido la veracidad de esta información y comprobando su utilidad en el desarrollo de los modelos analíticos.

El analista encargado que nos acompañó durante este proceso, estuvo a cargo de validar cada tablero y hacer las recomendaciones y cambios sobre los tableros según su criterio sobre que le aportaba mas valor a estos tableros, por lo que nos solicita adicionar nuevos filtros para conocer la descripción y ciclo de cada asignatura que se desea consultar, este cambio se reali-

za en todos los tableros para facilitar esta información en todo momento que el cliente la necesite, también solicita tener un filtro de suma de cantidad de estudiantes en cada árbol de flujo de las variables principales, asociar la cantidad de inscritos a los docentes que van a dictar las clases, esto con el fin de conocer la capacidad de atención de los docentes y como ultima solicitud nos pide tener en el tablero de control las gráficas de retiros y motivos de los estudiantes por asignatura.

Para el tablero de predicciones se valida el asertividad en cuanto a la inscripción de asignaturas en periodos anteriores con el fin de comprobar los resultados, los periodos de validación están comprendidos entre el ciclo "1810" y "1830"

1	* Los datos presentados corresponden solo a estudiantes c						
2	ASIGNATURAS NÚCLEO DE FORMACIÓN FUNDAMENTA						
3	ID	Cr	ASIGNATURA	Departamento	1810	1820	1830
4	004183	2	Análisis y Diseño O.O.	Ingeniería de Sistemas	39	-	45
5	004186	3	Bases de Datos	Ingeniería de Sistemas	48	-	50
6	004071	2	Introducción Ing. de Sistemas	Ingeniería de Sistemas	83	-	40
7	004206	3	Pensamiento Algorítmico	Ingeniería de Sistemas	91	-	53
8	004210	3	Programación de Computadores	Ingeniería de Sistemas	50	-	78
9	004204	3	Lenguajes de Programación	Ingeniería de Sistemas	29	-	28
10	004208	2	Prog. Orientada Objetos	Ingeniería de Sistemas	57	-	55
11	004064	2	Gerencia y Gestión Informática	Ingeniería de Sistemas	28	-	35
12	004070	4	Ingeniería de Software	Ingeniería de Sistemas	31	-	34
13	003194	2	Análisis de Algoritmos	Ingeniería de Sistemas	40	-	43
14	004185	3	Arquitectura de Software	Ingeniería de Sistemas	22	-	27
15	004190	4	Comunicaciones y Redes	Ingeniería de Sistemas	33	-	41
16	004196	3	Estructuras de Datos	Ingeniería de Sistemas	42	-	55

1810	25093	Hoja de Cálculo II	Sección Inscripción	Teorico Práctico	1	2	131	325	300,58
1810	19589	Hoja de cálculo nivel avanzado	Sección Inscripción	Teorico Práctico	1	2	131	21	31,56
1810	19588	Hoja de cálculo nivel básico	Sección Inscripción	Teorico Práctico	1	2	131	396	375,80
1810	4070	Ingeniería de Software	Sección Inscripción	Teorico	1	4	131	30	28,34
1810	4189	Intro. Computacion Grafica	Seccion Inscripcion	Teorico Practico	1	4	131	18	48,54
1810	22586	Intro. Sistemas Distribuidos	Sección Inscripción	Teorico Práctico	1	2	131	30	36,53
1810	4197	Introducción a Bases de Datos	Sección Inscripción	Teorico Práctico	1	2	131	32	41,71
1810	22587	Introducción a la Computación	Sección Inscripción	Teorico Práctico	1	2	131	20	26,61
1810	4071	Introducción Ing. de Sistemas	Sección Inscripción	Teorico	1	4	131	91	72,07

Figura 34. Validación de las inscripciones de asignaturas vs la predicción del modelo de analítica periodo 1810

ASIGNATURAS NÚCLEO DE FORMACIÓN FUNDAMENTAL

ID	Cr	ASIGNATURA	Departamento	1810	1820	1830
004183	2	Análisis y Diseño O.O.	Ingeniería de Sistemas	39	-	45
004186	3	Bases de Datos	Ingeniería de Sistemas	48	-	50
004071	2	Introducción Ing. de Sistemas	Ingeniería de Sistemas	83	-	40
004206	3	Pensamiento Algorítmico	Ingeniería de Sistemas	91	-	53
004210	3	Programación de Computadores	Ingeniería de Sistemas	50	-	78
004204	3	Lenguajes de Programación	Ingeniería de Sistemas	29	-	28
004208	2	Prog. Orientada Objetos	Ingeniería de Sistemas	57	-	55
004064	2	Gerencia y Gestión Informática	Ingeniería de Sistemas	28	-	35
004070	4	Ingeniería de Software	Ingeniería de Sistemas	31	-	34
003194	2	Análisis de Algoritmos	Ingeniería de Sistemas	40	-	43
004185	3	Arquitectura de Software	Ingeniería de Sistemas	22	-	27
004190	4	Comunicaciones y Redes	Ingeniería de Sistemas	33	-	41

