



UNIVERSIDAD JORGE TADEO LOZANO
FACULTAD DE CIENCIAS NATURALES E INGENIERÍA
MAESTRÍA EN INGENIERÍA Y ANALÍTICA DE DATOS
MIAD

ANÁLISIS DESCRIPTIVO Y PREDICTIVO
CON LA INFORMACIÓN DE
CARACTERIZACIÓN DE LOS ESTUDIANTES
UNAD

Mauricio Leon Ayala
MSc(c) Maestría Ingeniería y Analítica de Datos

Universidad Jorge Tadeo Lozano
Facultad de Ciencias Naturales e Ingeniería
Bogotá, Colombia
December 15, 2020

UNIVERSIDAD JORGE TADEO LOZANO

FACULTAD DE CIENCIAS NATURALES E INGENIERÍA
MAESTRÍA EN INGENIERÍA Y ANALÍTICA DE DATOS
MIAD

ANÁLISIS DESCRIPTIVO Y PRESCRIPTIVO
CON LA INFORMACIÓN DE
CARACTERIZACIÓN DE LOS ESTUDIANTES
UNAD

TRABAJO DE GRADO PRESENTADO COMO REQUISITO PARA OPTAR AL TÍTULO EN :
Magíster en Ingeniería y Analítica de Datos

Director:
PhD. Olmer Garcia Bedoya ¹

Universidad Jorge Tadeo Lozano
Facultad de Ciencias Naturales e Ingeniería
Bogotá, Colombia December 15, 2020

¹Docente de tiempo completo email olmer.garciab@utadeo.edu.co

Dedicatoria...

A mi Esposa:
Carmen Jimena Ortiz Delgado,

e Hijos:

Camilo Andrés León Ortiz y
Maria José León Ortiz
Por su Apoyo, Comprensión y Paciencia...

El crecimiento nunca es por mera casualidad; es el Resultado de fuerzas que trabajan juntas”.

James Cash Penney

Agradecimientos

Infinitas gracias doy a Dios Por esta Oportunidad que la vida me ofrece, por darme fuerza y capacidad para culminar esta etapa tan importante la cual me llena de orgullo no solamente a mi si no a los mios también.

Como no agradecer a mis padres y hermanas la cuales siempre me brindaron su apoyo durante este proceso, a mis compañeros de estudio con los cuales se compartieron momentos de dificultad, pero que fueron sorteados correctamente y que ayudan ahora a ver reflejado ese esfuerzo.

Resumen

El análisis de la información generada por los estudiantes a través de la plataforma MOODLE, se convierte en un opción muy importante a la hora de conocer patrones académicos, con el objetivo de establecer estrategias que conlleven a una mejor interacción entre los actores involucrados en este proceso, ayudando a determinar las mejores opciones en cuanto a franjas horarias y días de la semana por ejemplo, para el desarrollo de actividades escolares dentro de las cuales se puede incrementar el nivel de acceso por parte de los estudiantes y de esta forma mejorar la calidad de los cursos ofrecidos y el aprendizaje.

Estos datos son analizados desde las perspectiva de las analíticas descriptiva que permite conocer la situación actual con la información generada por cada estudiante dentro de la plataforma MOODLE, la analítica predictiva que muestra que puede suceder de acuerdo a la información anterior y prescriptiva con la cual se puede sugerir acciones a tener en cuenta, convirtiendo así estos datos en información de valor con respecto, tanto al conocimiento de los datos históricos, como la generación de posibles escenarios que pueden suceder con base en lo que ha venido pasando en periodos académicos anteriores, esto con el fin de encontrar patrones comportamentales desde la perspectiva académica para mejorar la interacción entre las partes.

Los sistemas de información *Learnig Management System* (LMS) han evolucionado notoriamente con el paso del tiempo, y esto a su vez a permitido un crecimiento bastante importante en la educación virtual tanto en Colombia como a nivel mundial. Teniendo en cuenta los análisis realizados en este trabajo se puede tener un mayor conocimiento acerca de las características de los estudiantes y generar pronósticos con el objetivo de mejorar la interacción académica entre alumnos y la comunidad educativa en general.

Contents

Agradecimientos	3
Resumen	4
1 Introducción	10
2 Estado del Arte	13
3 Marco teórico	16
3.1 Análisis de Series de Tiempo	16
3.2 Redes Neuronales LSTM	16
3.3 Herramientas de Visualización	18
4 Análisis Descriptivo	19
4.1 Análisis de los logs	19
4.2 Análisis de la herramienta MOODLE	19
4.3 Cargue de información suministrada de logs suministrada por la Unad	21
4.4 Generación de Sentencias SQL para el para la construcción de los indicadores propuestos.	22
4.4.1 Tiempo Total gastado por estudiante en el CMS Tp	22
4.4.2 Tiempo promedio que un estudiante en particular gasta en una activi- dad particular Ef	23
4.4.3 Número de visitas a los materiales del curso Vm	24
4.4.4 Número de logs generados en los días hábiles de la semana Ls	25
4.4.5 Número de logs generados en los fines de semana Lf	26
4.4.6 Radio IP Pa	27
4.4.7 Número de sesiones promedio por semana	28
4.4.8 Tiempo promedio semanal en foros	29
4.4.9 Número de diferentes días que el estudiante se logéa en el campus . .	30
4.4.10 Numero de eventos en que modifica el sistema en una semana	31
4.4.11 Cantidad de cursos a los que el estudiante accede durante una semana	32
4.5 Análisis sociodemografico	33
4.6 Correlación de las fuentes	34
5 Análisis Predictivo	46
5.1 Análisis de la serie de tiempo	47
5.2 Modelo de Redes Neuronales	49

6	Discusión	53
	Bibliografía	55

List of Figures

1.1	Tipos de Analítica	11
3.1	RNN - Redes Neuronales Recurrentes	17
3.2	LSTM - Redes Neuronales Recurrentes LSTM	17
4.1	Modelo Relacional Queos	21
4.2	información Cargada en MYSQL	22
4.3	Tiempo Total gastado por estudiante en el CMS	23
4.4	Tiempo promedio que un estudiante en particular gasta en una actividad particular Ef	24
4.5	Número de visitas a los materiales del curso Vm	25
4.6	Número de logs generados en los días hábiles de la semana Ls	26
4.7	Número de logs generados en los fines de semana Lf	27
4.8	Radio IP Pa	28
4.9	Número de sesiones promedio por semana	29
4.10	Tiempo promedio semanal en foros	30
4.11	Número de diferentes días que el estudiante se loguéa en el campus	31
4.12	Numero de eventos en que modifica el sistema en una semana	32
4.13	Cantidad de cursos a los que el estudiante accede durante una semana	33
4.14	Ingresos a la Plataforma por Genero	35
4.15	Tiempo en la Plataforma Según Dependencia Económica	36
4.16	Promedio de Accesos a la Plataforma por Zona de Residencia	37
4.17	Ingresos a la Plataforma según el Estado Civil	38
4.18	Medios de Acceso de los Estudiantes	39
4.19	Promedio de Accesos Por Edad	40
4.20	Porcentaje de Accesos Por Programa Académico	41
4.21	Tendencia por franjas horarias y Genero	42
4.22	Tiempo por franjas horarias y Genero	43
4.23	Tendencia por Día de la Semana y Genero	44
4.24	Tiempo en la Plataforma por Día de la Semana y Genero	45
5.1	Gráfica de Series de Tiempo por Variable	47
5.2	Prueba de Causalidad	48
5.3	Gráfica de Correlación	49
5.4	Gráfica de Separación de Datos	50
5.5	Arquitectura del Modelo	51
5.6	Gráfica de Predicción	52

6.1	Promedio de Accesos a la Plataforma por Genero y Día de la Semana	54
-----	---	----

List of Tables

2.1	Comparativo Modelos de Estudios Por Tipo de Algoritmo	15
4.1	Input indicators (per week)	34

Chapter 1

Introducción

Para la Universidad Nacional Abierta y a Distancia (UNAD) seccional Antioquia, es muy importante analizar el nivel de deserción escolar que se esta presentando en la misma, así como los motivos principales por los cuales se esta generando. Este indicador es de vital importancia a la hora de tomar decisiones al respecto sobre los programas, metodologías, herramientas utilizadas por parte de la universidad así como el plantear alternativas de solución para estudiantes que aunque quisieran se les puede hacer difícil continuar con sus estudios.

Como es de conocimiento general, el acceso a la educación virtual en Colombia esta aumentando rápidamente, lo que lleva a prestar especial atención a estos procesos de formación, para los cuales se requiere hacer diferentes tipos de investigaciones como lo menciona el artículo "Elementos para la comprensión del fenómeno de la deserción universitaria en Colombia. Más allá de las mediciones" [1], y por esto se hace importante conocer la información histórica y los fenómenos que generan estos altos índices con el fin de analizar la misma, generando minería de datos con la cual se pueda tener un punto de partida para encontrar posibles soluciones a este problema que tanto aqueja a la comunidad estudiantil.

Si tenemos en cuenta el crecimiento y aplicación que esta teniendo la analítica de datos, podremos observar grandes alternativas para encontrar patrones que permitan hacer análisis con los cuales se puedan generar pronósticos basados en la información existente, teniendo en cuenta la historia y el comportamiento en términos académicos de los estudiantes. Para este caso las líneas de tiempo multivariadas en la cuales se incluyen varios indicadores para un mejor pronóstico, como se describe en el artículo "Relación entre los modelos de series de tiempo univariadas con los modelos de series multivariadas" [2], ayudan a predecir cual puede ser el comportamiento que a futuro que pueden tener los estudiantes con el fin de determinar políticas educativas que permitan mejorar el rendimiento académico en conjunto.

Existen tres tipos de análisis que se pueden hacer a la información, estos corresponden a la analítica descriptiva, analítica predictiva y analítica prescriptiva, cada una tiene un valor muy importante al momento de realizar minería de datos hasta la predicción y prescripción de la mismas. La analítica descriptiva permite conocer el estado actual de las cosas en un estado preliminar, con el fin de obtener información importante para su posterior análisis. Después de hacer el análisis descriptivo podemos pasar a la analítica predictiva que ofrece la posibilidad de saber que puede suceder en un momento determinado según la información

con la que se cuente, y la analítica prescriptiva brinda la posibilidad de tomar decisiones determinando que puede pasar en cada una de ellas, usando en la mayoría de ocasiones técnicas de simulación. El valor de cada una está directamente ligado a las capacidades que ofrecen las mismas, teniendo en cuenta que de esto depende el grado de complejidad de las mismas para su desarrollo e implementación.

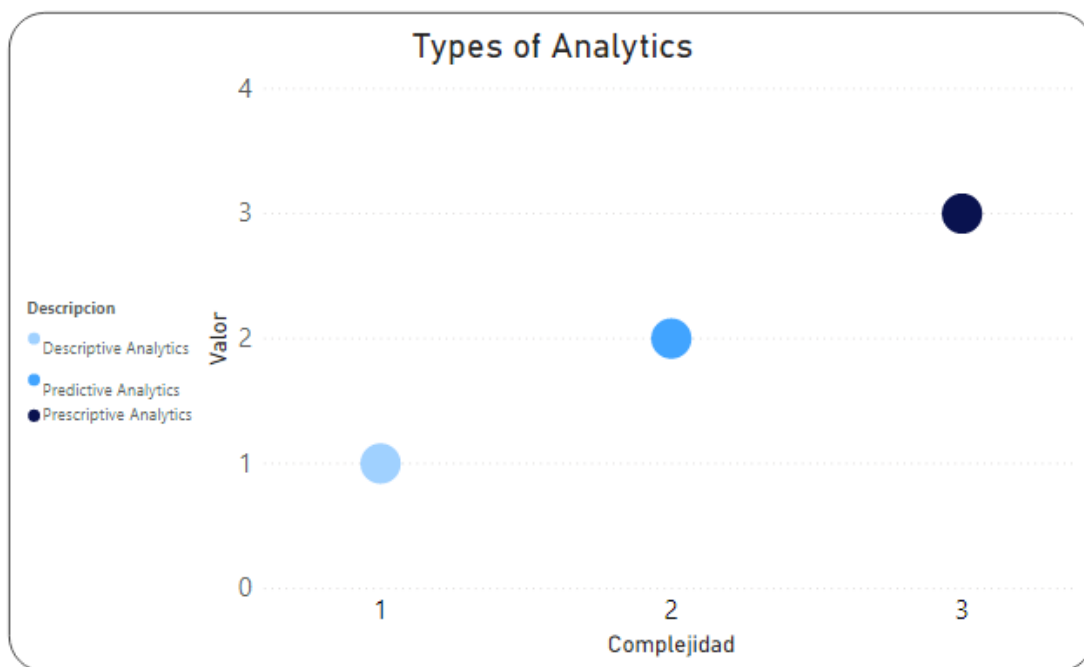


Figure 1.1: Tipos de Analítica

En la figura 1.1 se puede observar como los tres tipos de analíticas están relacionados de forma directa, teniendo como base las preguntas: ¿qué ha ocurrido?, ¿qué va a pasar? y cuál es la mejor estrategia, siempre en relación con los hechos ocurridos con la información materia de estudio.

En este proyecto es propuesto que el análisis se haga con la información de los datos generados a partir de los LOGs que los estudiantes tengan en la plataforma MOODLE y las acciones que se realicen allí por parte de los mismos, posteriormente se generan 22 indicadores que pueden ser relevantes para la toma de decisiones y con base en estos y en encuestas de caracterización se desarrollan modelos analíticos para predecir el comportamiento futuro de los mismos, y su vez proveer una plataforma en la cual se tenga acceso a esta información para la toma de decisiones como lo desarrollado a dentro del marco del proyecto de Minciencias para este fin. La herramienta permite hacer el análisis por estudiante y por semana académica, y de igual manera permite conocer a través de clusterización de características similares de los estudiantes y clasificación por zona de residencia (metropolitana y/o no metropolitana) usando algoritmos de K-MEANS, y XG BOOST.

El presente trabajo de grado de maestría se basa en las técnicas de analítica descriptiva y predictiva; desde la analítica descriptiva se obtiene información importante la cual permite conocer a los estudiantes desde la óptica de su desempeño académico en función de las actividades que estos desarrollan sobre la plataforma MOODLE, la cual es el canal oficial

para el desarrollo y avance de las actividades propuestas por los docentes. Adicionalmente, se tiene información sobre las características socio económicas de los mismos con las cuales se puede hacer un mejor perfilamiento de los patrones comportamentales según su propia condición y generar pronósticos de cantidad de conexiones que se puedan generar a futuro.

Con base en este conocimiento se procede a hacer una analítica predictiva a través de series de tiempo multivariadas usando técnicas de redes neuronales recurrentes, la cual permita conocer alternativas en el uso de los horarios y/o actividades que pueden ser propuestas teniendo en cuenta los temas que se deban tratar de acuerdo al aprendizaje planeado para el curso en el cual se deban enfatizar temas o por ejemplo el manejo de foros. Con lo anterior se busca obtener una participación mas activa por parte de los estudiantes teniendo en cuenta que este se basa en los horarios en los cuales se puede obtener un mayor beneficio para el aprendizaje.

Con estos análisis se busca mejorar la interacción escolar que tanto afecta a las entidades educativas pero mas a los estudiantes que al no encontrar en algunos casos este tipo de ayudas se ven obligados a abandonar sus estudios, no solo por el factor económico si no también por los horarios que se pueden cruzar con las actividades que tiene que realizar cotidianamente.

Transversalmente la visualización es una de las técnicas usadas en el desarrollo de este proyecto, con lo cual se busca que desde la parte gráfica se pueda obtener un conocimiento general del comportamiento de los estudiantes según los datos obtenidos desde la herramienta de MOODLE y los indicadores creados para este análisis, así como la encuesta realizada a los estudiantes.

Chapter 2

Estado del Arte

El crecimiento y auge que está teniendo la inteligencia artificial en esta era tecnológica está otorgando la posibilidad de desarrollar proyectos para muchos tipos de necesidades puntuales, a nivel mundial se han desarrollado diferentes estudios buscando respuestas a los diferentes factores que causan esta dificultad en cuanto a términos académicos se refiere. Para la universidad Nacional Abierta y a distancia UNAD, es muy importante este indicador ya que la misma es una de las pioneras en este tipo de educación virtual en Colombia, es por esto que en su estudio, "Una reflexión de la modalidad como solución a los requerimientos del siglo XXI" [3] busca mecanismos tecnológicos para mejorar la calidad de este tipo de educación ofreciendo mejores alternativas tanto a profesores como estudiantes creando entornos de aprendizaje cada vez mas dinámicos, y de la misma manera buscando mejores alternativas en cuanto a acompañamiento necesario y de manera permanente en cuanto a los estudiantes se refiere como lo busca el estudio Entornos Educativos Saludables [4] el cual fue realizado con estudiantes y practicantes de psicología.

Los factores que conllevan a la deserción escolar universitaria tiene múltiples connotaciones, la universidad nacional Daniel Alcides Carrión de España desarrollo una investigación [5] en cual se busca determinar la influencia de los factores socio económicos como las posibles y mayores causas de este fenómeno en esta universidad. Los factores pueden ser variados, para el caso de la universidad Autónoma del Carmen y como se explica en su estudio denominado Reprobación, síntoma de deserción escolar [6] se presenta una relación directa entre el grado de reprobación versus los índices de deserción.

