ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ciencias Naturales y Matemáticas

"Selección de candidatos para encuestas mediante técnicas de machine learning"

PROYECTO INTEGRADOR

Previo la obtención del Título de:

INGENIERO EN ESTADÍSTICA INFORMÁTICA

Presentado por:

Samuel Diaz Romero

&

Wladimir Sanyer Mosquera

GUAYAQUIL - ECUADOR

Año: 2021

DEDICATORIA

Le dedico este proyecto, en primer lugar, a Dios, ya que me ha demostrado en muchísimas ocasiones que él siempre está ahí ayudándome. Por consiguiente, le dedico este proyecto también a mi amada esposa y a mi familia quienes siempre me han apoyado y brindado ánimos.

Diaz Romero, Samuel Ismael

AGRADECIMIENTOS

En primer lugar, agradecerle a Dios, pues me ha brindado la fuerza y la fortaleza para seguir adelante y no flaquear, agradecerle por en todo momento estar para mí, aunque no siempre me acuerde de él.

Agradecerle a una persona que ha estado conmigo desde que tengo uso de razón, desde que estaba en la escuela, hasta ahora culminando esta etapa universitaria, a mi mejor amiga, y que, gracias a Dios, ahora es mi esposa, gracias por todo el apoyo brindado durante todos estos aaños.

Gracias a mis padres, quienes han sido un pilar fundamental en cada etapa de mi vida, a mis hermanos a mis sobrinos, y por supuesto a toda mi familia de Samborondón, quienes me han demostrado que siempre estarán para celebrar los logros y acompañarme en las derrotas.

Gracias a la institución, a los profesores que la conforman, pues gracias a sus enseñanzas, estoy cumpliendo una gran meta en mi vida, agradecimiento a la tutora Jessica, por habernos brindado

todo el apoyo durante la realización del presente proyecto.

Gracias a mi compañero de tesis Wladimir, por su arduo empeño en la realización de este proyecto.

Diaz Romero, Samuel Ismael

DEDICATORIA

A Dios, quien me da las fuerzas y la sabiduría que necesito. A mi familia, que a pesar de las adversidades me brindaron todo su apoyo desde el inicio de esta etapa y creyeron en mí.

Sanyer Mosquera, Wladimir Alejandro

AGRADECIMIENTOS

Gracias a la ESPOL, gracias a todos los profesores y en especial a la profesora Jessica, quien fue una gran tutora, quien estuvo dispuesta a ayudarnos y guiarnos en todo momento.

Agradecimiento a mi compañero de tesis Samuel por su apoyo y colaboración durante la ejecución del proyecto.

Sanyer Mosquera, Wladimir Alejandro

DECLARACIÓN EXPRESA

“Los derechos de titularidad y explotación, nos corresponde conforme al reglamento de propiedad intelectual de la institución; Samuel Ismael Diaz Romero y Wladimir Sanyer Mosquera damos nuestro consentimiento para que la ESPOL realice la comunicación pública de la obra por cualquier medio con el fin de promover la consulta, difusión y uso público de la producción intelectual"

Samuel Ismael Diaz Romero

Wladimir Sanyer Mosquera

EVALUADORES

Firmado electrónicamente por:
JESSICA ALEXANDRA MENENDEZ CAMPOS

Ph.D. Sandra García Bustos

PROFESOR DE LA MATERIA

M.Sc. Jéssica Menéndez Campos

PROFESOR TUTOR

RESUMEN

Las personas encargadas de tomar datos a través de la aplicación de cuestionarios en una encuesta tienen en su poder una gran responsabilidad, dado que, si se presentan errores en la toma de los datos, afectarían el posterior análisis de los resultados, es por esto que muchas empresas dedicadas a la recolección, procesamiento, análisis y presentación de resultados, buscan contratar gente que les permite obtener el menor error posible en el levantamiento de datos a través de las encuestas. Para esto se enfocan en el adecuado proceso de selección del personal, el cual incluye el cumplimiento de requisitos de base, todo tipo de herramientas que les ayude a disminuir el riesgo de seleccionar al personal inadecuado. Debido a esta necesidad, se propone la implementación de modelos de machine learning, que sirvan de soporte al departamento de recursos humanos para la contratación del nuevo personal. Para lograr esta implementación se trabajó con varias variables, por ejemplo, sociodemográficas, también se usó los Key Perfomance Indicators que la empresa utiliza para medir la eficiencia de sus colaboradores, y se implementaron algoritmos de machine learning tales como el Random Forest y el algoritmo K Nearest Neighbors. En el presente proyecto se demostrará que el modelo K Nearest Neighbors es el modelo adecuado para predecir observaciones, de acuerdo con las métricas que se usaron para evaluar la efectividad de este. Este modelo puede ser útil al momento de la toma de decisiones en el departamento de recursos humanos.

Palabras Clave: Precisión, Validación Cruzada, Random Forest, KNN

ABSTRACT

The people in charge of taking data through the application of questionnaires in a survey have a great responsibility in their power, since, if errors occur in the data collection, they would affect the subsequent analysis of the results, that is why Many companies dedicated to the collection, processing, analysis and presentation of results seek to hire people who allow them to obtain the least possible error in data collection through surveys. For this, they focus on the adequate personnel selection process, which includes compliance with basic requirements, all kinds of tools that help them reduce the risk of selecting the wrong personnel. Due to this need, the implementation of machine learning models is proposed, which serve as support to the human resources department for the hiring of new personnel. To achieve this implementation, several variables were worked on, for example, sociodemographic variables, the Key Performance Indicators that the company uses to measure the efficiency of its collaborators were also used, and machine learning algorithms such as Random Forest and the K algorithm were implemented. Nearest Neighbors. In this project, it will be demonstrated that the K Nearest Neighbors model is the appropriate model to predict observations, according to the metrics that were used to evaluate its effectiveness. This model can be useful when making decisions in the human resources department.

Keywords: Accuracy, Cross Validation, Random Forest, KNN

ÍNDICE GENERAL

ABREVIATURAS

INEC	Instituto Nacional de Estadística y Censos
ETL	Extract, Transform, Load
KPI	Key Perfomance Indicators
VP	Verdaderos Positivos
FP	Falsos Positivos
FN	Falsos Negativos
VN	Verdaderos Negativos
KNN	K Nearest Neighbors
LOOCV	Leave One Out Cross-Validation

SIMBOLOGÍA

K Número de particiones del set de datos.

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPÍTULO 1

1. INTRODUCCIÓN

Se conoce que el éxito de toda empresa, organización o institución, depende en gran parte del desempeño de sus trabajadores, debido a que es el talento humano quienes ayudan al cumplimiento de sus objetivos. Siendo el proceso de la selección del personal adecuado una de las etapas más importantes dentro de la institución y uno de los grandes desafíos del departamento de recursos humanos.

La selección del personal, consiste en encontrar a la persona idónea para ocupar el puesto adecuado, con la finalidad de obtener mejores resultados y sumando de una manera positiva al crecimiento y mejora de la organización. (Chiavenato, 2011).

La productividad laboral es un buen indicador que permite medir la eficiencia del personal que forma parte de una organización, así como también la relación que existe entre la cantidad de trabajo producido por el individuo y el uso de los recursos que intervinieron para su obtención. Cabe mencionar que la productividad se encuentra fuertemente relacionada con el incremento económico y los logros de la institución. (Morales, 2020).

Por lo tanto, las empresas e/o investigadores, dedican su tiempo y recursos para determinar cuáles son las características que debería cumplir un aspirante a un puesto determinado, en base a esto, se han desarrollado diferentes métodos, técnicas y algoritmos que ayuden a hacer la selección adecuada para ocupar un cargo.

Las empresas dedicadas a la labor del levantamiento de información mediante encuestas, tienen como finalidad, proveer información adecuada y veraz de los datos levantados, para esto, como antes mencionado, debería existir un grupo de personas altamente capacitadas y

calificadas para realizar esta tarea, por ejemplo en el caso del INEC, es una institución que cuenta con indicadores que son capaces de medir la productividad de sus colaboradores, convirtiéndose así en un instrumento a favor de la mejora continua de la organización.

El Instituto Nacional de Estadística y Censos (INEC) es la institución encargada de presentar las estadísticas oficiales del Ecuador, dedicándose a la recolección, registro y al procesamiento de información, la cual será usada para la toma de decisiones del país. Dentro de las actividades más importantes se encuentra el censo de población y vivienda, que se realiza cada 10 años y permite conocer algunas características (parámetros) de interés de la población. Adicionalmente INEC realiza encuestas importantes como; Encuesta Nacional de Empleo, Desempleo Subempleo – ENEMDU, Encuesta Nacional de Salud y Nutrición – ENSANUT, Encuesta Nacional de Desarrollo Infantil – ENDEIN, Encuesta de Superficie y Producción Agropecuaria Continua – ESPAC, entre otras. La calidad y veracidad de los resultados que se esperan obtener de la aplicación de encuestas se verán influenciadas por la correcta ejecución de los deberes y funciones del personal contratado siendo estos: el encuestador, supervisor, digitador, crítico, revisor, asistente estadístico, analista entre otros.

En base a lo expuesto, el presente proyecto tiene por finalidad poder realizar una mejor selección del personal encargado de realizar las encuestas, se utilizará como base la información del personal que realizó las encuestas anteriores, mediante las herramientas y métodos estadísticos, para así poder lograr una adecuada selección del personal encuestador, mediante el estudio de los indicadores de desempeño (KPI) proporcionados por la empresa.

1.1 Descripción del problema

En la última década el uso del dato ha crecido de manera exponencial en las organizaciones, especialmente al momento de tomar decisiones, es por eso que en la actualidad es de suma importancia basarse en datos que avalen las decisiones tomadas y en lo posible

predigan las consecuencias de las mismas. Toda esta serie de procesos inicia con la toma del dato.

Es importante cuidar los datos desde un primer momento, puesto que, si son erróneos, se estarían haciendo malas inferencias de la población ecuatoriana, por lo tanto, todo estudio basado en ellos, no serían confiables.

Provocando que la reputación de la empresa encargada en levantar las encuestas, se vea afectada de manera negativa, por ende, la búsqueda de métodos y soluciones para hacer posible la obtención del dato desde su primer filtro, es decir con los encuestadores, sea necesaria, para ello se debería recurrir a una evaluación de la calidad del trabajo realizado.

Por lo que el presente proyecto se centrará en dar soluciones a esta problemática, si bien es cierto los errores del pasado no se pueden corregir, pero se puede aprender de la experiencia, y esté será el tema principal del proyecto, pues en el campo estadístico existen muchas técnicas, modelos y algoritmos, que bien empleados, podrían ayudar a la contratación del personal.

En este caso lo que se tiene para el estudio es la data histórica de las mediciones de la calidad del trabajo de los encuestadores, junto con las características de los mismos, es decir la información brindada a la empresa para tenerlas en sus bases de datos, por ejemplo, las variables edad, sexo, nivel de instrucción entre otras. Se procederá a entrenar modelos de machine learning y se comprobará la eficiencia del modelo mediante métricas, y una vez que se tenga el modelo que haga la mejor predicción, será el modelo que brindará soporte al departamento de recursos humanos, en la selección del personal.

1.2 Objetivos

1.2.1 Objetivo General

Clasificar a un postulante laboral mediante el uso de técnicas machine learning, analizando su KPI y características del personal contratado.

1.2.2 Objetivos Específicos

- Analizar posibles similitudes entre los individuos, tomando en cuenta el desempeño laboral por medio de las variables de estudio.
- Estudiar la relación entre los KPI laborales por parte del histórico personal civil del INEC, por medio de análisis de correspondencia multivariado.
- Construir un Dashboard con el análisis multivariado para la clasificación automatizada de futuros postulantes mediante técnicas de machine learning.