Ciclo	IDCurso	Descripción	TipoClase	Componente	Sesion	NroHoras	Dias	TotalInscripciones	
1810	4055	Administración Bases de Datos	Sección Inscripción	Teorico Práctico	1	3	131	20	48,69
1810	4181	Administración Básica Linux	Sección Inscripción	Teorico Práctico	1	2	131	21	30,22
1810	3194	Análisis de Algoritmos	Sección Inscripción	Teorico Práctico	1	4	131	40	48,03
1810	4183	Análisis y Diseño O.O.	Sección Inscripción	Teorico Práctico	1	4	131	41	60,07
1810	4185	Arquitectura de Software	Sección Inscripción	Teorico Práctico	1	4	131	21	20,07
1810	4186	Bases de Datos	Sección Inscripción	Teorico Práctico	1	4	131	49	50,67
1810	4190	Comunicaciones y Redes	Sección Inscripción	Teorico Práctico	2	12	262	68	93,70
1810	4192	Desarrollo Multimedial	Sección Inscripción	Teorico Práctico	1	4	131	17	48,19
1810	4196	Estructuras de Datos	Sección Inscripción	Teorico Práctico	1	4	131	46	18,27
1810	4198	Fundamentos Redes e Internet	Sección Inscripción	Teorico Práctico	1	2	131	18	26,91
1810	4064	Gerencia y Gestión Informática	Sección Inscripción	Teorico	1	4	131	28	35,40

Figura 35. Validación de las inscripciones de asignaturas vs la predicción del modelo de analítica periodo 1810

ASIGNATURAS NÚCLEO DE FORMACIÓN FUNDAMENTA

ID	Cr	ASIGNATURA	Departamento	1810	1820	1830
004183	2	Análisis y Diseño O.O.	Ingeniería de Sistemas	39	-	45
004186	3	Bases de Datos	Ingeniería de Sistemas	48	-	50
004071	2	Introducción Ing. de Sistemas	Ingeniería de Sistemas	83	-	40
004206	3	Pensamiento Algorítmico	Ingeniería de Sistemas	91	-	53
004210	3	Programación de Computadores	Ingeniería de Sistemas	50	-	78
004204	3	Lenguajes de Programación	Ingeniería de Sistemas	29	-	28
004208	2	Prog. Orientada Objetos	Ingeniería de Sistemas	57	-	55
004064	2	Gerencia y Gestión Informática	Ingeniería de Sistemas	28	-	35
004070	4	Ingeniería de Software	Ingeniería de Sistemas	31	-	34
003194	2	Análisis de Algoritmos	Ingeniería de Sistemas	40	-	43
004185	3	Arquitectura de Software	Ingeniería de Sistemas	22	-	27
004190	4	Comunicaciones y Redes	Ingeniería de Sistemas	33	-	41
004196	3	Estructuras de Datos	Ingeniería de Sistemas	42	-	55

1810	4071	Introducción Ing. de Sistemas	Sección Inscripción	Teorico	1	4	131	91	72,07
1810	4079	Introducción Seg. Informática	Sección Inscripción	Teorico	1	3	131	13	46,02
1810	4204	Lenguajes de Programación	Sección Inscripción	Teorico Práctico	1	4	131	29	36,41
1810	18805	Matemática Computacional	Sección Inscripción	Teorico	1	4	131	9	37,24
1810	4072	Minería de Datos	Sección Inscripción	Teorico	2	6	262	48	93,62
1810	4205	Negocios en Internet	Sección Inscripción	Teorico Práctico	1	3	131	6	49,20
1810	4206	Pensamiento Algorítmico	Sección Inscripción	Teorico Práctico	1	4	131	352	323,96
1810	4075	Pensamiento Sistémico	Sección Inscripción	Teorico	2	7	262	41	88,76
1810	31339	Planeacion del Proyecto Final	Sección Inscripción	Teorico Práctico	1	2	131	18	54,91

Figura 36. Validación de las inscripciones de asignaturas vs la predicción del modelo de analítica periodo 1810

Posteriormente se evidencia que la aproximación de la predicción ejecutada, esta alineada al cumplimento de los objetivos específicos planteados, y se obtiene la validación con el cliente el cual nos informa de la importancia de los datos que estamos presentando, se solicitan nuevos ajustes con los cuales se da cumplimento a los objetivos y lo requerido por el cliente final.