Adicionalmente a los factores antes mencionados y por lo cuales se puede causar la deserción de un estudiante, también se pueden presentar sentimientos de frustración del mismo como se concluye en el estudio: Factores que promueven la deserción del aula virtual [7] al encontrar situaciones como: Los estudiantes se sienten abandonados cuando no obtiene respuesta inmediata a sus correos, no reciben atención personalizada, Se molestan cuando sus trabajos no son revisados con profesionalismo, No admiten el punto de vista de otros, Carecen de habilidades para la lectura de comprensión, No asumen el papel de alumnos y rechazan las observaciones de los tutores, mencionadas en el mismo.

Como parte de estas necesidades también se han desarrollado diferentes análisis desde la perspectiva tecnológica con los cuales se han podido encontrar algunas respuestas como por ejemplo en el artículo publicado por la universidad industrial de Santander con el nombre de: "Comparación de técnicas de minería de datos para identificar indicios de deserción estudiantil, a partir del desempeño académico" [8] en el cual se usaron algoritmos de Árboles

de decisión, regresión logística y Naive Bayes, encontrando que los tres muestran un buen nivel de predicción y la generación de alertas con el fin de mitigar la deserción. Existen también estudios de técnicas de minería de datos en las cuales con las cuales se puede demostrar que las condiciones psicológicas, económicas y de educación son una gran fuente de información y si están son usadas con volúmenes de datos muy grandes como lo expresa el documento: "Data Mining Applied in School Dropout Prediction" [9] se pueden obtener muy buenos resultados en la predicción.

Adicional a los estudios anteriores también existen el uso de técnicas de clusterización con las cuales se busca encontrar las seis clases de grupos de acuerdo al puntaje obtenido tomando como referencia el sistema de The Beck Depression Inventory, el cual consta de 21 preguntas con las cuales determinar estos niveles de depresión que pueden conducir a la deserción académica [10].

Para la universidad pedagógica nacional de Colombia es importante conocer cual puede ser el desempeño académico que pueden tener los estudiantes al momento de ingresar a la universidad ya que este se puede ser un pilar fundamental a la hora de tomar la decisión de abandonar las aulas, teniendo en cuenta condiciones sociodemográficas de los estudiantes como lo son los estudios realizados, los datos de estudio, datos laborales, salario de los padres, datos de vivienda y comodidades de vivienda a los cuales se les aplico técnicas de Machine Learning de acuerdo a lo expuesto en el documento "Técnicas de Machine Learning para la predicción de desempeño académico en el desarrollo del espacio proyectivo del pensamiento espacial" [11] obteniendo como resultado que estas condiciones afectan o mejoran los resultados académicos.

Para el sistema educativo mexicano es muy importante predecir la deserción académica, ya que es uno de los mas grandes problemas que se tiene en cuanto a formación escolar se refiere, la importancia de este estudio radica en que, como se ha mencionado anteriormente los patrones de deserción con los que llegan los estudiantes universitarios se convierten en una de las características que se deben tener en cuenta para un posible abandono futuro, adicionalmente generando otras variables materia de estudio las cuales pueden ser tenidas en cuenta para nuevos estudios, como lo es la posible deserción de acuerdo a la generación a la cual pertenecen los estudiantes como se menciona en el estudio Plataforma para la predicción de deserción escolar utilizando técnicas de inteligencia artificial [12].

Otros estudios realizados sugieren la clasificación de los factores que afectan a los estudiantes según su naturaleza en endógena que son los propios y asociados a la persona como tal, y los de naturaleza exógena que corresponden a las variables propias del entorno en el cual se desenvuelve cada persona. En este caso el estudio "Técnicas de Data Mining para extraer perfiles comportamiento académico y predecir la deserción universitaria" [13], como su nombre lo indica, se aplica técnicas de minería de datos usando los modelos de arboles Chaid que corresponden a una variante de los arboles de decisión, lo cual lleva a sugerir que se deben tener en cuenta factores como el familiar, laboral, social y cultural.

Las series de tiempo se usan en múltiples entornos de analítica como lo podemos ver dentro del estudio "Multi-Sequence LSTM-RNN Deep Learning and Metaheuristics for Electric Load Forecasting" [14], con el cual el departamento de ciencias de la computación y sistemas de ingeniería de la universidad Al Ain de Canadá, analiza el comportamiento del consumo de las redes eléctricas con base en el tiempo, con el fin de determinar cual debería ser la generación de la misma, logrando así optimizar los costos de producción basándose en la historia y promedios de periodos anteriores.

En la tabla 2.1 se clasifican los estudios anteriormente mencionados teniendo como referencia el tipo de algoritmo según el tipo de modelo aplicado para cada caso y la referencia Bibliográfica.

Artículo	Tipo de Algoritmo
Comparación de técnicas de minería de datos para identificar indicios de deserción estudiantil, a partir del desempeño académico. Ref [7]	-Árboles de decisión -regresión logística -Naive Bayes
Data Mining Applied in School Dropout Prediction. Ref [8]	Árboles de decisión
Application Method of Data Mining Using the K means Algorithm for the Determination of Stress Level in High School Students Using the Beck Depression Inventory. Ref [9]	K-means
Técnicas de Machine Learning para la predicción de desempeño académico en el desarrollo del espacio proyectivo del pensamiento espacial. Ref [10]	Cadenas de Markov
Técnicas de Data Mining para extraer perfiles comportamiento académico y predecir la deserción universitaria. Ref [11]	Árboles de decisión
Multi-Sequence LSTM-RNN Deep Learning and Metaheuristics for Electric Load Forecasting. Ref [12]	Árboles CHAID

Table 2.1: Comparativo Modelos de Estudios Por Tipo de Algoritmo

Como se puede observar, los estudios aplican diferentes técnicas de analítica de datos, de aquí la importancia de conocer muy bien las características de los estudiantes para realizar un aporte con el fin de pronosticar cuales pueden ser los mejores horarios para la programación de actividades, teniendo como punto de referencia las actividades a la plataforma en las diferentes franjas horarias. Las fuente de datos para alimentar el modelo, a partir de los registros (Logs), que se extraen de la plataforma MOODLE y que son generados cada vez que el estudiante ingresa y genera algún tipo de actividad dentro de las misma. En un estudio publicado en la revista Australasia Journal of Educational Technology se menciona como la analítica ayuda a predecir si un estudiante puede o no perder un examen basado en técnicas informáticas necesario para la graduación [15].

Esto con la finalidad de usar estos resultados como herramienta de generación de alertas tempranas para la toma de decisiones en búsqueda de mejorar los altos índices de deserción escolar.

Chapter 3

Marco teórico

3.1 Análisis de Series de Tiempo

Las series de tiempo en analítica de datos, hacen referencia al análisis de los datos históricos registrados cronológicamente en un periodo de tiempo determinado y con una frecuencia determinada por ejemplo años, meses, semanas, días, horas, minutos etc. [16]

Las series de tiempo deben tener una característica especial para un mejor análisis y predicción de los datos; la estacionalidad, indica que la serie de tiempo tiene tanto la media y la varianza de los datos analizados constantes en el tiempo y gráficamente se puede observar como conservan las mismas tendencias [16] (cuando la serie no tiene una estacionalidad en algunas de las variables a analizar se puede aplicar un proceso matemático para simular la misma.) [17]

Para el análisis de series de tiempo existen dos grandes tipos de modelos: los modelos arima [18] que son series de tiempo que se usan para predecir una variable con respecto a otra, y los modelos de series multivariadas [19] que son los mas usados en este momento y que permiten hacer el análisis de predicción de una variable con respecto a mas de una variable que pueden tener incidencia en la misma, para este caso de estudio se trabaja con serie de tiempo multivariada.

3.2 Redes Neuronales LSTM

Las Redes neuronales recurrentes (RNN) permiten que la información se mantenga durante el paso de el tiempo o las épocas como es conocido para el trabajo con series de tiempo [20], esta información se mantiene mediante listas temporales, sin embargo esta forma de trabajo es muy costosa en cuanto a procesamiento de maquina sobre todo cuando se trata de altos volúmenes de datos, este proceso se desarrolla a través de capas, las cuales tiene el mismo valor a la hora de desplegar la información, el funcionamiento en términos generales consiste en que cada neurona conecta a su vecina en cuanto a espacio de tiempo se refiere y otorgando dos entradas a cada una y esta ofreciendo una salida única como se muestra en la siguiente figura 3.1 [21]

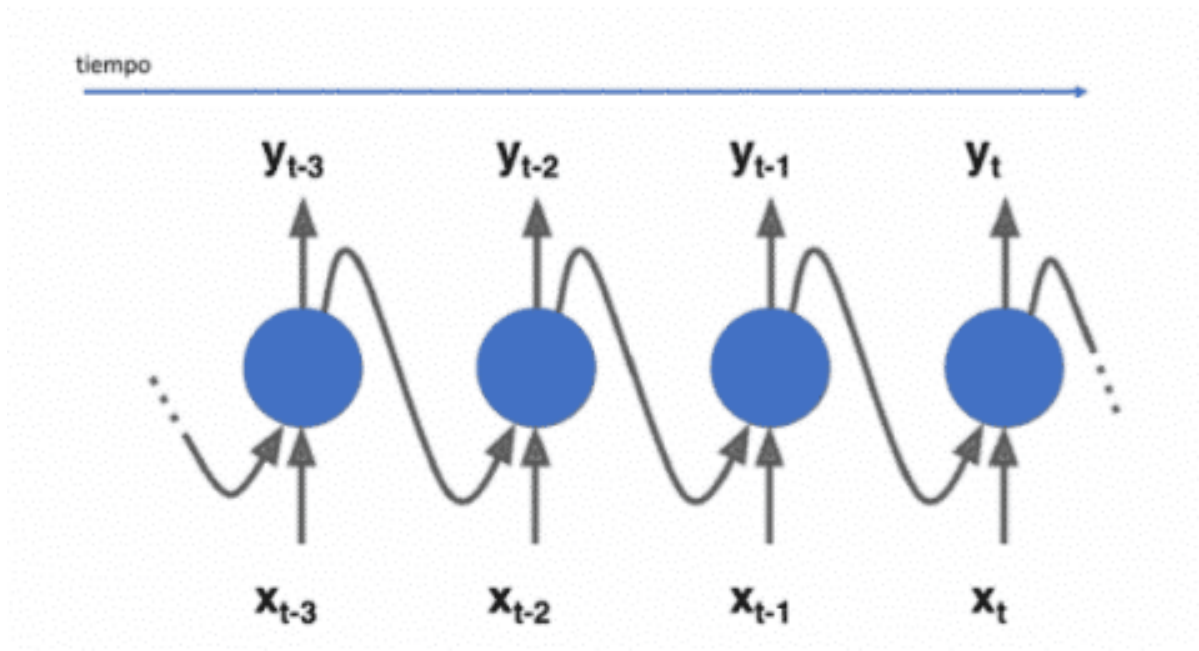


Figure 3.1: RNN - Redes Neuronales Recurrentes

Con el fin de mejorar el rendimiento de las mismas estas pueden usar las redes de memoria largo-corto plazo (LSTM) que por su arquitectura pueden ayudar al procesamiento de grandes volúmenes de datos con mejor calidad en la predicción ya que este modelo cuenta con una puerta de olvido, que ayuda a depurar mejor la información y a ofrecer mejores resultados en la predicciones en las cuales se usan con mucha frecuencia.

En la Figura 3.2 [22] se observa el comportamiento de la red lstm la cual cuenta con una capa de olvido que permite trabajar con la información que pueda ser reelevante para la predicción del modelo,

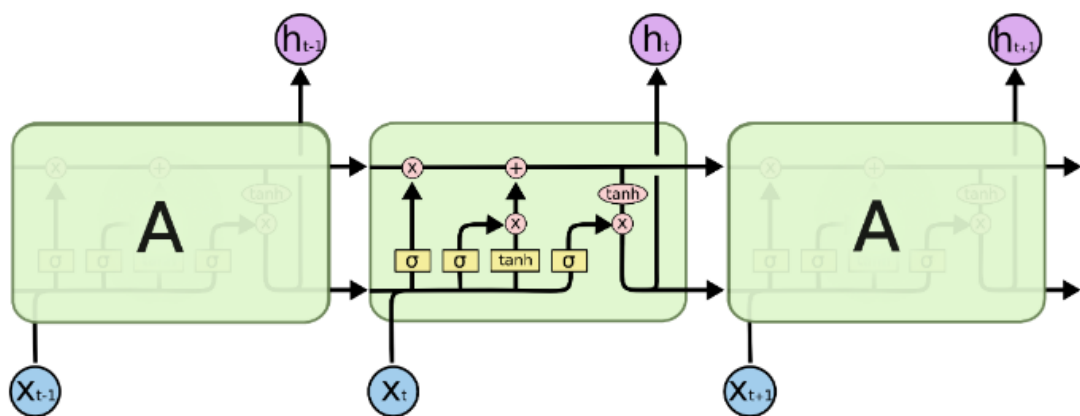


Figure 3.2: LSTM - Redes Neuronales Recurrentes LSTM

Los sistemas de predicción que usan las LSTM deben tener en cuenta que estas necesitan

ciertos parámetros de entrada como lo son: las unidades que seria la cantidad de neuronas que se usaran para el entrenamiento del modelo, las épocas por los cuales se quiere hacer el procesamiento del mismo y el tamaño del lote que indica la cantidad de datos a procesar por cada iteración que el mismo tenga.

3.3 Herramientas de Visualización

POWER BI es una herramienta Microsoft, la cual brinda la facilidad de crear tableros de control desde diferentes fuentes de datos, ofreciendo una fácil y rápida opción a los interesados en el conocimiento de la información la cual puede ser escalable y de muy fácil actualización como sus principales características.

POWER BI se puede usar para la generación de las predicciones en cuanto a analítica predictiva y prescriptiva se refiere facilitando la toma de decisiones a los docentes y personal administrativo de la universidad.

La visualización es una manera de representar los datos de manera gráfica, al usar elementos como cuadros , gráficos de barra, de líneas, de dispersión y mapas entre otros. Este trabajo se apoya en este tipo de herramientas visuales según los conceptos de storytelling con los cuales se apoya para la narrativa de la historia que se cuenta a través de los datos eligiendo el orden en que se muestra la información así como los elementos visuales para la misma, buscando generar impacto y recordación [23]

En el mercado existen muchas herramientas de visualización, las mas usadas son tableau, Qlik, y POWER BI. Según el cuadrante de Gartner la mejor posicionada en el mercado es POWER BI.

POWER BI es una herramienta de autoservicio, por su facilidad para utilizar y lo intuitivo de la misma. Esta herramienta tiene una versión de escritorio en la cual se puede hacer conexión a mas todos los tipos de fuentes que sea necesario para desarrollar los tableros de control.

Por otra parte esta herramienta permite compartir los tableros desarrollados muy fácilmente, mediante una publicación en linea y/o publicando los tableros de forma embebida dentro de una aplicación, ofreciendo una actualización personalizada de los datos a través de un gateway el cual se programa para la automatización de esta tarea.

Como las gran mayoría de estas herramientas POWER BI ofrece la posibilidad de tener disponibles los tipos de gráficos mas usados y adicionalmente una actualización de la misma, con una periodicidad mensual. Para los casos en que se requiera hacer uso de gráficos mas personalizados también cuenta con una tienda en linea en la cual se pueden descargar otras opciones para los mismos.

Por lo anteriormente mencionado esta fue la herramienta seleccionada para el desarrollo de los análisis gráficos de este trabajo con el fin facilitar la toma de decisiones.

Chapter 4

Análisis Descriptivo

4.1 Análisis de los logs

Con el fin de analizar el comportamiento académico de los estudiantes para determinar posibles patrones de comportamiento asociados a la deserción o al bajo rendimiento académico y que permitan mejorar la interacción de la comunidad estudiantil, para adelantar en este proceso se desarrollan las siguientes actividades

Parte de la información de este informe es basado en la documentación del desarrollo del software ELA que es la herramienta desarrollada desarrollado mediante la articulación los grupos de investigación de la UNAD (UBUNTU, SIGCIENCY, CIDLIS) y la Universidad Jorge Tadeo Lozano (ID&SI) y con la alianza estratégica con la empresa Queos S.A.S. y el auspicio de Minciencias Convocatoria 804 de 2018, “Convocatoria Regional proyectos de I+D que contribuyan al fortalecimiento de la formación virtual en el Departamento de Antioquia, Occidente y que se pone a disposición de los docentes y administrativos de la universidad en el se tiene la información descriptiva y predictiva analizada con el fin de facilitar la toma de decisiones y mejorar la interacción con los estudiantes mejorando los objetivos académicos, así como del análisis de requerimiento hecho en Trello por todo el grupo de trabajo en el cual participe durante el desarrollo del proyecto.

