UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE ECONOMÍA Y EMPRESA
CARRERA DE NEGOCIOS INTERNACIONALES

TEMA:

Estudio de la demanda laboral de las PYMES y grandes empresas de Guayaquil.

AUTOR:
Elías Veliz, Emilio Xavier

Trabajo de titulación previo a la obtención del título de
LICENCIADA EN NEGOCIOS INTERNACIONALES

TUTOR:
PhD. Carrera Buri, Félix Miguel Mgs.

Guayaquil, Ecuador
21 de agosto del 2025

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE ECONOMÍA Y EMPRESA

CARRERA DE NEGOCIOS INTERNACIONALES

CERTIFICACIÓN

Certificamos que el presente trabajo de titulación fue realizado en su totalidad por **Elías Veliz, Emilio Xavier**, como requerimiento para la obtención del título de **Licenciado en Negocios Internacionales.**

TUTOR:

f. _____
PhD. Carrera Buri, Félix Miguel Mgs.

DIRECTOR DE LA CARRERA:

f. _____
Hurtado Cevallos, Gabriela Elizabeth

Guayaquil, a los 21 del mes de agosto del año 2025

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE ECONOMÍA Y EMPRESA

CARRERA DE NEGOCIOS INTERNACIONALES

DECLARACIÓN DE RESPONSABILIDAD

Yo, **Elías Veliz, Emilio Xavier**

DECLARO QUE:

El Trabajo de Titulación, **Estudio de la demanda laboral de las pymes y grandes empresas de Guayaquil.,** previo a la obtención del título de **Licenciada en Negocios Internacionales**, ha sido desarrollado respetando derechos intelectuales de terceros conforme las citas que constan en el documento, cuyas fuentes se incorporan en las referencias o bibliografías. Consecuentemente este trabajo es de mi total autoría.

En virtud de esta declaración, me responsabilizo del contenido, veracidad y alcance del Trabajo de Titulación referido.

Guayaquil, a los 21 del mes de agosto del año 2025

LOS AUTORES

f. _____

Elías Veliz, Emilio Xavier

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE ECONOMÍA Y EMPRESA

CARRERA DE NEGOCIOS INTERNACIONALES

AUTORIZACIÓN

Yo, **Elías Veliz, Emilio Xavier**

Autorizo a la Universidad Católica de Santiago de Guayaquil a la **publicación** en la biblioteca de la institución del Trabajo de Titulación, **Estudio de la demanda laboral de las pymes y grandes empresas de Guayaquil.** Cuyo contenido, ideas y criterios son de mi exclusiva responsabilidad y total autoría.

Guayaquil, a los 21 del mes de agosto del año 2025

LOS AUTORES

f. _____

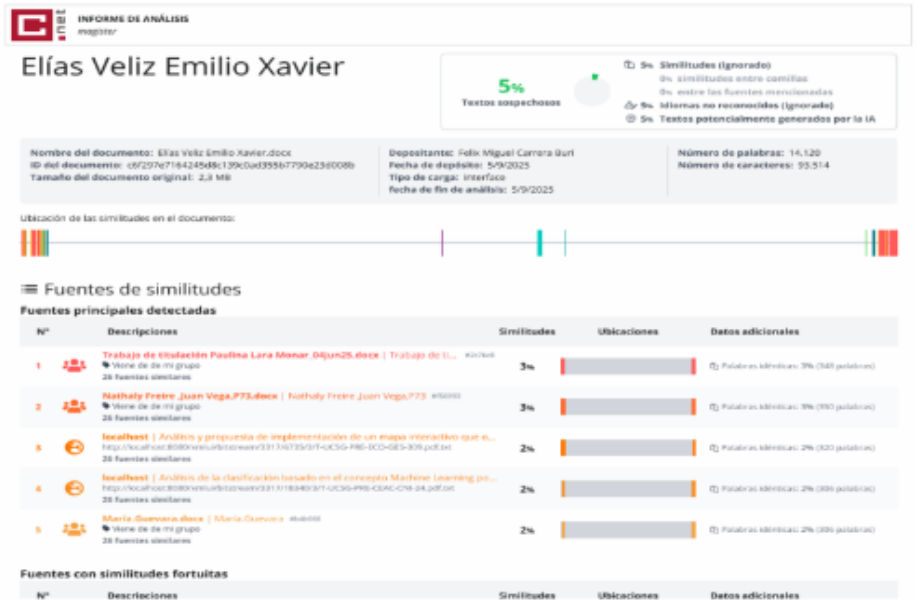
Elías Veliz, Emilio Xavier

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE ECONOMÍA Y EMPRESA

CARRERA DE NEGOCIOS INTERNACIONALES

Reporte Compilatio



INFORME DE ANÁLISIS

Elías Veliz Emilio Xavier

5%
Textos sospechosos

5% Similitudes (Ignorado)
0% similitudes entre comillas
0% entre las fuentes mencionadas
5% Idiomas no reconocidos (Ignorado)
5% Textos potencialmente generados por la IA

Nombre del documento: Elías Veliz Emilio Xavier.docx
ID del documento: c6f297e7164245d8c139c0ad3955b7790e23d008b
Tamaño del documento original: 2,8 MB

Depositante: Félix Miguel Carrera Buri
Fecha de depósito: 5/9/2025
Tipo de carga: interfaz
fecha de fin de análisis: 5/9/2025

Número de palabras: 14.120
Número de caracteres: 93.514

Ubicación de las similitudes en el documento:

Fuentes de similitudes

Fuentes principales detectadas

Nº	Descripciones	Similitudes	Ubicaciones	Datos adicionales
1	Trabajo de titulación Paulina Lara Monar_04jun25.docx | Trabajo de ti... Viene de mi grupo 26 fuentes similares	3%		Palabras idénticas: 3% (549 palabras)
2	Nathaly Freire_Juan Vega,P73.docx | Nathaly Freire Juan Vega,P73 Viene de mi grupo 26 fuentes similares	3%		Palabras idénticas: 3% (550 palabras)
3	localhost | Análisis y propuesta de implementación de un mapa interactivo que e... http://localhost:8080/xmlui/bitstream/3313/6735/3/T-UCSG-PRE-ECO-GES-309.pdf.txt 26 fuentes similares	2%		Palabras idénticas: 2% (320 palabras)
4	localhost | Análisis de la clasificación basado en el concepto Machine Learning po... http://localhost:8080/xmlui/bitstream/3313/18340/3/T-UCSG-PRE-CSAC-CTR-24.pdf.txt 26 fuentes similares	2%		Palabras idénticas: 2% (306 palabras)
5	María_Guevara.docx | María_Guevara Viene de mi grupo 26 fuentes similares	2%		Palabras idénticas: 2% (306 palabras)

Fuentes con similitudes fortuitas

Nº	Descripciones	Similitudes	Ubicaciones	Datos adicionales

Ing. Hurtado Cevallos, Gabriela Elizabet
DIRECTORA DE CARRERA

TUTOR: Ing. Carrera Buri, Félix Miguel Mgs.

AGRADECIMIENTOS

Quiero expresar mi más profundo agradecimiento a quienes han sido pilares fundamentales durante el desarrollo de esta tesis.

En primer lugar, a mis padres, quienes, con su amor incondicional, sacrificio y guía han sido la base sobre la cual he construido mi formación personal y académica. Su constante apoyo, sus consejos y la fe que siempre depositaron en mí han sido la mayor fuente de motivación en este proceso. Gracias por enseñarme con su ejemplo que la perseverancia, la responsabilidad y la honestidad son los valores más importantes para alcanzar cualquier meta. Esta tesis no es únicamente un logro personal, sino también el reflejo del esfuerzo, paciencia y dedicación que siempre han tenido hacia mí.

De manera especial, expreso mi sincero reconocimiento a mi tutor de tesis. Su compromiso, paciencia y acompañamiento fueron indispensables para llevar a cabo este trabajo. Cada orientación, corrección y sugerencia no solo contribuyeron a mejorar el contenido académico de este proyecto, sino que también me ayudaron a crecer como investigador y como estudiante. Su apoyo constante y su disposición para guiarme en cada etapa marcaron una diferencia significativa en la culminación de este proceso.

A ellos, mis padres y mi tutor de tesis, dedico con gratitud estas páginas, como muestra de reconocimiento a la invaluable influencia que han tenido en mi vida académica y personal.

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE ECONOMÍA Y EMPRESA

CARRERA DE NEGOCIOS INTERNACIONALES

TRIBUNAL DE SUSTENTACIÓN

f. _____

PhD. Carrera Buri, Félix Miguel Mgs.

TUTOR

f. _____

Ing. Hurtado Cevallos, Gabriela Elizabeth Mgs.

DIRECTORA DE CARRERA

f. _____

(NOMBRES Y APELLIDOS)

COORDINADOR DEL ÁREA O DOCENTE DE LA CARRERA

ÍNDICE

ÍNDICE VIII

Introducción 2

Problemática 6

Justificación 10

Objetivo general: 15

Objetivos específicos: 15

CAPÍTULO 1 16

Marco Teórico 16

Teoría de la Localización 16

Metodología del Justo a Tiempo (Just in Time – JIT) 22

Modelo Random Forest 25

Planteamiento del modelo Random Forest 28

Modelo Árbol de decisión 31

Planteamiento del modelo Árbol de decisión 34

Marco Referencial 37

Marco Conceptual 39

La Demanda 39

Demanda de mano de obra 40
Random Forest 41
Árbol de decisión..... 42
MARCO LEGAL..... 42
Metodología 45
R studio..... 45
Tipo de investigación 45
Métodos 46
Árbol de decisión 47
Radom Forest 47
Librerias Implementadas..... 48
Script del modelo desarrollado mediante el framework Rstudio 50
Visualización dentro del software 61
RESULTADOS..... 61
Distribución general 63
Diferencias específicas 64
Clases minoritarias..... 64
DESARROLLO DEL ÁRBOL 64
1. Estructura básica..... 64
2. Principales resultados observados 65

Interpretación de reglas del Árbol 66

Conclusión de los resultados 68

DISCUSION 69

CONCLUSIONES 72

RECOMENDACIONES 73

REFERENCIAS 74

RESUMEN

La presente investigación estudia la demanda laboral de las PYMEs y grandes empresas de Guayaquil mediante la aplicación de la ciencia de datos como herramienta analítica. El trabajo parte del contexto ecuatoriano caracterizado por elevados índices de desempleo y subempleo, así como por un desajuste entre las competencias de la población y los perfiles que requieren las organizaciones. Para el análisis se emplearon datos públicos del Instituto Nacional de Estadística y Censos (INEC), a través de la encuesta ENEMDU, procesados con el software R Studio. Se aplicaron modelos de aprendizaje supervisado, específicamente árboles de decisión y bosques aleatorios, alcanzando una precisión del 95%. Los resultados evidencian que las ocupaciones en los sectores “Privado”, “Externo” y “Obrero” concentran la mayor parte de la demanda, mientras que categorías como “Doméstico”, “No remunerado” y “Ayudante” presentan menor representación. La investigación concluye que la aplicación de modelos predictivos permite identificar patrones ocupacionales, anticipar tendencias de contratación y facilitar la toma de decisiones tanto para las empresas como para las políticas públicas en el ámbito laboral ecuatoriano.

Palabras clave: demanda laboral, PYMEs, Guayaquil, ciencia de datos, árbol de decisión, random forest, ENEMDU.

ABSTRACT

This research analyzes the labor demand of SMEs and large companies in Guayaquil, Ecuador, using data science as the main methodological tool. The study addresses the existing imbalance between the skills required by companies and the available workforce, in a context marked by high levels of unemployment and underemployment. Public data from the National Institute of Statistics and Census (INEC), specifically the ENEMDU survey, was processed and modeled using supervised machine learning techniques. Decision trees and Random Forest algorithms were applied in R Studio to classify and predict occupational demand, achieving an accuracy rate of 95%. Results show that the categories “Private,” “External,” and “Worker” concentrate most of the demand, while minor categories such as “Domestic,” “Unpaid,” and “Helper” remain underrepresented. The findings highlight the potential of predictive models to reduce recruitment mismatches, guide corporate strategies, and support evidence-based public policies for labor management in Ecuador.

Keywords: labor demand, SMEs, Guayaquil, data science, decision trees, random forest, ENEMDU.

Introducción

Para dar ímpetu, un tema controversial en el Ecuador ya que esto lleva años siendo la demanda laboral un contribuyente a la sociedad en un ámbito económico dentro del país. La cual se encarga de generar empleos para la sostenibilidad económica de una persona, En la ciudad de Guayaquil siendo unas de las ciudades con más productividad dentro del Ecuador es vista con muchos más conflictos sociales y políticos, el cual viene siendo tendencia hace ya mucho tiempo atrás, siendo esta ciudad tan diversa, tiene más relevancia la demanda laboral el comportamiento de esta causa. El análisis de la demanda laboral contribuye a la comprender como también es el funcionamiento de las pymes y grandes empresas de Guayaquil. Esta ciudad ha sido reconocida por la productividad que ha generado en los puestos de trabajo, muchos más enfrascado en la producción agrícola, pero sin dejar de lado los puestos administrativos donde se agradece a los emprendedores y empresas multinacionales situadas en la ciudad ya que gracias a estas es donde existe la oportunidad de superar obstáculos y el desarrollar mejores competencias a nivel mundial. Ecuador al ser un país diversificado entre ganaderos y más grupos rurales. para esta situación, sabemos que las pymes son un gran grupo representativo para la ciudad de Guayaquil, categorizándose así por su flexibilidad de trabajo, ya que aportan a la creación significante de empleos. pero este tema ha sido controversial a lo largo de los años, ya que siempre ha existido el dilema de que por qué entre las diferentes clases sociales existen una gran demanda laboral, o a su vez una decisión laboral notoria, por otro lado, las

grandes empresas o multinacionales están vinculadas en sectores estratégicos donde entran diferentes ámbitos laborales, ya sea la logística, telecomunicaciones, sector financiero económicos donde todas estas requieren de capacidades intelectuales mayores a la media además, demandan perfiles con mayor nivel de especialización dónde subes. Esta ofrece un empleo más estable, con mejores condiciones, salariales y acceso a muchos más beneficios. existen diferencias notorias dentro de las pymes y grandes empresas, ya que este crea único sistema laboral, heterogéneo dónde se conoce que cada uno de estos conviven distintos tipos de oportunidades y desafíos, ya que en las pymes enfrentan limitantes en cuanto a recursos y capitales. Por otro lado, las grandes empresas suelen ser todo lo contrario ya que estas velan por la innovación y la incorporación de talentos mucho más calificados, asimismo, existen factores externos dentro de las políticas gubernamentales, encargadas de fomentar el empleo dónde entran diferentes regulaciones actuales, como por ejemplo las transformaciones digitales, la cual impacta directamente a la capacidad de la contratación del personal dentro de una empresa. este año ha sido crucial dentro de la clasificación de un Personal capacitado para cumplir diferentes actividades y tareas que desempeñará el crecimiento de una empresa, sabiendo que existen diferentes controversias en cuanto a la capacidad de una persona en cumplir y destacar ciertos rasgos necesarios para la realización de tareas, se categoriza más por su experiencia, acompañado con sus estudios que únicamente la experiencia como tal, El desarrollo de nuevas áreas de trabajo son un segmento esencial donde se mide la flexibilidad y la capacidad de adaptación del empleador

y como este aporta dentro de la empresa. A su vez en las pymes, se categoriza por el aporte a la creación de empleo constantemente donde las pymes demandan, perfiles más diversos que respondan a consumos, servicios o comercios, donde este va de la mano a la absorción de la mano de obra juvenil, a su vez disminuyendo en cierta parte brechas de desempleo o su empleo. Para las grandes empresas se tomen en cuenta sectores estratégicos, tanto como la industria logística, telecomunicaciones finanzas, donde influye en un mercado laboral, mucho más amplio y estas demandan de una alta especialización donde se goza de oportunidades estables, las cuales van relacionadas con avances tecnológicos, y también a la globalización, donde se exige competitividad internacional, las grandes empresas también van de la mano con una actualización constante por las competencias. El gobierno ha brindado la recopilación y recaudación de datos estadísticos, donde dónde se puede visualizar cómo funciona la demanda laboral, o en qué se rige la deserción laboral, y como se analiza el empleo en el Ecuador, gracias al Instituto nacional de estadística y censos, informa la cantidad exacta de los diferentes grupos sociales, dentro del ámbito laboral, esta es conocida como ENEMDU encuesta nacional de empleo, desempleo y subempleo, el cual el INEC pone a disposición de los ciudadanos a los resultados anuales de los diferentes años y con esta información, los usuarios podrán analizar la situación del mercado laboral, pobreza, entre otras, con un grado mayor de desagregación, cómo puede ser a nivel provincial en un periodo más amplio, como es el anual, Los censos de población y vivienda son realizados cada 10 años en el Ecuador. Así lo establece

la Constitución y las recomendaciones internacionales para un periodo de censo. Sin embargo, el censo que fue previsto para el 2020 habría sido post Delgado por la pandemia del COVID-19 y éste se llevó a cabo dos años después, para la seguridad de todos los ciudadanos. reconociendo que existen diferentes situaciones las cuales no permiten a ciertas personas ser incluidas en el ámbito laboral, como son las personas que tienen alguna discapacidad, las leyes ecuatorianas exigen que por lo menos una persona que padezca alguna discapacidad tiene que ser empleador de alguna empresa. la ley orgánica para la justicia indica que la persona trabajadora con discapacidad, una vez que haya superado el periodo de prueba de 90 días pasará a formar parte de la empresa, de manera indefinida, sea esta una empresa privada o pública, la condición no debe ser un impedimento legal para ello es por esto que la ley lo promueve en la inclusión laboral para que estos puedan usar de los mismos derechos y estabilidad laboral que requiera.

