Estimación de la demanda máxima de cupos por asignatura usando Machine learning

Lorenzo Gutierrez
Juan Albis

September 2024

Introducción

La creación de modelos predictivos comienza con la recopilación de datos, los cuales varían considerablemente según el contexto del que provienen. A partir de estos datos, se lleva a cabo un análisis exploratorio con el fin de identificar comportamientos, patrones o señales implícitas.

El desarrollo de un modelo predictivo no solo permitirá a la Universidad responder de manera eficiente a las fluctuaciones en la demanda, sino que también optimizará la asignación de recursos docentes y administrativos, minimizando posibles desajustes entre la oferta y la demanda de cursos. El análisis de la demanda y el comportamiento de los cupos por asignatura en instituciones educativas es fundamental para optimizar la distribución de recursos y mejorar la planificación académica.

Una proyección adecuada de la demanda de cupos permite a la universidad ofrecer una mejor experiencia a los estudiantes, garantizando que haya suficientes plazas en los cursos más solicitados y evitando ineficiencias en aquellos con menor demanda. Dado el entorno cambiante de la demanda en la educación superior, es crucial comprender los factores que influyen en la inscripción de cada curso y cómo proyectar estos datos de manera efectiva para tomar decisiones informadas. Este estudio tiene como objetivo proporcionar un análisis del comportamiento histórico de los cupos por asignatura, identificar los factores que influyen en la demanda de los cursos, y, principalmente, crear un modelo predictivo que permita proyectar la disponibilidad de cupos para el próximo semestre.

Cifras y Magnitudes

Los datos, pertenecientes a la Universidad del Norte. Resulta complicado acceder a los resultados obtenidos por las instituciones de educación superior debido a la gran relevancia que este tema tiene en el ámbito local y a la limitada disposición para compartir dichos resultados.La información utilizada en la investigación abarca desde el año 2017. En la primera fase, los datos fueron segmentados según una selección de periodos preexistentes, específicamente los periodos 10 y 30.

	Matrícula Estimada	Proyectados Actual (S)	Meta Nuevos	Demanda Máx.	Grupos 40	Grupos 30	Grupos 25
count	8113	8113	8113	8113	8113	8113	8113
mean	57.47	69.95	9.31	85.21	1.58	0.19	0.66
std	104.59	153.76	63.56	160.64	3.63	1.08	3.07
min	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	9.00	10.00	0.00	19.00	0.00	0.00	0.00
50%	27.00	28.00	0.00	42.00	1.00	0.00	0.00
75%	71.00	71.00	0.00	89.00	2.00	0.00	1.00
max	2027.00	2813.00	1948.00	2744.00	55.00	68.00	125.00

Table 1: Resumen estadístico para los periodso 10

	Matrícula Estimada Hist.	Proyectados Actual (S)	Meta Nuevos	Demanda Máx.	Grupos 40	Grupos 30	Grupos 25
count	7910	7910	7910	7910	7910	7910	7910
mean	57.05	64.13	5.03	71.39	1.26	0.20	0.62
std	97.65	136.45	33.50	142.67	3.26	1.17	2.52
min	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	11.00	10.00	0.00	13.00	0.00	0.00	0.00
50%	28.00	28.00	0.00	32.00	0.00	0.00	0.00
75%	73.00	69.00	0.00	81.00	1.00	0.00	1.00
max	1763.00	2243.00	955.00	2243.00	56.00	62.00	43.00

Table 2: Resumen estadístico para los periodos 30

La Tabla 1 presenta un resumen estadístico de los datos correspondientes al periodo 10. Se observa que la matrícula estimada tiene un promedio de 57.47 estudiantes, con una proyección actual de 69.95 y una demanda máxima estimada de 85.21, destacando la mayor cantidad de grupos con 40 estudiantes. Por otro lado, la Tabla 2 refleja los datos del periodo 30, donde la matrícula histórica promedio es de 57.05, con una demanda máxima menor de 71.39, mostrando también una menor cantidad de grupos en comparación con el periodo anterior.