Resumen de la reunión						
Número total de participantes	3					
Título de la reunión	Validación Tablero de Control					
Hora de inicio de la reunión	12/11/2021, 9:38:34 a. m.					
Hora de finalización de la reunión	12/11/2021, 11:09:25 a. m.					
ID. de reunión	6744d038-3862-425c-9910-581c9e99548b					
Nombre completo	Hora en la q	Hora de salic	Duración	Correo elect	Rol	Id. de participante (UPN)
Pablo Miguel Nuñez Gonzalez	12/11/2021,	12/11/2021,	1 h 30 min	pablo.nunez	Organizador	pablo.nunez@javeriana.edu.co
Nelson Jovanny Rodriguez Bohorquez	12/11/2021,	12/11/2021,	1 h 30 min	ne.rodriguez	Asistente	ne.rodriguez@javeriana.edu.co
Sara Isabela Vergara Aguilar	12/11/2021,	12/11/2021,	1 h 30 min	vergara.sara	Asistente	vergara.sarai@javeriana.edu.co

Figura 37. Evidencia de las sesiones de validación de los tableros de control

7. CONCLUSIONES Y TRABAJOS FUTUROS

7.1. Trabajos futuros

Teniendo en cuenta la definición del alcance del proyecto planteado, el levantamiento de los requisitos y las entrevistas que se tuvieron con los interesados del Departamento de Ingeniería de Sistemas, se confirma el interés de continuar con el proyecto, ya que en la implementación de este sistema se estaría dando buen uso de la información almacenada y ayudaría a reducir el tiempo de ejecución de este proceso, por lo que se proponen unas mejoras sobre el trabajo realizado:

- Realizar la conexión entre el ERP (Peoplesoft) que maneja la universidad directamente como fuente principal de recolección de los datos utilizados en el algoritmo de predicción sin tener que depender del uso de archivos de texto planos (.xslx) para correr el modelo planteado
- Realizar un esquema de arquitectura exclusivo para el proceso de ETL del conjunto de datos que se debe manejar
- Incluir los tableros de visualización creados en Power BI como un módulo extra dentro del ARP que maneja la universidad, para asegurase de que los encargados de este proceso tengan acceso a toda la información necesaria para mejorar la toma de decisiones.
- Generar reportes a partir de las consultas realizadas en los tableros de control que permitan manipular la información resultante conforme a las necesidades del Departamento de Ingeniería de Sistemas
- Realizar un modelo complementario para el análisis de las predicciones de los posibles estudiantes que pierdan, se retiren o aprueben las asignaturas, teniendo en cuenta que sería necesario obtener información personal de cada estudiante.

7.2. Conclusiones

Durante el desarrollo del proyecto se evidenció la importancia de entender el negocio además de sus requisitos particulares. El esquema de arquitectura por procesos permitió identificar las oportunidades de mejora dentro del mismo, además de los actores, datos e información involucrada en la programación de clases para un periodo académico especifico dentro del departamento de ingeniería de Sistemas para el grado académico de pregrado.

Es necesario entender que los datos obtenidos de la identificación de los procesos involucrados en la programación de clases requieren de una trasformación previa antes de ser usados en el diseño de un modelo analítico, por lo tanto, se investigaron técnicas y herramientas para transformar estos datos; siendo el lenguaje de programación Python y algunas de sus librerías especializadas en la gestión de información las opciones seleccionadas. Una vez identificados los datos base se aplicaron sobre ellos herramientas de exploración limpieza e imputación de datos como Pandas y Skilearn. Es importante señalar que se dispuso de la creación de funciones específicas debido a los requerimientos puntuales planteados en los objetivos.