El documento consta de dos módulos principales Así:

1. Análisis de la herramienta MOODLE, la cual es la usada por la universidad para que los docentes y estudiantes realicen sus actividades académicas en la cual se puede hacer seguimiento a las mismas. con base esta herramienta se hace una descripción de las tablas mas importantes que van a ser usadas para extracción de información materia de este análisis.
2. Explicación del modelo de base de datos que es usada por la herramienta ELA para el cargue de información inicial de entrada así como los indicadores analizados para cada una de las necesidades planteadas.
3. Gráficos sugeridos como resultado de las sentencias SQL desarrolladas para efectuar los análisis correspondientes a cada uno de los indicadores materia de estudio.

4.2 Análisis de la herramienta MOODLE

La herramienta MOODLE es una plataforma diseñada para el aprendizaje, la cual incluye a todos los actores de la educación. esta plataforma se encuentra desarrollada sobre el motor

de base de datos relacional que normalmente es Postgres o MySQL. Para tener una mejor relación el motor de bases de datos utilizan un editor propio llamado xmldb el cual permite a través de archivos .xml generar la configuración necesaria para la instalación de la herramienta de una manera mas eficiente.

Aunque el proyecto se basa en la información de la tabla *MOODLE Logstore Standard Log* de MOODLE, para poder entender cuales usuarios del sistema son estudiantes fue necesario revisar las siguientes tablas:

MOODLE Context: Es la encargada de realizar todos los movimientos que se realicen sobre la plataforma, es por eso que es considerada la tabla principal. el contexlevel indica el tipo de movimiento que se esta realizando.

MOODLE Course: En esta tabla se almacena la información de todos los cursos que se creen con las fechas de inicio y finalización de los mismos.

MOODLE User: Esta tabla es la encargada de almacenar todos los usuarios que tienen acceso a la plataforma.

MOODLE Role: Permite almacenar los diferentes roles que se pueden tener dentro de la plataforma como por ejemplo: estudiantes, profesores, etc.

MOODLE Role Assignments: En esta tabla se guarda la información del role que puede tener asignado cada usuario

MOODLE Logstore Standard Log: En esta tabla se almacena la información de los registros generados en cada uno de los eventos registrados dentro de la plataforma. esta tabla contiene los contexid, courseid que permite hacer la relación con las tablas anteriormente mencionadas.

De las tablas mencionadas anteriormente se extrae la información que debe suministrarse a la empresa Queos S.A.S la cual cuenta con el siguiente modelo relacional para el desarrollo de la herramienta de consulta de los indicadores:

En figura 4.2 se muestra la información cargada en la base de datos Mysql usando como editor Workbench:

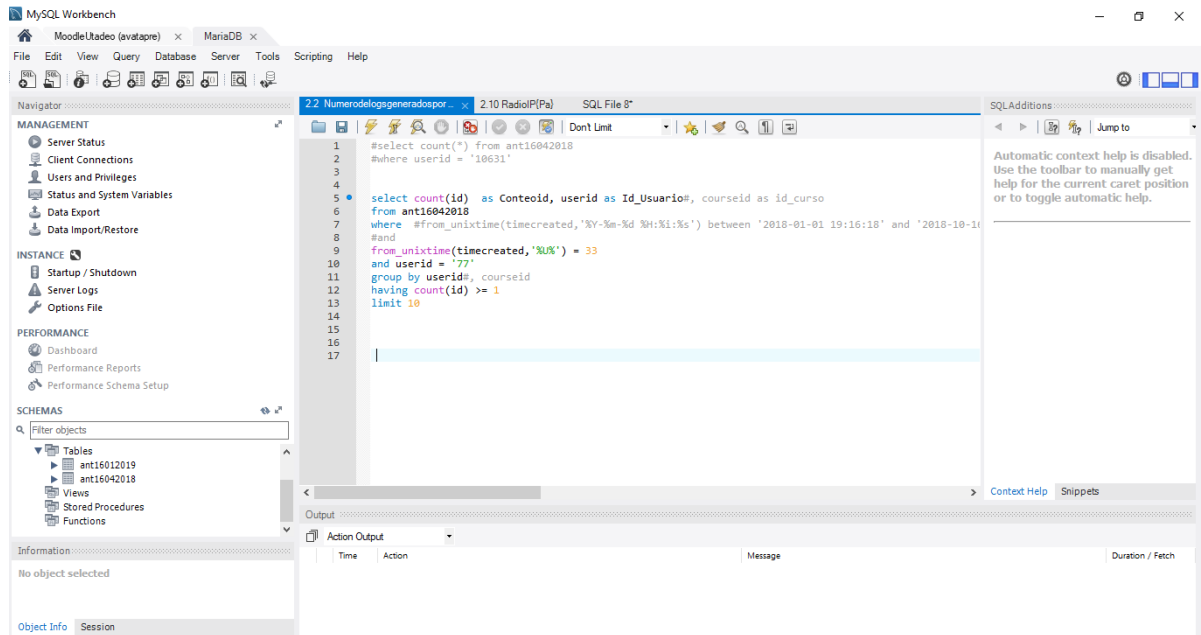


Figure 4.2: información Cargada en MYSQL

4.4 Generación de Sentencias SQL para el para la construcción de los indicadores propuestos.

Para algunos de los indicadores propuestos se han generado consultas obteniendo como resultados preliminares las siguiente información:

4.4.1 Tiempo Total gastado por estudiante en el CMS Tp

Conocer el tiempo total que un estudiante permanece conectado al CMS para Medir y Comparar el tiempo dedicado a la plataforma.

Se puede identificar que para la semana 33 un estudiante promedio permanece conectado a la plataforma al rededor de 220 minutos. este indicador pueda ayudar a tener claridad con respecto a la dedicación por estudiante a las actividades relacionadas con su rol dentro de la plataforma.

En la figura 4.3 podemos observar el resultado de esta consulta:

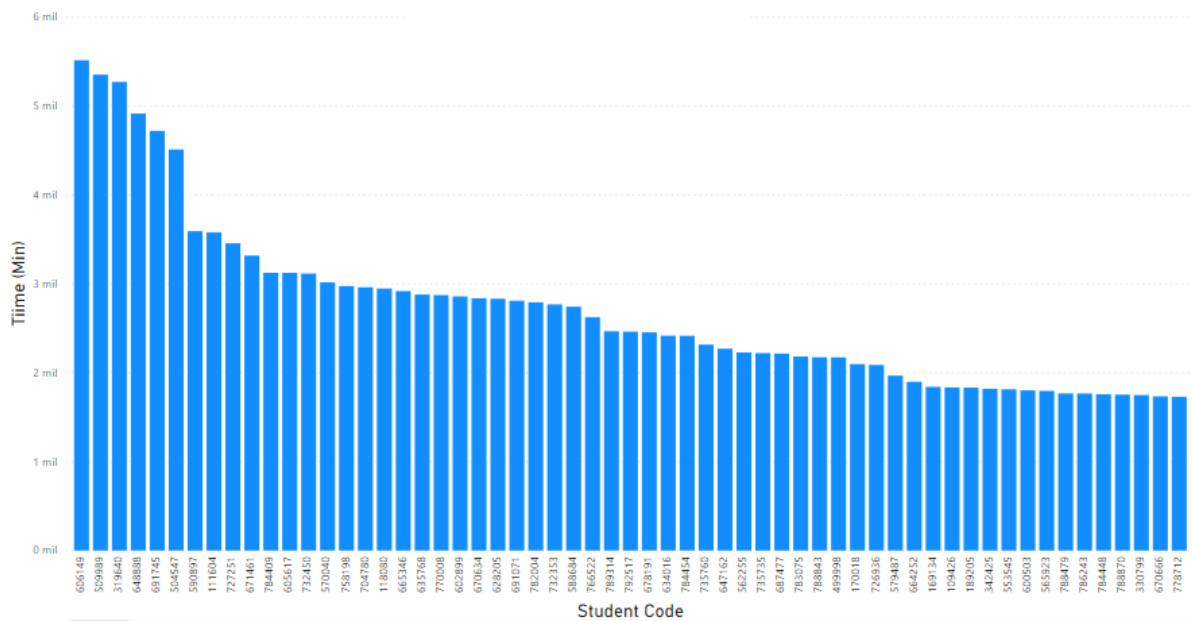


Figure 4.3: Tiempo Total gastado por estudiante en el CMS

4.4.2 Tiempo promedio que un estudiante en particular gasta en una actividad particular Ef

Se requiere conocer el número de veces que un estudiante interviene en una actividad para poder relacionar esta interacción con su nivel desempeño.

El desempeño como estudiante se puede ver reflejado según el tiempo que que cada estudiante dedique al desarrollo de cada actividad, con este indicador se puede establecer esta relación de manera directa como se muestra en la figura 4.4.

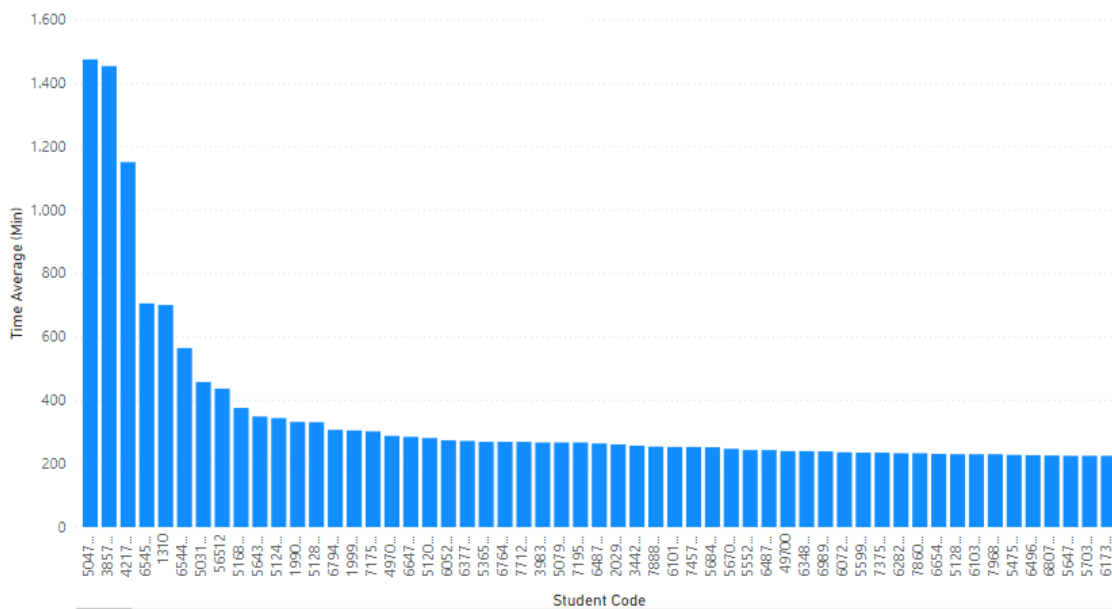


Figure 4.4: Tiempo promedio que un estudiante en particular gasta en una actividad particular Ef

4.4.3 Número de visitas a los materiales del curso Vm

Se requiere conocer el número de visitas a los materiales del curso para Medir o Comparar este indicador contra el desempeño del estudiante.

Con el calculo de este indicador se busca poder calcular, de acuerdo al numero de visitas registradas dentro de la plataforma y a un curso en particular, el interés que pueda tener cada uno de los estudiantes según el numero de visitas que este registre en el aplicativo.

Se encuentra que un estudiante en particular mantiene una cantidad diferente de visitas a cada curso, lo cual puede ser un motivo de estudio adicional, se podría analizar que esta generando este comportamiento, si por ejemplo tiene que ver con las actividades propuestas por cada uno de los docentes

En la figura 4.5 se puede observar el numero de visitas registradas por cada uno de los cursos seleccionando el estudiante al cual se le quiere hacer algún tipo de estudio, al igual que la franja de tiempo en que este genera las visitas a los cursos de su interés.

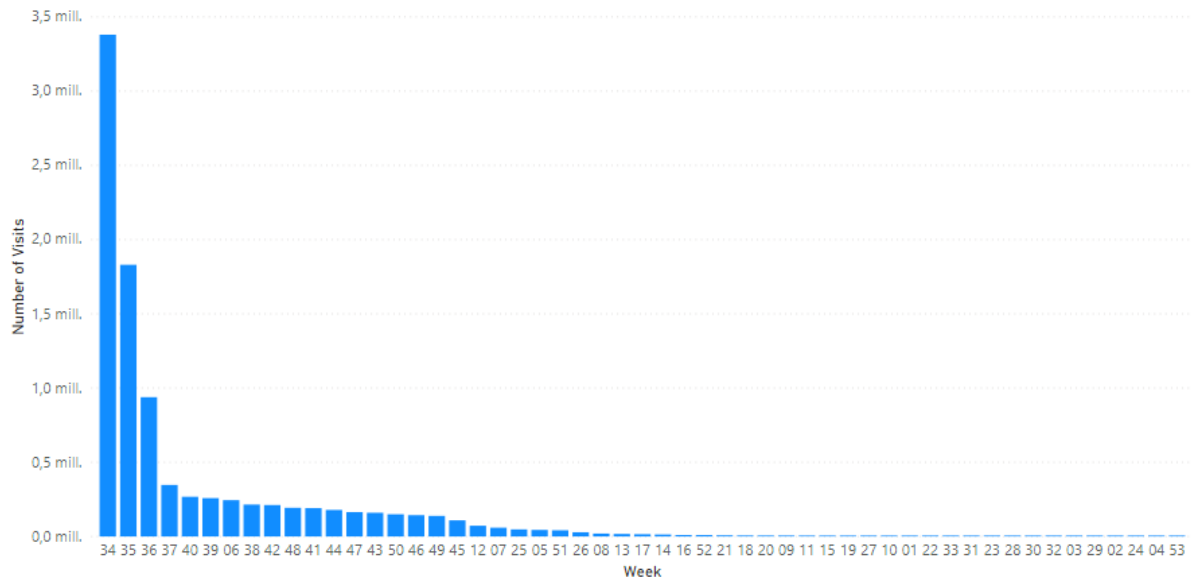


Figure 4.5: Número de visitas a los materiales del curso Vm

4.4.4 Número de logs generados en los días hábiles de la semana Ls

Se requiere conocer el número de logs generados de un estudiante para mi materia o mi programa durante los días hábiles de la semana.

En la figura 4.6 podemos observar el resultado de esta consulta:

Como resultado preliminar de esta consulta se puede asumir que los porcentajes de visitas dentro de la plataforma no es muy diferente para cada uno de ellos, sin embargo se nota una leve tendencia a que el día miércoles es en el que mas logs se generan.

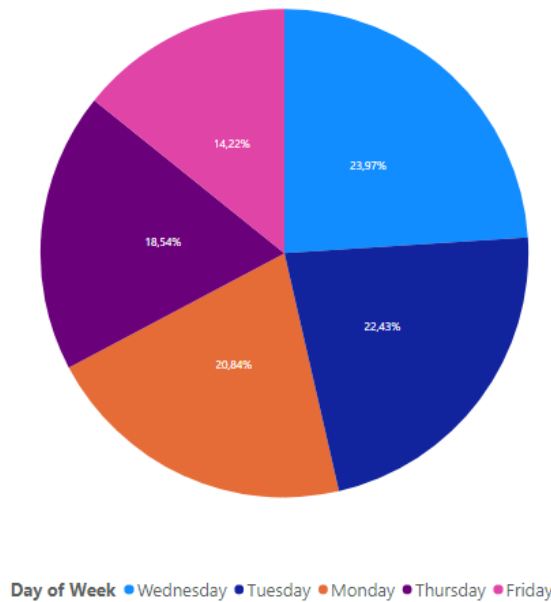


Figure 4.6: Número de logs generados en los días hábiles de la semana Ls

4.4.5 Número de logs generados en los fines de semana Lf

Se requiere conocer el número de logs generados por un estudiante para mi programa o algún curso durante los fines de semana.

Para los fines de semana la situación no es tan diferente a los días entre semana, no se ve una clara diferencia entre sábado y domingo con respecto a la cantidad de logs generados dentro de la plataforma, sin embargo el día sábado muestra una tendencia un poco mayor de visitas como se indica en la figura 4.7.

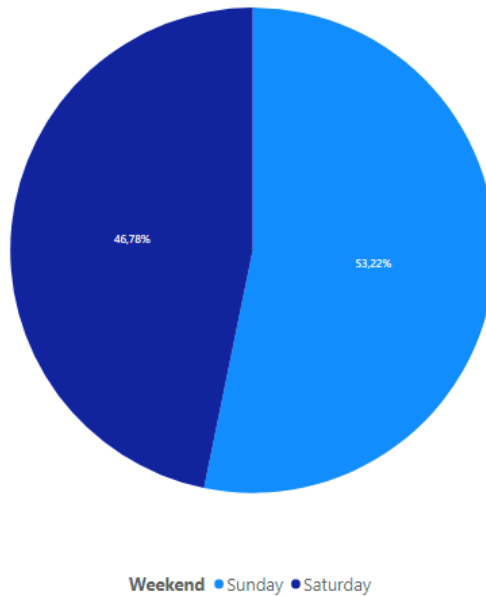


Figure 4.7: Número de logs generados en los fines de semana Lf

4.4.6 Radio IP Pa

Conocer el radio de IPs (número de diferentes IPs) Desde las que se conecta un mismo estudiante para conocer desde que áreas geográficas trabaja el estudiante.