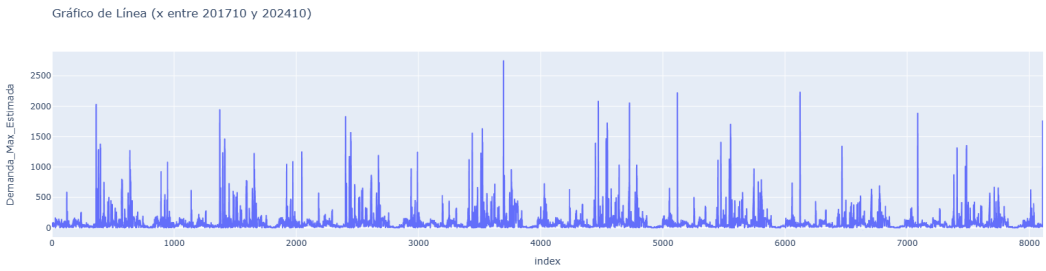

Figure 1: Comportamiento historico de la Demanda maxima estimada periodos 10

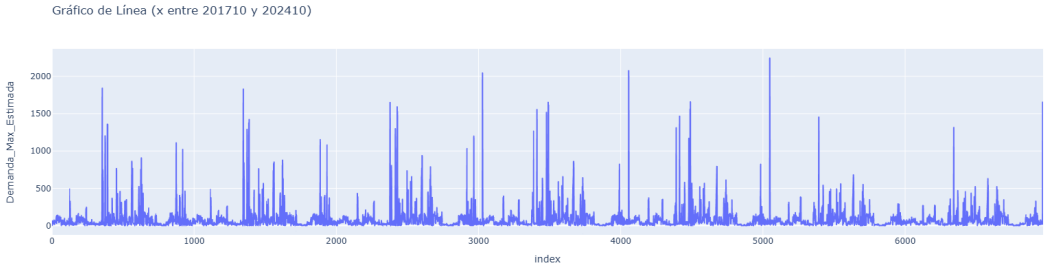

Figure 2: Comportamiento historico de la Demanda maxima estimada peridos 30

Para ambos comportamientos historicos se observa que la demanda alcanza picos significativos, superando los 2000, lo que indica momentos de alta demanda. La tendencia general sugiere variabilidad, con intervalos de baja demanda intercalados con picos, lo que podría reflejar patrones estacionales o eventos específicos que afectan la demanda. Luego de realizar la segmentacion por periodos surge otra, a partir de lo que determinaron los expertos como valores relevantes para generar grupos bajo los cuales se determina una proyeccion. En este caso se realiza una descripcion segun los periodos 30, para cada uno de estos grupos: Demanda maxima estimada < 40 y Demanda maxima estimada ≥ 40.

	Período	Código Dpto	Num. Periódicos	Matrícula Est.	Proyectados	Meta Nuevos	Demanda Máx.	Nro. Grupos
count	3481	3481	3481	3481	3481	3481	3481	3481
mean	202028.74	44.07	5.85	14.65	14.27	1.06	19.45	0.07
std	198.37	17.34	4.64	11.53	10.84	4.48	10.60	0.25
min	201730	0	0	0	0	0	1	0
25%	201830	34	2	2	4	0	12	0
50%	202030	49	5	15	13	0	19	0
75%	202230	52	9	23	22	0	28	0
max	202330	77	20	40	40	37	40	1

Table 3: Descripción para la demanda menor o igual a 40

	Período	Código Dpto	Num. Periódicos	Matrícula Est.	Proyectados	Meta Nuevos	Demanda Máx.	Nro. Grupos
count	3490	3490	3490	3490	3490	3490	3490	3490
mean	202024.61	41.52	8.78	100.63	117.65	8.96	142.40	2.78
std	198.71	17.65	5.50	127.17	187.12	46.90	192.15	4.46
min	201730	0	0	0	0	0	41	0
25%	201830	27	4	45	41	0	62	0
50%	202030	42	9	73	73	0	110	1
75%	202230	51	13	113	132	10	132	1
max	202330	77	21	1763	2243	955	2243	56

Table 4: Descripción para la demanda mayor a 40

Se puede observar que para el grupo con una demanda menor o igual a 40, el promedio del período es de 202028.74, con un desviacion estándar de 198.37. La matrícula estimada media es de 14.65, mientras que los proyectados tienen un promedio de 14.27. En cuanto a la demanda máxima, se registra un valor promedio de 19.45, con un valor máximo de 40.

Por otro lado, para la demanda mayor a 40, el promedio del período es de 202024.61, con una matrícula estimada significativamente mayor de 100.63 y una proyección media de 117.65. La demanda máxima promedio es de 142.40, con un máximo observado de 2243.
La influencia de estos datos es crucial: permiten identificar tendencias en la demanda de cursos específicos, facilitando una distribución más eficiente de recursos. Además, posibilitan la anticipación de necesidades futuras, lo que puede traducirse en una mejora en la experiencia educativa de los estudiantes y una optimización de los recursos institucionales. Sin embargo, se han seleccionado artículos relacionados con la metodología de estudios contemporáneos. Es importante señalar que el periodo de revisión de antecedentes abarca desde 2016 hasta 2024. Otros resultados, vistos desde una agrupacion por departamento, es que pocos de estos datos se comportan ajustados a una distribucion normal, y solo los que hacen parte de la imagen se comportan de esta manera, apoyado en el test de 'shapiro-wilk'.