Dentro del proceso de investigación del estado del arte se identificaron soluciones que tenían componentes en común al problema planteado, se usaron y se modificaron algunas de estas características para adaptarlas a la solución final. En las fases iniciales de desarrollo se plantearon entrevistas que permitieron obtener información relacionada con la programación de clases y que permitió principalmente diseñar un diagrama de grafos dirigido para esquematizar el plan de estudios del programa de ingeniería de sistemas, a través del uso de este diagrama se identificaron variables significativas para el desarrollo del modelo analítico, como por ejemplo los nodos (asignaturas) que son más importantes que otros, debido a que se consideran “cuellos de botella”.

Posterior a la creación de la vista minable sobre la cual realizar el diseño del modelo analítico, se decidió investigar acerca de las opciones disponibles para atender de forma más eficiente la predicción de la variable objetivo “Número de estudiantes inscritos para un periodo académico especifico.”, se dispuso del uso de las técnicas de SVM (Support Vector Machine) y Redes Neuronales para diseñar este modelo.

El conjunto de datos para entrenamiento y pruebas se dividió en periodos académicos específicos, es importante señalar que se usaron tres periodos para entrenamiento y uno para pruebas y así hasta formar 4 conjuntos de datos para experimentación, vale la pena resaltar que los tres primeros conjuntos corresponden a periodos de tiempo regulares en donde no se había presentado la emergencia social y sanitaria por Covid-19, por tanto los modelos entrenados usando periodos de tiempo posteriores al 2019 no son tan eficientes para predecir el número de estudiantes inscritos como los que fueron entrenados con datos previos a este periodo de tiempo.

Los modelos que tienen mejor rendimiento son los que fueron desarrollados usando SVM claro esta para periodos anteriores a la emergencia social y sanitaria por Covid-19, para el periodo de tiempo comprendido entre el año 2020 y el año 2021, el modelo que presenta mejor rendimiento es el que fue desarrollado usando redes Neuronales, claro está que el error

cuadrático medio es superior a cualquiera de los modelos SVM desarrollados para periodos académicos previos.

De los modelos que fueron desarrollados usando SVM el de mejor eficiencia promedio presenta para los 3 conjuntos de datos iniciales, fue el que se diseñó usando un núcleo linear, su RMSE promedio es de 26.66.

De los modelos que fueron desarrollados usando RN el de mejor eficiencia promedio presenta para los 3 conjuntos de datos iniciales, fue el que se diseñó usando 3 capas intermedias cada una con 50 neuronas y 300 épocas y un bacht size de 32, su RMSE promedio es de 35.

Entendiendo que la métrica usada para evaluar el desempeño de los modelos nos muestra una variabilidad en el mejor de los casos de 26.66, es necesario dar contexto a esta cifra en términos de negocio, en donde un error en la estimación de esta magnitud se debe diferenciar por el tamaño de los grupos, es decir si se trata de grupos grandes de estudiantes no representaría una falla importante y si se trata de grupos pequeños hay que revisar de forma específica las situaciones presentadas, porque esto no quiere decir que el modelo sobreestimará ampliamente el número de estudiantes para grupos pequeños, sino que por el contrario tratará de ajustarse lo mejor posible tanto a grupos grandes como a los más pequeños.

Gracias a la definición de la estructura de experimentación planteada por periodos académicos se identificó que los modelos diseñados son sensibles a la variabilidad de los datos de entrenamiento, razón por la cual los modelos piloto probados y entrenados con el último conjunto de datos (programación de clases en periodo de pandemia) presentaron una diferencia considerable en su desempeño respecto de los anteriores, hay que entender que la planeación de clases para un periodo atípico es diferente debido a que se consideran nuevos factores, para este caso particular la presencialidad y el modo de enseñanza dispuesto para atender las sesiones de clase. Además, hay otros factores clave, como por ejemplo la estructura cambiante del plan de estudios, es decir, momentos de transición de un plan anterior a un plan nuevo.

Gracias al proceso de diseño de los modelos planteados en este trabajo, se crearon herramientas que permitieron identificar en donde, en que formato, la confiabilidad, la frecuencia de actualización, la veracidad, la fuente y la exactitud de los datos necesarios para desarrollar un modelo piloto para la predicción de los estudiantes inscritos en una asignatura para un periodo específico.

8. REFERENCIAS

[1] C. E. y. T. M. Univesity, «Business intelligence,» p. 12.

[2] X. C. y. K. Siau, «Business Analytics/Business Intelligence and IT Infraestructure Impact on Organizational Agility,» *Journal of Organizational and End User Computing* , vol. 32, nº 4, pp. 138-161, 2020.