Este indicador permite geo referenciar a través de las direcciones ip desde donde se conecta cada uno de los estudiantes, con el fin de conocer las zonas en las que se conecta, convirtiéndose en un claro material de estudio ya que de esta clasificación puede obtenerse un análisis de deserción como se observa en la figura 4.8.

IP	Student Code	Week
\N	147370	38
10.1.1.133	792408	37
10.10.100.48	726573	37
10.172.52.60	658000	38
10.38.242.70	753113	16
10.8.115.252	756789	05
10.91.5.13	568954	41
104.129.196.187	504807	33
104.152.46.86	722493	33
104.156.210.130	755278	35
120.229.229.43	738719	33
128.90.108.217	666547	34
128.90.109.234	550047	37
128.90.115.148	678728	45
128.90.115.245	657259	34
138.117.84.234	507921	50
138.117.85.38	672215	34
138.121.13.98	726307	33
138.186.20.84	617747	47
138.186.23.74	665103	48
138.186.23.74	670186	34
138.219.140.28	637721	46
143.137.99.19	640984	48
147.75.123.66	512052	38
147.75.124.35	574847	34
147.75.127.143	763892	33
152.200.171.120	513941	38

Figure 4.8: Radio IP Pa

4.4.7 Número de sesiones promedio por semana

Determinar el promedio del numero de sesiones iniciadas ("loggedin") que el usuario realiza durante el periodo (semanal), para conocer la frecuencia de acceso del estudiante al CMS.

En la figura 4.9 podemos observar el resultado de esta consulta:

No todas los accesos a la plataforma son con algún fin específico en la misma, algunos estudiantes se loguean a hacer alguna revisión pequeña, lo cual muestra una clara tendencia a el interés que se pueda tener de estar enterados de los eventos que se puedan publicar en

la misma. Este indicador muestra claramente esta tendencia.

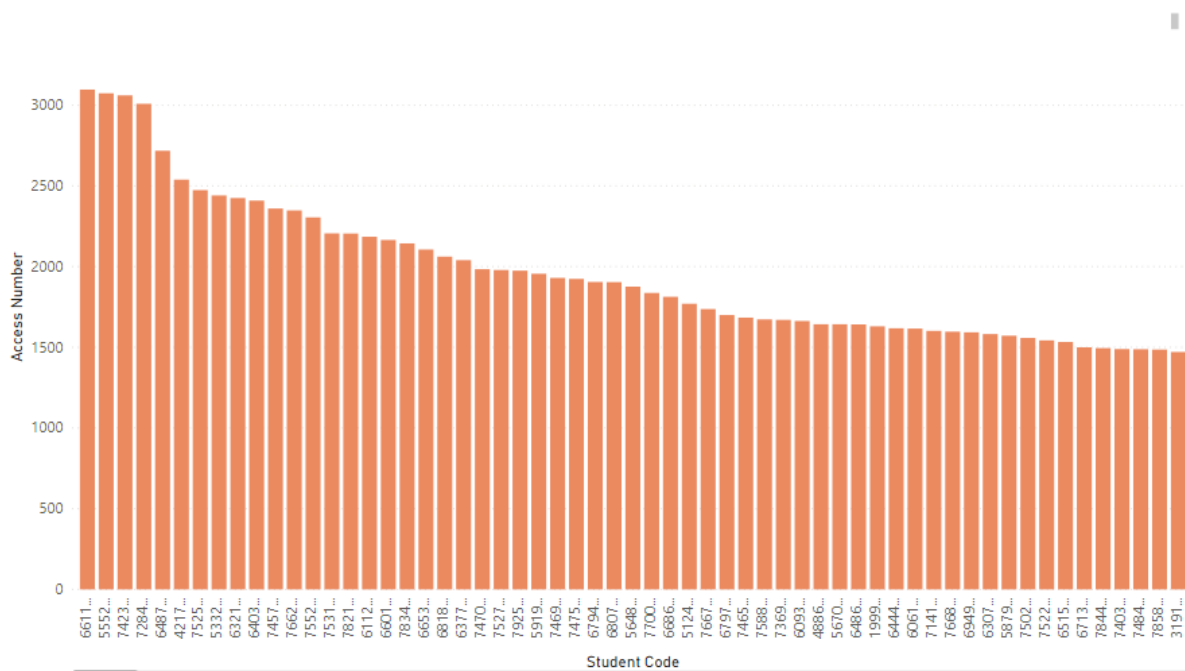


Figure 4.9: Número de sesiones promedio por semana

4.4.8 Tiempo promedio semanal en foros

Se requiere conocer el tiempo promedio durante la semana que un estudiante invierte en foros para Medir y Comparar el tiempo dedicado a la actividad.

La participación en foros permite conocer el nivel de interés que mantienen los estudiantes en compartir y conocer conocimientos y dudas que se puedan estar generando alrededor de un tema específico, como se menciona en el artículo "Clustering MOODLE data as a tool for profiling students". [24] como se muestra en la figura 4.10

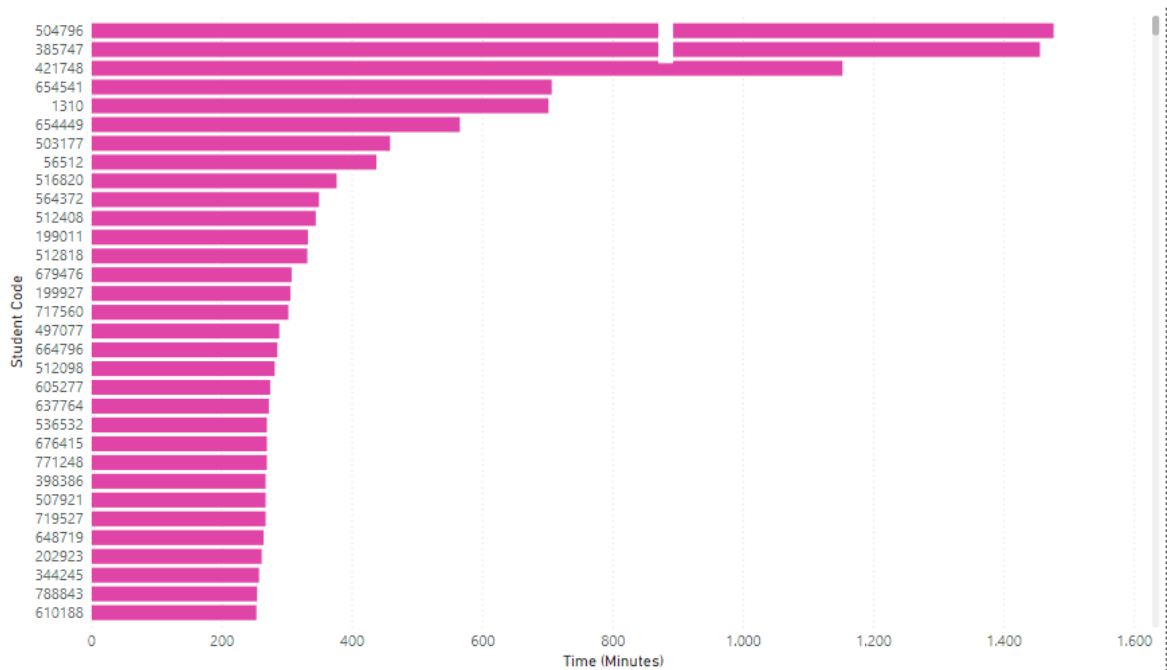


Figure 4.10: Tiempo promedio semanal en foros

4.4.9 Número de diferentes días que el estudiante se logéa en el campus

Conocer el número de días diferentes de la semana que un estudiante se conecta al CMS para Medir y Comparar los días en los cuales hay mayor actividad.

Para el ejemplo que se muestra en la figura 4.11 el estudiante materia de estudio muestra que se conecta todos los días de la semana, lo anterior y con la ayuda de la evaluación de este indicador evidencia el compromiso que se tiene con las actividades propuestas dentro de la plataforma como se puede ver en el analisis realizado para este tema en el articulo; "MOODLEMiner: Data Mining Analysis Tool for MOODLE Learning Management System" [25]



Figure 4.11: Número de diferentes días que el estudiante se loguéea en el campus

4.4.10 Numero de eventos en que modifica el sistema en una semana

Conocer el número de eventos de modificación del CMS en promedio en una semana para poder determinar el grado de interacción "activa".

En la figura 4.12 podemos observar el resultado de esta consulta:

Con este indicador se se logra identificar fácilmente cuales son los eventos que cada uno de los estudiantes generan en la plataforma por semana, lo cual puede facilitar el análisis de la actividades desarrolladas dentro de la misma, para ejemplo del gráfico el evento que mas se genera es el de viewed, en segundo lugar el de created.

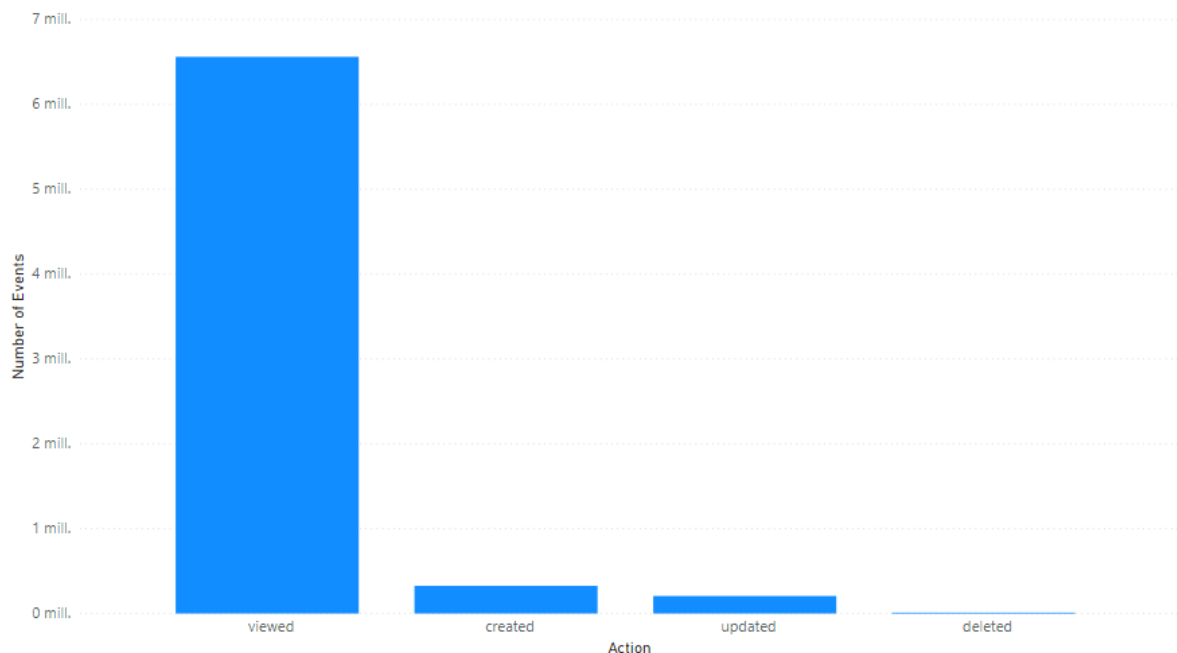


Figure 4.12: Numero de eventos en que modifica el sistema en una semana

4.4.11 Cantidad de cursos a los que el estudiante accede durante una semana

Los modelos deben poder ponderar los logs de acuerdo a la carga de trabajo que tiene el estudiante en LMS, esto porque los eventos de login no registran a que asignatura asistió.

En la figura 4.13 podemos observar el resultado de esta consulta:

Con este indicador se busca tener claridad a cuales cursos accede cada uno de los estudiantes durante una semana, lo cual permite tener conocer sobre cuales cursos tienen mayor grado de aceptación dentro de la comunidad estudiantil.

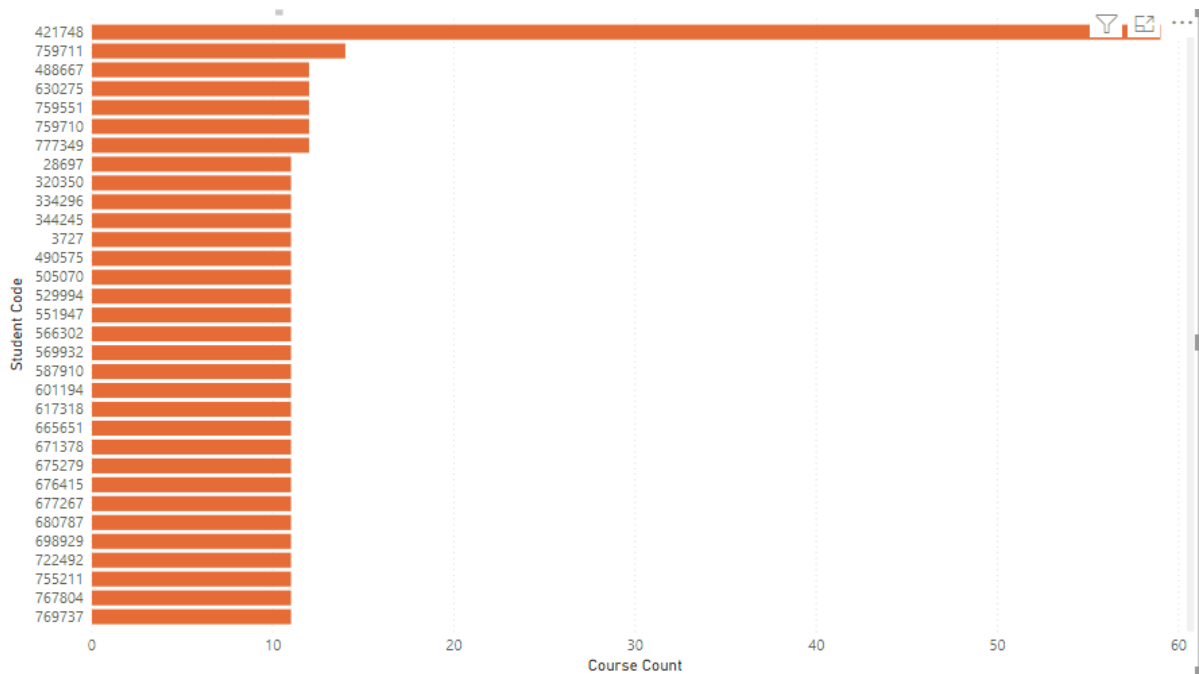


Figure 4.13: Cantidad de cursos a los que el estudiante accede durante una semana

4.5 Análisis sociodemografico

Para el análisis de esta información se han definido 22 indicadores principales, de los cuales tomamos principalmente los referentes a franjas horarias, (pro_2_8_F0, pro_2_8_F6, pro_2_8_F12, pro_2_8_F18) los cuales dividen las 24 horas del día en 4 partes, también se usaran los referentes a tiempos de permanencia en la plataforma (pro_2_3_Time), así como los referentes a días de la semana (pro_2_15_MON, pro_2_15_TUE, pro_2_15_WED, pro_2_15_THU, pro_2_15_FRI, pro_2_15_SAB, pro_2_15_SUN) como se muestra en la tabla 4.1

Indicator	Description
2.2	<i>Number of logs generated</i>
2.6	<i>Number of logs generated during weekdays</i>
2.7	<i>Number of logs generated on weekends</i>
2.8.1.F0	<i>Number of logs generated from 12am to 6am</i>
2.8.1.F6	<i>Number of logs generated from 6am to 12pm</i>
2.8.1.F12	<i>Number of logs generated from 12pm to 6pm</i>
2.8.1.F18	<i>Number of logs generated from 6pm to 12pm</i>
2.10	<i>Number of distinct IPs</i>
2.18	<i>Number of distinct courses accessed</i>
2.11	<i>Number of times a student logs in</i>
2.5	<i>Number of visits to course materials</i>
2.12	<i>Number of interactions with other course members</i>
2.13	<i>Number of sessions</i>
2.17	<i>Number of system modification events</i>
2.15_LUN	<i>Number of logs generated on a Monday</i>
2.15_MAR	<i>Number of logs generated on a Tuesday</i>
2.15_MIE	<i>Number of logs generated on a Wednesday</i>
2.15_JUE	<i>Number of logs generated on a Thursday</i>
2.15_VIE	<i>Number of logs generated on a Friday</i>
2.15_SAB	<i>Number of logs generated on a Saturday</i>
2.15_DOM	<i>Number of logs generated on a Sunday</i>
2.3	<i>Total time spent per student at CMS</i>

Table 4.1: Input indicators (per week)

4.6 Correlación de las fuentes

Relacionar franjas horarias y días de acceso con respecto a variables sociodemograficas como edad, sexo, trabajo y posiblemente medio de acceso(tablet, portatil o ...)

Para el análisis propuesto en este documento, comenzaremos por saber cuales son algunas de las características que se pueden conocer a través de los indicadores generados para tan fin y su comportamiento al relacionarse con las condiciones socioeconomicas de cada uno de los estudiantes así como su comportamiento en general.

Como parte inicial de este análisis se debe conocer que esta basado en una encuesta realizada a los estudiantes de la universidad UNAD matriculados para el año 2019 en la misma,y en la cual se evalúan diferentes factores, y se tienen los 22 indicadores generados para el análisis como se describe a continuación.