1.3 Marco Teórico

En esta sección se brindará información acerca de conceptos y teorías que son importantes a tener en cuenta para lograr un desarrollo eficiente de la problemática y entendimiento claro de las técnicas y conceptos que intervienen en todo el proceso.

1.3.1 Anonimización de datos

Anonimizar los datos en una base, consiste en transformar los datos de tal manera que, por medio de estos, no se puedan reconocer a los individuos, por tanto, la anonimización es un instrumento que contribuye a aminorar el peligro que conlleva el tratamiento masivo de datos personales, este proceso contribuye a la divulgación de la información sin que este afecte el derecho a la protección de los datos de las personas. (CEPAL, 2020).

En otras palabras, anonimizar una base de datos es precautelar la identidad de las personas, puesto que en la actualidad la obtención de datos se puede obtener de cualquier fuente y si se hace un mal uso de estos, puede llegar a perjudicar de alguna manera la integridad de las personas, por lo que es vital mantener la información que proporcionan bajo estricto margen.

Según la Agencia Española de Protección de Datos (2016) en su libro Orientaciones y garantías en los procedimientos de ANONIMIZACIÓN de datos personales, el objetivo principal o finalidad de anonimizar los datos es en primer lugar

mantener reservada la identidad de las personas, es decir que esta no pueda ser identificada una vez que los datos ya hayan sido anonimizados y conservando la veracidad de los resultados que se obtengan luego de cualquier tratamiento que se le aplique a los datos, dichos resultados obtenidos de usar una base de datos anonimizada no debería diferir de los resultados que se obtendrían si se usara la misma base de datos pero en este caso sin anonimizar.

1.3.2 Proceso ETL

El proceso ETL hace referencia a 3 fases que se presentan en todo proceso de análisis de datos, estas fases por las siglas que vienen del idioma inglés son: Extract que hace referencia a la extracción u obtención propia de los datos, de los cuales pueden encontrarse en diferentes fuentes, Transform que hace referencia a la parte más importante dentro de todo el proceso ETL, pues esta se encarga de adaptar o llevar los datos de acuerdo al objetivo planteado, de acuerdo a las necesidades que se desean cubrir, por último se tiene la fase de Load que hace referencia al proceso de carga de los datos, en donde los datos previamente ya transformados, serán almacenados o cargados en algún sistema o base de datos, el lugar donde será alojada dicha data dependerá de las especificaciones de la empresa u organización, conservando la finalidad del proceso ETL, que es el uso de los datos para obtener información valiosa.

1.3.3 KPI

Proviene del inglés Key Perfomance Indicators (KPI), se los conoce por diversos nombres, tales como; indicadores clave del desempeño, indicador clave de actuación, indicadores de desempeño, entre otros, sin embargo, en todas las formas en que se los denomina, los KPI están relacionados con los objetivos que se desean alcanzar dentro de una empresa, haciendo uso de indicadores que puedan cuantificar la productividad.

Los KPI se los conoce como aquellos indicadores que posibilita medir el éxito de toda acción en base a los objetivos anticipados. (Penguin, 2018).

Según Gonzales (2021) para lograr los objetivos establecidos dentro de una empresa, o dentro de un equipo de trabajo, lo más común es encontrar una manera de poder evaluar o medir la productividad, el rendimiento o la eficacia de las actividades que intervienen dentro del objetivo a alcanzar, dicha medición se la conoce como un indicador que pasa a ser un KPI, ya que estos hacen referencia a aquellos indicadores que miden el rendimiento o el desempeño, y que comúnmente se encuentran expresados en términos de porcentaje, teniéndolos como una herramienta de apoyo a la hora de tomar decisiones futuras.

1.3.4 Análisis Multivariante

El análisis multivariante se refiere a métodos que se utilizan para analizar, estudiar e interpretar el efecto de múltiples variables de forma simultánea, sobre un conjunto de individuos.

1.3.5 Machine Learning

Machine learning como se lo conoce en inglés y su traducción en español Aprendizaje automático, comprende métodos que sirven para tomar decisiones, los cuales están basado en los datos, es decir las decisiones se basan en lo aprendido en datos anteriores. Comprende todo algoritmo, modelo o método utilizado para el análisis de datos.

1.3.5.1 Datos de entrenamiento

Corresponde al conjunto de datos, los cuales se emplean para que el modelo de machine learning aprenda, es decir como lo dice su nombre, con este conjunto de datos, el modelo se va a entrenar tantas veces como el analista le indique.

Este conjunto de datos, es de vital importancia, debido a que, si se hace una mala elección del conjunto, el modelo no tendrá la suficiente capacidad para lograr predecir correctamente futuras observaciones.

1.3.5.2 Datos de test

Corresponde al conjunto de datos, los cuales son empleados para probar la eficiencia de los diferentes algoritmos usados dentro del campo del machine learning.

El conjunto de datos de test, o también llamados el conjunto de datos de prueba, ayuda a determinar de entre todos los modelos de machine learning, cuál es el que presenta un menor error al momento de realizar alguna predicción, o, en otras palabras, cual, de estos modelos al predecir una nueva observación, se acerca más al valor real del dato.

1.3.5.3 Aprendizaje Supervisado

El Aprendizaje Supervisado forma parte de uno de los métodos de machine learning, caracterizado a que la función principal radica en realizar la predicción de un dato, tomando en cuenta el estudio de datos anteriores, se conoce como el estudio del pasado para conocer el futuro. El Aprendizaje Supervisado posee dos clasificaciones al momento de hacer alguna predicción, la predicción como tal, se da cuando la observación que se quiere predecir es un número, y por otra parte cuando queremos clasificar una nueva observación en alguna categoría, a esto se lo conoce como problema de clasificación.

En este método de Aprendizaje Estadístico, lo que se pretende es encontrar el modelo que permita realizar la relación entre las características medidas de los datos, junto con una determinada etiqueta la cual se asocia a estos, posteriormente una vez obtenido el mejor modelo, este sirve para adherir etiquetas a nuevos datos. El Aprendizaje Supervisado comprende una división que consta entre realizar tareas de regresión o de clasificación, cuando la etiqueta es una cantidad continua se habla del caso de la regresión, mientras que cuando la etiqueta es una categoría, se habla del caso de la clasificación. (VanderPlas, 2016).

Como uno de los tipos del Aprendizaje estadístico, se tiene el Aprendizaje supervisado, el cual se caracteriza por etiquetar al set de datos de entrenamiento, para posterior realizar futuras predicciones. (Raschka & Mirjalili, 2020)

Una subcategoría del Aprendizaje Supervisado consiste en la predicción de una variable cuantitativa, es decir donde se desea obtener un resultado que sea continuo, a esta subcategoría se la conoce como análisis de regresión, el cual consiste en predecir un resultado mediante la relación que existe entre un conjunto de variables llamadas predictoras o explicativas y una variable que es de respuesta y que esta a su vez es continua. (Raschka & Mirjalili, 2020).

Otra subcategoría de este tipo de aprendizaje es la denominada clasificación, cuyo objetivo principal, predice la etiqueta de una nueva observación cuya clase es categórica. (Raschka & Mirjalili, 2020).

1.3.5.4 Aprendizaje No Supervisado

El aprendizaje no supervisado es uno de los tipos de machine learning, el cual, de acuerdo a su funcionalidad, de acuerdo al conjunto de algoritmos que se pueden emplear en este tipo de aprendizaje, se considera que este cabe dentro del marco de la exploración descriptiva de los datos, a diferencia del aprendizaje supervisado el cual tiene como finalidad principal predecir una variable de interés de acuerdo a las variables con que esta se encuentra asociada.

Se dice que el aprendizaje no supervisado entra en el campo de la exploración de lo datos, pues lo que se busca es hallar esas agrupaciones o patrones que se encuentran intrínsecas en los datos, y lo que no se busca es realizar predicciones a futuras observaciones, pues en este caso no se trabaja con ninguna variable de salida o variable de respuesta.

Se puede clasificar a los algoritmos o métodos que se emplean en este tipo de aprendizaje a los métodos de formación de clústeres, métodos de visualización, análisis de componentes principales (ACP).

En el método de Aprendizaje No Supervisado no se hace referencia de ninguna etiqueta a los datos, pues se busca modelar las características de estos, en otras palabras, es dejar que los datos hablen por sí solos. Siendo más específicos los modelos que intervienen en este método están relacionados a la reducción de la dimensionalidad y al agrupamiento de los datos. Los algoritmos los cuales permiten reducir la dimensionalidad ayudan a realizar una representación resumida de los datos, por otro lado, los algoritmos de agrupamiento, como el nombre bien lo indica, ayuda a determinar los posibles grupos en que se estén formando los datos. (VanderPlas, 2016).

En este tipo de aprendizaje al contrario del aprendizaje supervisado, no se cuenta con una etiqueta en los datos, no se cuenta con una variable de salida, o variable de respuesta, lo que se pretende con este aprendizaje es hallar estructuras que se encuentran en los datos, para obtener información que sea significativa. (Raschka & Mirjalili, 2020).

1.3.5.5 Sesgo

El concepto de Sesgo en el machine learning es similar al que se conoce en estadística, ya que, en este caso, sería la diferencia entre la predicción del modelo entrenado y el verdadero valor de la variable de respuesta de la nueva observación.

Un modelo puede presentar dos casos, tiene un sesgo alto o un sesgo bajo, el sesgo alto hace referencia a cuando las predicciones del modelo están muy alejadas al verdadero valor del dato, y caso contrario en el sesgo bajo es cuando el modelo si logra capturar de muy de cerca el verdadero valor del dato.

1.3.5.6 Varianza

La varianza se encarga de medir que tan dispersos están las predicciones que arroja el modelo.

Y de igual manera que con el sesgo, un modelo puede presentar una varianza alta o una varianza muy baja.

1.3.5.7 Sobreajuste

El sobreajuste u overfitting, como se lo conoce en el idioma inglés, es un concepto muy importante que se emplea dentro del campo del machine learning, y hace referencia si el modelo presenta sobreajuste o está sobreajustado.

El sobreajuste implica que el modelo utilizado para probar con el set de datos, se encuentra muy ajustado a los datos de entrenamiento, es decir que este modelo se ajusta bastante a la forma de los datos, debido a esto no se cuenta con la suficiente libertad para poder captar el ruido que presentan los datos, lo cual es muy importante a tener en cuenta si se quiere realizar una buena predicción.

El sobreajuste continuamente se presenta en los modelos que son considerados más complejos, y que son mucho más difíciles de interpretar, un ejemplo de este tipo de modelos es considerar los modelos no lineales, modelos tales como las redes neuronales, o en cuyo caso todos aquellos modelos que presentan o que tienen demasiados grados de libertad, otorgándoles la característica de que dichos modelos pierden rigidez, y por lo tanto estos son bastantes flexibles, dando como resultado que se adaptan mucho más a los datos de entrenamiento.

Se puede destacar entre las características que tiene un modelo que presente sobreajuste, que las predicciones del modelo estarán bastante cerca del verdadero valor real, en otras palabras, que estos modelos presentan un sesgo

bajo, cabe destacar que dichos modelos sobreajustados presentan una varianza muy grande.

De acuerdo a las características que se han mencionado y como consecuencia, estos modelos sobreajustados se los considera como modelos que realizan predicciones deficientes, es decir no van a ser tan buenas, esto debiéndose a que el modelo se encuentra sobreajustado al conjunto de datos de entrenamiento.