[3] A. S. y. S. S. P. Vassiliadis, «Conceptual Modeling for ETL Processes,» *National Technical University of Athens.*

[4] A. V. y. E. Z. J. Awiti, «From Conceptual to Logical ETL Design Using BPMN and Relational Algebra,» *Big Data Analytics and Knowledge Discovery,* vol. 11708, pp. 299-309, 2019.

[5] A. F. y. H. C. A. O. E. Quinteros, «Construcción de una Vista Minable para aplicar Mineria de Datos Secuenciales Temporales,» p. 10.

[6] M. A. A. ,. y. E. W. N. Ashish Kumar Jha, «A note on big data analytics capability development in supply chain,» *Decision Support Systems,* vol. 138, p. 113382, 2020.

[7] C. Camacho, «Regresión Lineal Simple,» *Universidad de Sevilla.*

[8] B. Lipo Wang - Springer, «Support Vector Machines: Theory and Applications,» 2005.

[9] J. M. A. y. A. Iqbal, «Introduction to Support Vector Machines and Kernel Methods,» p. 10.

[10] D. Y. Singh y A. S. Chauhan, «Neural Networks in data Mining,» *Journal of Theoretical and Applied Information Technology,* p. 6, 2005.

[11] H. Lu, S. R y H. Liu, «Effective Data Mining Using Neural Networks,» *IEEE Transactions on Knowledge and Data,* vol. 8, pp. 957-961, 1996.

[12] R. W. y. J. Hipp, «CRISP-DM Towards a Standard Process Model for Data Mining,» p. 11.

[13] H. W. D. S. y. S. I. S. Huber, «DMME: Data mining methodology for engineering applications - a holistic extension to the CRISP-DM model,» *Procedia CIRP*, vol. 79, pp. 403-408, 2019.

[14] T. X. y. M. v. d. S. J. Xu, «Personalized Course Sequence Recommendations,» *IEEE Trans. Signal Process*, vol. 64, nº 20, pp. 5340 - 5352, 2016.

[15] J. X. O. A. y. M. v. d. S. Y. Meier, «Predicting Grades,» *IEEE Trans. Signal Process.*, vol. 64, nº 4, pp. 959-972, 2016.

[16] R. A. e. al, «Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos,» p. 8.

[17] A. R. P. N. D. G. R. D. S. G. Instituto Politécnico Nacional, «Construcción e implementación de un modelo para predecir el rendimiento académico de estudiantes universitarios mediante el algoritmo de Naive Bayes,» *DSE*, nº 19, 2019.

[18] R. Molontay, N. Horvath y y. M. S. D. Szekrenyes, «Characterizing Curriculum Prerequisite Networks by a Student Flow Approach,» *IEEE Transactions on Learnin Technologies*, vol. 13, pp. 491-501, 2020.

[19] E. U. y. D. D. B. Sen, «Predicting and analyzing secondary education placement- test scores: A Data Mining Approach,» *Expert Systems with Application*, vol. 39, nº 10, pp. 9468-9476, 2012.

[20] A. D. R. H. M. C. y. S. P. R. R. Salazar, «Procederes de regresión lineal como soluciones al problema de la comparación de métodos. II. Errores analíticos constantes pero diferentes,» p. 13.

[21] A. E. W. C. S. y. R. G. B. A. S. Lan, «Sparse Factor Analysis for Learning and Content Analytics,» p. 50.

[22] E. Z. J. N. M. y. J. T. Z. El Akkaoui, «ABPMN - Based Design and Maintenance Framework for ETL Processes,» *International Journal odf Data Warehousing and Mining*, vol. 9, nº 3, pp. 46-72, 2013.

[23] N. C. y. B. Scholkopf, «Support Vector Machines and Kernel Methods: The New Generation of Learning Machines,» p. 12.

[24] C. Camacho, «Regresion Lineal Simple,» p. 44.

[25] A. M. E. F. M. R. R. y. C. M. S. S. Buckl, «Using Enterprise Architecture Management Patterns to Complement TOGAF,» *IEEE International Enterprise Distributed Object Computing Conference,* pp. 34-41, 2009.

[26] G. V. W. H. D. y. M. W. N. Ain, «Two decades of research on business intelligence system adoption, utilization and success – A systematic literature review,» *Decision Support Systems,* vol. 125, p. 113, 2019.

[27] M. Arif, K. A. Alam y M. Hussain, «Application of Data Mining Using Artificial Neural Network: Survey,» *Internacional Journal of Database Theory and Application,* vol. 8, pp. 245-270, 2015.