Es de anotar que se profundiza mas en el análisis de los indicadores referentes a las franjas horarias y días de las semana con el fin de ayudar en la decisión y establecimiento de mejores horarios y días para realizar actividades académicas.

En la figura 4.14 se establece un comparativo entre el porcentaje de acceso que tiene cada uno de los géneros al la plataforma según las franjas horarias establecidas, como se puede observar los porcentajes son muy parecidos en todas las franjas, lo cual indica que no tenemos una gran diferencia a la hora de analizar a hombres y mujeres.

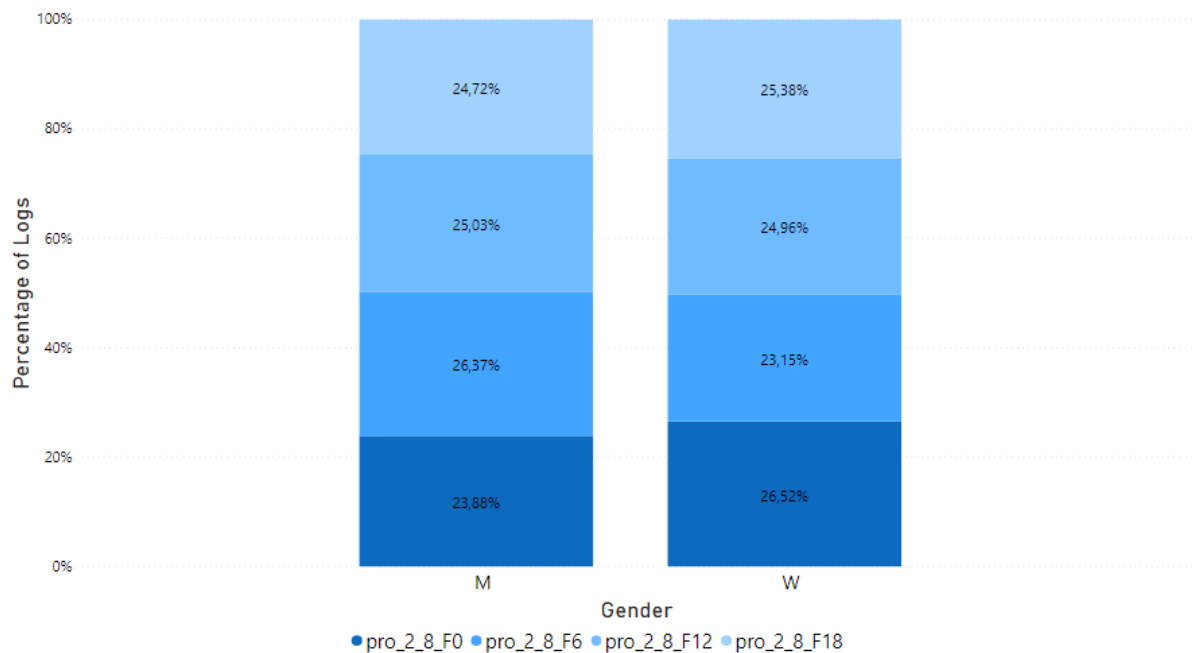


Figure 4.14: Ingresos a la Plataforma por Genero

La dependencia económica puede convertirse en un factor muy importante a la hora de analizar el comportamiento estudiantil, para los estudiantes que también deben trabajar para obtener sus ingresos, es posible que deban organizar sus horarios de estudio de acuerdo a los tiempos libres que estos tengan, según la gráfica siguiente los estudiantes que si dependen económicamente tienen un tiempo de permanencia en la plataforma de 1/3 con respecto a los estudiantes que obtienen sus propios ingresos, lo cual lleva a pensar que se tiene un mayor sentido de compromiso con la formación académica para los que deben pagar por si mismo sus estudios, sin embargo sus posibilidades de tiempo podrían ser un poco mas limitadas que el resto de los estudiantes.

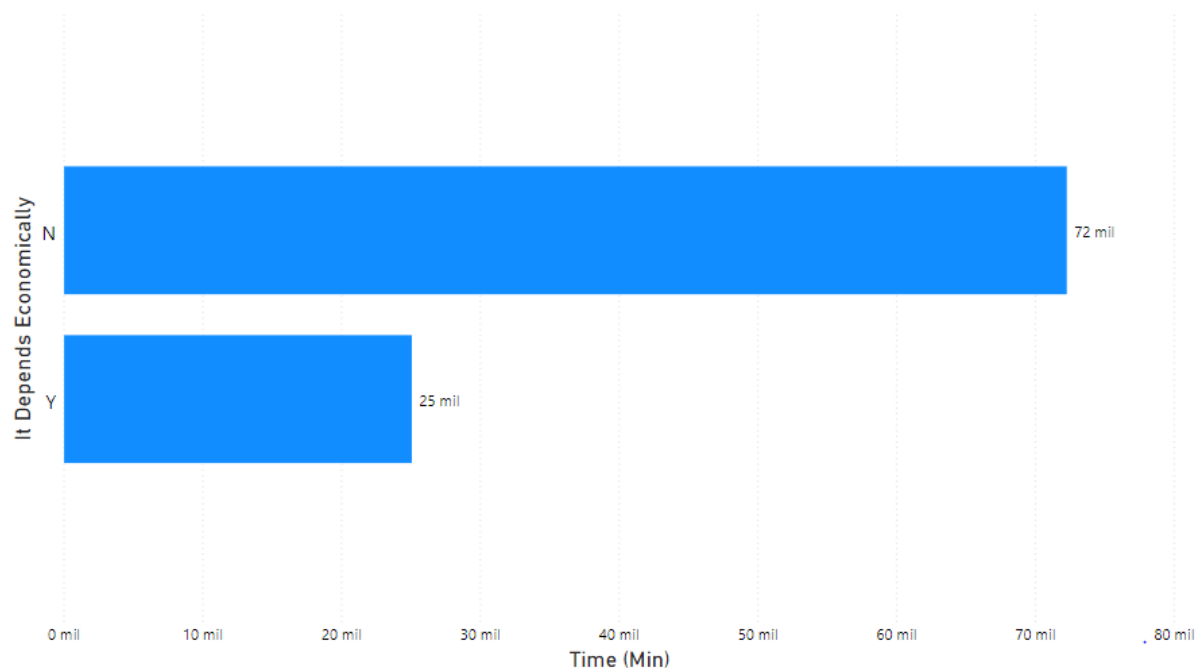


Figure 4.15: Tiempo en la Plataforma Según Dependencia Económica

La zona de residencia es una variable muy importante a la hora de conocer a los estudiantes y promover actividades para los mismos, en el figura 4.16 se ver una tendencia muy marcada con respecto a los que viven es zona rural, estos tienen mayor frecuencia de acceso en el horario de las 12 del día a las 6 de la tarde, en cambio para las personas que viven en zona urbana el acceso a la plataforma en los diferentes horarios analizados es muy similar en todas las franjas

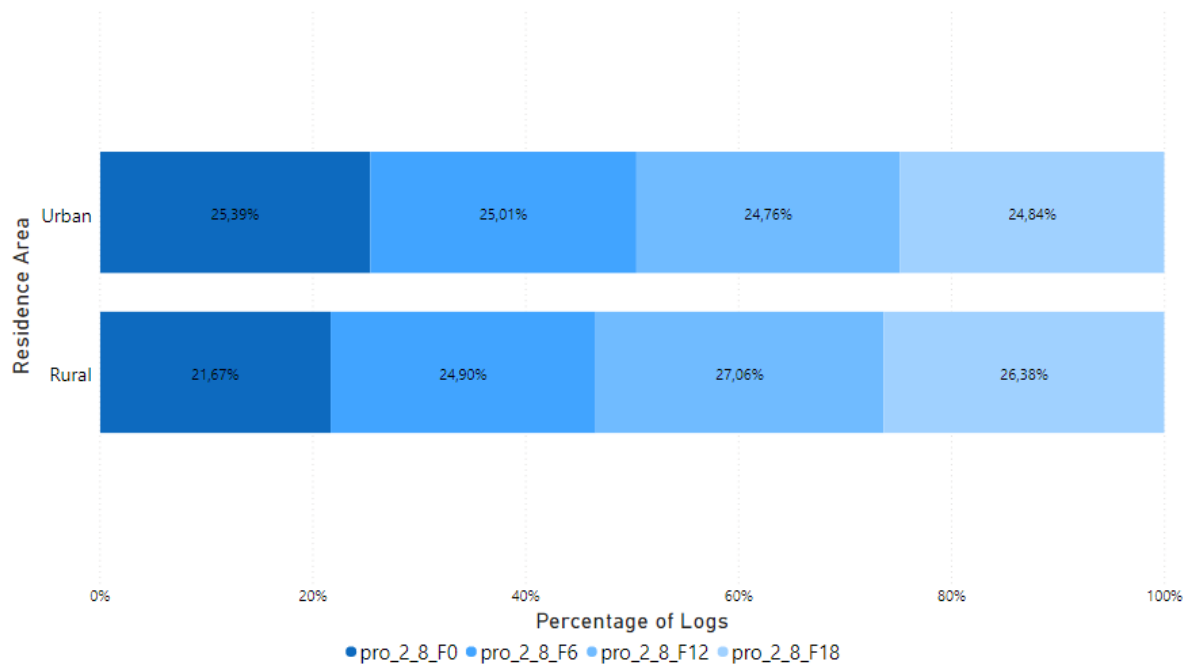


Figure 4.16: Promedio de Accesos a la Plataforma por Zona de Residencia

El estado civil del estudiante puede convertirse en un factor determinante a la hora de la dedicación de tiempo de estudio, como es sabido dependiendo de los compromisos familiares, de acuerdo a este análisis las personas separadas usan mas los horarios de la madrugada para sus actividades académicas, para el caso de los solteros repartido en porcentajes muy parecidos durante todo el día, para los estudiantes que viven es unión libre o casados el porcentaje en el cual se conectan con menor frecuencia es a la madrugada como se muestra en la figura 4.17:

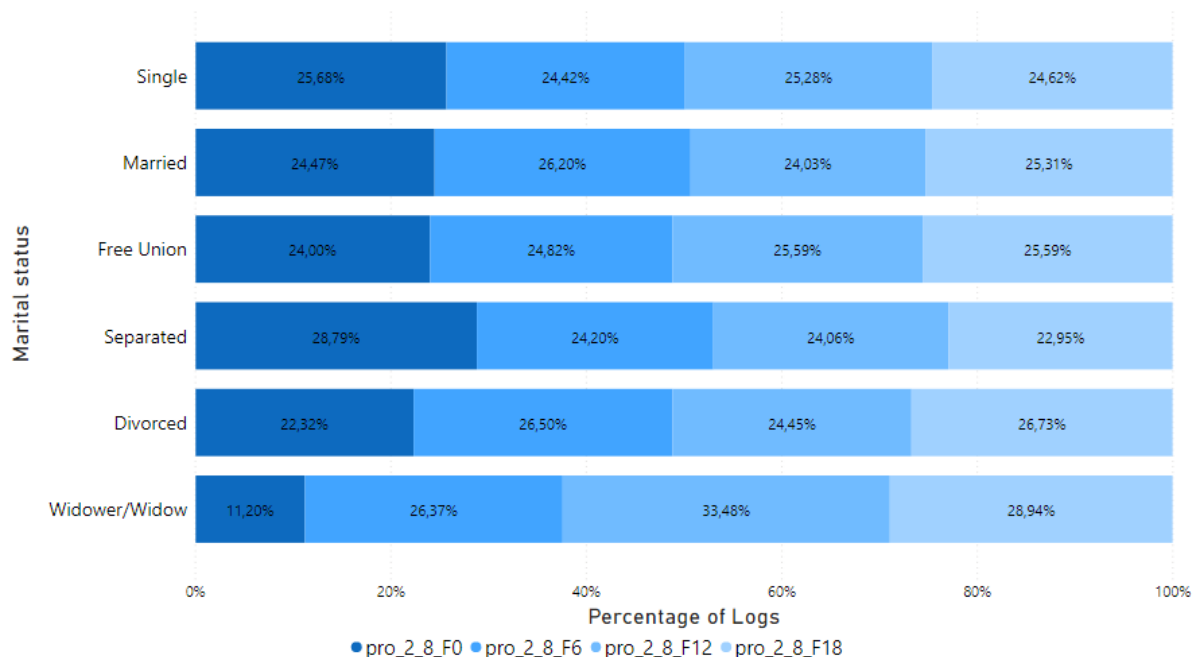


Figure 4.17: Ingresos a la Plataforma según el Estado Civil

Siempre es importante conocer la facilidad con que un estudiante cuenta para conectarse sobre todo a la hora de proponer actividades académicas, ya que un estudiante que cuenta con un equipo portátil y/o una tablet, puede conectarse fácilmente desde cualquier lugar, haciendo el análisis de las 349 personas que contestaron este ítem en la encuesta podemos observar que de estos 186 estudiantes no cuentan con computador de escritorio, pero de los mismos, 183 tienen portátil, y del mismo modo 27 de estos no tienen acceso a tablet, lo cual deja ver que su mayoría usa medios portátiles para hacer sus conexiones como se muestra en la figura 4.18.

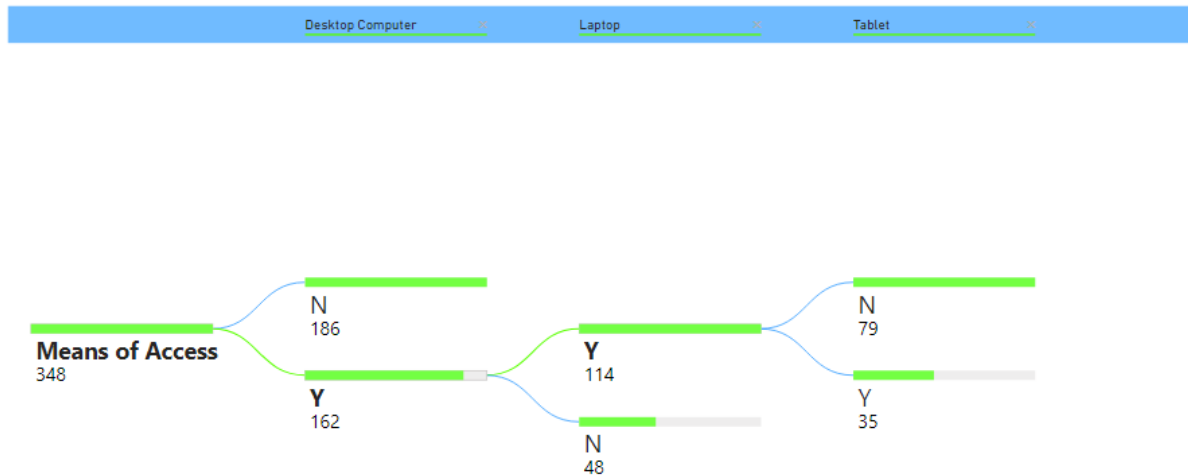


Figure 4.18: Medios de Acceso de los Estudiantes

Para proponer actividades académicas también es muy importante conocer el rango de edad de los estudiantes que se encuentran matriculados y los horarios de acceso que cada rango. de acuerdo la figura 4.19 el pico mas alto de acceso de los estudiantes lo esta en el rango entre 25 y 30 años, sin embargo es de anotar que si bien este es el pico mas alto también el mismo se mantiene para las cuatro franjas horarias que estamos analizando, también es muy notorio que a medida que avanza la edad la frecuencia de accesos va disminuyendo de manera considerable pero muy parecido con respecto a las franjas horarias.

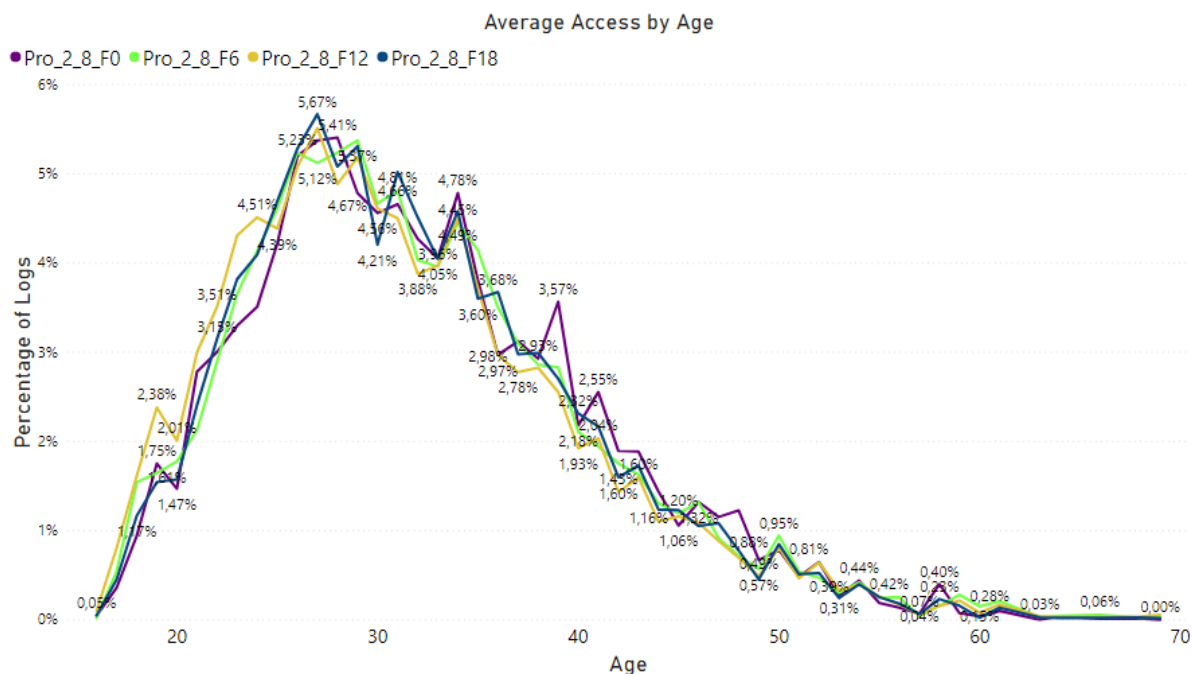


Figure 4.19: Promedio de Accesos Por Edad

En la tabla 4.20 se hace un análisis de de la frecuencia de accesos a la plataforma pro programa académico y franjas horarias. la carrera que presenta mayor porcentaje de accesos es psicología y siendo mayor este en el horario de las 12 de la noche a las 6 de la mañana, seguido muy de cerca por el horario de la tarde, entre 12 del día y 6 de la tarde. Esta información es útil a la hora de proponer actividades académicas desde la parte administrativa, ya que indica por carrera cual es el comportamiento de las mismas.