Problemática

en la actualidad, el Ecuador enfrenta distintos panoramas laborales, las cuales son caracterizadas por crecientes diversidades de los perfiles ocupacionales que requieren dentro de las empresas ya sea privado público, las áreas demandadas, es decir, los diferentes tipos de empleos que son más solicitados, se han convertido en un tema crucial para la demanda laboral, y como las competencias para ocuparlos, esto va a influir en un reto Constante visto como para los empleadores como para trabajadores y personal capacitado para la selección de estos. Uno de los principales problemas se ha visto reflejado en un cierto desequilibrio por parte de los ciudadanos del Ecuador, lo cual está derivando a varios índices de desempleo y subempleos, esto se puede ver reflejado en área solicitadas, como son las partes operativas, comerciales, productivas, económicas, incluso de atención al cliente. Esto ocasiona un cierto injusto a la validación de títulos profesionales otorgados por diferentes universidades del país. Se observa una falda de oportunidades enfocado en los jóvenes con formación académica, universitaria o técnica avanzada.

por otra parte, los cambios que se ha sufrido actualmente con la transformación digital y diferentes adversidades tecnológicas, han representado una dificultad adicional en la demanda ocupacional. Por esto cada vez, empresas públicas privadas, grandes y pequeñas, requieren personal capacitado para suplir estas actividades donde son las que más se encuentran falencias. Para esto, el Instituto nacional de estadística censo nos otorga y ofrece una base de datos exhaustiva para analizar la demanda laboral de Guayaquil el cual es la

encuesta nacional de empleo, desempleo y su empleo, donde se otorga las diferentes ocupaciones, demandadas en la actualidad.

con el uso de la ciencia de datos, se visualizará diferentes ocupaciones en cuanto a los ciudadanos de la ciudad de Guayaquil, para así obtener una información detallada de cuales categorías de ocupación son las más demandadas actualmente, y poder concluir a un análisis más asertivo. en otra parte, la ausencia de estudios estadísticos sobre la demanda ocupacional, dificulta el análisis dentro de políticas públicas, donde promueve el empleo sin una información detallada y actualizada de los sectores en expansión y áreas de trabajo. Resulta complejo crear estrategias beneficiadoras para la fomentación del trabajo, donde se incluya socialmente y mejoren la competitividad entre más empresas. Uno de los dilemas actuales en la ciudad de Guayaquil, es la ausencia de información sobre las áreas, con mayor contratación que produce la gran parte de la población y las altas tasas de desempleo y subempleo, esto provoca que las pymes y grandes empresas no encuentren un perfil adecuado y que esto afecte a los empleadores a acceder a un contrato definido o indefinido, siendo esto, uno de los causantes más controversiales actualmente en la pobreza, llevando consigo, problemas políticos y sociales. Las empresas sufren dificultades en la planificación empresarial, al no contar con diferentes estudios precisos de la demanda laboral que enfrentan actualmente. En la sociedad existirá diferentes problemas al planificar los procesos de selección y capacitación del personal, de igual manera impactará a la productividad y

competitividad del resto de organizaciones globalizando, un contexto tecnológico y multinacional,

Sin la información adecuada sobre el empleador, con mayor experiencia o profesionalismo, no podremos obtener una clara demanda de los puestos más cotizados. Para esto, se generará técnicas mediante las ciencias de datos para así obtener junto a modelos predictivos, como son árboles de decisión y Random Forest, un diseño claro para visualizar la demanda ocupacional, dependiendo a los habitantes de Guayaquil, a su vez, diseñando, programas de formación técnica y universitaria, lo cual lo volverá poco eficiente. Esto beneficiará de manera pública al gobierno donde se obtendrá brechas salariales donde no hay claridad sobre la demanda laboral, y esto permite que exista salarios muy bajos y así aprovechando la sobreexplotación de empleadores o trabajadores, para luego ser un tema político y social, que provoque controversia a lo largo de los años en los ciudadanos ecuatorianos, se sabe que los inconvenientes con el tema de desempleo, son más segmentados en referencia a los sectores o sitios con alta generación de empleo, pero persiste una falta de precisión en la identificación de cuales ocupaciones tienen mayor demanda y como evolucionan estos con el pasar del tiempo, donde no se reconoce muchas de las ocupaciones actuales, como podría ser un empleada doméstica, obrero, jornalero o simplemente una persona contratada para el jardín, o ya sea una persona externa contratada dentro de una empresa para prestar sus servicios, es por esto que se plantea una situación donde sea de mayor visualización, la demanda de las ocupaciones que presentan los ciudadanos en Guayaquil, se ponen mesa,

tomando en cuenta diferentes variables de cómo esto puede surgir mediante el pasar del tiempo, y como esto puede verse reflejado como una falta de desempleo en la ciudad y así tomar en cuenta las diferentes dificultades que permanecen en el ámbito laboral, dentro de la ciudad de Guayaquil. otro de los problemas recurrentes es que las estadísticas tradicionales para el empleo suelen ser descriptivas, lo que no permite anticipar cambios en las diferentes áreas de trabajo o ocupaciones que desempeñan las personas a lo largo del tiempo, es decir, para una persona externa se complica el hecho de saber cuánto tiempo indefinido sea aparte de un ente privado o público. gracias al apoyo y la ciencia de datos determinará una brecha entre la demanda, ocupación al real y la demanda ocupación del previo, ya que al realizar la comparación entre los datos reales, los resultados que arroje el modelo a diseñar, aparecerán discrepancias más significantes al momento de realizar dicha comparación, y es tu refrigerada que existen factores externos como políticas públicas, nos ayudará a determinar también el área más demandada, coyunturas, económicas y transformaciones tecnológicas, ya que no siempre se capturan los modelos estadísticos, limitando la exactitud en las predicciones a su vez, identifica las variables más determinantes, las cuales nos ayudarán a analizar múltiples factores como la edad, el nivel educativo, la experiencia, un salario esperado, el sector económico, entre entre muchas otras, y que por lo general no queda claro cuáles son las variables con mayor peso en la obra de la contratación empresarial laboral y es este vacío, dificultará generar propuestas de capacitación efectivas y enfocar realmente en los perfiles solicitados, así ayudando al reclutamiento por

parte del área de recursos humanos, a beneficiar de una manera congruente a un área pública o privada.

Justificación

La demanda laboral dentro de las pymes y grandes empresas es un sector altamente sensible debido a la naturaleza de los datos proporcionados, los cuales requieren condiciones específicas de recolección de información, lectura analítica de datos específicos además de entrevistadores capacitados para este ámbito. Éste contribuye a un ámbito crítico en el desarrollo económico social de la ciudad, las transformaciones en los patrones de empleo, la adaptación de tecnología a nuevas áreas de trabajo y la competitividad del mercado laboral. Hacen que la gestión que desarrolle el área de talento humano se convierte en un factor determinante para el bien de una empresa. Con esto, las empresas desafían, enfrentamientos, significantes desde el análisis de un perfil apto para las diferentes áreas, hasta el análisis de un perfil que puedan anticipar la evolución de las necesidades entre actividades administrativas y operativas. La aplicación de la ciencia de datos surge como una herramienta estratégica, la cual permite optimizar el proceso. Mediante análisis de información provenientes de fuentes como el Instituto nacional de estadística y censo, las pymes y grandes empresas podrán analizar una mejor oferta y demanda para la ocupación o área de trabajo de sus empleadores, como también prever tendencias de contratación y así adaptar políticas de reclutamiento y capacitación, lo cual impactará grandemente a las empresas también contribuyendo a reducir desajustes en el mercado laboral y competencias disponibles en la población. Podemos encontrar

principales beneficios para una propuesta de nuevos perfiles, a destacar dentro de las pymes o grandes empresas de la ciudad:

- ofrecer un horizonte real y nuevo sobre los sectores áreas o perfiles con mayor demanda en la ciudad de Guayaquil.
- reduciendo costos desarrollados por la variación de la demanda laboral o áreas de trabajo, evitando la rotación excesiva de empleadores y la sobreoferta en áreas poco requeridas.
- Facilitando la toma de decisiones basada en datos estadísticos, proporcionados por el gobierno que se han llevado tanto al nivel empresarial, como para políticas públicas de empleo.

de esta manera, la siguiente justificación no sólo limita a un interés únicamente académico, a su vez, este responde a una necesidad práctica y de manera urgente ya sea mejorar la similitud entre las demandas del mercado laboral y las capacidades que desempeñan los empleadores. Y de conocimiento público que una adecuada gestión de la demanda laboral, impactaría directamente a la competitividad y desarrollo de pymes y grandes empresas para la ciudad de Guayaquil, también, generando empleo, dignos y capacitaciones para seguir con una enseñanza de manera óptima, lo cual impactaría directamente en la generación del empleo y desarrollo económico sostenible para la ciudad de Guayaquil o el Ecuador. Al incorporarse técnicas a base de ciencia de datos y modelos predictivos. Se presenta una alternativa más innovadora para analizar ciertos fenómenos con mayor precisión, a través del proceso de información histórica, si es posible identificar patrones en la demanda

ocupacional en la contratación de empleadores y el análisis del área laboral más cotizado. la relevancia de este estudio radica la capacidad de contribuir tanto de sector empresarial como el ámbito académico gubernamental. Para las empresas va a representar cierta facilidad al momento de la selección, información y desarrollo del personal, no es así en general, propuestas, sustentables y sostenibles, donde se favorezca al desarrollo económico y la productividad empresarial, incluyendo la sociedad en el entorno de una constante transformación. la demanda laboral ocupacional indica uno de los principales factores para comprender el funcionamiento del mercado de trabajo, ya que no sólo refleja la cantidad de empleos requeridos por las empresas, sino también las características específicas de los empleadores o perfiles que se van a querer seleccionar.

En la ciudad de Guayaquil, donde influyen pymes y grandes empresas y en sectores estratégicos, como la agricultura y comercio servicios la logística del estudio de una demanda ocupacional va adquirir en la relevancia, suficiente, permitiéndonos y dimensionar con precisión cuales áreas de trabajo, son mayor representadas dentro de una empresa y cuáles presentarán diferentes riesgos o dificultades para poder emplearse.

en la práctica diferentes empresas desafían un constante problema de reclutamiento para talento humano, que se ajuste tanto a las exigencias de cada puesto, como lo que busca la empresa, mientras que en algunas pymes, principalmente demanda la mano de obra operativa, y un personal que sea más poli funcional, el cual puede adaptarse a diferentes tareas o actividades, las

grandes empresas necesitan profesionales especializados y con la capacitación avanzada, generando diversidad a un panorama heterogéneo que, si no es analizado adecuadamente, va a provocar desajuste en la asignación por parte de la fuerza laboral y este aumentará índice sobre empleo y desempleo en la ciudad.

gracias a la aplicación de la ciencia de datos y los modelos estadísticos realizados para el análisis de la demanda ocupacional, constituye a una herramienta innovadora que transformará la manera en las que se tomarán decisiones empresariales y gubernamentales, permitiéndonos así la facilidad del uso de nuevos datos estadísticos proporcionados por el árbol de sillón, acompañado con bosques aleatorios donde se realizará de una manera más sencilla y gráfica de datos que serán de mucha ayuda para la realización de la demanda ocupacional laboral permitiendo que se analicen hallazgos de nuevos perfiles asociados al área en la que se desempeñará sus labores, al igual que el uso de algoritmos predictivos donde será posible. Identificar mayor demanda en el corto mediano plazo, influyendo, factores como edad, experiencia o nivel de educación.

de tal manera, la justificación para la investigación, no sólo limitará un interés descriptivo, sino que responderá una necesidad para generar información estratégica Tales como:

- El fortalecimiento de una planificación empresarial, donde las pymes y grandes empresas, anticipen, requerimientos de personal o

empleadores, diseñando, políticas, efectivas, de contratación y capacitación.

- Reduciendo la desarticulación entre la fértil y la demanda laboral, contribuyendo a su vez a disminuir niveles de desempleo y su empleo en Guayaquil.
- Se apoya la toma de decisiones públicas y privadas, brindando evidencias, certeras sobre áreas de trabajo u ocupaciones demandadas.
- Permitiendo su vez la contratación de perfiles para mayor oportunidad de inserción y así reduciendo la incógnita del desempleo en el Ecuador.

Bajo la perspectiva académica, el análisis y apoyo de la ciencia de datos constituye para una acción estratégica, beneficiando y siendo de gran apoyo a empresas para su competitividad al igual que para la sociedad, en general, donde se va a destacar mayores oportunidades de empleo, fomentando un crecimiento económico para el desarrollo de nuevas fuentes de empleo

.

Objetivo general:

Adaptar la ciencia de datos para el análisis de la demanda laboral de las pymes y grandes empresas de Guayaquil.

Objetivos específicos:

- Analizar la demanda ocupacional de las pymes y grandes empresas en la ciudad de Guayaquil, y como estas incluyen en la Sociedad

- Aplicar la ciencia de datos y modelos estadísticos, específicamente árboles de decisión y bosques aleatorios mediante el software R Studio, para clasificar y predecir la demanda ocupacional entre las características del Instituto nacional de censo y estadística

- comparar la demanda ocupacional real, con la demanda ocupacional, Pre, dicha, evaluando la precisión de los modelos generados para la proyección empresarial y académica

CAPÍTULO 1

Para que esta investigación tenga una base académica rigurosa es necesario abordar ciertas teorías, conceptos, trabajos referenciales y leyes que resultan importantes para la realización de esta. En este caso, se ha realizado la revisión de la literatura de la Teoría de la localización, de la Gestión de la Cadena de Suministro, de la metodología Just in Time, y de los modelos predictores Random Forest y Árbol de decisión. Además, se repasó algunos conceptos y leyes para que de esta manera, se logre una base teórica que permita que este trabajo tenga una mayor relevancia académica.

Marco Teórico

Teoría de la Localización

La historia de los modelos de localización varía a lo largo de las épocas: la época arcaica (desde tiempos inmemoriales hasta el año 1908), en la que se escribieron contribuciones ocasionales, pero que no estaban relacionadas entre sí y no constituían una línea de investigación. A esta le siguió la época media, de los años comprendidos entre 1909 y 1963. Durante estos años se sentaron las bases: fue entonces cuando las distintas disciplinas iniciaron sus propias líneas de investigación como la geografía y la economía. La época clásica comenzó en el año 1964 y se caracterizó por la discusión de todos los tipos básicos de problemas: las medianas, los centros y los problemas de cobertura, junto con sus propiedades y técnicas de solución específicas para ellos (Marianov & Eiselt, 2024).