1.3.5.8 Subajuste

El subajuste o underfitting como conocido en el idioma inglés, es otro de los conceptos que son muy importantes dentro del campo del machine learning, en palabras generales es todo lo contrario al concepto relacionado con el sobreajuste de los modelos.

Al igual que en el caso de que un modelo presenta sobreajuste, no es considerado un buen modelo, pasa de igual manera para los que presentan subajuste, estos son considerados no tan buenos, ya que un modelo que se encuentre subajustado es un modelo muy simple y por tanto muy fácil de interpretar.

Estos modelos subajustados presentan la característica que poseen pocos grados de libertad, es decir que son muy rígidos, desembocando que no se ajusten muy bien al conjunto de datos de entrenamiento, es decir que, si se cambia el conjunto de datos de entramiento, el modelo no va a variar mucho.

1.3.5.9 Precisión del Modelo

Lo importante de la implementación de un modelo de machine learning es saber cómo logra predecir los datos a futuro, pues se requiere de alguna manera poder cuantificar la eficiencia o la precisión del modelo.

1.3.5.10 Error de Entrenamiento

Se conoce como el error de entrenamiento a aquel error promedio que comete el modelo de machine learning al predecir la variable de salida o variable de respuestas, pero en el conjunto de datos de entrenamiento.

El error de entrenamiento no es una buena estimación de la capacidad predictiva del modelo.

1.3.5.11 Error de test

Se conoce como el error de test a aquel error promedio que comete el modelo de machine learning al predecir la salida de nuevos datos, siendo diferentes del conjunto de datos de entrenamiento. Este tipo de error es un buen estimador de la precisión del modelo.

En el caso conciso de los problemas de clasificación que es el caso de este proyecto, la precisión del modelo puede ser estimado ya sea por el porcentaje de aciertos en la clasificación, por la matriz de confusión o por la curva ROC.

1.3.5.12 Matriz de confusión

La matriz de confusión es una herramienta dentro del campo del machine learning, que se usa para evaluar la precisión del modelo, mediante esta matriz se pueden obtener diferentes métricas que ayudan a determinar si el modelo es lo suficientemente bueno para poder predecir observaciones de manera correcta.

Matriz de confusión		Predicción del modelo	
		Positivos	Negativos
Datos Reales	Positivos	VP	FN
	Negativos	FP	VN

Tabla1.3: Matriz de confusión

Fuente: Elaboración propia

La tabla 1.3 muestra como está estructurada la matriz de confusión, lo cual se aprecia que con esta herramienta se puede comparar la cantidad de casos positivos y negativos que predice el modelo versus la cantidad de casos positivos reales y negativos reales, haciendo mención que los datos reales hacen referencia al conjunto de datos de test y que las predicciones del modelo son en base al conjunto de datos de test.

Dentro de la matriz tenemos:

VP: Hace referencia a los verdaderos positivos es decir la cantidad de casos en el que el modelo predijo correctamente que son positivos.

FN: Hace referencia a los falsos negativos, es decir la cantidad de casos en el que el modelo predijo que son negativos cuando en realidad son positivos.

FP: Hace referencia a los falsos positivos, es decir la cantidad de casos en el que el modelo predijo que son positivos, pero en realidad son negativos.

VN: Hace referencia a los verdaderos negativos, es decir la cantidad de casos en el que el modelo predijo correctamente que son negativos.

1.3.5.13 Exactitud

Como mencionado antes, para medir la capacidad predictiva del modelo, se necesitan métricas, entre ellas se tiene la exactitud o también llamado Accuracy, el cual está relacionado con el porcentaje de observaciones que fueron clasificadas correctamente con respecto al total de predicciones.

Cálculo de la exactitud (Accuracy) mediante la fórmula (1.3.1)

$$Exactitud = \frac{VP + VN}{VP + VN + FP + FN} \tag{1.3.1}$$

1.3.5.14 Estimación del error de test

La estimación del error de test en un problema de clasificación se la obtiene con $1 - Exactitud$, esto indica en cuanto se ha equivocado el modelo en predecir correctamente las observaciones.

1.3.5.15 Especificidad

La especificidad o también conocida como la tasa de verdaderos negativos, está relacionada con la proporción que existe entre los casos en el que el modelo predijo correctamente los casos negativos, con respecto al total de los negativos, la especificidad se la obtiene aplicando la fórmula (1.3.2)

$$Especificidad = \frac{VN}{FP + VN} \tag{1.3.2}$$

1.3.5.16 Sensitividad

La sensitividad también conocida como exhaustividad, hace referencia a la tasa de verdaderos positivos, la cual está relacionada con la proporción que existe entre los casos en el que el modelo predijo correctamente los casos positivos, con respecto al total de los casos positivos, por lo que la sensitividad se la obtiene aplicando la fórmula (1.3.3)

$$Sensitividad = \frac{VP}{VP + FN} \tag{1.3.3}$$

1.3.5.17 Curva Roc

La Curva Roc, es otra métrica usada para conocer qué tan bueno es el modelo con el que se trabaja, pues con esta curva se logra cuantificar las mediciones de los falsos positivos con los verdaderos positivos, en otras palabras, cuantificar la cantidad de casos en el que el modelo predijo correctamente que son positivos y la cantidad de casos en el que el modelo predijo que son positivos, pero

en realidad son negativos, donde se considera un valor óptimo si el valor de los falsos positivos fuese 0 y el de los verdaderos positivos fuese 1.

1.3.5.18 Métodos de Remuestreo

Los métodos de remuestreo consisten en tomar varias muestras de un mismo conjunto de datos, sin tomar en cuenta alguna, la distribución que estos poseen.

Bajo el marco del aprendizaje estadístico, los métodos de remuestreo son indispensables debido a que se obtendrán diversas muestras del conjunto de datos, resultando en varios conjuntos de datos de entrenamiento, por consiguiente, el modelo con que se trabaja podrá entrenarse con el conjunto de datos de entrenamiento que se obtuvo de aplicar alguna técnica de remuestreo.

Una manera de conocer la eficiencia del modelo, es el uso de técnicas de remuestreo, las cuales tienen una función similar, es decir que se va a tomar un conjunto de muestras que van a servir como conjunto de datos de entrenamiento y otra muestra que servirá como conjunto de datos de test, y este proceso será repetido varias veces, en lo que se diferencian cada técnica es en la forma como se toman las muestras. (Johnson & Kuhn, 2013).

1.3.6 Revisión Bibliográfica:

Martínez, Broche et al,(2013) Aplicaron métodos multivariantes tales como el análisis de componentes principales (ACP), análisis factorial (AF), análisis de clúster (AC), que les permita obtener información que se encuentra contenida en los datos, que no se puede apreciar de manera simple, consiguiendo de esta manera una base para la toma de decisiones al momento de seleccionar a estudiantes que pasarán a formar parte de los proyectos de software de la carrera de Ingeniería en Ciencias Informáticas, mediante el estudio de las calificaciones alcanzadas por los estudiantes del segundo año de la misma carrera. Con el análisis de componentes principales se redujo la dimensionalidad original de los datos, y con este resultado

se formaron 3 grupos mediante el análisis de clúster. Con el análisis factorial se logró identificar la estructura intrínseca latente en los datos, obteniéndose 3 factores. Con la obtención de los grupos se logró discriminar entre los estudiantes que tenían un alto desempeño en asignaturas específicas, pero que tienen un menor desempeño de forma general, siendo estos los estudiantes que conformaban tanto el grupo 1 como el grupo 2.

Koutra, Barbounaki et al (2017). En el trabajo que realizan, presentan una combinación de técnicas y modelos para investigar cuales son los criterios que se pueden aplicar para la selección del personal, que va a formar parte de la industria marítima en Grecia, recopilando datos de 14 empresas marítimas. Se hace uso del modelo multicriterio para la toma de decisiones, en este caso para la selección del personal, dicho modelo consiste en elegir la mejor opción o alternativa, y sobre los candidatos que se encuentran disponibles, tomar decisiones, quienes se encuentran determinados por diversos criterios. Este modelo multicriterio se apoya en el proceso de jerarquía analítica (AHP), el cual es escogido por el potencial que presenta con respecto a la automatización de complejos problemas de decisión multicriterio, también tiene como fuerte el modelar opiniones subjetivas. La técnica multivariante de Análisis de Correspondencia (AC), interviene en el proceso como una técnica nueva que permite clasificar los criterios que fueron más evaluados.

Gustavo y Castillo (2021) en su trabajo Machine learning en la mejora del proceso de selección del personal administrativo de la Corte Superior de Justicia de Lima, 2020, surge a través del problema en que, en los procesos de selección del personal administrativo hechos de forma tradicional, ocasionaban puestos desiertos o cancelados. En el trabajo se utilizó un enfoque cuantitativo para la toma de muestras y un enfoque cuasi experimental en donde abordaron el uso específico tres indicadores, tales como el índice de contratación, el índice de evaluación curricular y por último el índice de personal postulante, para esto, se contó con la experiencia de los expertos para la obtención de las observaciones, utilizando herramientas tales como el machine learning, el test de Wilcoxon, que les ayudan a comparar los resultados de

pretest y postest en la obtención de las observaciones. Los resultados que obtuvieron de los indicadores tuvieron una mejoría del 35% en el índice de personal postulante, una mejoría del 2% con respecto al índice de contratación, y por último una mejoría del 17% en el índice de evaluación curricular.

CAPÍTULO 2

2. METODOLOGÍA

Según Cortés e Iglesias (2004) en su libro Generalidades sobre Metodología de la Investigación, definen a la metodología como una ciencia, la cual se encarga de dar las directrices, o de guiar un determinado proyecto o proceso de una manera eficiente y eficaz para obtener los resultados esperados, teniendo como meta final o como objetivo principal el proporcionar las estrategias a seguir durante todo el proceso.

En otras palabras, lo que se pretende dar a conocer es que la metodología o marco metodológico, es la herramienta que muestra los pasos a seguir de una manera sistemática, con respecto al desarrollo o ejecución de un proceso, para luego realizar la interpretación de los resultados obtenidos y alcanzar el objetivo planteado.

En este capítulo se muestran las herramientas y los procedimientos que se aplicaron para estructurar la base de datos con la cual se trabajó, así como la implementación de los diferentes algoritmos que son usados dentro del campo del machine learning, que fueron de ayuda para alcanzar el objetivo general y los objetivos específicos planteados en el capítulo anterior.

2.1 Bases de datos

Se trabajaron con 42 bases de datos, proporcionadas por la institución con la cual se trabajó, estas bases contaban con el registro histórico del personal civil de campo que fue

contratado durante el desarrollo de la actualización cartográfica, que se usará posteriormente para el censo de población y vivienda.

Cada base registró la información proporcionada por las personas involucradas en la ejecución de dicha encuesta, en las mismas se encontraban variables de tipo sociodemográficas y variables que la institución usó para medir la eficiencia del trabajo realizado por el personal.

Cabe mencionar que la institución nos entregó cada base totalmente anonimizada, pues como se mencionó en el primer capítulo del marco teórico, la anonimización buscaba proteger la identidad de las personas, en este caso del personal que trabajó para la institución, haciendo de esta manera que cada persona que había participado en la realización de dicha encuesta estaba identificada por un ID, garantizando la anonimización de la persona y que cualquier tratamiento o resultado de algún análisis aplicado a los datos anonimizados iba a ser igual a trabajar con los datos no anonimizados.

2.2 Software y Librerías

2.2.1 Software

El software que se utilizó para poder realizar el proceso ETL, así como también la implementación de los diferentes algoritmos que nos brinda el machine learning, fue Rstudio, que es el IDE del lenguaje de programación R.