Academic Program	Pro_2_8_F0	Pro_2_8_F6	Pro_2_8_F12	Pro_2_8_F18
PSICOLOGIA (Resolución 3443)	19,11%	16,87%	18,31%	16,47%
ADMINISTRACION DE EMPRESAS	11,14%	11,13%	10,80%	12,33%
INGENIERIA DE SISTEMAS	6,91%	7,88%	6,76%	6,15%
TECNOLOGIA EN REGENCIA DE FARMACIA (RESOLUCION 08200)	6,06%	4,94%	5,68%	5,44%
INGENIERIA AMBIENTAL	5,92%	6,10%	6,12%	6,03%
TECNOLOGIA EN RADIOLOGIA E IMAGENES DIAGNOSTICAS	4,05%	3,30%	3,71%	3,35%
INGENIERIA INDUSTRIAL (Resolución 05867)	3,99%	4,61%	4,39%	4,46%
INGENIERIA ELECTRÓNICA (Resolución 13155)	3,76%	3,30%	2,79%	3,18%
ADMINISTRACION EN SALUD	3,47%	4,42%	4,63%	4,67%
LICENCIATURA EN INGLES COMO LENGUA EXTRANJERA	3,25%	2,10%	2,32%	2,41%
AGRONOMIA	2,77%	4,68%	5,04%	5,46%
INGENIERIA DE TELECOMUNICACIONES (Resolución 14518)	2,70%	3,07%	2,71%	2,78%
COMUNICACIÓN SOCIAL	2,62%	2,61%	2,81%	2,32%
ZOOTECNIA	1,91%	2,14%	2,24%	2,31%
TECNOLOGIA EN SEGURIDAD Y SALUD EN EL TRABAJO	1,45%	1,18%	1,39%	1,36%
INGENIERIA DE ALIMENTOS(Resolucion 9429 nuevos)	1,28%	0,94%	0,93%	0,95%
SOCIOLOGIA	1,19%	0,96%	0,87%	0,87%
LICENCIATURA EN PEDAGOGIA INFANTIL	1,12%	1,12%	1,34%	1,34%
LICENCIATURA EN MATEMATICAS	1,05%	1,30%	1,17%	1,41%
Total	100,00%	100,00%	100,00%	100,00%

Figure 4.20: Porcentaje de Accesos Por Programa Académico

En la figura 4.21 de dispersión, muestra cual es la tendencia de las conexiones según las franjas horarias en en las cuales esta repartido el día, este mismo análisis esta realizado por genero y en el que se puede ver fácilmente que las franjas de la tarde y noche son las que mayor cantidad des accesos tienen con respecto a las otras dos franjas horarias. También es importante ver que en cuanto al genero se mantienen tendencias muy parecidas.

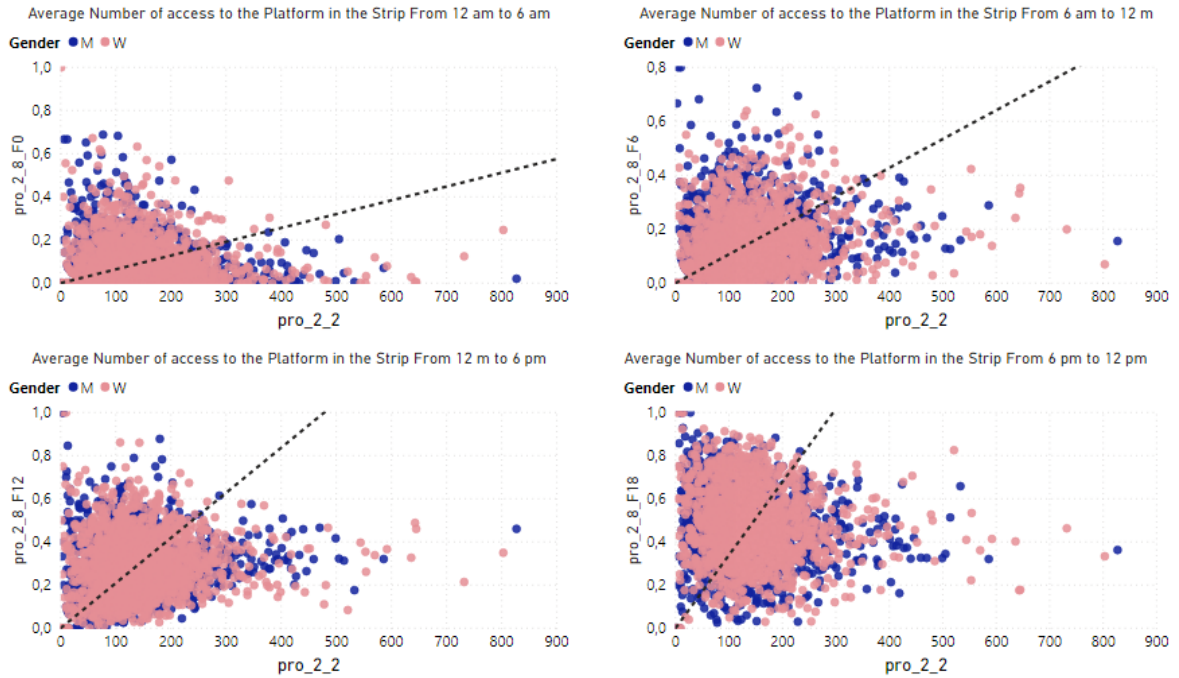


Figure 4.21: Tendencia por franjas horarias y Genero

De acuerdo a las franjas horarias establecidas para este análisis el tiempo de permanencia en las mimas muestra una tendencia muy significativa a ser mucho mas alto en las franjas de la tarde y de la noche, siendo la mas baja en la madrugada como lo indica el análisis de accesos a la plataforma por franjas horarias, lo cual indica que para este caso el tiempo de permanencia en la misma es proporcional a la cantidad de accesos como se ve en la figura 4.22.

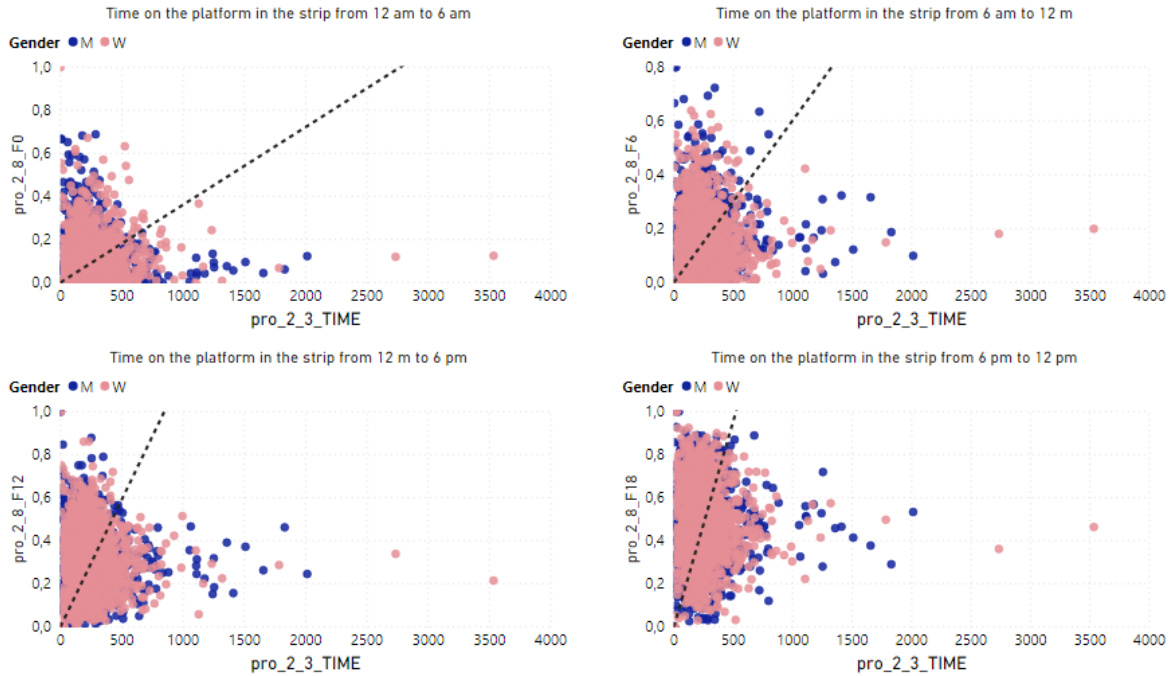


Figure 4.22: Tiempo por franjas horarias y Genero

Conocer la tendencia de accesos a la plataforma por días de la semana es importante ya que es de gran ayuda para el análisis que se esta realizando en el sentido de que sirve para determinar días específicos para proponer actividades académicas. los días martes, miércoles y jueves son los de mayor frecuencia de accesos, como se ha mencionado anteriormente estos podrían considerarse los días mas indicados para los trabajos que se deseen realizar, el día lunes tiene una menor frecuencia con respecto a los anteriores pero se debe tener en cuenta que hay muchos festivos en ese día, y por otra parte desde el viernes se nota una baja significativa que termina en sábado y domingo con frecuencias de acceso muy bajas con respecto a los demás días segun la figura 5.6.

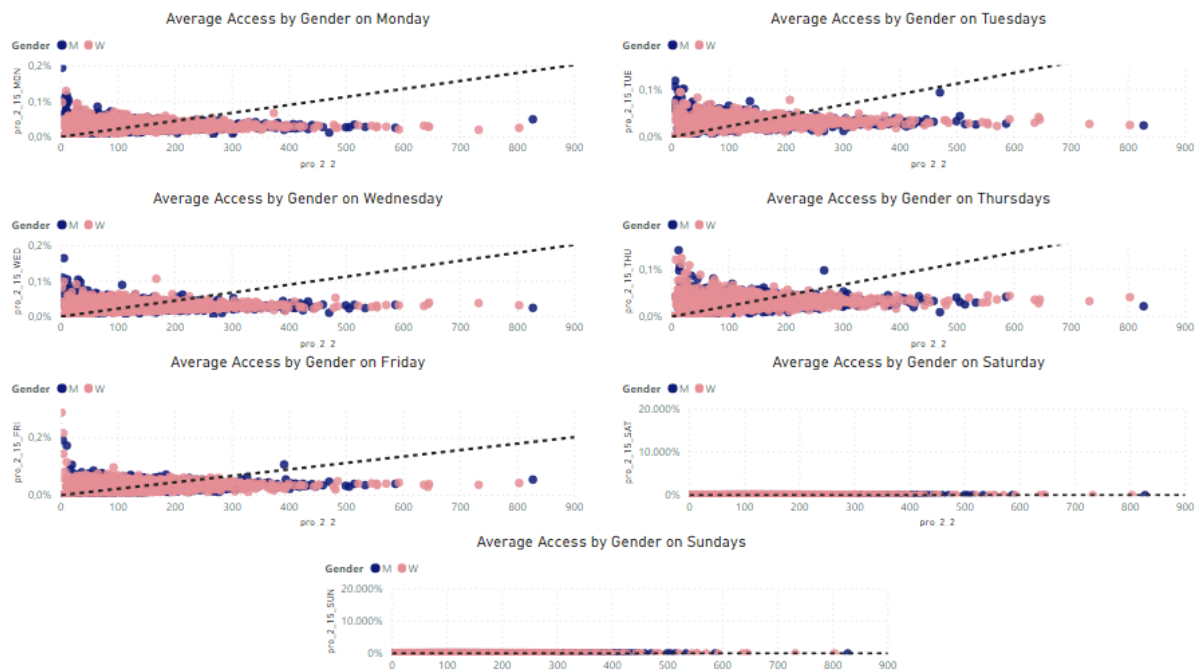


Figure 4.23: Tendencia por Día de la Semana y Genero

Conocer el tiempo que un estudiante permanece en la plataforma da un indicativo de las actividades que este realiza en la misma, para el caso de análisis los días que muestran mayor tiempo de conexión son los días Martes y Jueves y con menor frecuencia los relacionados al fines de semana como son Viernes, Sábado y Domingo como se muestra en la figura 4.24. Es importante aclarar que un estudiante puede tener cierta cantidad de accesos a la plataforma pero no en todos los casos es directamente proporcional a la tiempo que permanece en la misma desarrollando actividades.

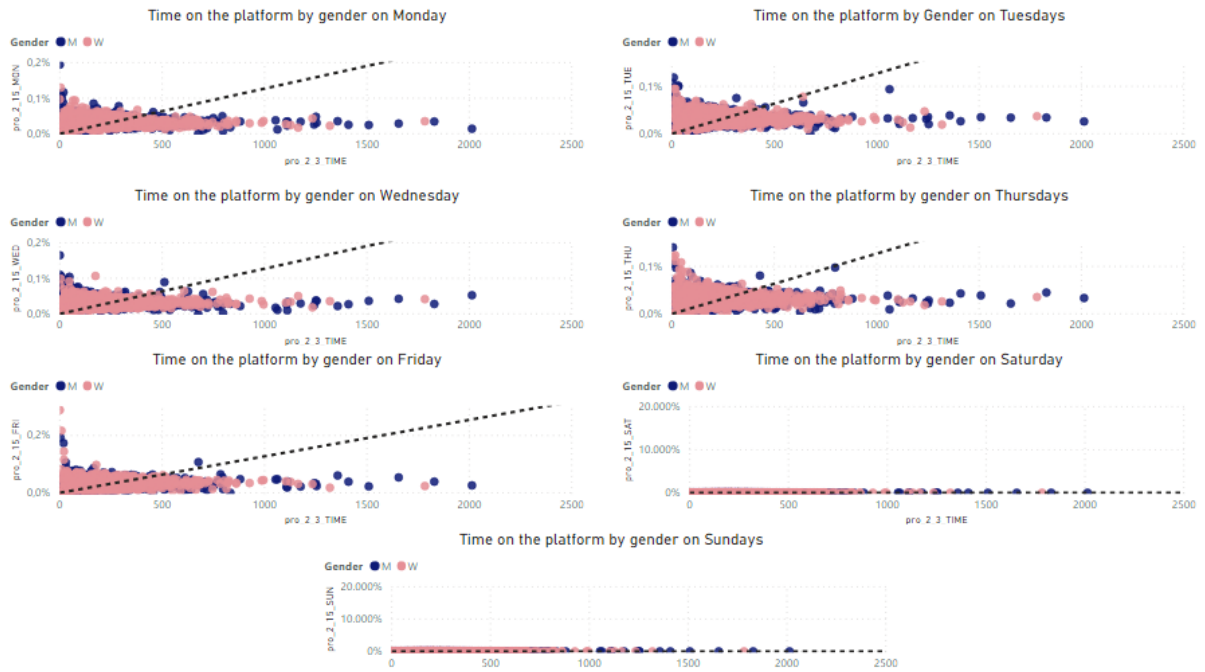


Figure 4.24: Tiempo en la Plataforma por Día de la Semana y Genero

Chapter 5

Análisis Predictivo

Posterior al análisis descriptivo y con el conocimiento que este ha proporcionado se puede proceder al análisis predictivo con el cual se busca generar respuesta a la pregunta de que puede pasar y con base en esta información tomar decisiones acordes con lo que la información suministrada sugerir que puede pasar con un alto grado de probabilidad.

Para trabajar en el análisis predictivo se toma como base los indicadores de franjas horarias 2.8.1.F0 que indica el numero de conexiones que se realizan entre las 12 de la noche y las 6 de la mañana, el 2.8.1.F6 que indica las conexiones entre las 6 am y las 12 del día, el 2.8.1.F12 que se mide entre las 12 del día y 6 de la tarde y por ultimo 2.8.1.F18 que se encarga de evaluar las conexiones entre las 6 pm y la media noche.