En este contexto, la época clásica de los modelos de localización terminó alrededor del año 1978, cuando comenzó ya la era moderna, en la que las contribuciones de la investigación comenzaron a centrarse en problemas específicos y se incluyeron características adicionales como por ejemplo los objetivos múltiples, consideraciones de equidad, entre otras. Por otro lado, si se considera esta línea temporal, la era arcaica de la teoría de la localización comenzó alrededor del año 1640, cuando Fermat planteó el problema a Torricelli (el inventor del barómetro), quien, o su alumno Viviani, lo resolvió. No obstante, la contribución de Johann Heinrich Von Thünen rompió el patrón de las contribuciones puramente matemáticas e introduce un modelo geográfico del uso del suelo en la discusión (Marianov & Eiselt, 2024).

En efecto, los famosos anillos de Von Thünen delimitaron zonas de uso del suelo alrededor de un lugar central en un plano idealizado sin rasgos distintivos, minimizando los costos de transporte al centro como objetivo. Esto dio inicio, aunque mucho más tarde, al campo de la teoría de la localización central. En este contexto, un impacto importante en la teoría de la localización se produjo a principios del siglo XX, cuando el matemático ruso Georgy Voronoi en el año 1908 trabajó en formas cuadráticas. En pocas palabras, un diagrama de Voronoi o polígono de Thiessen es una teselación del espacio que resulta de asignar todos los puntos del espacio dado al punto de partida más cercano. Los diagramas de Voronoi a su vez se basaron en trabajos anteriores de Descartes en la década de 1640 y Dirichlet en el año de 1850 (Onditi & Yates, 2021).

De la misma manera, durante la Edad Media, esta teoría experimentó un momento culminante en los estudios de un geógrafo como Weber en el año de 1909, cuando escribió su obra sobre la Ubicación de las industrias, que situó el problema puramente matemático de Fermat en el contexto de los puntos de oferta y demanda, y la minimización de los costes de transporte. Asimismo, el apéndice de Georg Pick formalizó los argumentos de Weber y describió una técnica de solución para la determinación de una nueva planta de procesamiento entre los puntos de oferta y demanda, con el fin de minimizar la suma de las distancias ponderadas entre los puntos de oferta y demanda y la planta de procesamiento. Como tal, puede considerarse una extensión ponderada del problema de Torricelli. Cabe destacar también que la técnica de Pick difiere de la de Launhardt del año de 1882 para el mismo problema (Drezner y Eiselt, H, 2024).

Posteriormente, el siguiente impacto de esta teoría provino de una dirección muy diferente. En el año de 1929, el estadístico convertido en economista Harold Hotelling describió un duopolio en un mercado lineal simple con clientes distribuidos uniformemente, en el que dos empresas compiten por la ubicación y los precios de un único bien homogéneo. En este caso, su conclusión fue que ambas empresas se ubicarían arbitrariamente cerca una de la otra en el centro del segmento de línea, lo cual, fue cuestionada posteriormente y se demostró que era falsa medio siglo después, ya que solo se cumple en condiciones muy restrictivas, como precios fijos e iguales (Mariotti et al., 2024).

En ese contexto, desde el mundo minorista surgió la contribución de Reilly en el año 1931, quien propuso la ley de gravitación minorista. Según esta ley, los centros comerciales atraen clientes en proporción directa a su tamaño e inversamente proporcional a la distancia elevada a una determinada potencia, lo que recuerda a la ley de gravitación de Newton. Esta ley determinista se concibió como una heurística para determinar las áreas de mercado de los diferentes centros de una región. Mucho más tarde, Huff en el año de 1964 presentó una versión probabilística de la ley de Reilly, ampliamente utilizada por economistas, así como en la teoría del marketing y la localización competitiva. Unos años más tarde, Wilson en el año de 1967 proporcionó una sólida base analítica al modelo de Huff, al obtener la fórmula mediante la maximización de la entropía (Correia, 2022).

En lo que respecta a Gilmore y Leonard (2020), estos autores señalaron que los economistas desarrollaron la teoría de la localización industrial a principios del siglo XX, centrándose en las empresas individuales y las variables que influyen en la selección de nuevos emplazamientos. Estas teorías económico-espaciales neoclásicas consideran a los empresarios como tomadores de decisiones racionales que poseen conocimientos y habilidades perfectos en un proceso de selección racional que conduce a los mejores resultados en términos de costes, ingresos y beneficios

Además, Stevens y Shearmur (2020) señalaron que antes de la década de 1960, el trabajo analítico de esta teoría consistía en interpretar la ubicación de plantas o industrias individuales con referencia al marco conceptual

proporcionado por la teoría neoclásica. El objetivo era buscar la ubicación ideal en un momento determinado, y el enfoque más adecuado era analizar la ubicación de industrias pesadas esenciales, como las siderúrgicas, que estaban a la vanguardia del progreso industrial contemporáneo. En efecto, el rápido crecimiento económico de la década de 1960 resultó en una inversión excepcional, y quizás única, en nuevos establecimientos manufactureros en Europa Occidental, Norteamérica y Japón, lo que generó un creciente interés académico y político en la toma de decisiones sobre la ubicación.

Este período marcó el inicio de la teoría conductual de la localización, que se centra en la geografía, el crecimiento y el comportamiento de las empresas, que no se consideran unidades racionales de toma de decisiones económicas, sino unidades regidas por objetivos contrapuestos, conocimiento y control ambiental limitados, percepciones y comportamientos irracionales, etc. Así pues, la teoría conductual de la ubicación explora factores internos como la antigüedad y tamaño que son importantes en el proceso de toma de decisiones y que llevan a una empresa a elegir una ubicación específica. Según la teoría del comportamiento, es más probable que un emprendedor que tenga que trasladar su empresa elija un lugar cercano, ya que le resulta más familiar o más fácil de imaginar que un lugar distante (Leonard, 2021).

En las décadas de 1970 y 1980, creció el interés por las instituciones culturales, los sistemas de valores y las innovaciones sociales. Estos nuevos patrones se adoptaron en enfoques institucionales, donde el comportamiento de ubicación era el resultado de negociaciones entre la empresa y diversas

entidades locales y nacionales. En el enfoque institucional, factores no materiales como la confianza y el capital social son elementos clave en todos los niveles económicos. En efecto, el comportamiento de ubicación de una empresa resulta de su interacción con proveedores, el gobierno, los sindicatos y otras instituciones (Kézai & Skala, 2024).

En lo que se refiere al enfoque más reciente, desarrollado desde principios de la década de 1990, es la teoría de la toma de decisiones que subyace a la economía evolutiva. Este enfoque evolutivo se basa en el comportamiento rutinario más que en la elección racional. Según la teoría evolutiva, las empresas no están dispuestas a cambiar de ubicación porque su competitividad está determinada por los conocimientos, las rutinas y la experiencia que han adquirido (dentro de un entorno local particular), que son difíciles de imitar para los competidores (Ma et al., 2023).

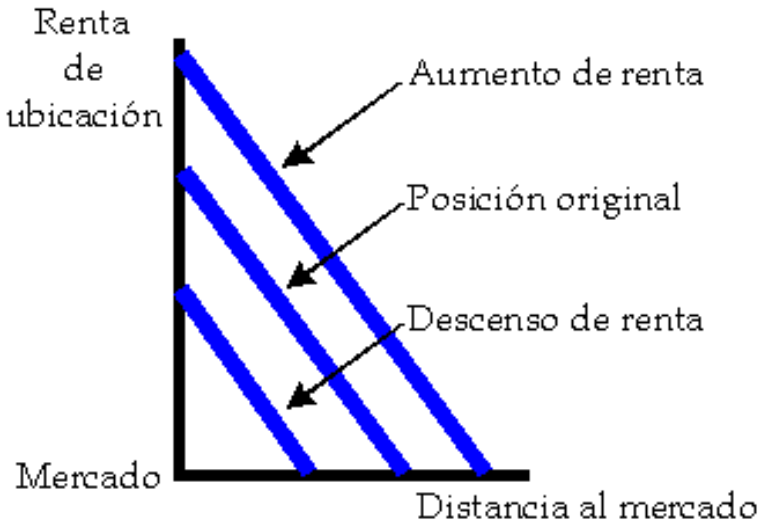


Figura 1Teoría de la Localización

Metodología del Justo a Tiempo (Just in Time – JIT)

Justo a Tiempo (JIT) es una metodología que se presenta como un sistema que controla los procesos técnicos y de recursos humanos en la organización según su uso. La desarrolló la empresa Toyota y su filosofía JIT busca eliminar todas las actividades que no son importantes y que no aportan valor añadido dondequiera que se encuentren. La definición de Justo a Tiempo establece que el sistema JIT es un sistema integral de gestión de inventario y fabricación donde las materias primas y diversas piezas se compran y producen en el momento en que se producen y se utilizan en cada etapa del proceso de producción o fabricación (Guo et al., 2020).

En este mismo contexto, el concepto del sistema *Just In Time* puede definirse como una serie de actividades de producción que utilizan inventario en forma de materias primas mínimas, y que luego se procesan para obtener productos terminados. Cabe recalcar que este concepto también se basa en el supuesto de que no se producen bienes hasta que se necesita su producción (Ralahallo, 2021).

De este modo, el objetivo de JIT en algunas ocasiones es proporcionar la información correcta a las personas adecuadas en el momento oportuno para que (la empresa y el proveedor) puedan responder inmediatamente al mercado tan pronto como se reciba la información (pedido). Por tanto, se requiere un equilibrio entre la oferta y la demanda de producción para lograr un proceso de producción continuo y estable. Por esta razón, la planificación agregada es

necesaria para equilibrar y determinar el nivel general de producción a corto o mediano plazo ante la fluctuación de la demanda (Xu et al., 2025).

En este sentido, Hussein y Zayed (2021) alegaron que existe una relación entre el sistema *Just In Time* y el rendimiento de la empresa dado que la medición del rendimiento organizacional se puede observar a través de los siguientes indicadores: la capacidad para alcanzar la cuota de mercado, la capacidad para crear nuevos productos, la capacidad de la empresa para operar con el máximo rendimiento, altos niveles de productividad y la capacidad para satisfacer las necesidades de los clientes.

Por otro lado, Li et al. (2023) manifestaron que JIT está estrechamente relacionado con el rendimiento financiero y de mercado. Así pues, la menor estructura de costos generada por la implementación del JIT resultará en un mejor rendimiento financiero relativo en comparación con la competencia, así como en un aumento del retorno de la inversión (ROI). Además, el JIT también tiene una correlación positiva con un mejor rendimiento del mercado, ya que los indicadores de mercado, como el crecimiento de las ventas y la cuota de mercado, también aumentan cuando se implementa con mayor frecuencia. En definitiva, se alegó que existe un estudio realizado donde se afirma que la aplicación del JIT tiene un efecto en el rendimiento de la empresa.

En un sistema JIT, si el ensamblaje en sitio progresa según lo programado, todos los componentes se entregarán justo a tiempo para las actividades de ensamblaje, sin causar desperdicio de inventario. Sin embargo, la entrega de componentes según lo programado causaría inventario adicional para

los contratistas generales cuando el progreso se desvía del plan original (generalmente más tarde). Por consiguiente, la compleja cadena de suministro, la mala comunicación y la coordinación insuficiente podrían causar desperdicios como inventario adicional, transporte, doble manipulación y posibles daños cuando ocurren variaciones. Además, el tiempo perdido entre actividades impacta significativamente el tiempo total de todo el proceso. Así pues, un inventario grande y la falta de cuidado suficiente son los principales problemas de la gestión de existencias en sitio (Xu et al., 2025).

De la misma forma, el modelo JIT identifica el problema del inventario como la actividad sin valor agregado más crítica. Por tanto, la capacidad de una fábrica para entregar puede verse afectada por múltiples tipos de incertidumbres, como pérdida de productividad, fallos de equipos, variaciones en los tiempos de preparación, defectos de trabajo inesperados y condiciones relacionadas con los materiales. Si una fábrica experimenta una interrupción inesperada de la producción, se requieren horas extra para cumplir con la fecha de entrega. En efecto, la peor situación es cuando el ensamblaje de los componentes producidos en tiempo extra se retrasa, lo que genera desperdicio para ambas partes (Xie et al., 2022).

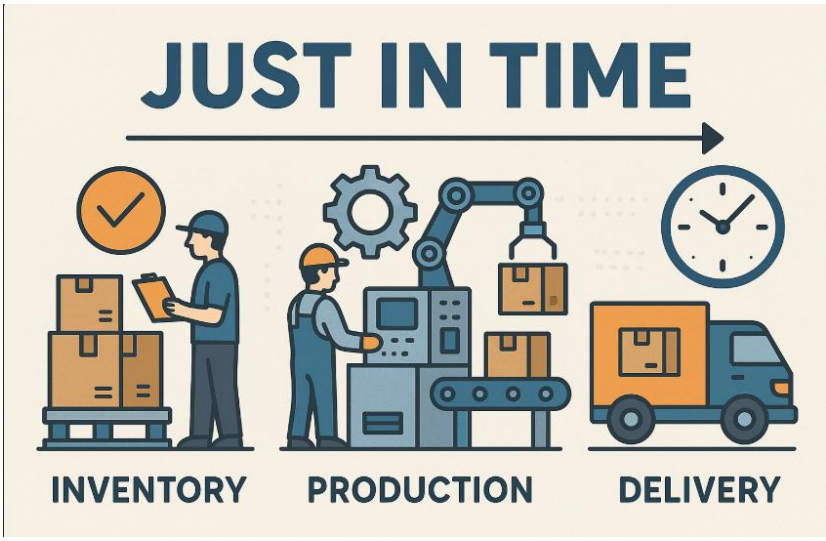


Figura 2 Just In Time

Modelo Random Forest

El desarrollo de modelos de predicción con datos longitudinales mediante enfoques estadísticos o de aprendizaje automático también es crucial. Uno de los métodos más avanzados y actuales para el desarrollo de modelos de predicción es el algoritmo de bosque aleatorio o *Random Forest* . En este sentido, este modelo se trata de un enfoque no paramétrico que puede adaptarse a diferentes tipos de respuestas, como resultados categóricos o cuantitativos y tiempos de supervivencia. Además, el *Random Forest* puede funcionar con predictores de diversas escalas o distribuciones y es adecuado para aplicaciones en entornos de alta dimensión, donde el número de predictores puede ser mayor que el número de observaciones. Por lo tanto, es muy adecuado para analizar datos complejos que a menudo son de alta dimensión (Hu & Szymczak, 2023).

De la misma manera, señaló que los métodos basados en árboles forman subgrupos de muestras basados en datos, lo que puede ser beneficioso para la

estratificación de datos. Así pues, mediante las denominadas medidas de importancia variable, el método también puede destacar la relevancia de cada predictor. En efecto, el modelo demuestra que su algoritmo de frecuencias proporciona una predicción comparable o mejor de la volatilidad del dato, en comparación con métodos convencionales como los mínimos cuadrados parciales y las máquinas de vectores de soporte (Milanovic et al., 2021).

Sin embargo, al igual que con otros métodos de *Machine Learning* (ML), el algoritmo de *Random Forest* asume que las observaciones se muestrean de forma independiente de una población. Así pues, al realizar análisis estadísticos sobre datos longitudinales sin considerar la dependencia entre observaciones podría conducir a inferencias sesgadas debido a errores estándar subestimados en modelos lineales. De la misma manera, la identificación de subgrupos espurios y la selección inexacta de variables en métodos basados en árboles. Además, los métodos de clasificación que se ajustan a la estructura de los datos y, por lo tanto, gestionan adecuadamente la correlación debida a mediciones repetidas, tienen un mejor rendimiento de predicción (Balyan et al., 2022).

Para realizar predicciones con *Random Forest*, una observación recorre todos los árboles de decisión del bosque. Así pues, la predicción final de la observación del *Random Forest* se realiza por votación mayoritaria o promediando, basándose en los resultados de todos los árboles de decisión del bosque. Dado que el algoritmo *Random Forest* utiliza muestras *Bootstrap* para el crecimiento de cada árbol de decisión, algunas observaciones se omiten en la construcción de un árbol determinado. Al tratar estas muestras fuera de la bolsa

como observaciones necesarias para la predicción, se puede proporcionar una estimación del error de predicción del bosque construido (Prashil et al., 2020).

En este contexto, para cada predictor se puede obtener la denominada medida de importancia de la variable, que mide su relevancia para la predicción. Por lo tanto, para conjuntos de datos de alta dimensión, como los datos ómicos, son posibles los procedimientos de selección de variables basados en la medida de importancia de la variable y la referencia correspondiente para una descripción y comparación de diversos procedimientos de selección de variables basados en la medida de importancia de la variable (Han et al., 2020).