Este lenguaje de programación es usado para realizar todo tipo de análisis estadístico, pues brinda una gama variada de herramientas relacionadas con metodologías estadísticas, este lenguaje de programación es usado frecuentemente cuando se trabaja en la resolución de una problemática, que basará sus posibles soluciones haciendo uso de la implementación de algoritmos de machine learning.

2.2.2 Librerías

La tabla 2.1 indica las librerías utilizadas en el lenguaje de programación R son las siguientes:

Tabla 2.1. Librerías que se usaron para la elaboración del proyecto

Librería	Descripción
readxl	Empleada para la obtención de datos Excel
data.table	Empleada para la manipulación de gran volumen de datos.
caret	Empleada para el uso de modelos robustos con validaciones
purrr	Empleada para la manipulación de datos no estructurados
dplyr	Empleada para la manipulación de datos
recipes	Empleada para el uso de validaciones de modelos

2.3 Proceso ETL

2.3.1 Limpieza de datos

Como una parte muy importante que se da al inicio de todo análisis estadístico, es la relación con la limpieza de los datos, la cual forma parte del proceso ETL ya mencionado en el marco teórico.

La limpieza que se realizó a las diferentes bases de datos, fue de manera exhaustiva y se encontró que existían registros que poseían solo una ID, más, sin embargo, el resto de columnas o el resto de variables carecían de información alguna, tenían valores nulos, motivo por el cual se procedió a eliminarlos, puesto que lo que se deseaba era utilizar datos con toda la información completa.

2.3.2 Transformación de datos

- Para la variable Titulo se hallaron varios valores con 0, por lo cual a los valores con 0 se le asigno el factor No, el cual hace referencia que no posee título.
- Para la variable *Nivel_Instruccion* se hallaron varios valores con 0, por lo cual a los valores con 0 se le asigno el factor Bachiller, el cual hace referencia al mínimo nivel que debe poseer un postulante a encuestador.
- Se filtro por el factor de interés Personal de Campo (Encuestador - Supervisor) de la variable *Cargo_Aplica*, puesto que nuestro estudio se centra solamente para dicho cargo.
- Puesto que se deseaba solo a los trabajadores que fueron contratados se procedió a quitar de la data los que no fueron contratados para cada proceso de selección.
- Se eliminó las variables que no serán de nuestro interés para el estudio en cada base de datos.
- Se procedió a eliminar valores de ID duplicados en cada base de datos por separado.

2.4 Variables

En esta sección se van a detallar las variables que fueron objetos de estudio y con las cuales se procedió a dar un primer vistazo a como se están comportando los datos, estas variables seleccionadas fueron:

2.4.1 Sexo

Corresponde a la distinción del personal de campo (encuestadores), entre hombres y mujeres

Clasificación de sexo:

Femenino

Masculino

2.4.2 Edad

Esta variable corresponde a la edad que el encuestador posee al momento del levantamiento de la encuesta.

2.4.3 Estado Civil

El Estado Civil hace referencia a la situación de cualquier persona de acuerdo al Registro Civil del Ecuador, con respecto si al mismo posee o no pareja y la posición legal con respecto a esto.

Clasificación de estado civil:

Soltero/a

Casado/a

Divorciado/a

Unión libre

Viudo/a

2.4.4 Nivel de Instrucción

El Nivel de Instrucción hace alusión al nivel más alto obtenido por la persona dentro del sistema formal de educación. (INEC, 2013).

El Nivel de Instrucción de una persona nos indica los estudios realizados, cuyo nivel es el más alto que ha obtenido, esto independiente de si ha terminado, este en proceso de terminación, o en cuyo caso haya habido abandono del estudio (INEC, 2013)

Clasificación de Nivel de Instrucción:

Primaria

Bachillerato

Egresado

Estudiante Universitario

Técnico Superior

Tecnología

Tercer nivel

2.4.5 Experiencia

Esta variable hace referencia a si el individuo ha trabajado anteriormente para la institución.

Categorías de la variable Experiencia.

Sí

No

2.4.6 Sector

Esta variable hace referencia al total de edificios que le fue asignado a cada persona encargada de levantar la encuesta.

Se puede definir a un edificio como “toda construcción o estructura que puede estar constituida por una o varias viviendas, establecimientos económicos, instituciones públicas o privadas, que ocupa un espacio determinado. Ejemplo: casas, escuelas, iglesias, garajes, bodegas, etc.” (INEC, 2021).

2.4.7 Total de errores

La variable Total de errores corresponde a la suma de todos los errores que el encuestador ha cometido según los criterios de evaluación que contempla la institución.

2.4.8 KPI Logístico

Esta variable se forma dividiendo la variable Total de errores para la variable Sector, y el resultado se muestra en porcentajes, de esta manera se conoce cuál fue el porcentaje de errores que el encuestador ha cometido.

Una vez que se obtuvo este valor del KPI, para efectos del estudio se procedió a hacer una categorización de los resultados obtenidos, puesto que con una revisión de la parte descriptiva de los datos se pudo apreciar que existe valores muy altos de este KPI en este caso correspondía a que el encuestador ha cometido muchísimos más errores de lo que se podría esperar, como también sucedió el caso contrario, es decir que se presentaban casos en donde había errores muy bajos y errores nulos, para esos casos ya se tenía una idea de que sí hubo encuestadores que hicieron un excelente trabajo.

Debido a la consideración anterior, se llegó a establecer una regla de decisión, la cual consiste en que si el porcentaje de errores es decir el valor del KPI logístico sobrepasaba el 50%, a este encuestador se lo procedió a clasificar como “No Aceptable”, caso contrario el encuestador se lo clasificó como “Aceptable”, se tomó este umbral para la regla de distinción, por cuanto se esperaría que el encuestador cometa pocos errores que sería el caso en que obtenga un KPI menor de 50%, que equivaldría a que el encuestador se equivocó un 50%.

Por el otro extremo los encuestadores quienes obtuvieron mayor al 50% es porque alcanzaron dicho porcentaje de error, pues se encontraron casos de encuestadores que tenían un porcentaje de error de más de100%, de más de 200%, etc.

2.4.9 Nota final

Esta variable hace referencia a la puntuación total que obtuvieron los encuestadores en el proceso de selección, pues se les aplicaron algunas pruebas y esta variable representa el total de la calificación que lograron obtener, teniendo como rango esta variable que va de 0 a 100

Luego del respectivo proceso ETL y luego de haber seleccionado las variables antes mencionadas, cabe indicar que no fueron todas estas variables con las que se entrenaron a los modelos de machine learning usados para este proyecto, previo a eso se realizó la implementación de dos técnicas que permiten conocer la importancia de estas variables

predictoras, dichas técnicas se basan en la implementación del algoritmo Random Forest, lo cual se dará una más amplia explicación más adelante en este capítulo.

2.5 Metodología General

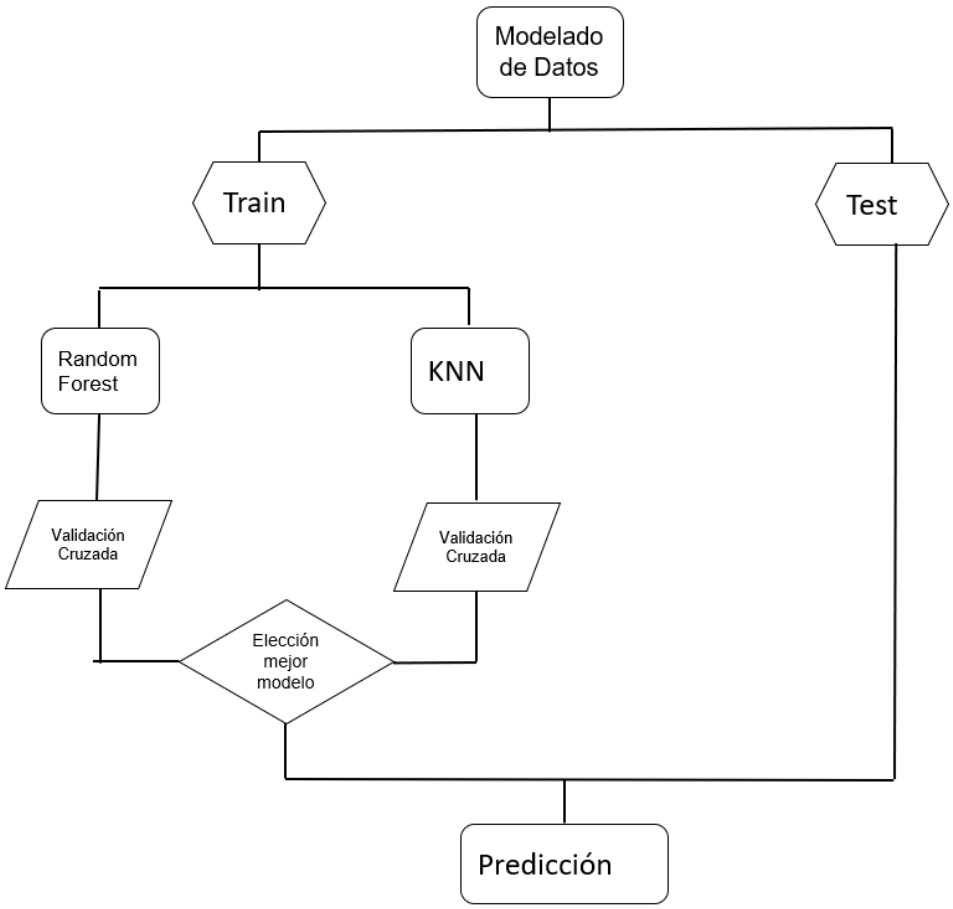


Gráfico 2.2: Diagrama de flujo de selección mejor modelo

Se puede apreciar mediante el grafico 2.2 un esquema que detalla cómo se procedió a tratar a los datos para su uso en los modelos, se tomaron de los datos originales un conjunto que sirvió para entrenar el modelo y otro para predecir, luego el conjunto de datos de entrenamiento se los usó para cada uno de los modelos para que estos sean entrenados, luego se aplicó la técnica de validación cruzada para que se pueda obtener

una mejor estimación de que tan bien lograron predecir los modelos en nuevas observaciones, usando el conjunto de datos de test.

2.6 Análisis de Correspondencia

El análisis de Correspondencia forma parte de los métodos que intervienen dentro del análisis multivariante, pues el análisis de correspondencia busca reducir las dimensiones de un conjunto de datos, manteniendo en lo posible la menor cantidad de pérdida de la información.

Esta técnica hace su estudio sobre las tablas de contingencia o tablas de doble entrada, por lo que esto indica que se estará trabajando con variables categóricas.

El objetivo del análisis de correspondencia es medir si es que existe o no; la relación que existe entre filas y columnas, y por otro lado la medir la relación o la asociación que existe entre las modalidades de una sola fila o una sola columna.

En este estudio se aplicó esta técnica multivariante, para conocer cómo se encontraban asociadas las variables categóricas que conformaron el conjunto de datos originales.

2.7 Hiperparámetros

En el campo del machine learning existen varios modelos de los cuales necesitan de hiperparámetros, y que estos no es que se pueden aprender o no es que se obtienen del conjunto de datos de entrenamiento, es decir que hay que darle a conocer al modelo cual es el valor de los hiperparámetros que se necesita, este valor lo decide la persona quien realiza el respectivo análisis.