Con lo anterior se busca predecir cual es el la franja horaria en que mayor cantidad de estudiantes se conectan durante el día, lo cual puede ayudar a predecir como sera el comportamiento en la semana siguiente a la analizada y de este forma promover actividades académicas en procura de que la participación por parte de los estudiantes sea la mayor posible, mejorando así la interacción entre estudiantes y docentes.

Para lograr este objetivo se genero un modelo de series de tiempo multivariadas el cual permite establecer como sera el comportamiento de un estudiante en la semana siguiente a la analizada teniendo en cuenta factores los diferentes herramientas que este modelo permite como lo es el de generación de series por separado para las variables materia de estudio con el fin de determinar si son series con tendencias similares como se muestra en la figura 5.1.

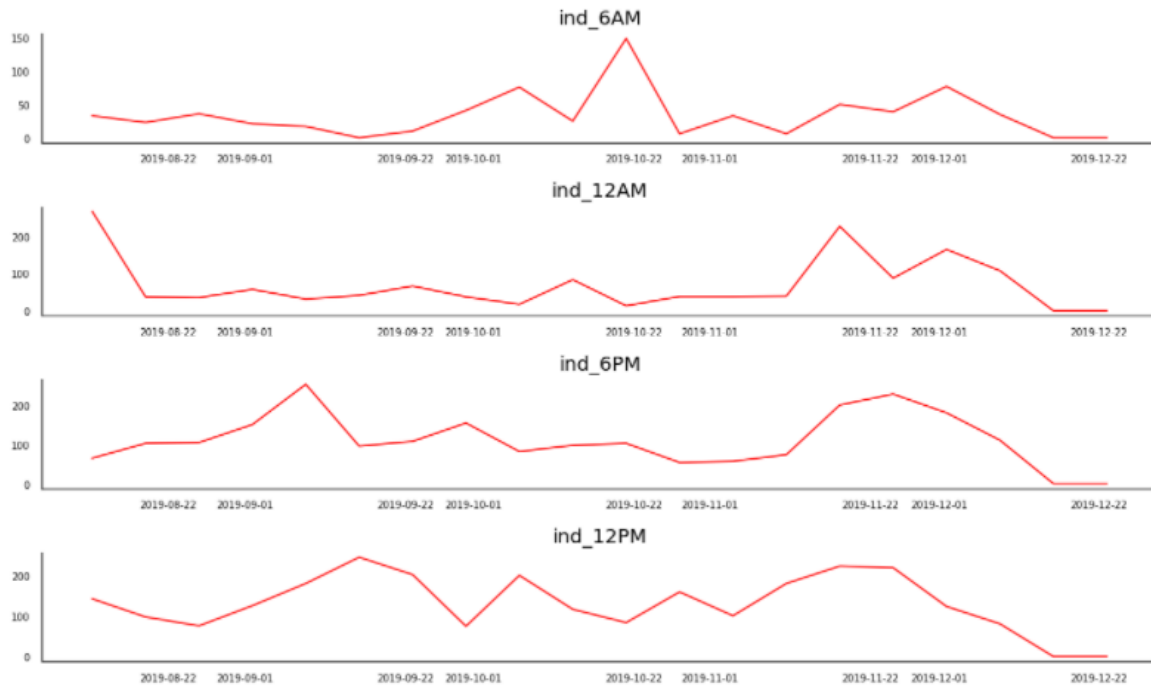


Figure 5.1: Gráfica de Series de Tiempo por Variable

5.1 Análisis de la serie de tiempo

En esta sección el objetivo es describir el proceso que se llevo a cabo para la predicción a través del modelo de series de tiempo como base para la toma de decisiones. Una de la pruebas que se pueden aplicar a este tipo de modelos son la de causalidad y correlación, estas ayudan a determinar la relación que pueden tener las variables a usar en el modelo [26] [27]. Sin embargo es importante conocer que pueda que exista cierta correlación entre las variables lo cual no indica que también exista causalidad entre las mismas.

La prueba de causalidad de Granger [28] se realiza para comprobar que una variable A puede ser la causante de las variaciones en los datos de la variable B, y esta se determinaría como causalidad unidireccional, pero si de la misma forma la variable B puede causar las variaciones en los valores de la variable A se puede decir que tienen causalidad bidireccional. Esta causalidad se determina a través del nivel de significancia, cuando este es menor que 0.05 entonces se puede descartar la hipótesis nula.

Para el caso de las variables materia de análisis podemos observar que la relación mas fuerte esta determinada por la variable de las 12 de la noche a las 6 de la mañana como se muestra en la tabla 5.2.

	ind_6AM_x	ind_12AM_x	ind_6PM_x	ind_12PM_x
ind_6AM_y	1.0000	0.5637	0.4351	0.5966
ind_12AM_y	0.9405	1.0000	0.2790	0.0057
ind_6PM_y	0.3228	0.4279	1.0000	0.0313
ind_12PM_y	0.5701	0.3463	0.1534	1.0000

Figure 5.2: Prueba de Causalidad

La Correlación es otra prueba elemental que se debe hacer con el fin de establecer si las dos variables analizadas tienen un crecimiento lineal simétrico, de ser así este sería una correlación directa y su coeficiente sería 1 o muy cercano a 1, de lo contrario la correlación se puede decir que es menor en la medida de que este coeficiente sea más lejano a 1. Por otra parte, existen las correlaciones espúreas que pueden mostrar una relación que no tienen sentido o no poseen relación lineal alguna como se expresa en el artículo "El coeficiente de correlación y correlaciones espúreas [29]"

Para el caso de las variables de franjas horarias de la serie 2.8.1.F0 con respecto a 2.8.1.F6 no muestran una correlación, se nota más dispersión en la figura 5.3 que una correlación directa.

El análisis de series de tiempo ayuda a tener una predicción de cuál puede ser el comportamiento futuro de una medida con el fin de tomar acciones que ayuden en la mejora de la programación, y planeación de los periodos académicos futuros.

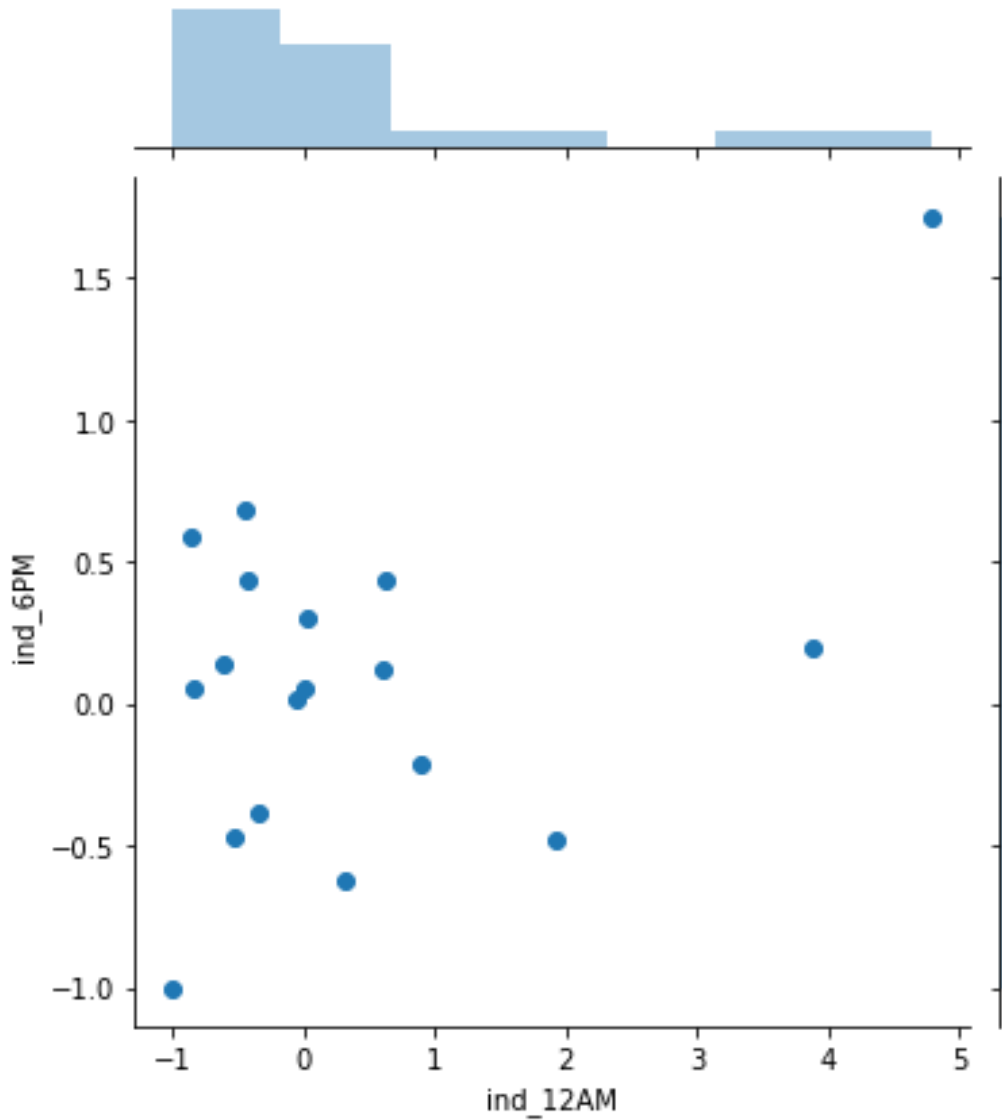


Figure 5.3: Gráfica de Correlación

5.2 Modelo de Redes Neuronales

el uso de la redes neuronales recurrentes fortalecen el modelo de series de tiempo ya que a través de ellas se logran identificar comportamientos que se han presentado y se pueden presentar dentro de las secuencias analizadas por su arquitectura en la distribución de su interconexión y procesamiento de datos [30]

El modelo debe ser dividido entre datos de entrenamiento y datos de pruebas, para este análisis se tomo como datos de entrenamiento las 14 primeras semanas de las 20 disponibles en el set de datos, y el resto para los datos de pruebas. Esto corresponde a un 70 por ciento para entrenamiento y 30 por ciento para pruebas, como se muestra a en la figura 5.6.

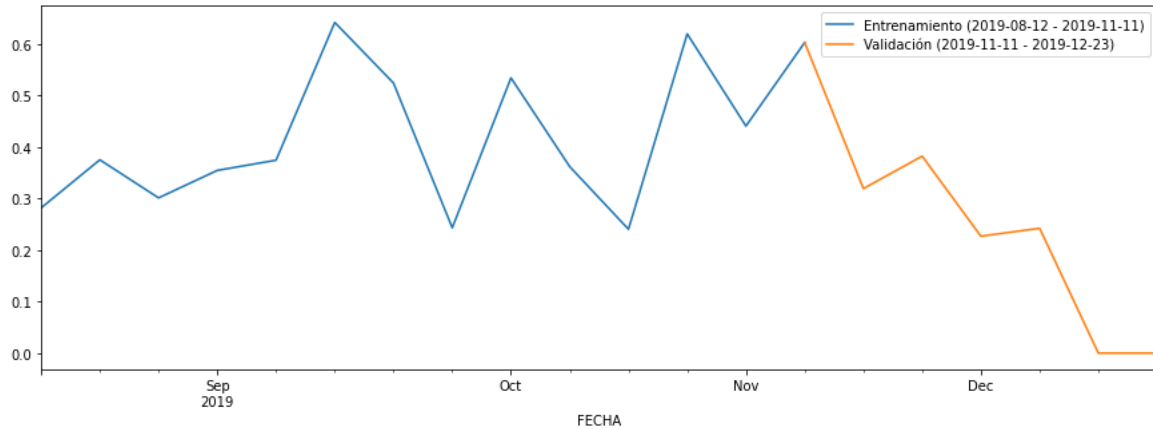


Figure 5.4: Gráfica de Separación de Datos

Como parte definitiva del modelo y después de llevar a cabo los pasos anteriores, se realiza la normalización de datos, usando la función `MinMaxScaler` con una escala entre 0 y 1, esto con el objetivo de tener todos los datos del set de entrenamiento en una escala común, con el fin de facilitar el pronóstico dentro del modelo y posteriormente proceder a la predicción la cual debe servir para la toma de decisiones, para este caso pronosticamos tomando como hiperparametros 20 capas de entrada, 500 épocas y 30 unidades como tamaño del lote [31].

Listing 5.1: Modelo LSTM con keras

```
dim_salida = 1
na = 20

modelo = Sequential()
modelo.add(LSTM(units=na, input_shape=dim_entrada))

modelo.add(Dense(units=dim_salida))
modelo.compile(optimizer='adam', loss='mse')

modelo.fit(X_train, Y_train, epochs=500, batch_size=30)

x_test = set_validacion.values
x_test = sc.transform(x_test)

X_test = []
for i in range(time_step, len(x_test)):
    X_test.append(x_test[i-time_step:i, 0])
X_test = np.array(X_test)
X_test = np.reshape(X_test, (X_test.shape[0], X_test.shape[1], 1))

prediccion = modelo.predict(X_test)
prediccion = sc.inverse_transform(prediccion)
```

La capa de entrada reciben la información de la fuente llamadas también capas densas y la llevan dentro de la red , para este caso se usaron 20, las capas ocultas son las que se encargan del procesamiento de la información para este caso tenemos dos arreglos de capas ocultas de 10 neuronas cada una siendo estas las capas de lstm, y las capas de salida son las que entregan el resultado del modelo, para este caso tenemos una, como se muestra en la figura 5.5.

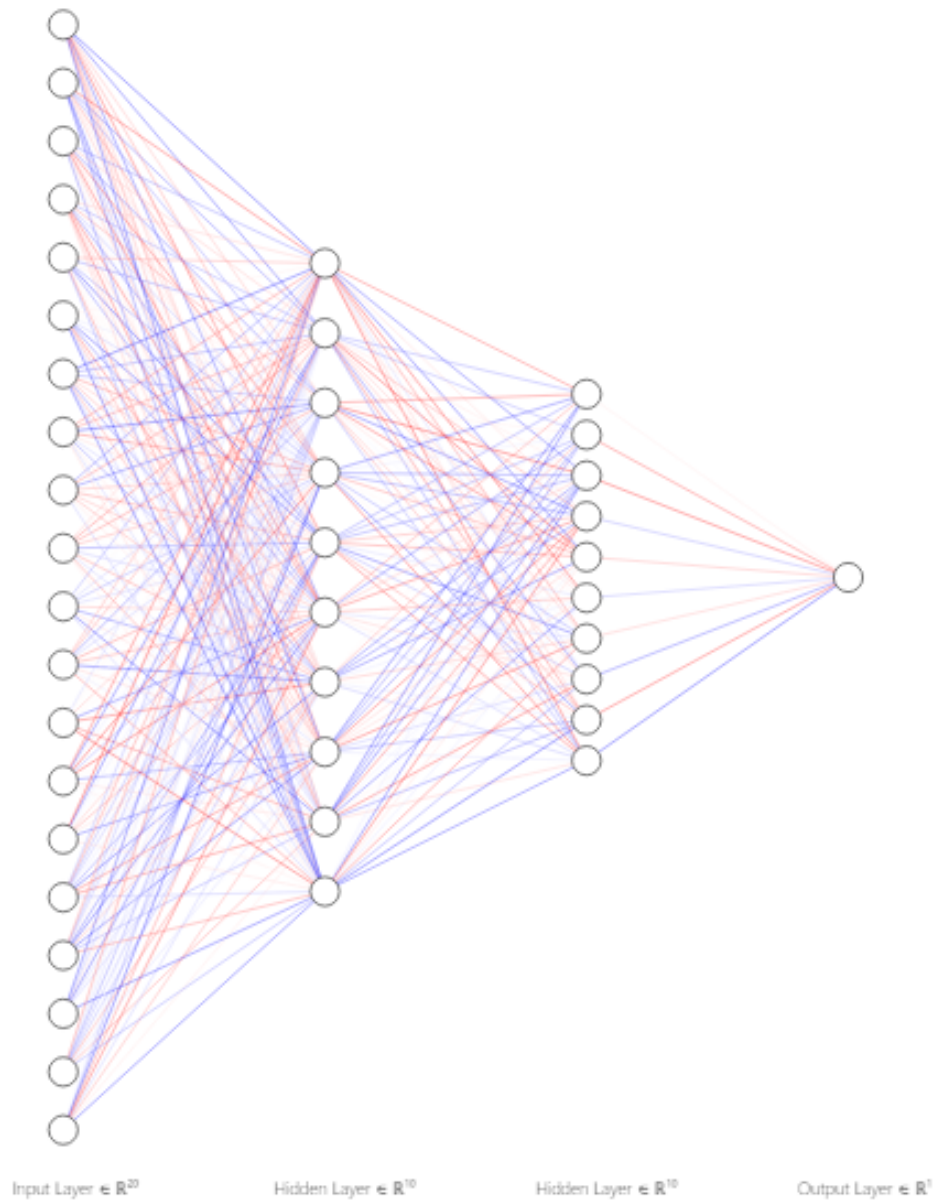


Figure 5.5: Arquitectura del Modelo

Lo anterior ayuda a predecir cual seria la cantidad de accesos a la plataforma en el horario de la noche, correspondiente a la variable de 2.8.1.F18 mostrando como resultado una tendencia positiva como se ve en la figura 5.6.