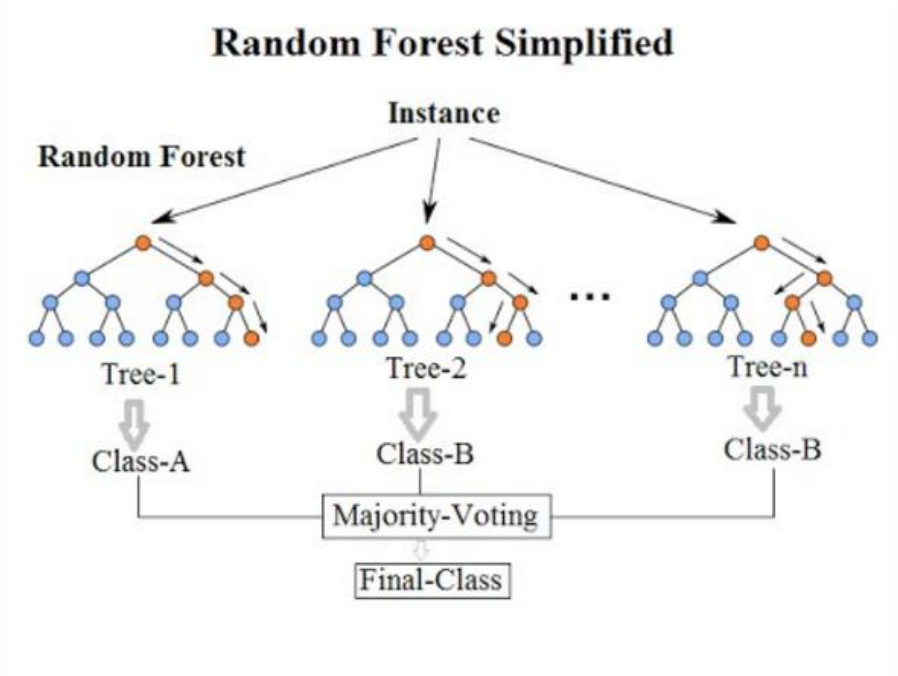


Figura 3 Modelo Random Forest

Planteamiento del modelo Random Forest

Variable Dependiente

$\mathbf{y_{it}}$ representa un indicador de eficiencia en la logística dado el periodo *t* para la ciudad *i* (ejemplo puede ser tiempo de entrega del medicamento, costo operativo, entre otros).

Variables Independientes

$\mathbf{X_{it}}$ representa a la matriz de características que incluye:

1. Predictores de la demanda de medicina veterinaria
    a) D_{it} Demanda histórica de medicinas.
    b) E_{it} Variables económicas (PIB, inflación, etc.).
    c) C_{it} Condiciones climáticas (lluvias fuertes, temperatura).

2. Variables de rutas
    a) R_{it} representa al tráfico, distancia.
    b) L_{it} representa a la localización de centros de distribución.

3. Gestión del inventario
    a) S_{it} niveles de actual de inventario.
    b) T_{it} tiempo de reposición de medicinas.

Por consiguiente, la fórmula del modelo *Random Forest* queda de la siguiente manera:

$y_{it} = f(X_{it}) + €_{it}$, donde *f* es un ensemble de *B* árboles (T_1, …T_n)

En donde:

€ representa el error aleatorio (valor real observado – valor predicho)

** Cada árbol T_b se entrena con un *bootstrapping* y *features* ($m \leq p$).

Submodelo 1 (Predicción de demanda)

$$\hat{D}_{it} = \frac{1}{B} \sum_{b=1}^{B} T_b(X_{it})$$

En donde X_{it} está en función de D_{it} -1, E_{it} C_{it}

Además, la métrica del error es:

$$RMSE = \sqrt{\left[\frac{1}{N} \sum (D_{it} - \hat{D}_{it})^2 \right]}$$

Submodelo 2 (Optimización de rutas)

$$\min_{R_{it}} \sum_{i=1}^{N} \left(\text{Costo}(R_{it}) \times \mathbb{I}(T_b(X_{it}^{\text{ruta}}) \right)$$

En donde:

Input: es $\hat{D}_{it}$ restricciones geográficas de las rutas

Output: Ruta optimizada (menor tiempo, menor distancia)

Submodelo 3 (Gestión de inventarios)

$$S_{it}^{*} = \max\left(\widehat{D}_{it}, S_{it-1} - \alpha \cdot \text{Lead Time}\right)$$

α representa el factor de seguridad calibrado con *Random Forest*.

Métricas de evaluación del modelo

- Error de las Muestras Excluidas (OOB) error:

$$\text{OOB} = \frac{1}{N}\sum_{i=1}^{N}(y_{it} - \hat{y}_{it}^{\text{OOB}})^2$$

- Importancia de variables

$$I_j = \frac{1}{B}\sum_{b=1}^{B}(\text{Reducción de impureza por } X_j \text{ en } T_b)$$

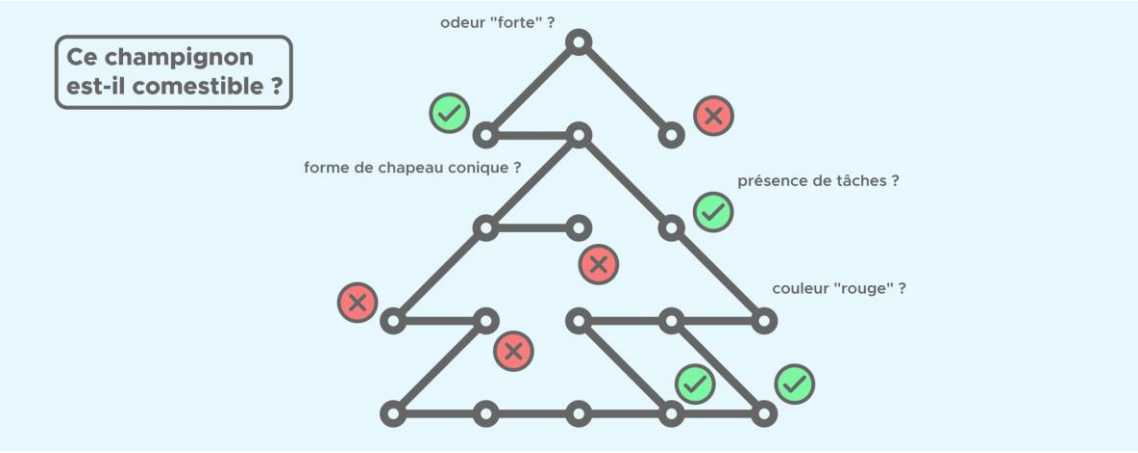


Figura 4 Modelo Bosque Aleatorio

Modelo Árbol de decisión

Los árboles de decisión (DT) son modelos predictivos en aprendizaje supervisado, conocidos no solo por su indiscutible utilidad en una amplia gama de aplicaciones, sino también por su interpretabilidad y robustez. La investigación sobre este tema se mantiene vigente tras casi 60 años desde su inicio, y en la última década, varios investigadores han abordado temas clave en el campo. Si bien se han publicado numerosos estudios relevantes, existe una brecha, ya que ninguno abarca la última década del campo en su conjunto (Costa & Pedreira, 2023).

Por otro lado, en un árbol de decisión, el proceso de división de nodos consiste en seleccionar una variable de división y determinar la regla de división. Así pues, el principio rector de la división de nodos es minimizar la impureza de los valores de respuesta en cada nodo, lo cual a menudo se mide mediante el índice de Gini si la variable de respuesta es categórica o mediante la varianza si es cuantitativa. Asimismo, el crecimiento de cada árbol de decisión finaliza si los nodos a dividir ya son puros (todas las muestras dentro del nodo provienen de la misma clase o tienen el mismo valor de respuesta) o si se cumplen otras reglas de parada predeterminadas. En efecto, los nodos en la capa final de un árbol se denominan hojas y se utilizan para la predicción de nuevas observaciones (Hu & Szymczak, 2023).

Según Lee et al. (2022) en el análisis predictivo se utiliza a menudo el árbol de decisión, ya que se considera una herramienta predictiva intuitiva que permite a los usuarios interpretar fácilmente los datos. El árbol de decisión es un

algoritmo de aprendizaje automático supervisado que se centra en deducir la clase o el valor de las variables objetivo según el orden de aprendizaje automático entrenado con los datos de entrenamiento. De la misma manera, este enfoque es fácil de usar e interpretar con matemáticas sencillas, sin necesidad de conocimientos estadísticos ni fórmulas complejas. Además, este modelo tiene un enfoque intuitivo, ya que los datos necesarios se preparan fácilmente sin realizar cálculos complejos. Además, una vez definidas las variables, se requiere menos intervención en la optimización de los datos.

Acorde con Sarker (2021) el árbol de decisión es un tipo de herramienta de clasificación supervisada fácil de interpretar. En este caso, el árbol de decisión se señala como una herramienta consolidada que puede utilizarse sin conocimientos estadísticos y no requiere fórmulas complejas. Del mismo modo, el árbol de decisión es intuitivo y sus resultados se pueden interpretar fácilmente en comparación con otras herramientas de aprendizaje automático supervisado que requieren conocimientos estadísticos, como Naive Bayes (NB), regresión logística (LR), máquina de vectores de soporte (SVM) y *Random Forest* (RF).

Acorde con Karalis (2020) señaló que en muchos casos, el significado de la información se relaciona erróneamente con el sentido de los datos o con la noción de conocimiento. Existe una secuencia crucial de pasos antes de que la información se convierta en conocimiento, y el valor de los datos depende de su existencia para producir conocimiento. El método más común para producir conocimiento a través de datos se basa en el análisis de datos y, principalmente, en la interpretación de los resultados. Esta es la forma en que los humanos toman

decisiones, basándose en su conocimiento existente, y por lo tanto, utilizando el método de árbol de decisión, intentan simular varias herramientas de decisión artificiales.

Por consiguiente, los árboles de decisión son una de estas herramientas de decisión artificiales. Según Kumar y Kumar (2020) el principal objetivo de éstos consiste en el análisis automático o semiautomático de big data, así como en la creación de nuevos patrones. En efecto, los árboles de decisión se pueden aplicar en diversos campos científicos, como la bioinformática. De esta manera, las aplicaciones más utilizadas de los árboles de decisión son la minería de datos y la clasificación de datos. En definitiva, diversos investigadores de diversos campos y trayectorias han considerado la extensión de un árbol de decisión a partir de los datos disponibles, como estudios de máquinas, reconocimiento de patrones y estadística.

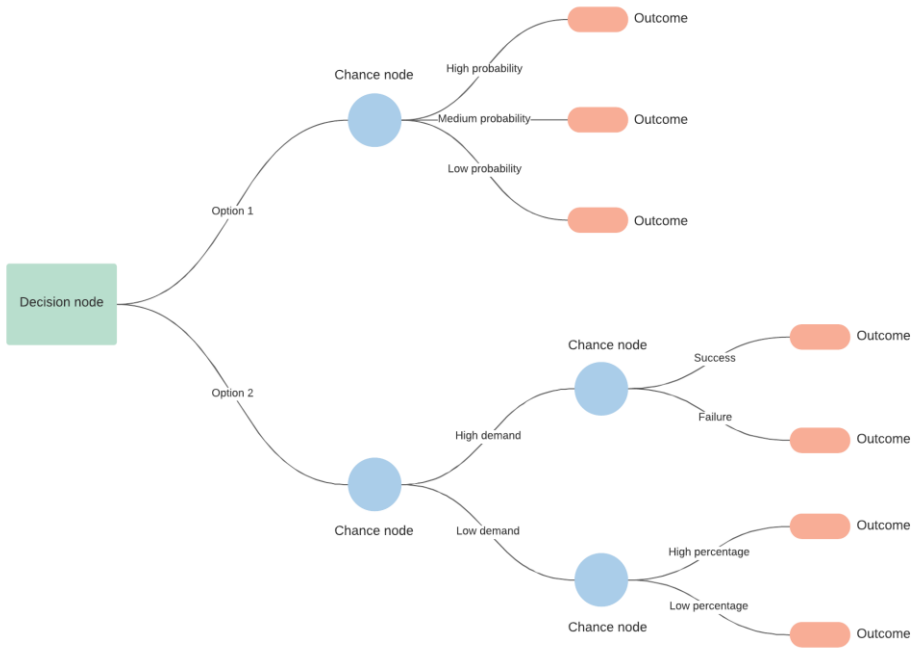


Figura 5 Modelo Árbol de decisión

Planteamiento del modelo Árbol de decisión

Variable Dependiente

$\mathbf{y_{it}}$ representa un indicador de eficiencia en la logística dado el periodo *t* para la ciudad *i* (ejemplo puede ser tiempo de entrega del medicamento, costo operativo, entre otros).

Variables Independientes

$\mathbf{X_{it}}$ representa a la matriz de características que incluye:

4. Predictores de la demanda de medicina veterinaria
    d) D_{it} Demanda histórica de medicinas.
    e) E_{it} Variables económicas (PIB, inflación, etc.).
    f) C_{it} Condiciones climáticas (lluvias fuertes, temperatura).

5. Variables de rutas
    c) R_{it} representa al tráfico, distancia.
    d) L_{it} representa a la localización de centros de distribución.

6. Gestión del inventario
    c) S_{it} niveles de actual de inventario.
    d) T_{it} tiempo de reposición de medicinas.

Por consiguiente, la fórmula del modelo Árbol de decisión queda de la siguiente manera:

$$y_{it} = f(X_{it}) + \epsilon_{it},$$ donde ***f*** es un único árbol de decisión.

Entonces:

Para la división de nodos se eligen variables X_j que maximizan la disminución de impureza (se utiliza Gini o Entropía para clasificar y MSE para la regresión). Asimismo, para el criterio de parada se utiliza la profundidad máxima , el tamaño mínimo del nodo , etc.

Árbol Único (Predicción de demanda)

$$\widehat{D}_{it} = T(X_{it}^{\text{demanda}})$$

Entonces, para la división optimizada de cada nodo se elige X_j y umbral *S* que minimizan:

$$\textbf{MSE} = \sum_{i \in \text{Izquierda}} (D_{it} - \bar{D}_{\text{Izquierda}})^2 + \sum_{i \in \text{Derecha}} (D_{it} - \bar{D}_{\text{Derecha}})^2$$

Métricas de evaluación del modelo

- Error de predicción:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_{it} - \hat{y}_{it})^2$$

- Importancia de variables

I_j representa el número de veces que X_j se utilizaría para dividir $x_{número}$ de reducción de impureza.

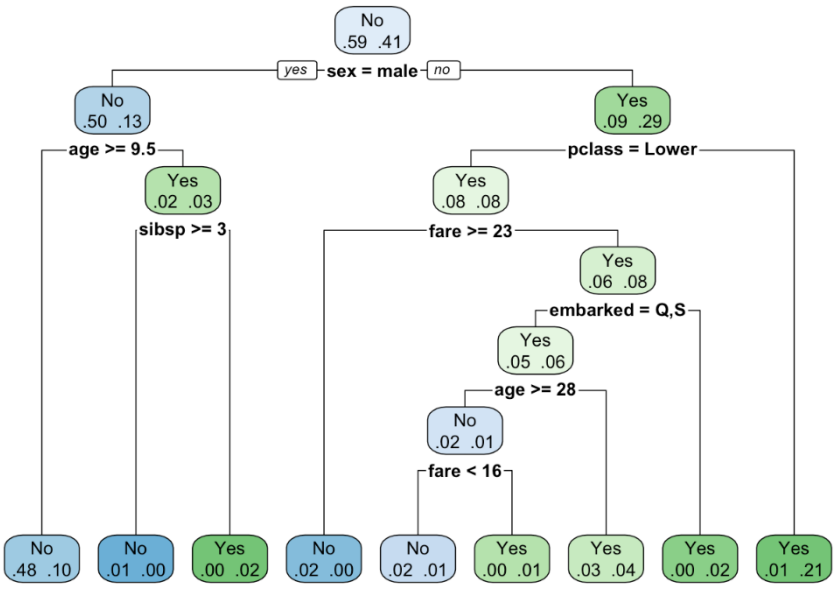


Figura 6 Planteamiento Árbol de decisión en R

Marco Referencial

Un trabajo de gran aporte académico y que se utilizó de referencia fue el de Raba (2021) en el cual se trata sobre la implementación de una solución asistida por computadora para abordar el problema de la demanda laboral enfocada en la demanda ocupacional, lo cual guardaría relación con esta investigación. Usando la data generada por las empresas ganaderas y realizó un proceso de analítica de datos en el cual se creó un algoritmo para calcular la demanda y los días restantes de estimación de stock usando una serie temporal ordenada por fecha ascendente con pesos estimados que incluyen la última lectura disponible realizando un muestreo por un período de tres días. En definitiva, con machine Learning se realizó un pronóstico de la demanda futura. En conclusión, el modelo realizado permitió evolucionar hacia una configuración más optimizada, debido a la mejora y actualización de cada componente de la cadena de suministro.