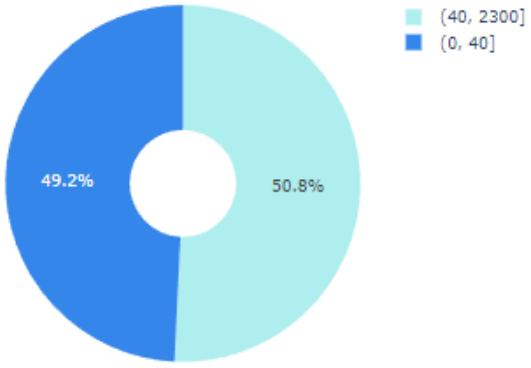

Figure 3: Distribución de la Demanda Máxima Estimada

Antecedentes

En primer lugar, el artículo titulado "Estudio comparativo sobre la planificación del estacionamiento vehicular en campus universitarios: Caso Bogotá, Colombia" (2016) [2], compara los métodos tradicionales de planificación de estacionamientos y su impacto en la oferta de cupos.

En segundo lugar, el artículo "Modelo para definir la capacidad instalada de la Universidad de los Andes" (2016) [4],

desarrolla una herramienta de apoyo para la toma de decisiones relacionadas con el crecimiento de la Universidad de los Andes, considerando el número de estudiantes, profesores y la infraestructura, en alineación con los objetivos estratégicos de mediano y largo plazo de la institución.

En tercer lugar, el artículo "Una mirada a la demanda ocupacional de la carrera de Contabilidad y Auditoría de la Unidad de Educación a Distancia de la Universidad Nacional de Loja" (2019) [5], analiza la demanda insatisfecha de la carrera de Contabilidad y Auditoría, identificando las competencias profesionales adquiridas por los graduados y su capacidad para resolver problemas económicos y financieros en el país.

En cuarto lugar, el artículo "Análisis comparativo de modelos tradicionales y modernos para pronóstico de la demanda: enfoques y características" (2022) [3], examina diferentes tipos de pronósticos de demanda, proponiendo modelos útiles para pequeñas y medianas empresas, basados en datos históricos de ventas.

En quinto lugar, el artículo "Modelos para estimar la demanda en sistemas de distribución" (2022) [1], explora las características del problema de la estimación de la demanda en sistemas de distribución. Posteriormente, describe los diversos modelos empleados para estimar la demanda, así como las metodologías determinísticas y estadísticas utilizadas para calcular los parámetros, resaltando las limitaciones de cada una.

Finalmente, el artículo "Aplicación de algoritmos de Machine Learning para predecir la deserción estudiantil en alumnos de primer y segundo semestre en universidades públicas del Ecuador" (2023) [6], presenta el desarrollo de un modelo matemático basado en redes neuronales artificiales (RNA) para predecir la deserción estudiantil, clasificando a los alumnos entre aquellos en riesgo de deserción y aquellos que no lo están.

Este conjunto de estudios refleja un enfoque contemporáneo hacia la investigación en diversas áreas educativas, con una combinación de métodos y tecnologías modernas para abordar problemáticas como la deserción estudiantil y la planificación universitaria.

Objetivos

Objetivo general

Analizar el comportamiento histórico de los cupos por asignatura, identificar los factores clave que influyen en la demanda de los cursos y desarrollar un modelo para proyectar los cupos necesarios, con el fin de optimizar la planificación académica y la asignación de recursos.

Objetivos especificos

- Describir el comportamiento histórico de los cupos por asignatura.
- Identificar qué factores relevantes explican la demanda de los cupos de los cursos.
- Modelar la proyección de los cupos por asignatura para el próximo semestre.

Objetivos Específicos	**Metodologías**	**Productos**
1. Describir el comportamiento histórico de los cupos por asignatura.	1. Segmentación de datos, identificación de variables y clasificación de tendencias.	1. Informe detallado sobre el comportamiento histórico de los cupos por asignatura.
2. Identificar los factores relevantes que explican la demanda de cupos en los cursos.	2. Visualización de datos para encontrar patrones relevantes.	2. Análisis estadístico y visualizaciones que explican los factores relevantes que afectan la demanda.
3. Modelar la proyección de los cupos por asignatura para el próximo semestre.	3. Identificación de un modelo adecuado al problema que mejore el rendimiento histórico de las predicciones.	3. Modelo predictivo con proyección de cupos por asignatura, acompañado de un informe de resultados.