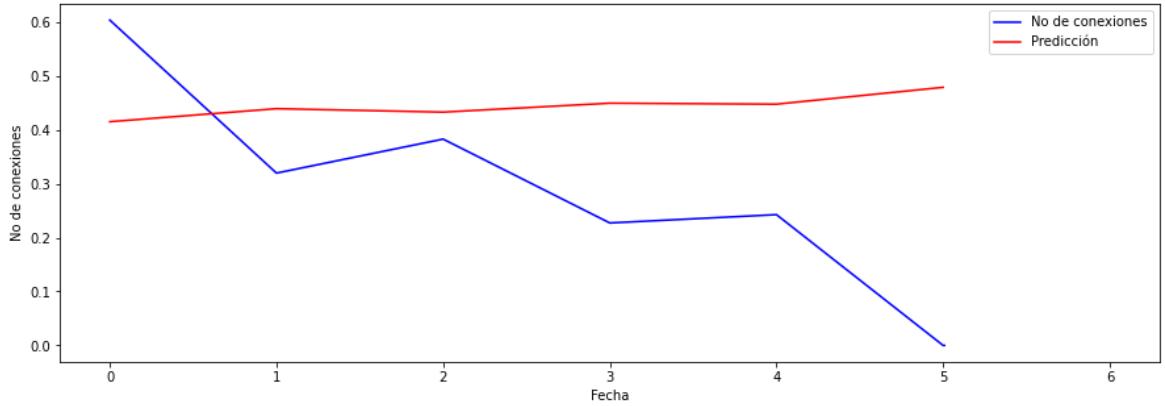


Figure 5.6: Gráfica de Predicción

Para determinar la eficacia del modelo se incluye el método MSE que permite conocer cual puede ser el posible valor de errores de pronostico del modelo a través del calculo del coeficiente de determinación R^2 .

Este valor se calcula conociendo el numero de periodos de la serie y el grado del primer polinomio (n). [32] , es importante anotar que para series de tiempo el set de datos es muy importante a la hora de medir la eficacia del modelo, si la magnitud del mismo es mayor la predicción sera de mejor calidad, en este caso particular solo se entreno con veinte entradas correspondientes a veinte semanas académicas, por este motivo la predicción del mismo no es la mas precisa que se pueda esperar, se sugiere que seguir entrenado el modelo con base en el crecimiento que el set de datos pueda tener en la medida que se tenga mas información de otros periodos académicos posteriores.

Para el desarrollo de esta metodología se debe tener en cuenta que se pueden incluir otros tipos de variables como por ejemplo la cantidad de conexiones por día de la semana que pueden ayudar a mejorar el entrenamiento del modelo. También se puede usar el método de la ventana a través de lstm, que consiste en determinar cuales pasos recientes dentro de la serie temporal se pueden usar para calcular el próximo paso temporal.

Chapter 6

Discusión

El Conocer la información y su estado actual, es muy importante para tener una visión global del comportamiento de los estudiantes en términos académicos, gracias al auge de la tecnología y conectividad que se ofrece en la actualidad, existen diferentes mecanismos para hacer analítica de información.

La analítica de información en la educación abre varios horizontes para lograr una educación mas adaptable a las necesidades del estudiante. En este proyecto basándonos en la información de los Logs de la plataforma MOODLE se pudo observar como se puede segmentar la información a través de diferentes variables socio económicas para identificar patrones en el comportamiento que los estudiantes tienen en el sistema de educación virtual. Dentro de lo encontrado tenemos que, por ejemplo los estudiantes que usan el fin de semana para realizar sus actividades académicas prefieren el día domingo para el desarrollo de las mismas, y a su vez se encuentra que las personas que no tienen una dependencia económica son las que mayor numero de conexiones a la plataforma generan, encontrando una coherencia del por que se escoge el día domingo para la dedicación a temas relacionados con su entorno estudiantil, esto abre un panorama de cuales días son una buena opción para proponer actividades y hacer seguimiento a las mismas [33].

En los casos anteriores en que se ha visto reflejado por ejemplo cuales días de la semana y cual franja horaria los estudiantes usan mas para conectarse a la plataforma. Esta información puede usarse en búsqueda del beneficio académico para los estudiantes como para los docentes. Haciendo uso de la analítica prescriptiva y basados en la herramienta ELA con la cual se puede gestionar la información generada a través de los modelos matemáticos implementados dentro de la misma, y teniendo en cuenta todo el análisis desarrollado con la información generada por la herramienta MOODLE, se plantea por ejemplo, la posibilidad proponer talleres, foros y otras actividades académicas los días martes, miércoles y jueves en las franja horaria de las 6 de la tarde en adelante, que son los días y horarios en los cuales el análisis muestra mejores valores en cuanto a índices de conectividad a la plataforma se refiere, esto con el fin de beneficiar el aprendizaje ofreciéndole al estudiante alternativas cada vez mas asociadas a sus patrones de dedicación al estudio, y así mismo la calidad de los cursos ofrecidos puede verse mejorada con la participación activa de mas alumnos dentro de los mismos, ver figura 6.1.

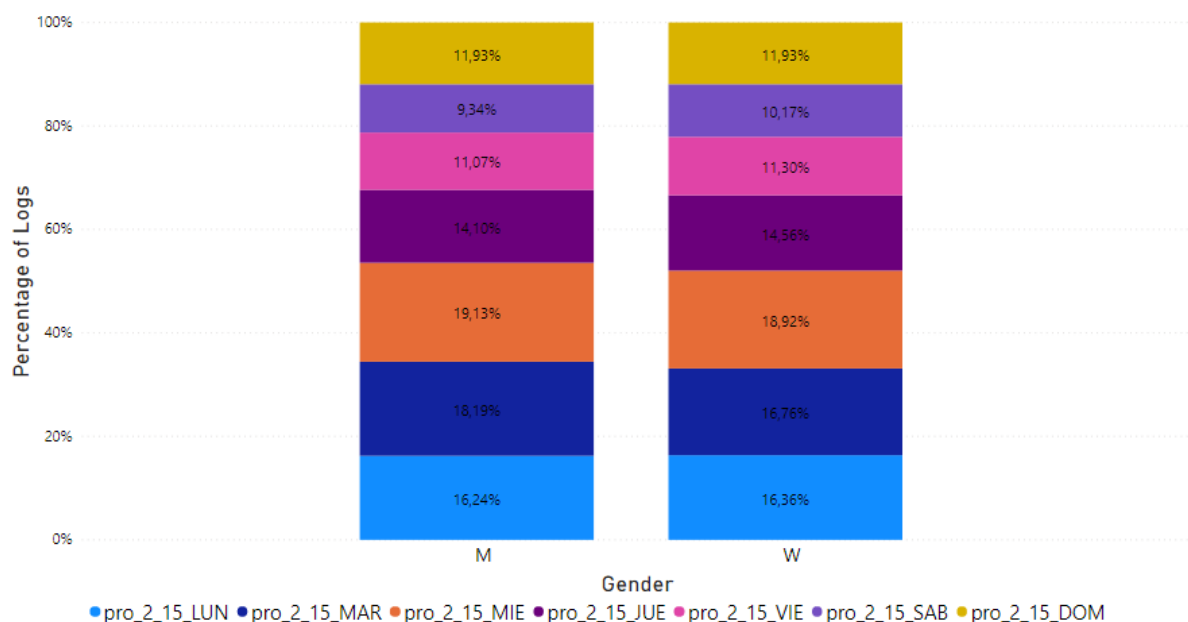


Figure 6.1: Promedio de Accesos a la Plataforma por Genero y Día de la Semana

Adicionalmente, este trabajo propone como predecir que puede pasar con respecto al comportamiento de una semana en relación a la siguiente, teniendo en cuenta la cantidad de accesos a la plataforma y tiempo de permanecía en la misma. Este modelo ayudaría a analizar cual seria una buena alternativa para mejorar estos indicadores de dedicación a su estudio tanto para los estudiantes, así como para los docentes y parte administrativa en términos de programación y preparación de los cursos.

Para realizar esta predicción se usaron modelos de redes neuronales LSTM que permite mejores predicciones en el modelo de series de tiempo multivariadas a través de las compuertas de memoria largo y corto plazo que permiten mejorar la predicción con respecto a otros modelos como por ejemplo los de regresión [34].

Aunque el rendimiento de los modelos LSTM no fue el esperado debido a que la ventana de tiempo analizada fue muy corta, a futuro se espera poder mejorar el resultado del algoritmo basado en una mayor cantidad de información histórica disponible para analizar, ya que este tipo de modelos pueden mejorar el nivel de predicción en la medida que los datos brinden el ciclo completo del estudiante. Por ejemplo se puede agregar datos de los indicadores generados en otros periodos académicos lo cual posibilitara mejores entradas para el modelo e identificar las posible variaciones que pueda tener un estudiante en relación a periodos académicos anteriores ya que sus condiciones pueden variar con el paso del tiempo.

El uso de herramientas de visualización facilitan la toma de decisiones ya que con el apoyo de estas se puede tener un panorama muy claro de la situación actual y futura, y con el apoyo de modelos analíticos como el usado podemos tener alternativas para una mejor planeación y ejecución de actividades académicas que benefician la interacción entre los estudiantes y los docentes así como con todas las áreas de la universidad.

Se propone que a futuro se siga fortaleciendo la herramienta ELA con base en el entrenamiento de los modelos con set de datos de mayor volumen, con el fin de fortalecer el conocimiento que se adquiere a través de estos, de igual forma con los logs generados dentro de MOODLE se pueden definir indicadores adicionales que permitan llegar a otro tipo de análisis a los desarrollados dentro de este trabajo al igual que el fortalecimiento de las encuesta de caracterización realizadas a los estudiantes en cada periodo académico.

Bibliography

- [1] L. M. S. Fonseca, “La evaluación y su incidencia en la deserción escolar:¿ falla de un sistema, de las instituciones educativas, del docente o del estudiante?” *Educación y humanismo*, vol. 18, no. 31, pp. 313–325, 2016.
- [2] J. Sánchez, “Relación entre los modelos de series de tiempo univariadas con los modelos de series multivariadas,” Ph.D. dissertation, Universidad de Nariño, 2013.
- [3] N. Medina Vela, P. J. Urbano Lozano *et al.*, “Estudio de educación superior a distancia en colombia: una reflexión de la modalidad como solución a los requerimientos del siglo xxi.”
- [4] N. E. Parra Acevedo *et al.*, “Entornos educativos saludables en los programas ofertados por la universidad nacional abierta ya distancia unad-ccav dosquebradas.”
- [5] A. Romualdo Rosario, “Factores socioeconómicos en la deserción estudiantil universitaria de la universidad nacional daniel alcides carrión, pasco 2018,” 2020.
- [6] Á. E. Torres-Zapata, J. R. Domínguez, P. F. López, M. d. P. G. Reyes, and D. A. C. Trejo, “Reprobación, síntoma de deserción escolar en licenciatura en nutrición de la universidad autónoma del carmen,” *RIDE Revista Iberoamericana para la Investigación y el Desarrollo Educativo*, vol. 10, no. 20, 2020.
- [7] J. La Madriz, “Factores que promueven la deserción del aula virtual,” *Orbis. Revista Científica Ciencias Humanas*, vol. 12, no. 35, pp. 18–40, 2016.
- [8] B. R. Pérez-Gutiérrez, “Comparación de técnicas de minería de datos para identificar indicios de deserción estudiantil, a partir del desempeño académico,” *Revista UIS Ingenierías*, vol. 19, no. 1, pp. 193–204, 2020.
- [9] A. Vilorio, J. G. Guliany, W. N. Núñez, H. H. Palma, and L. N. Núñez, “Data mining applied in school dropout prediction,” in *Journal of Physics: Conference Series*, vol. 1432, no. 1. IOP Publishing, 2020, p. 012092.
- [10] F. d. J. N. Cárdenas, E. A. Hernández, A. E. G. Zenil, A. M. F. Redondo *et al.*, “Application method of data mining using the k means algorithm for the determination of stress level in high school students using the beck depression inventory,” *Ciencia Huasteca Boletín Científico de la Escuela Superior de Huejutla*, vol. 8, no. 15, pp. 1–8, 2020.
- [11] J. G. López Martínez and Ó. A. Méndez Aguirre, “Técnicas de machine learning para la predicción de desempeño académico en el desarrollo del espacio proyectivo del pensamiento espacial,” 2020.

- [12] D. G. Díaz, J. P. Alcaraz, M. G. Martínez, and C. H. Jacome, “Nodes: Plataforma para la predicción de deserción escolar utilizando técnicas de inteligencia artificial.”
- [13] N. Bedregal-Alpaca, D. Aruquipa-Velazco, and V. Cornejo-Aparicio, “Técnicas de data mining para extraer perfiles comportamiento académico y predecir la deserción universitaria,” *Revista Ibérica de Sistemas e Tecnologías de Informação*, no. E27, pp. 592–604, 2020.
- [14] S. Bouktif, A. Fiaz, A. Ouni, and M. A. Serhani, “Multi-sequence lstm-rnn deep learning and metaheuristics for electric load forecasting,” *Energies*, vol. 13, no. 2, p. 391, 2020.
- [15] C.-F. Tsai, C.-T. Tsai, C.-S. Hung, and P.-S. Hwang, “Data mining techniques for identifying students at risk of failing a computer proficiency test required for graduation,” *Australasian Journal of Educational Technology*, vol. 27, no. 3, 2011.
- [16] J. Villavicencio, “Introducción a series de tiempo,” *Puerto Rico*, 2010.
- [17] C. H. A. Irgoin *et al.*, “Análisis de series de tiempo,” *Contribuciones a la Economía*, no. 2011-02, 2011.
- [18] J. Contreras, R. Espinola, F. J. Nogales, and A. J. Conejo, “Arima models to predict next-day electricity prices,” *IEEE transactions on power systems*, vol. 18, no. 3, pp. 1014–1020, 2003.
- [19] H. Urbisaia and J. Brufman, “Análisis de series de tiempo univariadas y multivariadas,” *Cooperativas, Buenos Aires*, 2001.
- [20] J. A. Pérez-Ortiz, “Modelos predictivos basados en redes neuronales recurrentes de tiempo discreto,” 2002.
- [21] J. Torres, “Redes neuronales recurrentes,” Sep 2019. [Online]. Available: <https://torres.ai/redes-neuronales-recurrentes/>
- [22] g. lstm, “Understanding lstm networks.” [Online]. Available: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>
- [23] Y. Zhang, “Converging data storytelling and visualisation,” in *International Conference on Entertainment Computing*. Springer, 2018, pp. 310–316.
- [24] A. Bovo, S. Sanchez, O. Héguy, and Y. Duthen, “Clustering moodle data as a tool for profiling students,” in *2013 Second International Conference on E-Learning and E-Technologies in Education (ICEEE)*, 2013, pp. 121–126.
- [25] G. AKÇAPINAR and A. BAYAZIT, “Moodlemineer: Data mining analysis tool for moodle learning management system,” *İlköğretim Online*, vol. 18, no. 1, pp. 406–415, 2019.
- [26] C. Batanero, M. M. Gea, P. Arteaga, and G. Cañadas, “Interpretando la correlación,” 2016.
- [27] J. F. B. Redondo, “Prueba de causalidad y determinación de la nairu,” *El Trimestre Económico*, pp. 663–692, 1999.

- [28] J. C. Alonso *et al.*, “Estimación de modelos var, prueba de causalidad de granger y función impulso respuesta empleando easyreg,” Universidad Icesi, Tech. Rep., 2011.
- [29] E. Lahura *et al.*, *El coeficiente de correlación y correlaciones espúreas*. Pontificia Universidad Católica del Perú, Departamento de Economía, 2003, vol. 218.
- [30] J. A. P. Ortiz, “Modelos predictivos basados en redes neuronales recurrentes de tiempo discreto,” *Universidad de Alicante. Departamento de lenguaje y sistemas informáticos.*[Documento en línea http://www.conicyt.cl/573/Modelos_predictivos_basados_en_redes_neuronales_recurrentes_de_tiempo_discreto.pdf][22/07/08], 2002.
- [31] M. Leon, “series de tiempo multivariadas.” [Online]. Available: <https://gitlab.com/olmerg/moodle-unad/-/blob/dd52ac6591fdc75b44104ac8cf649521b9d463c3/Seriesdetiempofranjashorarias.ipynb>
- [32] M. Acosta-Cervantes, M. Villarreal-Marroquín, and M. Cabrera-Ríos, “Estudio de validación de un método para seleccionar técnicas de pronóstico de series de tiempo mediante redes neuronales artificiales,” *Ingeniería, investigación y tecnología*, vol. 14, no. 1, pp. 53–63, 2013.
- [33] M. G. Gomez Zermeño, “Sep prepárate, modelos de educación a distancia en el nivel medio superior: Estudio descriptivo sobre los modelos innovadores de educación a distancia para adolescentes en condiciones de rezago.” *Revista Q*, vol. 6, no. 11, 2011.
- [34] D. F. Herrera Cofre *et al.*, “Predicción para el mercado de acciones con redes neuronales lstm,” 2020.