Otro trabajo referencial de suma importancia fue el de Gonzales y Pérez (2025), en el cual tuvo como objetivo principal el de conocer la relación entre la gestión logística y el control de inventarios en una veterinaria en Perú.Se utilizó un modelo correlacional con un diseño no experimental de corte transversal y un instrumento de recolección de datos compuesto por un cuestionario de medición de inventarios y de medición logística. Asimismo, la muestra estuvo compuesta por 25 colaboradores de la empresa veterinaria. En definitiva, se determinó que existe una relación directa y significativa entre la gestión logística y el control de

inventarios, por lo cual, los empleados cumplen con los estándares mínimos esperados por la empresa.

El trabajo de Figueroa (2023) fue de gran aporte a este estudio debido a que se realizó una investigación sobre analítica de datos para optimizar procesos de comercialización de dos empresas colombianas de venta de carne. Para este efecto, se utilizó una metodología de investigación no experimental de enfoque mixto. Asimismo, se realizaron técnicas de recolección de datos como las entrevistas semiestructuradas (realizadas a los dos gerentes generales de estas dos empresas estudiadas) y se elaboraron matriz de uso y no uso de datos. En definitiva, los resultados del estudio muestran que las empresas analizadas muestran cierto nivel de analítica de datos, pero deben implementar modelos prescriptivos y predictivos para mejorar los procesos de comercialización.

Otro trabajo referencial importante fue el de Herrera y Romero (2024) en el cual abordaron la implementación de un modelo de logística con la finalidad de utilizar herramientas de pronósticos de series temporales en una empresa de comercialización de medicamentos veterinarios. Para este objetivo se propuso un modelo que sigue un enfoque integral que mezcla técnicas para predecir la demanda y el modelo de la Cantidad Económica de Pedido con alertas Kanban. En conclusión, este estudio presenta resultados tales como que la implementación del modelo mejora el cumplimiento de las entregas de manera sustancial alcanzando una efectividad del 90%. De la misma manera, se mejoró la gestión de inventarios dado que se redujo la obsolescencia de las medicinas.

Finalmente, otra investigación de mucha utilidad fue le de Anagumbla (2024) en la cual se realizó un sistema de gestión de inventarios en una clínica veterinaria en Ecuador. Con la ayuda de una metodología inductiva de tipo no experimental de alcance descriptivo. Además, se utilizaron técnicas de recolección de datos tales como la observación directa, revisión bibliográfica y revisión documental. Luego, con una muestra de 381 productos se los clasificó con el método ABC, de los cuales 157 productos representan el 80% del total del inventario. En definitiva, los resultados del estudio mostraron que la clínica veterinaria no cuenta con un control físico de inventarios, por consiguiente, el control de entrada/salida de medicinas es deficiente, por lo que se recomienda implementar el sistema propuesto y realizar un control físico de inventarios.

Marco Conceptual

La Demanda

Se denomina demanda al acto procesal de parte, mediante el cual se ejerce el derecho constitucional de acción, a través de una pretensión concreta de parte. La demanda es el acto procesal de la parte actora que inicia el proceso y que constituye una manifestación de voluntad formalmente expresada por escrito y dirigida a un órgano jurisdiccional con el fin de solicitar que se inicie el proceso, se desarrolle y culmine con una decisión que acoja su pretensión procesal. (Dr. Sergio Artavia B., *Aspectos generales sobre la demanda*, 2018., p. 1)

Demanda de mano de obra

La demanda de mano de obra hace referencia a los factores que determinan la cantidad de trabajadores que las empresas, instituciones u organizaciones desean contratar, así como la forma en que se utiliza esa fuerza laboral, especialmente en lo que concierne al número de horas trabajadas. Este concepto no solo contempla a los trabajadores como un grupo homogéneo, sino que también distingue entre ellos a partir de características demográficas específicas, así como por el nivel de habilidades y competencias que poseen.

Dentro de los principales desafíos analizados en la literatura económica se encuentra el comprender de qué manera los empleadores responden a las variaciones en los costos asociados al empleo, tanto en términos generales como en lo que respecta a determinados grupos de trabajadores. En este contexto, resulta fundamental identificar cómo la reducción en los costos de contratar a un perfil específico puede motivar a los empleadores a sustituir a unos trabajadores por otros, y de qué manera los costos relativos —ya sea por trabajador o por hora laboral— influyen en las decisiones de sustitución entre personas y horas trabajadas.

La teoría económica que sustenta estas dinámicas proporciona un marco de análisis para comprender dichos fenómenos y permite evaluar el grado de conocimiento actual en torno a estas cuestiones. La evidencia empírica disponible se utiliza, además, para examinar el impacto que tienen distintas

políticas laborales en el comportamiento de la demanda de trabajo. Entre estas políticas destacan el establecimiento de salarios mínimos, las regulaciones que restringen o sancionan las horas extraordinarias, y la aplicación de impuestos sobre la nómina.

Un aspecto clave dentro de esta temática es que los ajustes en la demanda de mano de obra no se producen de manera inmediata, sino que requieren tiempo. La velocidad y trayectoria de dicho ajuste dependen en gran medida de las restricciones derivadas de los costos de contratación y despido, así como de la influencia de las normativas gubernamentales. En consecuencia, el análisis de la evidencia empírica resume cómo estos costos, junto con las regulaciones estatales, alteran no solo el nivel de empleo disponible, sino también la forma en que este evoluciona a lo largo del tiempo Hamermesh, D. S. (2001).

Random Forest

Este modelo es un conjunto de árboles de decisión, donde cada árbol se construye a partir de una versión *Bootstrap* del conjunto de datos de entrenamiento. Cada árbol se desarrolla mediante el principio de partición repetitiva, donde, a partir del nodo raíz, se aplica el mismo procedimiento de división de nodos repetidamente hasta que se cumplen ciertas reglas de parada. De este modo, su poder de predicción proviene de la agregación de muchos aprendices más débiles (árboles de decisión). El rendimiento es especialmente

bueno si las correlaciones entre los árboles del bosque son bajas. Por tanto, se pueden encontrar descripciones y análisis más detallados sobre este modelo (Hu & Szymczak, 2023).

Árbol de decisión

Según Costa & Pedreira (2023) los árboles de decisión son modelos predictivos en el aprendizaje supervisado, conocidos no solo por su incuestionable utilidad en una amplia gama de aplicaciones sino también por su interpretabilidad y robustez. Por otro lado, Sarker (2021) señaló que los árboles de decisiones son un tipo de herramienta de clasificación supervisada fácil de interpretar dado que es una herramienta consolidada que puede utilizarse sin conocimientos estadísticos y no requiere fórmulas complejas al ser un modelo intuitivo y sus resultados se pueden interpretar fácilmente en comparación con otras herramientas de aprendizaje automático supervisado que requieren conocimientos estadísticos.

MARCO LEGAL

Para el marco legal de la investigación, partimos del Plan Nacional de Protección de Datos Personales (PNPDP) 2025-2029, plan de acción jurídico emitido por Superintendencia de Protección de Datos Personales (SPDP). El PNPDP persigue consolidar un ecosistema ético y competitivo en el tratamiento de la información y dispones de cuatro ejes articulados, los cuales son: gobernanza digital, universalización del derecho, cultura ciudadana e innovación

segura. Como el modelo de esta investigación es basado en ciencia de datos este gestionara el registro logístico, lo que hará que el diseño concuerde con este plan para asegurar que cada etapa del ciclo de vida de los datos (recopilación, entrenamiento de datos y predicción de la demanda) debe ser regida por los estándares de seguridad y seguimiento requeridos a escala nacional.

La SPDP, establecida con base en el artículo 213 de la constitución, cumple funciones de supervisión, auditoria y sanción, y requiere evidencias de responsabilidad proactiva y demostrada de todos los encargados del tratamiento. Para la demanda de las pymes y grandes empresas de Guayaquil, desempeñara esta función documentando evaluaciones de impacto, matrices de riesgos y reclutamiento de datos reales, de manera que la autoridad pueda confirmar no solo el acatamiento de las regulaciones sino también la justicia de las proyecciones que respaldan las decisiones operativas.

En la parte jurídica interna, a Ley Orgánica de Protección de Datos Personales (LOPDP) extiende su alcance con todo el manejo de datos contenidos en algún medio como lo dice el artículo 2 y establece, en su artículo 10, implementar acciones técnicas, físicas y organizativas que van dirigidas a reducir riesgos. Para la actual situación, esto abarca la anonimización de identificadores directos de órdenes y rutas antes de su comparación mediante árbol de decisión y bosques aleatorios, restringir variables cuando únicamente son requeridas y se mantiene un registro de actividades al día que evidencie la legalidad y el propósito del procesamiento.

En la parte constitucional, el artículo 66 numeral 19 que menciona y reconoce el derecho fundamental a la protección de datos y limita cualquier tratamiento a la autorización del titular o a mandato legal. Por otro lado, el artículo 385 impulsa la investigación científica y la innovación tecnológica, permitiendo que proyectos, como el que actualmente maneja la institución de estadística y censo, donde recluta datos reales de los ciudadanos ecuatorianos para la alimentación de una base de datos y el estudio de varias variables, que en todo momento se respete la dignidad y la privacidad de las personas. De esa manera, lo propuesta de esta investigación conjuga el mandato de fomentar ciencia aplicada con la obligación paralela de salvaguardar la información personal.

A modo de conclusión, el propio PNPDP requiere protocolos de seguridad específicos, supervisión de algoritmos y alianzas con estándares internacionales para evitar decisiones discriminatorias y brechas de lealtad. El instituto nacional de estadística y censo (INEC) debe adoptar procesos donde permita a los ciudadanos tener la seguridad de que los datos recolectados tendrán únicamente la finalidad del estudio que realizara una mejor demanda de cualquier ámbito dentro del país, disminuyendo la demanda de desempleo o subempleo. En consecuencia, el proyecto se va a insertar en un marco legal que paralelamente protege los derechos de protección de datos e incentiva la innovación digital de manera responsable.

Metodología

Introducción

En el capítulo se desarrollará los métodos que consistirán en como diseñar un modelo para que se pueda estimar la demanda laboral que existe en las pymes y grandes empresas de Guayaquil, con el fin de dar una solución a un dilema tan codiciado por la población ecuatoriana enfocado en la ciudad de Guayaquil, donde con la ayuda de la ciencia de datos se determinara diferentes incógnitas planteadas por la sociedad de por qué existe tanto desempleo y a su vez porque hay puestos tan cotizados y como estos son utilizados y demás. Se realizará una serie de diseños mediante el software R studio acompañado con una base de datos pública (INEC) donde arrojará resultados más asertivos.

R studio

Software diseñado para trabajar con un lenguaje de programación conocido como R, especializada para la estadística y análisis de datos para facilitar la escritura y organización de la base de datos ya mencionada anteriormente, se visualizará de una manera estructurada y limpia el diseño de gráficos, cálculos estadísticos donde se pueden exportar diferentes formatos como EXCEL, CSV, PDF, WORD, entre otras.

Tipo de investigación

El método utilizado para el desarrollo de esta investigación donde se analizará la demanda laboral de las pymes y grandes empresas de Guayaquil se

generará mediante un árbol de decisión y bosques aleatorios también conocido como random forest

Técnicas e instrumentos de recolección de datos

Mediante una investigación exhaustiva de datos apropiados para trazar un diseño limpio de datos y calcular la demanda laboral en Guayaquil se encuentra la entidad rectora de la estadística titulada como el Instituto Nacional de Estadística y Censo más conocido como “INEC” la cual es la encargada de recopilar la mayor información de datos sean en aspectos sociales, económicos de la población, se proporciona datos de manera gratuita donde puedes ser importados mediante Excel, CSV, PDF, dependiendo de para que sea desarrollado, de igual manera esta se comparte con diferentes ramas de investigación, donde se busca hallar concretamente la demanda laboral. Existe un apartado dentro del inec el cual es denominado como “ENEMDU” esta se refiere a la Encuesta Nacional de Empleo, Desempleo y Subempleo, la necesaria para el análisis de la demanda laboral de Guayaquil incluso el ecuador, fundamental para el diseño de nuevas políticas claves enfocada en el desarrollo económico y social.

Métodos

Los métodos utilizados para el análisis de la demanda laboral que existe en las pymes y grandes empresas de Guayaquil fue utilizado mediante el software R studio utilizando el la ciencia y análisis de datos mediante diferentes técnicas y usos de los datos, como lo es el árbol de decisión y random forest (bosques aleatorios) la cual lleva consigo varias librerías las cuales serán de

mucha ayuda para lo que necesitamos. Los Arboles de decisión dentro de la ciencia de datos, son usados para demostrar de una manera más sencilla y visual el análisis de los resultados arrojados.

Árbol de decisión

Este es un modelo de predicción donde es más comúnmente utilizado en la ciencia de datos, Éste se basa un modelo predictivo donde se organiza una información en forma de un árbol jerárquico, está compuesto por una decisión ramas y hojas donde el nudo de la raíz representa una variable inicial, la cual se comienza a dividir la información, luego tenemos los nódulos internos donde se notan los puntos de decisión y se plantea de condiciones sobre variables explicativas. Las ramas representan posibles respuestas a las condiciones y las hojas corresponden a los resultados o predicciones finales. Es un método, el cual permite clasificar o predecir resultados de una manera sencilla, visual e interpretable Éstas son valorados por su facilidad, interpretación y su capacidad para manejar tanto variables categóricas como numéricas.

Radom Forest

lo busques aleatorios, al igual que el árbol de decisión son momentos de aprendizajes supervisados que pertenece a la sociedad de datos, Machine Learning este está compuesto por múltiples árboles de decisión independientes, las cuales van a ser entrenados, de manera aleatoria a partir de subconjuntos de

datos y variables para producir una clasificación o predicción más robusta, selección de manera aleatoria a partir de la muestra de los datos originales, sabiendo que cargo genera una predicción. El bosque aleatorio toma la decisión final mediante una votación mayoritaria este tiene mucho más precisión y estabilidad. En comparación con un único árbol también manejan variables categóricas como numéricas, permitiendo si calcular la importancia de cada variable, mostrando cuáles influyen más en un resultado Final.

Librerias Implementadas

library(caret) #entrenamiento de datos

(Classification And REgression Training) esta librería es el encargada de entrenar los datos y evaluar modelos predictivos en este caso sobre la demanda ocupacional, para la precisión de las predicciones, este paquete integra un conjunto amplio de herramientas para la construcción de modelos en el Machare, facilitando el proceso de selección de variables, entrenamiento y validación de modelos

library(rpart) # para poder hacer el árbol

(Recursive Partitioning and Regression Trees) Éste es utilizada para la construcción de árboles de decisión, ya sea este para clasificación o regresión, va a generar clasificar a los trabajadores en grupos, ya sea alta media o baja en la demanda laboral, también con variables significativas como experiencia, edad,

sector, etc. este método es uno de los más significantes y más demostrativos para el uso de esta herramienta.

library(rpart.plot)

este es un paquete, el cual se complementa con el R Part, ya que este no permite visualizar, gráficamente a los árboles de decisión, y a su vez facilitará la demostración de los resultados del árbol, acotando los factores más influyentes en la demanda ocupacional.

library(randomForest)

la implementación del algoritmo de bosques, aleatorios o Random Forest, el cual combina múltiples árboles de decisión para una mejora y precisión de todas las predicciones arrojadas así con la demanda laboral, comparando la demanda, ocupación real, con la predicha y a su vez, encontrando variables más influyentes.

library(ggplot2)

Uno de los paquetes más utilizados en el software R para visualización de datos, al igual este permite la creación de gráficos de barras, dispersión, líneas, y muchas otras más con una alta personalización y comprensión, donde se verá reflejada la distribución de la demanda ocupacional real versus la pre, dicha, facilitando así el análisis visual de los resultados

library(readxl)

sencillamente el paquete utilizado para importar archivos de Excel XLS y XLSX, directamente hacia el software. R, donde se hallara cargada la base de datos laboral para el análisis de aquella set.seed (2018)

El comando establece un número fijo denominado como semilla, el cual genera números aleatorios en el software R utilizada para controlar la aleatoria edad, dentro de cualquier proceso que usen números aleatorios. Selecciona aleatoria de filas para entrenamientos y pruebas y la elección aleatoria de variables en un bosque aleatorio.

na.omit :

la función que elimina filas con valores en blanco, dentro de la base de datos, ayuda a la limpieza de las observaciones que se encuentren vacías o con valores perdidos, asegurando que los modelos de predicción no fallen por datos incompletos

Script del modelo desarrollado mediante el framework Rstudio

1**) Empezamos con la carga de librerías y datos dentro del software R**

```
library(openxlsx)
DO <-
read.xlsx("C:/Users/Usuario/Desktop/tesis/Data_DemandaLabroal.xlsx")
DO <- na.omit(DO)
```

openxlsx: permite leer archivos Excel sin depender de Java/Perl.

read.xlsx: carga tu base DO desde Excel.

na.omit: elimina filas con valores faltantes en cualquier columna (deja el dataset “limpio” para modelar).