Trascendencia y Consecuencias

El análisis de la demanda de cupos por asignatura en instituciones educativas tiene implicaciones significativas para la planificación académica y la optimización de los recursos. La creación de un modelo predictivo no solo permitirá a la universidad responder de manera eficiente a las fluctuaciones en la demanda, sino que también optimizará la asignación de recursos docentes y administrativos, reduciendo posibles desajustes entre la oferta y demanda de cursos.

Además, los resultados de este estudio podrían sentar un precedente para futuras investigaciones que busquen aplicar técnicas predictivas en otros ámbitos de la administración educativa, tales como la proyección de matrículas, la planificación de infraestructura, o la distribución de recursos en diferentes programas académicos.

Es importante destacar que los resultados de este proyecto pueden tener un impacto directo en la optimización de la oferta académica en la Universidad del Norte, beneficiando tanto a estudiantes como a administradores. La adecuada proyección de los cupos por asignatura evitará la sobrecarga de cursos o la falta de plazas en asignaturas de alta demanda, lo cual mejorará la satisfacción estudiantil y ayudará a evitar retrasos en la progresión académica. A largo plazo, estos resultados pueden influir en las políticas de planificación académica de otras instituciones educativas que enfrentan desafíos similares. Asimismo, sugieren la posibilidad de desarrollar sistemas de gestión basados en predicciones que aborden otras áreas de la administración universitaria.

Metodologia

Para abordar el primer objetivo, *"Describir el comportamiento histórico de los cupos por asignatura"*, se implementó una metodología basada en la visualización de datos utilizando la herramienta **Streamlit en Python**. En primer lugar, se realizó una recopilación y limpieza de los datos históricos de cupos por asignatura, garantizando la calidad y consistencia de la información. Posteriormente, se desarrollaron gráficos interactivos y tableros dinámicos en Streamlit que permitieron identificar patrones y tendencias relevantes en la asignación y demanda de cupos. Para analizar el comportamiento histórico de los cupos por asignatura, se desarrollaron las siguientes visualizaciones clave:

- **Matriz de correlación:** Se utilizó para identificar relaciones significativas entre variables, como la demanda de cupos por asignatura y factores temporales (semestres, periodos) o características específicas de las asignaturas.
- **Gráficos de líneas de tiempo:** Estas visualizaciones mostraron la evolución histórica de la asignación y ocupación de cupos, permitiendo detectar tendencias, picos de demanda y estacionalidades.
- **Gráficos de barras:** Utilizados para comparar la cantidad de cupos asignados y utilizados entre distintas asignaturas, proporcionando una perspectiva clara sobre las asignaturas más demandadas.
- **Gráficos interactivos:** Implementados en Streamlit, permitieron a los usuarios explorar dinámicamente los datos, filtrando por asignatura, periodo o cualquier otra variable relevante.

Para abordar el segundo objetivo, *"Identificar qué factores relevantes explican la demanda de los cupos de los cursos"*, se utilizó un enfoque analítico basado en técnicas estadísticas y de aprendizaje automático. Se hizo de la mano del análisis exploratorio de los datos para seleccionar variables clave. Posteriormente, se aplicaron modelos de regresión, debido al contexto de los datos, que permitieron determinar el peso e influencia de cada factor en la demanda de cupos, facilitando así la identificación de patrones significativos y relaciones causales.

Para abordar el tercer objetivo, *"Modelar la proyección de los cupos por asignatura para el próximo semestre"*, se implementaron diferentes modelos de aprendizaje supervisado para seleccionar el que mejor se ajustara a los datos históricos y proporcionara predicciones precisas. Los modelos utilizados incluyeron **K-Nearest Neighbors (KNN)**, **XGBoost**, **Linear Regressor**, **Ridge**, **Lasso** y **Support Vector Regressor (SVR)**. Se inició con la implementación de modelos generales fue para medir el comportamiento de los datos reales, teniendo en cuenta que los valores de la demanda, por sí solos, presentan variabilidad descontrolada. Este paso permitió comprender la naturaleza de los datos y preparar un enfoque adecuado para su modelado. Los datos

fueron preprocesados mediante normalización y se dividieron en conjuntos de entrenamiento y prueba. Cada modelo fue evaluado utilizando métricas como el error cuadrático medio (MSE) y el coeficiente de determinación (R^2) para determinar su rendimiento. Además, se realizaron ajustes de hiperparámetros mediante validación cruzada, con el fin de optimizar las predicciones y seleccionar el modelo más adecuado para proyectar la demanda de cupos en el próximo semestre.