No existe una ley o una regla que indique cual o como calcular el valor de los hiperparámetros en los diferentes modelos de machine learning, sin embargo se recurre a varias estrategias para poder lograr obtener el valor adecuado, que ayude al modelo ser bueno, entre uno de estos, está el de probar con varios valores, pero esto podría resultar muy costoso

computacionalmente dependiendo del modelo usado, comúnmente para poder probar el valor del mejor hiperparámetro se utiliza la técnica de K-Fold Cross Validation

2.8 Árboles de decisión

Los modelos de árboles de decisión se encuentran entre los métodos más populares que se usan dentro del machine learning, como característica principal que son basados en que el espacio de las variables explicativas o predictoras se particiona en un conjunto de regiones, y posteriormente lo que se busca es una aproximación dentro de cada una de estas regiones.

Este modelo de machine learning representa a sus regiones a través de diagramas de árbol, por lo que se lo conoce como dendograma.

2.8.1 Random Forest

Random Forest o también conocidos como Bosques Aleatorios es una metodología en donde se presenta la agrupación de muchos árboles de decisión, con el fin de mejorar de cierta manera el desempeño del modelo.

Este método es una mejora del método de Bagging, pues busca reducir la varianza cuando se realiza el promedio de los modelos, esto mediante la disminución de la correlación entre los diversos árboles que se han ajustado.

Dentro del método de Bagging se considera que puede existir correlación en las muestras que se obtienen mediante el muestreo con reemplazo o Bootstrap, ya que existe la posibilidad de que haya conjunto de muestras que pueden tener las mismas observaciones, pero en posiciones distintas.

El modelo de Bagging también busca reducir la varianza del modelo, pues usa como idea que sacando el promedio de un conjunto de observaciones de esta manera se logra reducir la varianza.

Para Rodrigo (2020) la idea del método de Bagging es “una forma de reducir la varianza y aumentar la precisión de un método predictivo es obtener múltiples muestras

de la población, ajustar un modelo distinto con cada una de ellas, y hacer la media (la moda en el caso de variables cualitativas) de las predicciones resultantes)”

Es decir que lo busca Bagging es promediar la estimación del modelo con diferentes conjuntos de datos de entrenamiento.

Como se mencionó anteriormente, ciertos modelos de machine learning necesitan de valores iniciales de los hiperparámetros que su propio modelo necesita, el Random Forest, presenta hiperparámetros tales como el número mínimo que debe tener un nodo para que este pueda ser dividido, por consiguiente, si este es mayor por tanto el modelo será menos flexible, deberá considerar el número de observaciones como mínimas que debe tener el nodo terminal.

La diferencia del Random Forest con el método Bagging, es cómo está construido el árbol, es decir que para cada división de la variable predictora se va a tomar el criterio de seleccionar únicamente un subconjunto aleatorio de predictores que serán seleccionados para dividir el espacio.

2.8.1.1 Importancia de los predictores

El algoritmo de Random Forest también es usado para hacer análisis exploratorio, este modelo busca reducir la varianza, pero tomando la idea de usar Bootstrap al igual que el modelo de Bagging, para tener una mejor capacidad predictiva.

Se han desarrollado estrategias que permitan cuantificar de alguna manera la importancia de las variables predictoras, entre estas se tiene la importancia por permutación y la impureza de nodos.

Esta técnica reconoce que tanto influye cada predictor sobre alguna métrica con la cual se haya evaluado el modelo, sin olvidar que estas se obtienen ya sea por validación cruzada o por OOB.

OOB (Out-Of-Bag) hace referencia al conjunto de datos que queda fuera al realizar muestreo con reemplazo es decir al aplicar la técnica de bootstrap

El gráfico siguiente muestra un algoritmo de cómo funciona esta técnica de importancia por permutación, que presenta Amat (2020).

1. Crear el conjunto de árboles que forman el modelo.
2. Calcular una determinada métrica de error (*mse*, *classification error*, ...). Este es el valor de referencia ($error_0$).
3. Para cada predictor j:
   - Permutar en todos los árboles del modelo los valores del predictor j manteniendo el resto constante.
   - Recalcular la métrica tras la permutación, llámese ($error_j$).
   - Calcular el incremento en la métrica debido a la permutación del predictor j.

$$\%\text{Incremento}_j = (error - j - error_0)/error_0 * 100$$

Figura 2.8: Funcionamiento de la técnica Importancia por permutación.

Fuente: Tomado de (Amat. J, 2020)

Se esperaría que el modelo tienda a aumentar el error, en caso de que el predictor que se permutó realmente ha contribuido al modelo. Por ende, se considera que, si el porcentaje de error del modelo ha aumentado debido a la permutación del predictor, entonces se puede interpretar que realmente el modelo esta influenciado por dicho predictor.

2.8.1.2 Impureza de nodos

Este método pretende cuantificar ese incremento de la pureza, que es ocasionado debido a las particiones en las que el predictor interviene.

La forma en cómo funciona es ligeramente sencilla, puesto que se almacena el descenso que se consigue en la métrica que se usa como criterio de la división de los árboles, estas medidas pueden ser ya sea la media cuadrática del error, o con índice Gini, posteriormente se calcula el promedio de los descensos que se han obtenido en cada

uno de los conjuntos de los árboles, dando como resultado en que será mayor la contribución del predictor para el modelo, si el promedio tiene valores altos.

Para este proyecto, el paso previo a la implementación del Random Forest, se hizo una inspección de cuáles eran las variables que mejor aportaban al modelo, y esto se logró obtener utilizando las técnicas previamente mencionadas, importancia por permutaciones e impureza de los nodos, dando como resultado que las variables a usar en el modelo fueron tres: Nota final, Edad y Estado civil del encuestador.

2.9 KNN (K NEAREST NEIGHBORS)

K NEAREST NEIGHBORS o también llamado K vecinos más cercanos, en términos sencillos, lo que hace es asignar las diferentes clases de acuerdo a las similitudes que presente con respecto al historial de los datos.

Es decir que, para la predicción de un nuevo dato, en este caso bajo el marco de la clasificación como está orientado el presente proyecto, lo que hace este algoritmo es ver el comportamiento de la variable de entrada y compararla con los datos con los que ya fue entrenado, y le asigna la variable de respuesta de estos, ya que ve la similitud en el comportamiento.

El parámetro que necesita este método es únicamente K, que representa cuantos vecinos más cercanos se utilizan para la predicción.

Es de mencionar que el conjunto de entrenamiento estará conformado por vectores en un espacio multidimensional, la variable de respuesta posee varios atributos, varias dimensiones.

Por lo que al momento de hacer nueva predicción el algoritmo utiliza una distancia para poder hallar los atributos más similares entre los atributos de entrada y los atributos que ya posee el modelo debido a los datos de entrenamiento, esta distancia generalmente es la distancia euclidiana.

$$x_i = (x_{1i}, x_{2i}, x_{3i}, \ldots\ldots, x_{pi}) \in X$$

Donde:

x_i: Representa el vector de la observación que forma parte del conjunto de entrenamiento y con $1 < i < n$, siendo n la última observación del conjunto de entrenamiento.

p: Hace referencia a los atributos que se encuentra en cada observación del conjunto de entrenamiento, es decir hace referencia al atributo total, al atributo edad y al atributo estado civil.

X: Es la matriz que se forma con todos los vectores resultantes del conjunto de entrenamiento

$$d(x_i, x_i) = \sqrt{\sum_{r=1}^{p} (x_{ir} - x_{ij})^2}$$

(2.8)

La fórmula (2.8) es la distancia euclidiana, en este caso menciona que la distancia entre la observación x_i y x_i, se define de esa manera, es decir tal como se mencionó anteriormente, lo que se busca es encontrar el elemento que tenga las características más parecidas a los datos de entrada, es decir que tengan la menor distancia posible con los vectores formados por el conjunto de entrenamiento.

2.10 Validación de Algoritmo

2.10.1 Validación Cruzada

La finalidad de todo modelo qua forma parte del conjunto de algoritmos usados dentro del campo del machine learning específicamente dentro del aprendizaje supervisado, es predecir.

Como mencionado, para evaluar la eficiencia del modelo al predecir los datos futuros, una buena estimación es estimar el error de test, la validación cruzada ayuda a tener una mejor estimación de este error.

La técnica de validación cruzada permite tener una mejor estimación de que tan bien predice el modelo con el cual se trabaja, existen diversas técnicas de validación cruzadas, entre las que se tiene K-Fold Cross Validation, LOOCV (Leave One Out Cross-Validation, Repeated K-Fold Cross Validation, Bootstrapping).

En los modelos que se usaron en el presente proyecto, se aplicaron estas técnicas de remuestreo o de validación cruzada como también se las conoce, en concreto los métodos de validación cruzada que se usaron fueron K-Fold Cross Validation y LOOCV (Leave One Out Croos-Validation).

2.10.2 K-Fold Cross Validation.

Este método o técnica de validación cruzada consiste en descomponer al conjunto de datos originales en K partes de tamaños aproximadamente iguales. Lo que hace este método es escoger $K-1$ partes de la división y tomarlas como el conjunto de datos de entrenamiento, y la parte restante se la toma como el conjunto de datos de test, es decir que el modelo el cual se está usando para modelar los datos se va a entrenar K veces y en cada una de las veces se aplicará este método.

Para obtener una estimación mucho más cercana del verdadero valor del error de test lo que se hace es sacar un promedio entre cada uno de los errores de test que se obtienen en cada una de las veces en que se entrena el modelo.

Según la literatura y de manera general el número de particiones de los datos originales, es decir el valor de K puede estar en 5 a 10 particiones, debido a que con estos valores se consigue un mejor equilibrio o un mejor balance entre el sesgo y la varianza que presente el modelo.

El conjunto de datos que se utilizó en el presente proyecto se dividió en $K=10$ partes y para poder realizar una división de las partes en casi iguales tamaños, lo que se hizo fue dividir el número de registros de la base de datos es decir los registros, para el número de partes dando como resultado en que cada una de las partes en que se dividió el conjunto de datos originales estuvo compuesto por un aproximado de 20 registros.

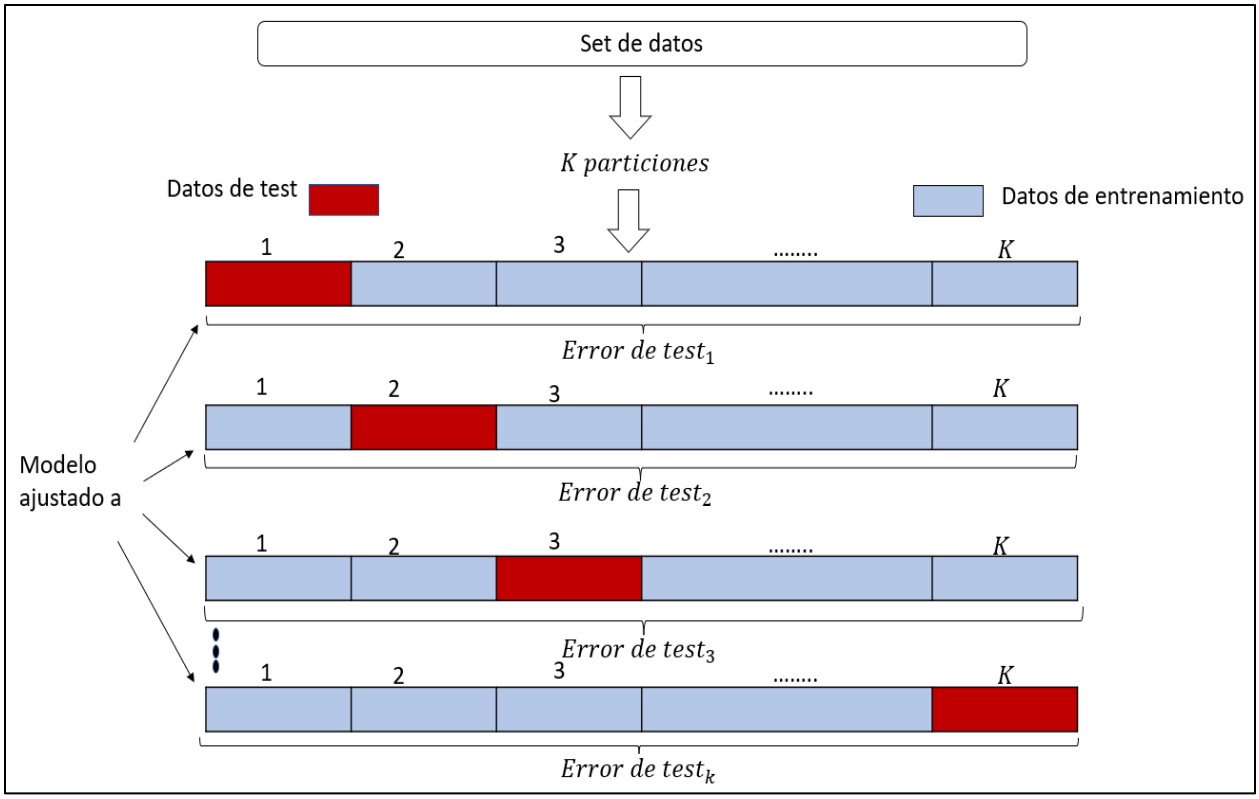


Gráfico 2.10.1: K-Fold Cross Validation

Fuente: Creación propia

El gráfico 2.10.1 ayuda a entender de mejor manera cómo funciona la técnica de k-Fold Cross Validation, y como se mencionó antes, para cada iteración, se calcula la estimación del test, y para tener una mejor estimación del verdadero valor del error de test, al final se hace un promedio de todos estos errores obtenidos en las iteraciones, esto mediante la fórmula