2) Usamos las siguientes librerías para ML y árboles de decisión

```
library(caret)
library(rpart)
library(rpart.plot)

table(DO$Demanda_Ocupacional)
```

caret: utilidades de preprocesamiento, partición estratificada, métricas, etc.

rpart / rpart.plot: entrenamiento y visualización del árbol de decisión.

table(...): vistazo rápido a la distribución de la variable objetivo.

3)Procedemos con las Etiquetas de las variables objetivo

```
DO$Demanda_Ocupacional <- factor(DO$Demanda_Ocupacional, levels
= c(0,1,2,3,4,5,6,7,8,9),
                                 labels = c("Gobierno","Privado", "Externo", "Obrero",
                                            "patrono", "Propio", "Domestico",
```

"No remunerado", "Ayudante", "Empleado"))

Esta Convierte los códigos (0–9) a etiquetas legibles para interpretar resultados.

Y asegura que el modelo trate la variable como factor (clasificación).

4) Plantamos semilla y partición estratificada de los datos

set.seed(123)

dataentrenamiento <- createDataPartition(DO$Demanda_Ocupacional, p=0.70, list = FALSE)

set.seed(123): donde garantiza reproducibilidad (mismas particiones y resultados al re-ejecutar).

createDataPartition: creamos índices del 70% para entrenamiento manteniendo la proporción de clases (estratificación). El 30% restante queda como test.

5) Entrenamiento de un árbol de decisión con rpart

arbol <- rpart(Demanda_Ocupacional~., data = DO[dataentrenamiento,], method = "class")

arbol

Usamos Fórmula Demanda_Ocupacional ~ . donde usa todas las demás columnas como predictores.

method="class": tarea de clasificación (no regresión).

Al imprimir arbol vemos tamaño/nodos, variables de partición y criterios.

Gráfico del árbol

rpart.plot(arbol, type = 1, digits = -1, extra = 0, cex = 0.7, nn= TRUE,

fallen.leaves = TRUE)

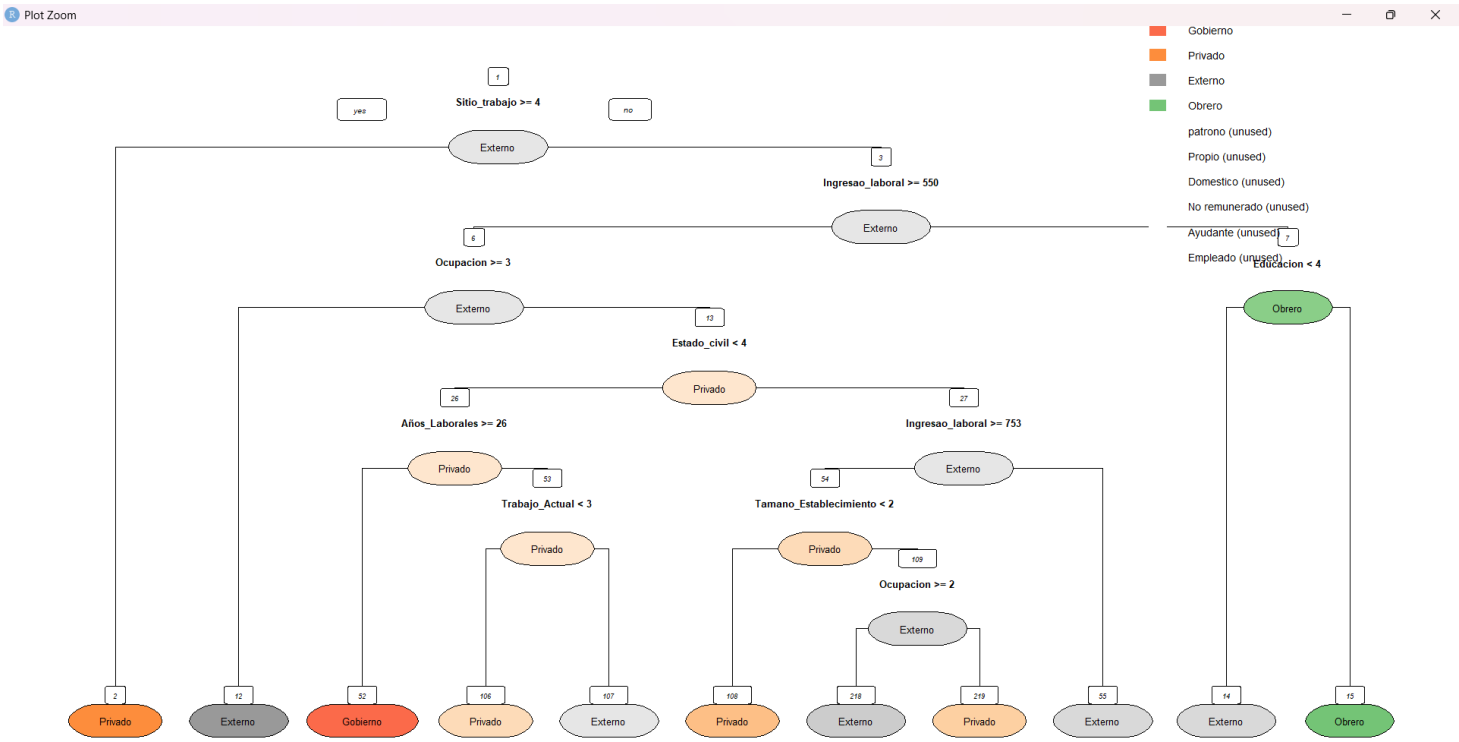


Figura 7 ScreenShot Árbol de decisión R studio

type = 1: **etiquetas en las split labels (condiciones en ramas).**

digits = -1: **controla los decimales mostrados (aquí usa opción por defecto del objeto).**

extra = 0: información mínima por nodo (solo clase).

nn = TRUE: numera nodos.

fallen.leaves = TRUE: **hojas "caídas" en la base del gráfico (para una mejor lectura).**

6)Damos paso al entrenamiento de un Bosque Aleatorio

```
library(randomForest)

RandomTreeModel <-  randomForest(x=DO[dataentrenamiento, 1:10],
                                 y=DO[dataentrenamiento,11],
                                 ntree = 3000, keep.forest = TRUE)
RandomTreeModel
```

randomForest: **ensamble de muchos árboles para mejorar generalización y reducir varianza.**

Luego aquí entrenamos con:

x = columnas 1:10 (predictores),

y = columna 11 (clase objetivo),

ntree = 3000: número de árboles (alto ⇒ más estable pero más costo computacional).

RandomTreeModel impreso muestra error OOB, importancia de variables (si la pides), y una confusion matrix OOB interna.

7) Predicción en el conjunto de prueba

```
Prediccion <-  predict(RandomTreeModel,DO[-dataentrenamiento,])
Prediccion
```

Se generarán etiquetas predichas para el 30% de test (filas que no están en dataentrenamiento).

8) Continuamos con la matriz de confusión en test

```
Matriz <- table(DO[-dataentrenamiento,"Demanda_Ocupacional"],
Prediccion,
                dnn = c("Actual", "Predicho"))
```

Matriz

Donde se compara clase real vs clase predicha en test.

dnn pone nombres a las dimensiones (“Actual”, “Predicho”).

Con esta tabla podemos calcular exactitud, recall por clase, etc. (ver “Mejoras”).

9) Damos paso a la predicción para todo el dataset y guardado en columna

DO[,"Prediccion"] <- predict(RandomTreeModel,DO)

Se añade una columna Prediccion con etiquetas para todas las filas (train + test).

Útil para luego graficar o comparar distribuciones globales.

10) Pasamos con los Gráficos de frecuencias (real vs predicción)

library(ggplot2)

mk_freq <- function(x, etiqueta){

  df <- as.data.frame(table(x), stringsAsFactors = FALSE)

```
  names(df) <- c("Clase","Freq")
  df$Tipo <- etiqueta
  df[order(-df$Freq), ]
}

freq_real <- mk_freq(DO$Demanda_Ocupacional, "Real")
freq_pred <- mk_freq(DO$Prediccion,          "Predicción")
```

mk_freq: **esta función auxiliar que**:

calcula tabla de frecuencias,

la lleva a data.frame con columnas Clase, Freq,

agrega el campo Tipo,

ordena de mayor a menor frecuencia.

freq_real: frecuencias de la variable objetivo real.

freq_pred: **las frecuencias de la columna de predicción (sobre todo el dataset).**

Gráfico 1: Real

```
ggplot(freq_real, aes(x = reorder(Clase, -Freq), y = Freq)) +
  geom_col() +
  geom_text(aes(label = Freq), vjust = -0.3) +
  labs(title = "Frecuencias - Demanda Ocupacional (Real)",
       x = "Clase", y = "Frecuencia") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Se ordenan las Barras de mayor a menor.

geom_text **muestra el conteo encima de cada barra.**

Eje X rotado a 45° para legibilidad.

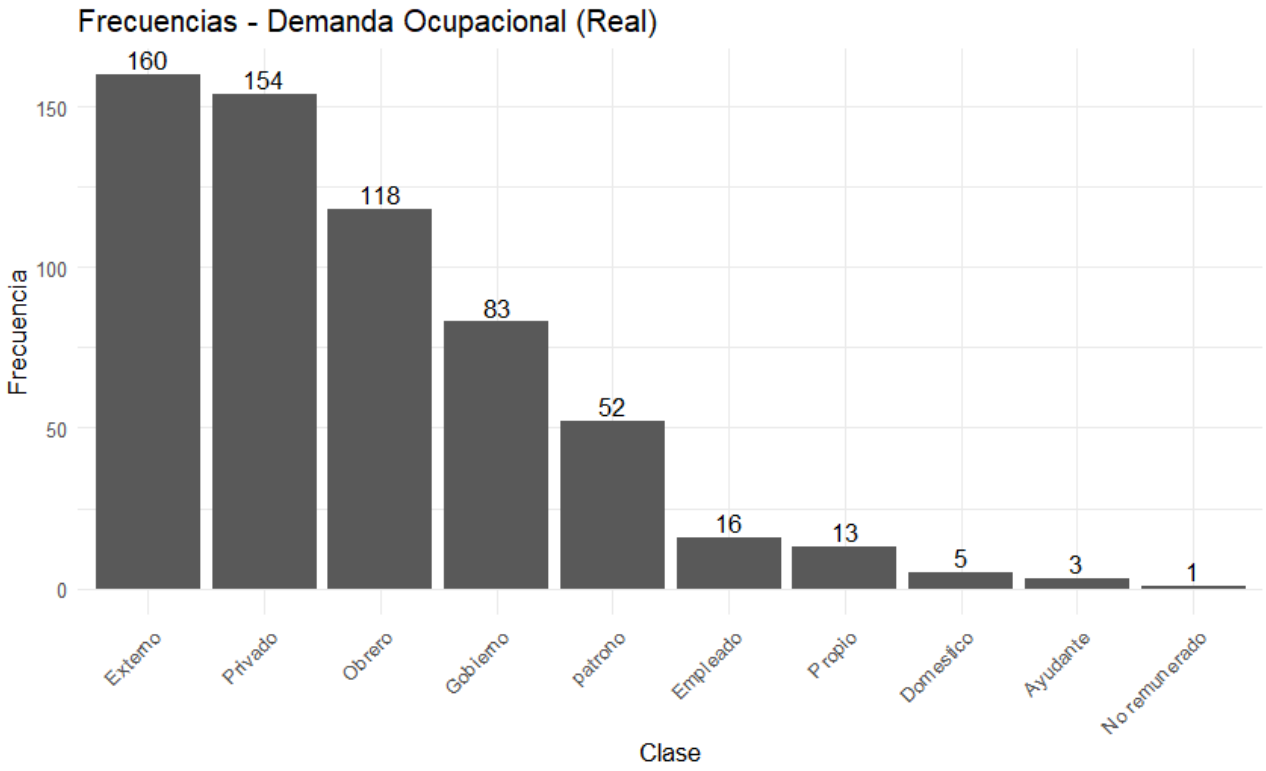


Figura 8 Gráfico de barras Demanda real

```
ggplot(freq_pred, aes(x = reorder(Clase, -Freq), y = Freq)) +
  geom_col() +
  geom_text(aes(label = Freq), vjust = -0.3) +
  labs(title = "Frecuencias - Demanda Ocupacional (Predicción)",
       x = "Clase", y = "Frecuencia") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Mismo formato, pero para las frecuencias predichas.

La comparación visual entre ambos gráficos te permite ver si el modelo sobre-representa o sub-representa ciertas clases respecto a la realidad.