Durante la implementación de los modelos, se observó que los resultados obtenidos no se ajustaban correctamente a los datos reales, lo que se reflejó en bajos valores del coeficiente de determinación (R^2) y altos errores en las métricas, como el error cuadrático medio (MSE). Las gráficas de residuos mostraron una clara dependencia, indicando que los modelos no lograban capturar adecuadamente las tendencias y patrones presentes en los datos. Esto puede atribuirse a la alta variabilidad y descontrol de los valores de la demanda, lo que dificultó la capacidad de los algoritmos para generalizar y generar predicciones precisas.

Análisis de Segmentación

Segmento 1: Demanda menor o igual a 40

Para mejorar el ajuste de los modelos, se decidió segmentar los datos según los niveles de demanda. El primer grupo incluye únicamente asignaturas con una demanda menor o igual a 40 cupos. Esta segmentación permitió reducir la variabilidad de los datos y ajustar los modelos a un rango más manejable. Se realizaron nuevamente las pruebas de los modelos **KNN**, **XGBoost**, **Linear Regressor**, **Ridge**, **Lasso** y **SVR**, calculando las métricas de evaluación como el error cuadrático medio (MSE) y el coeficiente de determinación (R^2) para este segmento. Además, se aplicó validación cruzada por cada periodo para garantizar la robustez de los resultados y evaluar cómo los modelos se comportan en predicciones basadas en datos de baja demanda. Esta estrategia permitió observar un desempeño más estable en comparación con el análisis inicial, aunque aún se identificaron áreas para mejorar.

Segmento 2: Demanda mayor a 40

Para el segundo segmento, que incluye asignaturas con una demanda mayor a 40 cupos, fue necesario realizar una subsegmentación adicional para capturar mejor las variaciones dentro de este rango elevado. Esta subsegmentación se llevó a cabo utilizando técnicas de clustering, como ***K-means***, para agrupar las asignaturas en subgrupos con patrones similares de demanda. Estos clústeres permitieron analizar los datos de manera más detallada y ajustar los modelos específicamente para cada grupo, mejorando así el desempeño y la precisión de las predicciones.

Los cuatro clusters identificados reflejan distintos niveles de demanda:

- **Cluster 0 (Demanda Baja)**: Este grupo contiene una mayoría de cursos con demanda moderada, con un promedio de aproximadamente 89 estu-

diantes y un rango de demanda entre 41 y 212. Este cluster representa la mayoría de los cursos, con un total de 6613 clases.

- **Cluster 1 (Demanda Muy Alta)**: Este grupo es el más reducido, con solo 82 clases, pero es significativo por agrupar cursos con demanda extrema, alcanzando en promedio 1488 estudiantes y con un máximo de hasta 2744.
- **Cluster 2 (Demanda Alta)**: Este cluster contiene 238 cursos con un promedio de demanda de 733 estudiantes y un máximo de 1107, sugiriendo la necesidad de planificación para evitar cuellos de botella.
- **Cluster 3 (Demanda Moderada)**: Con 675 cursos, agrupa aquellos con una demanda moderada, en promedio 336 estudiantes, y un máximo de 533.

Para profundizar en el análisis, realizamos un mapeo temporal de la demanda en cada cluster a lo largo del tiempo. Este análisis sugiere que los cursos de demanda baja, moderada y alta (clusters 0, 2 y 3) muestran estabilidad, mientras que los de demanda muy alta (clusters 1) presentan variaciones significativas. Estos cambios podrían deberse a diferentes factores, podemos observar que para cada semestre no hay una tendencia clara. Lo cual justifica para un monitoreo continuo para el cluster.