$$Estimación\ del\ error\ de\ test = \sum_{i=1}^{k} \frac{Error\ de\ test_i}{k} \qquad (2.7.1)$$

Donde K como bien se ha mencionado antes, representa las partes en que se va a dividir el conjunto de datos originales.

2.10.3 LOOCV (Leave One Out Croos-Validation)

Este método de validación cruzada consiste en tomar como datos de entrenamiento todo el conjunto de datos originales exceptuando una observación o registro, y lo que queda fuera de este conjunto de datos de entrenamiento, pasa a formar parte del conjunto de datos de test, es decir que por medio de este método el conjunto de datos de test está formado por una única observación.

El proceso que interviene en este método se lo realizará tantas veces como observaciones posee el conjunto de datos, dejando como conjunto de datos de test una observación diferente en cada repetición, y el modelo que se emplee se lo ajusta en cada repetición de este proceso.

Para conocer la precisión del modelo lo que se emplea es sacar un promedio entre la estimación de la precisión del modelo de cada una de las iteraciones.

LOOCV (Leave One Out Cross-Validation), presenta una desventaja por el coste computacional, debida a que el modelo se ajusta tantas veces como observaciones hay en el conjunto de datos.

Uno de los problemas a considerar cuando se usa la técnica de LOOV, es que debido a que para entrenar al modelo solo se le aplica una única observación y como conjunto de datos de entrenamiento, prácticamente casi todo el conjunto de datos originales, esto desemboca en que el modelo presente una alta varianza y esto a su vez se cae en el problema de sobreajuste. (Aggarwal, 2020).

Es decir que debido a que solo se deja una observación afuera en cada iteración, el modelo no va a variar casi en nada al momento de ajustarse al conjunto de datos de entrenamiento, pues lo único que varía en este conjunto de datos de entrenamiento es la observación que forma parte del dato a testear, esto conduce a que haya una reducción en el sesgo, y por otra parte debido a que será muy dependiente del conjunto de datos

que se posee, entonces la varianza tenderá a aumentar, y es exactamente de lo que se habla cuando un modelo presenta sobreajuste.

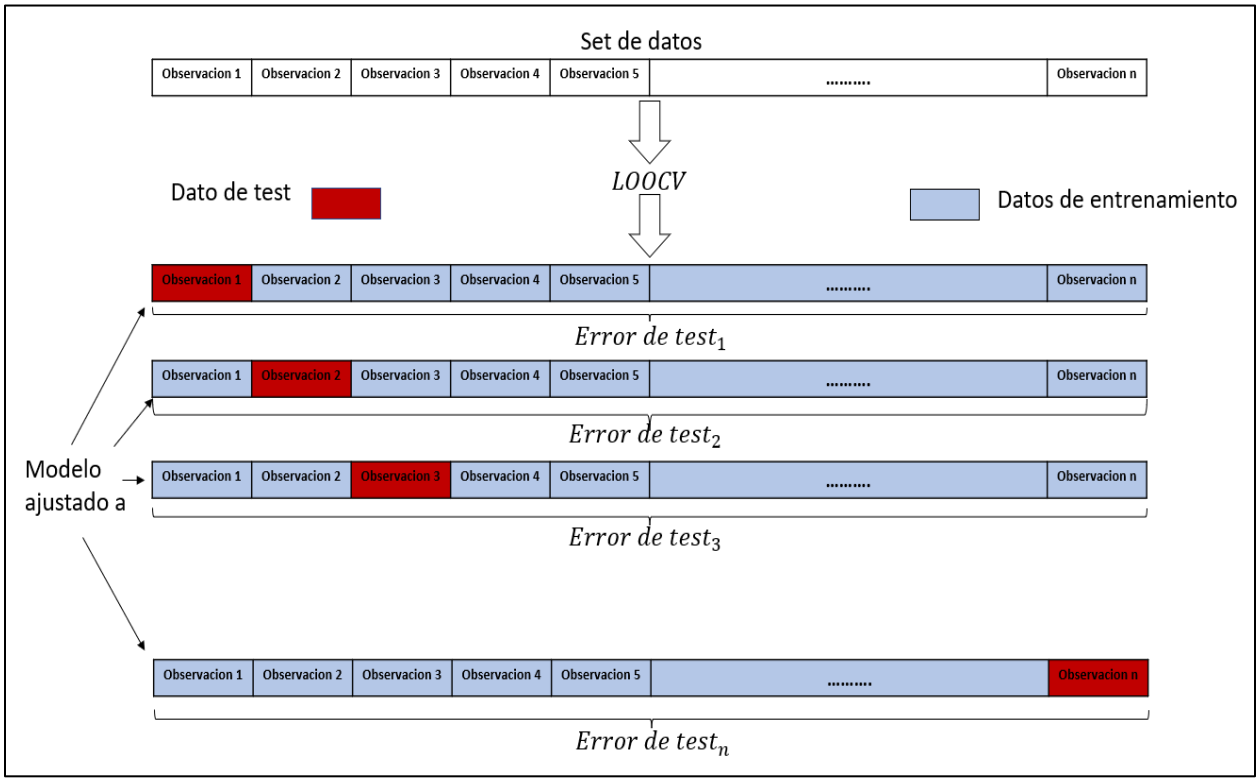


Gráfico 2.10.2: LOOCV (Leave One Out Cross Validation)

Fuente: Elaboración propia

El gráfico 2.10.2 ayuda a comprender de una mejor manera cómo funciona la técnica de LOOCV, pues como bien se mencionó consiste en este caso en dejar solo una observación como conjunto de datos de test, y considerar el resto de las observaciones como el conjunto de datos de entrenamiento, por lo que se puede apreciar de un mejor modo, que habrá tantos modelos como observaciones tenga el set de datos originales.

La estimación del error test final se la obtendría con la siguiente formula:

$$Estimación\ del\ error\ de\ test = \sum_{i=1}^{n} \frac{Error\ de\ test_i}{n} \qquad (2.7.2)$$

Donde n corresponde al tamaño total de los datos originales.

CAPÍTULO 3

3. RESULTADOS Y ANÁLISIS

En este capítulo se muestran los resultados obtenidos del análisis exploratorio de los datos, así como la aplicación de la metodología de KNN como de Random Forest, para obtener el mejor clasificador para nuestra variable de respuesta Kpi Logístico, y a su vez las distintas medidas que se tomaran para poder seleccionar el mejor clasificador y en base a los criterios en que nos hemos basado.

3.1 Análisis Exploratorio

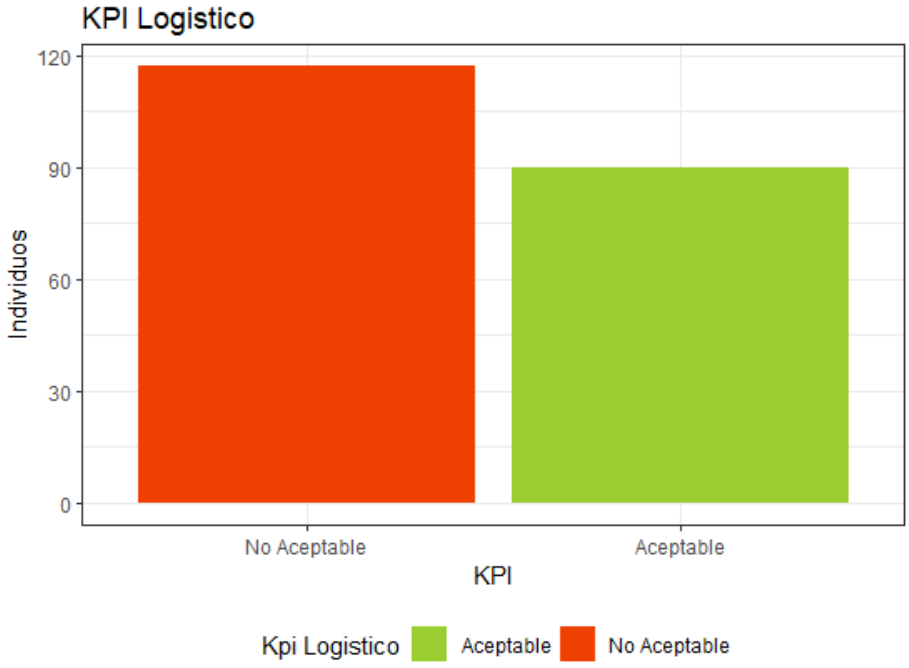


Gráfico 3.1.1: Diagrama de barras de la variable Kpi Logístico

Fuente: Elaboración propia

Se apreció de acuerdo al gráfico 3.1.1 que la cantidad de encuestadores con un Kpi Logístico considerable como “No Aceptable” no superaba por mucho a la cantidad de

encuestadores considerados como “Aceptables”, se obtuvo un pequeño vistazo de como estaría equilibrada la data, y que si tendríamos cantidad de datos representativos para poder realizar el entrenamiento con los algoritmos de machine learning.