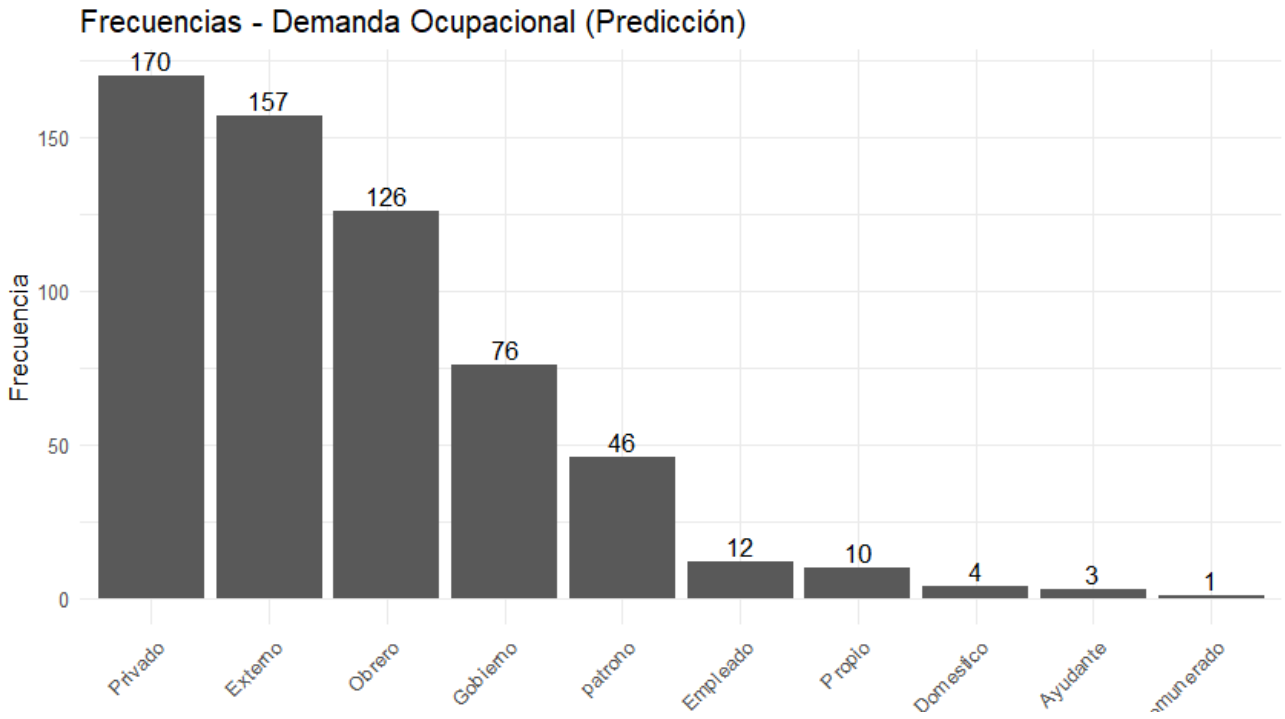


Figura 9 Gráfico de Barras Demanda Predicha

Visualización dentro del software

Esto dentro del R studio se visualizaría de esta manera

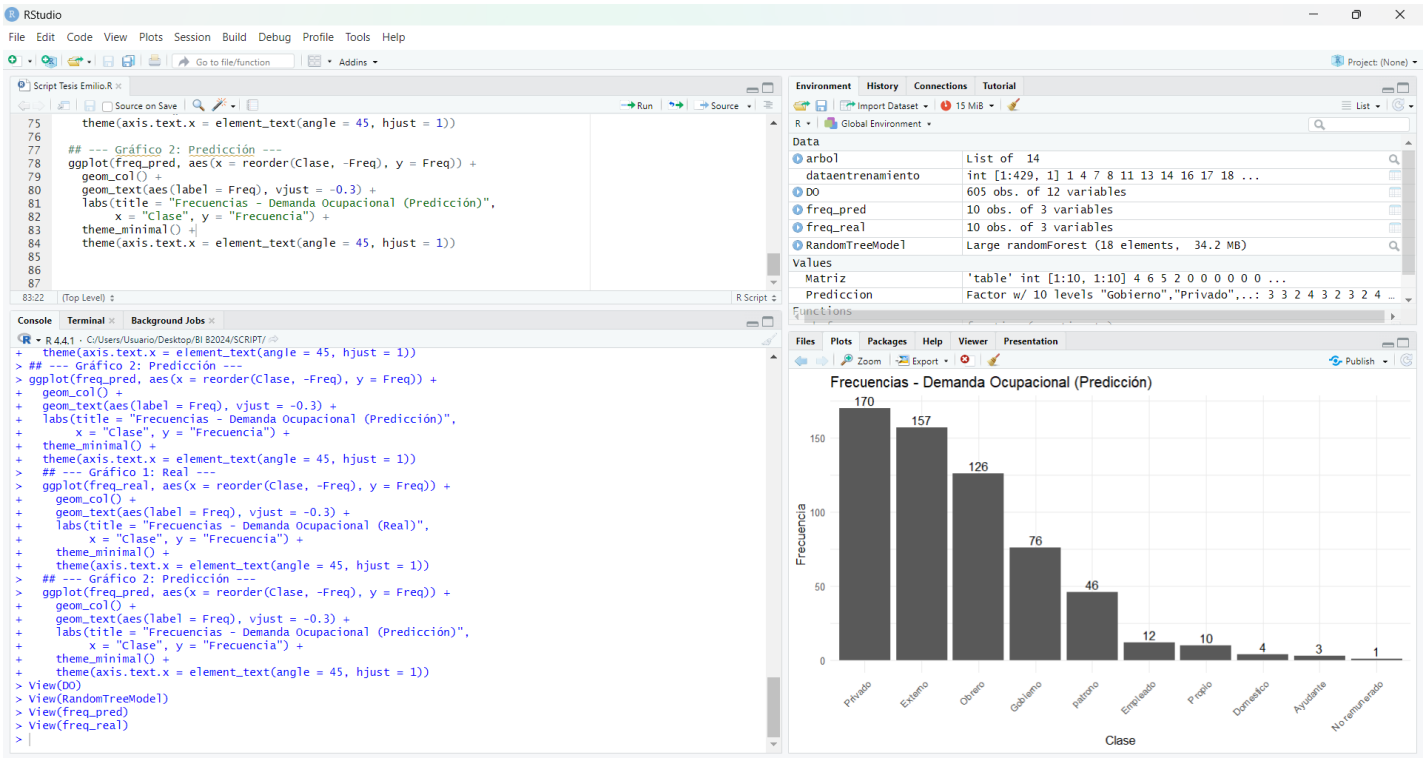


Figura 10 Visualización de Panel R STUDIO

RESULTADOS

después de una exhaustiva realización de modelos gráficos mediante el software R podemos notar los siguientes resultados, la comparación que existe entre la distribución real de la demanda ocupacional y la predicha. En el modelo diseñado podemos evidenciar los siguientes hechos.

dentro de estos dos gráficos de barras, se logra apreciar que existe muy poca diferencia entre la demanda ocupacional real versus la demanda

ocupacional predicho, es decir, esta mantiene lo mismo perfiles, ocupacionales o áreas de trabajo, podemos notar que en la ciudad de Guayaquil, no hay una tendencia palpable de la demanda de nuevos tipos de especializaciones o ocupaciones.

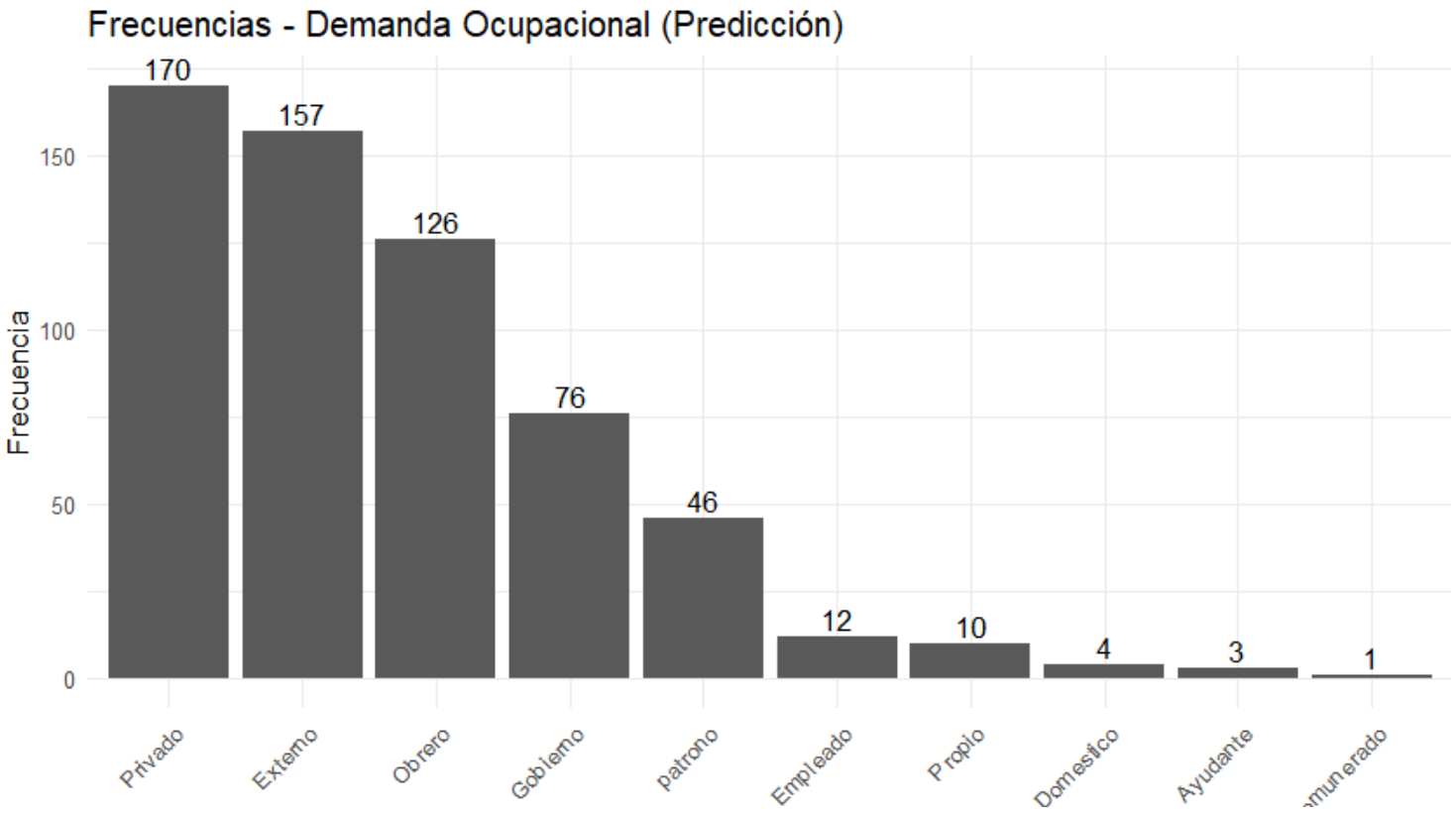


Figura 11 Frecuencias demanda ocupacional Predicha

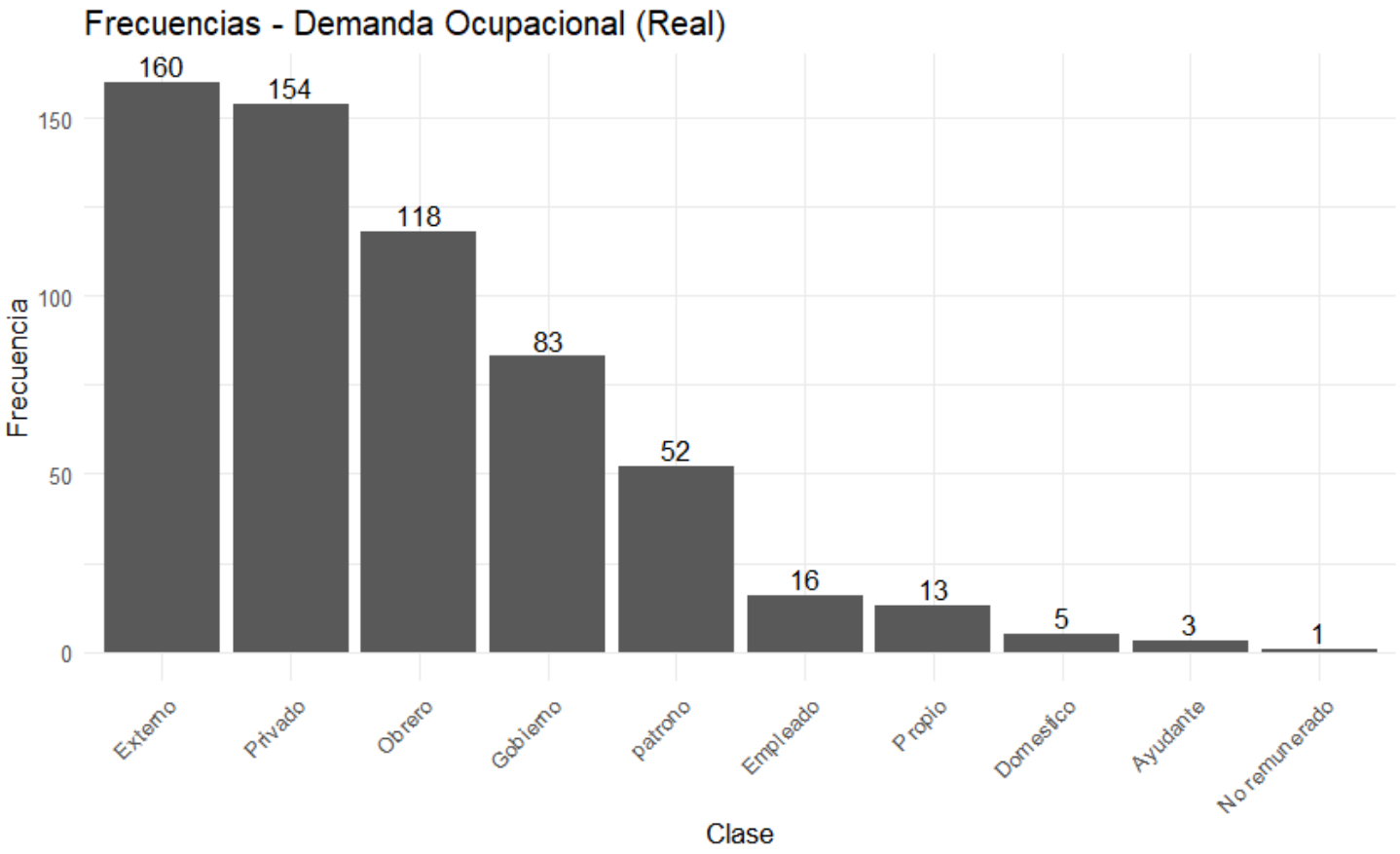


Figura 12 Frecuencia demanda ocupacional Real

Distribución general

- ambas gráficas muestran que las ocupaciones como privado, externo y obrero, se concentra la mayor parte de la demanda, tanto en la realidad como en la predicción
- esto indica que el modelo captó la tendencia principal de las áreas de trabajo o las ocupaciones dónde esas tres categorías son las más frecuentes dentro de la base de datos desarrolladas por el Instituto nacional de estadística y censos, al igual que lo fue con las predicciones

Diferencias específicas

- El modelo predijo un número ligeramente mayor de observaciones, que la realidad es la categoría del sector privado 170 vs. 154
- por otro lado, sucedió lo contrario, en la categoría del sector externo, el modelo predijo menos 157 vs. 160
- en la categoría del obrero, el modelo los sobreestimó 126 vs. 118
- Y por consiguiente, las categorías como gobierno patrono, se pueden dar leves diferencias Gobierno 76 vs. 83 reales; Patrono 46 vs. 52 reales

Clases minoritarias

- Las categorías con muy baja frecuencia, como Doméstico, Ayudante y No remunerado, aparecen en ambos gráficos con valores muy bajos 1 a 5 observaciones
- En estos casos, el modelo suele tener más dificultad porque hay pocos ejemplos en los datos para aprender. Sin embargo, lo positivo es que el modelo sí identificó estas clases.

DESARROLLO DEL ÁRBOL

1. Estructura básica

- **Nodo raíz:**

  La primera variable que aparece (*Sitio_trabajo >= 4*) es la que mejor separa los datos según la demanda ocupacional.

- Si **sí** cumple la condición (>=4) → se clasifica hacia la izquierda (principalmente como *Externo*).

- Si **no** cumple (<4) → el análisis sigue hacia la derecha, donde entran en juego otras variables como *Ingresao_laboral*, *Ocupacion*, *Estado_civil*, etc.

- **odos intermedios:**
  Representan reglas condicionales basadas en las variables (años laborales, estado civil, ingresos, tamaño del establecimiento, nivel educativo, etc.).
- **Hojas (nodos terminales los óvalos de colores):**
  Muestran la **clase final asignada** (Privado, Externo, Gobierno, Obrero…). Cada hoja indica la categoría que el modelo predice para los casos que cumplen las reglas del camino.

2. Principales resultados observados

- **Clases más frecuentes:**
  Se observa que las predicciones terminan mayoritariamente en **Privado**, **Externo** y **Obrero**, lo que coincide con los gráficos de frecuencias (esas son las clases dominantes en los datos).

- **Gobierno:**

  Aparece en una hoja intermedia, pero con **menos casos** que las categorías anteriores.

- **Clases minoritarias (patrono, propio, doméstico, no remunerado, ayudante, empleado):**

  Están listadas en la leyenda, pero en el árbol aparecen como **unused**, lo que significa que el modelo **no encontró suficientes patrones estadísticamente fuertes** para crear reglas específicas que separen estas categorías.

  En la práctica, esto ocurre porque esas clases tienen muy pocos registros en la base de datos.

Interpretación de reglas del Árbol

- **Si Sitio_trabajo >= 4 → Externo.**

  Es decir, el lugar de trabajo (quizás tipo o ubicación) mayor o igual a 4 se asocia directamente con empleos externos.

- **Si Sitio_trabajo < 4 y Ingresao_laboral >= 550 → Externo.**

  Los ingresos más altos tienden a relacionarse con trabajos externos.

- **Si Sitio_trabajo < 4, Ingresao_laboral < 550 y Educacion < 4 → Obrero.**

  Es decir, menores ingresos y menor educación se asocian con trabajos de tipo obrero.

- **Si Años_Laborales >= 26 → Privado.**

  Mucha experiencia laboral se relaciona con empleos en el sector privado.