Resultados y Conclusiones

Rango	**Modelo**	**RMSE**	**MAPE**	$\mathbf{R}^2$	**Ljung-Box p-value**
≤ 40	KNN	1.49	5.774	0.980	0.26
$41 - 213$	KNN	7.7022	4.0012	0.9543	0.642
$213 - 533$	$RandomForest$	13.025	1.23	0.971	0.2237

Intervalo ≤ 40 El modelo **KNN** muestra un **RMSE** bajo de **1.49** y un **MAPE** de **5.774**, lo que indica un desempeño sólido en este rango. El coeficiente de determinación ($\mathbf{R^2}$) es alto (**0.98**), lo que confirma la capacidad del modelo para explicar la variabilidad de los datos en este intervalo. **Intervalo** $41 - 213$ El modelo **KNN** mantiene un buen desempeño con un **RMSE** de **7.7022** y un **MAPE** reducido a **4.0012**. Aunque hay un incremento en los errores absolutos en comparación con el primer intervalo, el modelo sigue demostrando una alta precisión con un $\mathbf{R^2}$ de **0.9543**. **Intervalo** $213 - 533$ En este intervalo, el modelo **Random Forest** toma la delantera, logrando un **RMSE** de **13.025** y un **MAPE** bajo de **1.23**. Este resultado destaca la capacidad de **Random Forest** para manejar datos más complejos en este rango, con un excelente $\mathbf{R^2}$ de **0.971**.

En términos de pruebas estadísticas, los valores de **Ljung-Box p-value** y **Jarque-Bera p-value** confirman que los modelos ajustados son consistentes y

no presentan problemas significativos de autocorrelación o distribución anómala en los residuos.

Rango	**Modelo**	**RMSE**	**MAPE**	$\mathbf{R^2}$	**Ljung-Box p-value**
534_1107	KNN	111.7147	4.028	0.18	0.926
1107 <	$RandomForest$	116.5303	5.4971	0.0	0.0

Para los últimos dos intervalos (**534-1107** y **1107**), los resultados muestran diferencias en el desempeño de los modelos:

Intervalo 534-1107 El modelo **KNN** obtiene un **RMSE** elevado de **111.7147**, acompañado de un **MAPE** moderado de **4.028** y un coeficiente $\mathbf{R^2}$ bajo de **0.18**. Aunque los valores del **Ljung-Box p-value** (**0.926**) y **Jarque-Bera p-value** (**0.038**) sugieren que los residuos son aceptables, la precisión del modelo en este intervalo es limitada.

Intervalo 1107 El modelo **Random Forest** muestra un **RMSE** de **116.5303**, ligeramente más alto que el modelo KNN en el intervalo anterior, junto con un **MAPE** de **5.4971**. Sin embargo, los valores de **Ljung-Box p-value** y **Jarque-Bera p-value** son **0**, lo que puede indicar un ajuste menos robusto.

Estos resultados destacan que, aunque los modelos explorados ofrecen cierta capacidad predictiva, no logran superar significativamente los errores observados con métodos tradicionales. Por el momento, se recomienda seguir utilizando enfoques conservadores en estos intervalos mientras se investigan alternativas que puedan mejorar la precisión y estabilidad de las predicciones.

References

[1] Gladys Caicedo Delgado, Carlos A Lozano, Angélica María Bahamón, and Llefry Arias Ochoa. Modelos para estimar la demanda en sistemas de distribución. *Energia y computación*, 11(1):35–45, 2002.

[2] Fredy Leandro Espejo-Fandiño and Daniel Pérez-Rodríguez. Estudio comparativo sobre la planificación del estacionamiento vehicular en campus universitarios: Caso bogotá, colombia. *Estudios de Transporte*, 20(1), 2016.

[3] César Ángel Fierro Torres, Velia Herminia Castillo Pérez, and Claudia Irene Torres Saucedo. Análisis comparativo de modelos tradicionales y modernos para pronóstico de la demanda: enfoques y características. *RIDE. Revista Iberoamericana para la Investigación y el Desarrollo Educativo*, 12(24), 2022.

[4] Juliana Gómez Sarmiento and Antonio Elías Ochoa Parra. Dinámica modelo para definir la capacidad instalada de la universidad de los andes. 2016.

[5] Whinzon Patricio Cuenca Herrera, Amparito del Rosario Zhapa Ama, Lucia Armijos Tandazo, Jimena Elizabeth Benítez Chiriboga, and José Luis Ríos Zaruma. Una mirada a la demanda ocupacional de la carrera de contabilidad y auditoría de la unidad de educación a distancia de la universidad nacional de loja. *Revista Metropolitana de Ciencias Aplicadas*, 2(2):92–99, 2019.

[6] Cristóbal Alejandro Rodríguez Vásconez. Aplicación de algoritmos de machine learning para predecir la deserción estudiantil en alumnos de primer y segundo semestre en universidades públicas del ecuador. Master's thesis, Universidad Técnica de Ambato. Facultad de Ingeniería en Sistemas ..., 2023.