Lo antes mencionado se logró apreciar mucho con la tabla 3.1 al obtener las proporciones de las clases de la variable, se evidenció que la clase o categoría “No aceptable” se encuentra en una proporción mayor en los datos pues esta se encuentra en el 57% de los datos, es decir que el 57% de los encuestados cayeron dentro de esta categoría, y el 44% cayeron dentro de la categoría de “Aceptable”

Tabla 3.1: Proporción de las clases de la variable Kpi Logístico

Fuente: Elaboración propia

No Aceptable	Aceptable
0.57	0.43

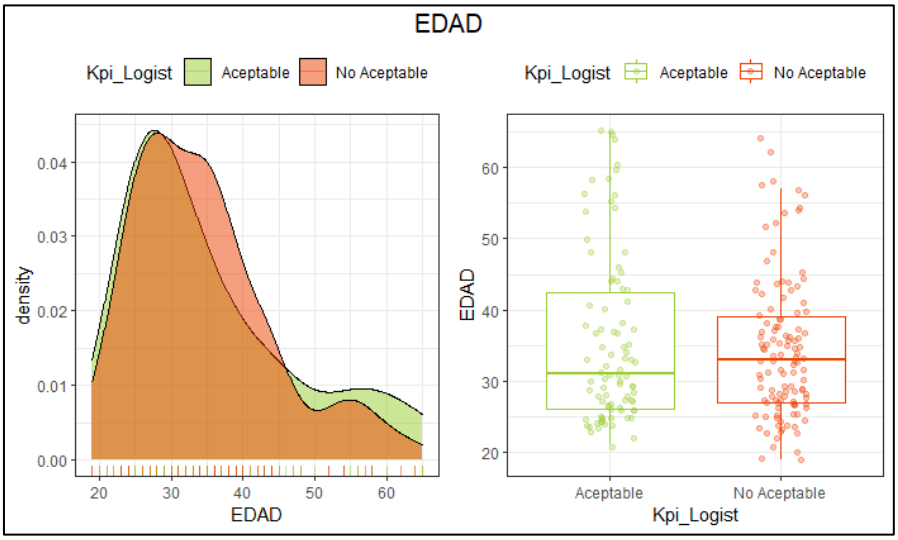
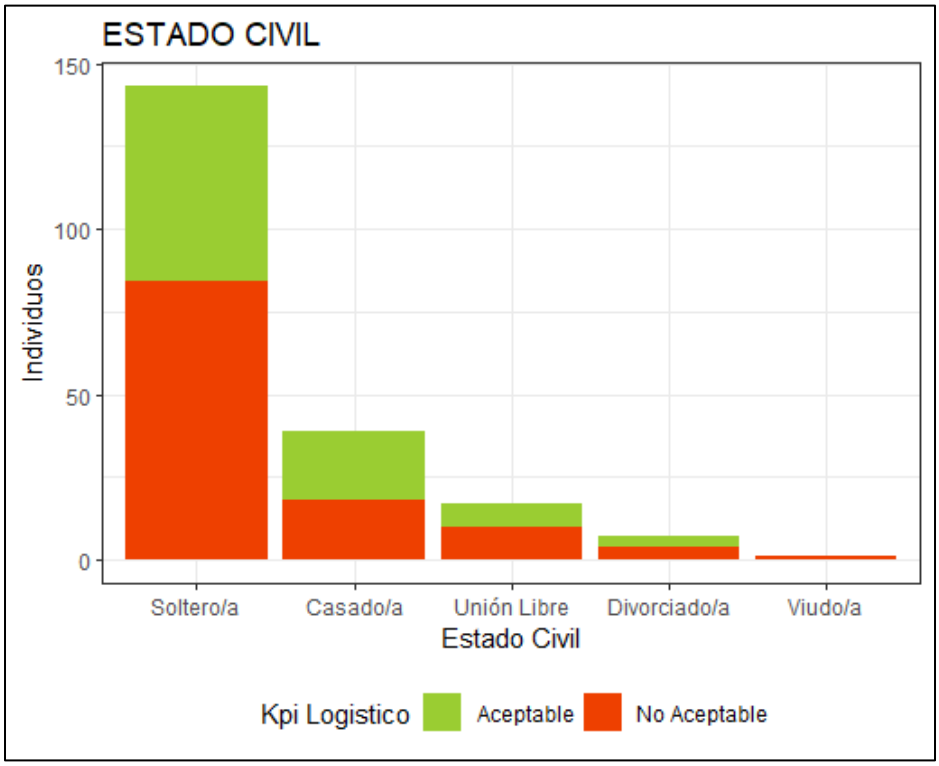


Gráfico 3.1.2: Densidad y boxplot de la variable Edad de acuerdo a las categorías de la variable Kpi Logístico

Fuente: Elaboración propia

La distribución de densidad de las categorías de la variable Kpi Logistico son similares dentro de un rango de edad,



aproximadamente hasta los 30 años se pudo apreciar que son similares, más de 30 se notó un incremento en la proporción de encuestadores que son considerados como “No Aceptables”, hasta un poco menos de los 50 años, y pasado esta edad se nota una predominancia de los encuestadores que son considerados como “Aceptable”, de manera similar se puede ver en el diagrama de cajas que existe mayor cantidad de encuestadores clasificados como “No Aceptable”, en un rango de entre 30 a 40 años, diferente a los clasificados como “Aceptable”, de esta manera se puede interpretar el Gráfico 3.1.2

Gráfico 3.1.3: Cantidad de encuestadores por Kpi Logístico y Estado Civil

Fuente: Elaboración propia

Según el gráfico 3.1.3 se observó que existe mayor cantidad de encuestadores que son solteros y que de estos existen mayor cantidad que son categorizados como “No Aceptables” que como “Aceptable”, caso contrario sucede para los que son casados que son pocos los encuestadores y de estos existe casi la misma cantidad para ambas categorías del Kpi Logístico

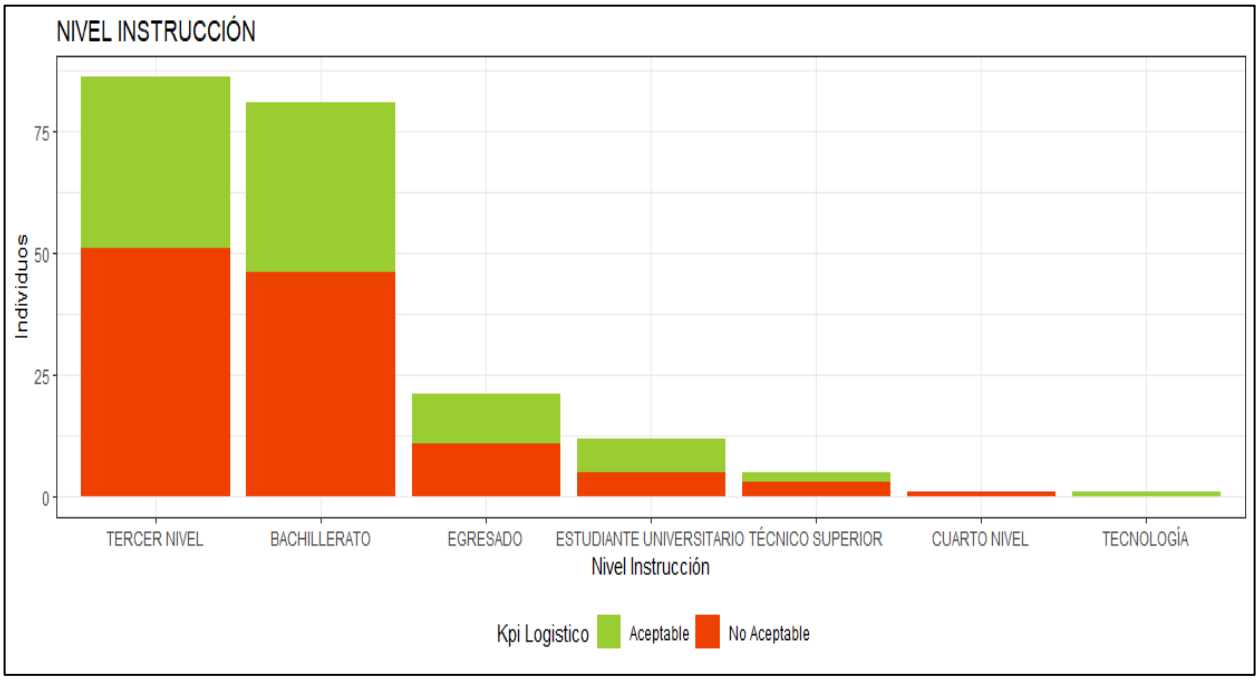


Gráfico 3.1.4 Cantidad de encuestadores por Kpi Logístico y Nivel de Instrucción

Fuente: Elaboración propia

El gráfico 3.1.4 reflejó una clara idea de la cantidad de encuestadores que posee la base por nivel de instrucción, se notó que existen mayor cantidad de bachilleres y encuestadores que poseen un tercer nivel y que en ambos existen casi la misma cantidad de encuestadores que

fueron categorizados como “No Aceptable” y como “Aceptable”, vemos el otro extremo de esto, con respecto a los que tienen un cuarto nivel de educación, que la cantidad es casi mínima.

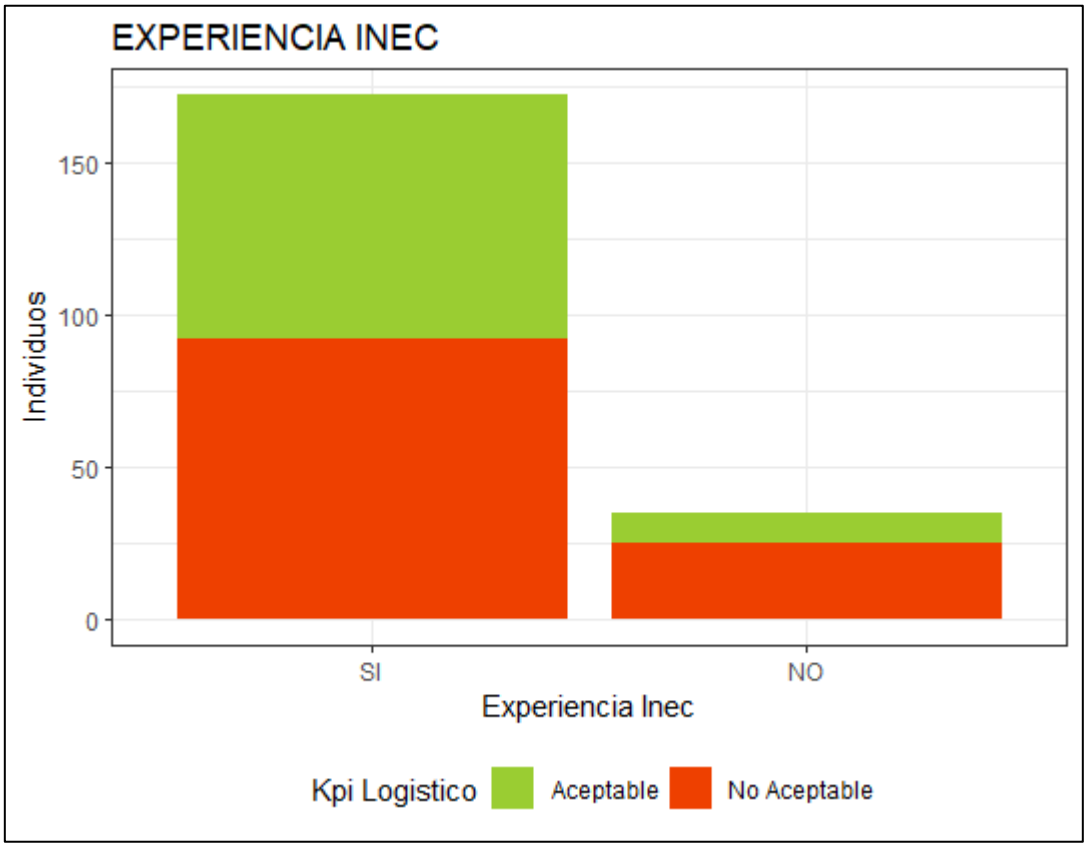


Gráfico 3.1.5: Cantidad de encuestadores por Kpi Logístico y Experiencia INEC

Fuente: Elaboración propia

En la gráfica 3.1.5 mostró que existen más cantidad de encuestadores que ya han tenido alguna experiencia en el INEC más sin embargo se tenía la hipótesis de que si algún encuestador

ya tenía experiencia en el INEC iba a tener un mejor rendimiento, pero la realidad es que existe igual cantidad de quienes fueron clasificados como “No Aceptable” así como de “Aceptable”, y caso contrario hay poca cantidad de encuestadores que no han tenido experiencia en el INEC, y se esperaba que no iban a tener un buen rendimiento, lo que se puede confirmar en el gráfico