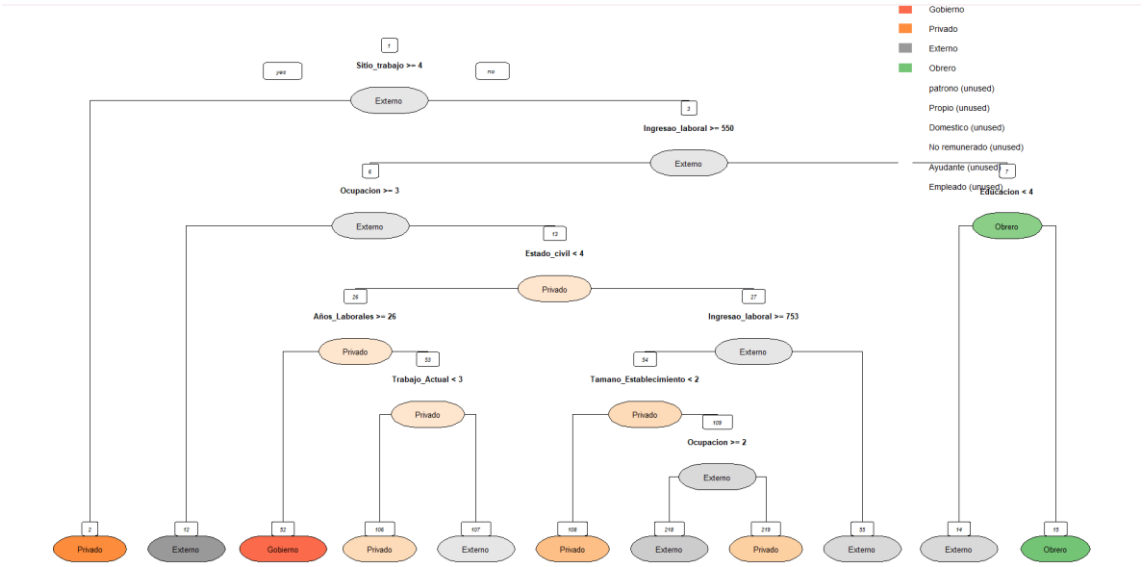


Figura 13 Modelo Visual Árbol de decision R studio

Conclusión de los resultados

La comparación entre la distribución real de la demanda ocupacional y la predicha por el modelo evidencia que las categorías Privado, Externo y Obrero concentran la mayor parte de las observaciones en ambos casos, lo que refleja la capacidad del modelo para captar la tendencia principal del mercado laboral. No obstante, se observan ligeras diferencias en el número de casos por clase, especialmente en la categoría Privado, donde se registra una sobrestimación, y en Gobierno, con una subestimación. En las categorías con menor representación (Doméstico, Ayudante y No remunerado), el modelo mantiene conteos cercanos a la realidad, aunque su baja frecuencia limita la capacidad de predicción. En general, el modelo logra reproducir la estructura de la demanda ocupacional con un alto grado de correspondencia respecto a los datos reales.

El árbol de decisión construido muestra que variables como sitio de trabajo, ingreso laboral, años laborales y nivel de educación son los principales determinantes en la clasificación de la demanda ocupacional. El modelo predice principalmente en las categorías Privado, Externo y Obrero, coincidiendo con la distribución real observada en los datos. En contraste, categorías con menor representación como Doméstico, Ayudante o No remunerado no generan ramas independientes en el árbol, reflejando su baja incidencia en el conjunto de datos. En términos generales, el árbol permite identificar reglas interpretables que vinculan características socioeconómicas con la pertenencia a un determinado tipo de ocupación, aportando claridad y transparencia al modelo predictivo

DISCUSION

Éste proyecto está centralizado en calcular la demanda laboral que existe en las pymes y grandes empresas de Guayaquil, usando un modelo The Machine Learning y Deep Learning, utilizando Random Forest y árboles de decisión, el cual nos clasifique la demanda ocupación real por la demanda ocupacional que se predijo, gracias a la base de datos que nos proporcionan instituto nacional de censos y estadística del Ecuador permitiéndonos así categorizar diferentes ocupaciones o áreas de trabajo, ya sea privado, externo, obrero, etc. y obtener distribuciones de frecuencias bastante cercanas a la realidad permitiéndonos así conocer y visualizar, mediante gráficos y resultados los diferentes problemas socioeconómico y políticos que existen en la ciudad de Guayaquil. Gracias al uso de la ciencia de datos en este proyecto. por otra parte, existe un artículo más relevante en cuanto a la demanda laboral, utilizando Random Forest, cuál es un estudio de pronóstico de dinámicas del desempleo (un Proxy cercano a la demanda laboral) dónde también fue empleado Random Forest y varios otros modelos Machine Learning, con datos macroeconómicos y del mercado laboral, donde desempeña más el Random Forest, el cual tuvo un rendimiento más sobresaliente para capturar patrones no lineales en los datos laborales y económicos. Este artículo creado por Kyungsu Kim en Georgia Institute of Technology School of Engineering, Atlanta GA 30332, USA (2015) dónde se desempeñó la predicción del desempleo en Estados Unidos, superando así, revisiones lineales y a métodos complejos como LSTM estuvo en muestra que tanto en contexto locales, como es en la ciudad de Guayaquil como

internacionales, los métodos de ensamblaje y árboles son altamente efectivos para fenómenos laborales, dónde la demanda laboral de las pymes y grandes empresas de Guayaquil, trabaja con datos micro de carácter socio demográfico y laboral empleados por el INE y aplicado a la realidad ecuatoriana. El artículo. Utiliza indicadores macroeconómicos como PIB, inflación, cuentas externas, etc. para proyectar así la tasa de desempleo nacional en Estados Unidos. Donde podemos darnos cuenta que el proyecto realizado se centra en la estructura interna del mercado laboral, local, mientras que el artículo aborda al nivel agregado de desempleo en una economía mucha más avanzada. alguna de las implicaciones prácticas del proyecto, como la para la ciudad de Guayaquil, los hallazgos permiten a las empresas y autoridades a identificar sectores con mayor demanda laboral u ocupacional, y así detectar posibles desajustes entre la oferta y la predicción, permitiendo que los empresarios utilicen a sus empleadores, encargados de recolectar al personal sofisticado y capacitado, para que no exista una una brecha en la demanda laboral por otro lado para Estados Unidos el estudio de este artículo sugiere que las instituciones económicas puedan anticipar cambios en la tasa de desempleo y ajustar políticas monetarias o fiscales de una manera más preventiva, gracias a la ciencia de datos para ambos escenarios. Demostró un potencial claro para mejorar la planificación y reducir incertidumbre dentro de los mercados laborales. De esta manera tenemos convergencia de nuestras conclusiones, ambas investigaciones se pueden concluir en que los modelos de ensamble como Random Forest o bosques, aleatorios y árboles de decisión son mucho más confiables que los métodos

lineales tradicionales usados actualmente y si bien es cierto la utilidad que que desarrollan estos modelos no sólo se radica en la precisión, sino que también en su versatilidad para distintos contextos, como tanto en un mercado laboral, emergente y segmentado, como es en el país de Ecuador, como es un entorno macroeconómico, desarrollado como lo es la potencia mundial, Estados Unidos.

CONCLUSIONES

El análisis de los conceptos teóricos permitió comprender la relevancia de aplicar técnicas de inteligencia artificial en la demanda laboral de las pymes y donde empresas de Guayaquil, así que estas impactan directamente en la sociedad y el gobierno reduciendo así este conflicto político que lleva años, siendo controversial para el país gracias al uso de la ciencia de datos. Podemos darnos cuenta de cuáles son las verdaderas ocupaciones demandadas Ecuador, específicamente en la ciudad de Guayaquil utilizando métodos como árbol de decisión, nos podemos dar cuenta de cuáles fueron realmente las ocupaciones relevantes dentro del modelo estadístico el cual nos ayude a interpretar de una manera más organizada los datos relevantes para así poder luego traspasarlos a un modelo de bosques aleatorios o random forest, por otra parte, el modelo de Random Forest o bosques aleatorios, nos permitió tener una gráfica más diseñada de los datos, usando un gráfico de barras, texturizado por el modelo en R. Studio permitió comparar entre los datos reales contra los datos predichos, utilizando la base de datos del Instituto nacional de censo y estadística, gracias a esto nos pudimos dar cuenta de la importancia que tiene la ciencia de datos para calcular no solamente la demanda laboral, sino cualquier aspecto gubernamental o que sea de alto interés para el beneficio y mejora de cualquier ámbito empresarial, social económico Dándonos así, resultados de sumo interés, tanto como para investigaciones académicas, proyectos empresariales o temas gubernamentales.

RECOMENDACIONES

El modelo propuesto de árboles de decisión y bosques, aleatorios o Random Forest nos va a permitir predecir la demanda laboral de las pymes y grandes empresas de Guayaquil, permitiéndonos así desarrollar mediante análisis de datos y estadística, la demanda ocupacional o el área de trabajo, donde más suelen haber congruencias hasta este año Sin embargo, para lograr garantizar que la efectividad del modelo, sea recomendable, implementar herramientas de analítica avanzada y software especializado sobre modelos de clasificación mediante ciencia de datos El proyecto ya construye un flujo completo de análisis predictivo con las recomendaciones, puede darse más rigurosidad académica, para mostrar que el modelo es confiable y resaltar el valor práctico para empresas y políticas laborales permitiéndonos así tener una calidad de datos asegurándonos de documentar el proceso de limpieza entre las variables que tenían números en blanco porque se eliminaron cuántos registros quedaron, etc. mientras que gracias al árbol de decisión es un modelo base y fácil de interpretar donde gracias a Rand Forest, podemos ajustarlos y parámetros y reportar cómo afectan esto en la precisión de los datos donde gracias también a la matriz infusión este explica no sólo los aciertos totales sino que categorías en el modelo donde más se confunde

REFERENCIAS

Bahrami, M., Xu, Y., Tweed, M., Bozkaya, B., & Pentland, A. (2022). Using gravity model to make store closing decisions: A data-driven approach. *Expert Systems with Applications, 205*, 117703. https://doi.org/10.1016/j.eswa.2022.117703 PubMed

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324 ACM Digital Library

Costa, V. G., & Pedreira, C. E. (2023). Recent advances in decision trees: An updated survey. *Artificial Intelligence Review, 56*(5), 4765–4800. https://doi.org/10.1007/s10462-022-10275-5 ACM Digital Library

Drezner, Z., & Eiselt, H. A. (2024). Competitive location models: A review. *European Journal of Operational Research, 316*(1), 5–18. https://doi.org/10.1016/j.ejor.2023.10.030 IDEAS/RePEc

Drezner, Z., & Hamacher, H. W. (Eds.). (2002). *Facility location: Applications and theory*. Springer. SpringerLink

Huff, D. L. (1963). A probabilistic analysis of shopping center trade areas. *Land Economics, 39*(1), 81–90. https://doi.org/10.2307/3144521 JSTOR

Krugman, P. (1991). Increasing returns and economic geography. *The Quarterly Journal of Economics, 106*(2), 407–437. https://doi.org/10.1162/qjec.1991.106.2.407 Google Books

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News, 2*(3), 18–22. https://cran.r-project.org/doc/Rnews/ R Project

Marianov, V., & Eiselt, H. A. (2024). Fifty years of location theory—A selective review. *European Journal of Operational Research, 318*(3), 701–718. https://doi.org/10.1016/j.ejor.2024.01.036

Okabe, A., Boots, B., Sugihara, K., & Chiu, S. N. (2000). *Spatial tessellations: Concepts and applications of Voronoi diagrams* (2.ª ed.). Wiley. https://doi.org/10.1002/9780470317013 Wiley Online Library

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science, 2*, 160. https://doi.org/10.1007/s42979-021-00592-x SpringerLink

Stevens, L., & Shearmur, R. (2020). The end of location theory? Some implications of micro-work, work trajectories and gig-work for conceptualizing the urban space economy. *Geoforum, 111*, 155–164. https://doi.org/10.1016/j.geoforum.2020.02.010 Richard Shearmur

Wilson, A. G. (1967). A statistical theory of spatial interaction. *Transportation Research, 1*(3), 253–269. https://doi.org/10.1016/0041-1647(67)90035-4

DECLARACIÓN Y AUTORIZACIÓN

Yo, **Elías Veliz, Emilio Xavier**, con C.C: **# 0953840469** autores del trabajo de titulación: **Estudio de la demanda laboral de las PYMES y grandes empresas de Guayaquil.** previo a la obtención del título de **Licenciado en Negocios Internacionales** en la Universidad Católica de Santiago de Guayaquil.

1.- Declaro tener pleno conocimiento de la obligación que tienen las instituciones de educación superior, de conformidad con el Artículo 144 de la Ley Orgánica de Educación Superior, de entregar a la SENESCYT en formato digital una copia del referido trabajo de titulación para que sea integrado al Sistema Nacional de Información de la Educación Superior del Ecuador para su difusión pública respetando los derechos de autor.

2.- Autorizo a la SENESCYT a tener una copia del referido trabajo de titulación, con el propósito de generar un repositorio que democratice la información, respetando las políticas de propiedad intelectual vigentes.

Guayaquil, **21** de **agosto** de **2025**

f. _____

Nombre: **Elías Veliz, Emilio Xavier**

C.C: 0953840469

Presidencia de la República del Ecuador

Plan Nacional de Ciencia, Tecnología, Innovación y Saberes

SENESCYT
Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación

REPOSITORIO NACIONAL EN CIENCIA Y TECNOLOGÍA

FICHA DE REGISTRO DE TESIS/TRABAJO DE TITULACIÓN

TEMA Y SUBTEMA:	Estudio de la demanda laboral de las PYMES y grandes empresas de Guayaquil.		
AUTORA	Elías Veliz, Emilio Xavier		
REVISOR(ES)/TUTOR(ES)	Carrera Buri, Félix Miguel		
INSTITUCIÓN:	Universidad Católica de Santiago de Guayaquil		
FACULTAD:	Facultad de Economía y Empresa		
CARRERA:	Negocios Internacionales		
TITULO OBTENIDO:	Licenciada en Negocios Internacionales		
FECHA DE PUBLICACIÓN:	21 de agosto de 2025	**No. DE PÁGINAS:**	74 p.
ÁREAS TEMÁTICAS:	Economía del trabajo, Política de empleo, Empleo, Estudio de mercado.		
PALABRAS CLAVES/ KEYWORDS:	Demanda Laboral, PYMES, Guayaquil, Ciencia de Datos, Árbol de Decisión, Random Forest, ENEMDU.		

RESUMEN/ABSTRACT (150-250 palabras):

La presente investigación estudia la demanda laboral de las PYMEs y grandes empresas de Guayaquil mediante la aplicación de la ciencia de datos como herramienta analítica. El trabajo parte del contexto ecuatoriano caracterizado por elevados índices de desempleo y subempleo, así como por un desajuste entre las competencias de la población y los perfiles que requieren las organizaciones. Para el análisis se emplearon datos públicos del Instituto Nacional de Estadística y Censos (INEC), a través de la encuesta ENEMDU, procesados con el software R Studio. Se aplicaron modelos de aprendizaje supervisado, específicamente árboles de decisión y bosques aleatorios, alcanzando una precisión del 95%. Los resultados evidencian que las ocupaciones en los sectores “Privado”, “Externo” y “Obrero” concentran la mayor parte de la demanda, mientras que categorías como “Doméstico”, “No remunerado” y “Ayudante” presentan menor representación. La investigación concluye que la aplicación de modelos predictivos permite identificar patrones ocupacionales, anticipar tendencias de contratación y facilitar la toma de decisiones tanto para las empresas como para las políticas públicas en el ámbito laboral ecuatoriano.

ADJUNTO PDF:	☒ SI	☐ NO
CONTACTO CON AUTOR/ES:	**Teléfono:** +593 994426735	**E-mail:** emilioeliasve@gmail.com
CONTACTO CON LA INSTITUCIÓN (C00RDINADOR DEL PROCESO UTE)::	**Nombre:** Freire Quintero, Cesar Enrique	
	Teléfono: +593 990090702	
	E-mail: cesar.freire@cu.ucsg.edu.ec	

SECCIÓN PARA USO DE BIBLIOTECA	
Nº. DE REGISTRO (en base a datos):	
Nº. DE CLASIFICACIÓN:	
DIRECCIÓN URL (tesis en la web):	