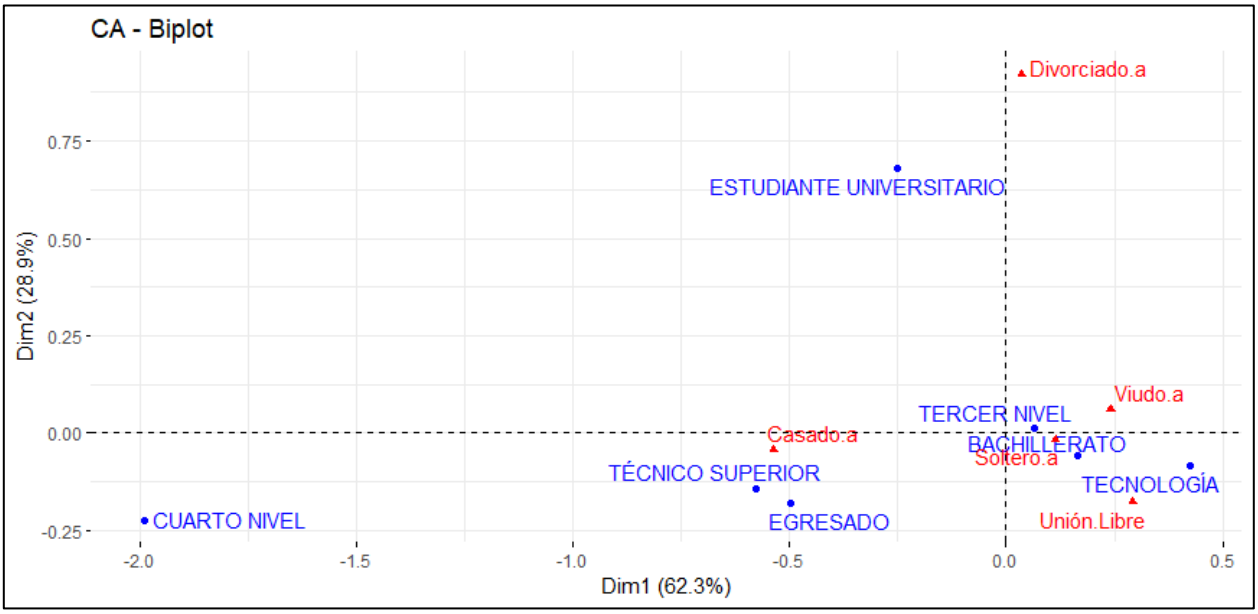


Gráfico 3.1.6: Biplot del Análisis de Correspondencia Simple entre Estado civil vs Nivel de Instrucción de acuerdo a la Nota final.

Fuente: Elaboración propia

En la gráfica 3.1.6 se puede observar la relación que existen en las variables Estado Civil y Nivel de Instrucción en base a la variable Nota final, se tomaron dichas variables puesto que para el análisis de correspondencia simple se necesitaba dos variables cualitativas que contengan un número mayor a 2 factores. Se observa cómo existe una fuerte relación entre una similar nota total obtenida de pre selección con los factores soltero/a y el Nivel de Instrucción de Tercer Nivel.

Por otro lado, en la esquina inferior izquierda se observa que existe relación entre la nota final de pre selección obtenida por individuos con estado civil casado/a con nivel de técnico superior y egresados, se puede visualizar que muy discriminados de estos grupos se encuentran los individuos con estado civil Divorciado/a, y los estudiantes universitarios con notas finales de pre selección muy distintas a los demás.

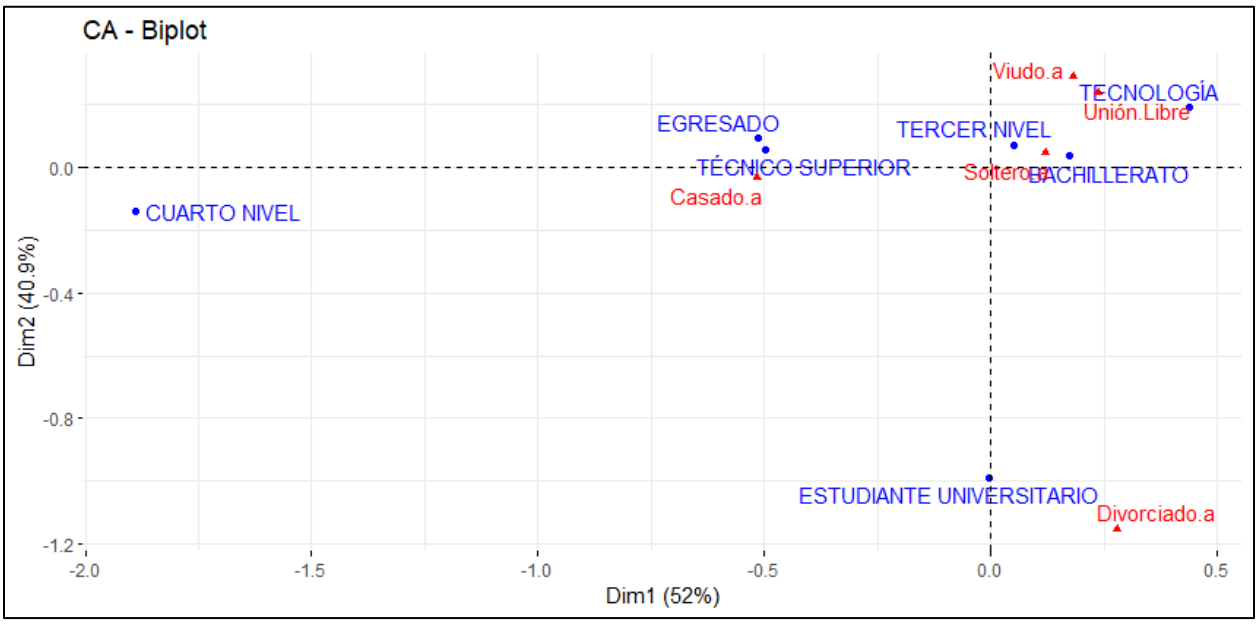


Gráfico: 3.1.7 Biplot del Análisis de Correspondencia Simple entre Estado civil vs Nivel de Instrucción de acuerdo a la Edad.

Fuente: Elaboración propia

Se observa en el gráfico 3.1.7 una similitud entre la edad de individuos con el nivel de instrucción de egresado y técnico superior con el factor casado/a, y a su vez sí se puede observar una fuerte relación entre la edad de soltero/a con el nivel de instrucción bachillerato, así mismo muy cercana a la edad de individuos de unión libre y con nivel de instrucción de tercer nivel, curiosamente el análisis de correspondencia nos indica que no existe relación entre la edad con un nivel de instrucción de estudiante universitarios y un estado civil divorciado/a.

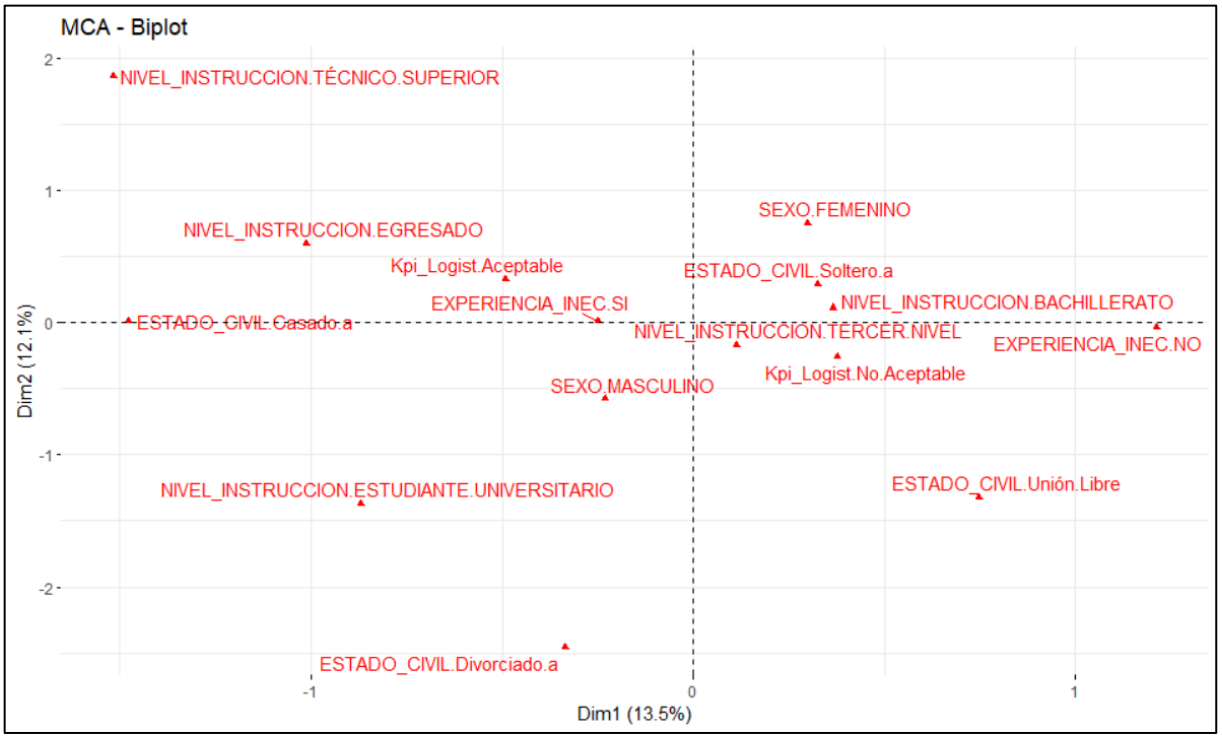


Gráfico 3.1.8: Biplot del Análisis de Correspondencia Múltiple

Fuente: Elaboración propia

En el gráfico 3.1.8 se puede observar el análisis de correspondencia múltiple realizado a todos los factores de la data, donde se observa que existe una posible relación con Kpi Logístico Aceptable con Estado Civil soltero/a, Estado Civil Casado/a, y personas que tuvieron experiencia en INEC.

A su vez, también se refleja como un Kpi Logístico No Aceptable, está posiblemente relacionado a Estado Civil Unión Libre, sexo masculino, sin experiencia en INEC, de forma

discriminante vemos a Estado Civil de Unión Libre y sexo femenino, lo que significa que posiblemente no posean una similitud cercana con los demás factores de estudio.

3.2 Variables predictoras

Para la selección de las variables predictoras a utilizar en los modelos se utilizó un método muy extendido de selección de la importancia de variables, esta estrategia consiste en cuantificar la importancia de los predictores que hacen de los modelos bagging (Random Forest), una metodología de suma potencia, no solo para predecir sino a su vez para análisis exploratorios. Las medidas que se usaron a consideración como criterio para escoger a las variables predictores más importantes fueron por permutación e impureza de nodos.

En la tabla 3.2 se pueden observar los resultados de la precisión de manera descriptiva donde se determinó, que los predictores más significativos para el modelo fueron total, edad y estado civil, a su vez si analizamos la pureza de las variables al formar nodos en las que participa el predictor podemos observar que tienen mayor influencia nota final, edad y nivel de instrucción.

Por lo cual se decidió tomar las variables que en ambas medidas han tenido mayor influencia para ser seleccionadas como las predictoras del modelo. Sin embargo, se

destaca que esta metodología indica la influencia que tienen dichos predictores con el modelo y no confundir con la relación con la variable de respuesta.

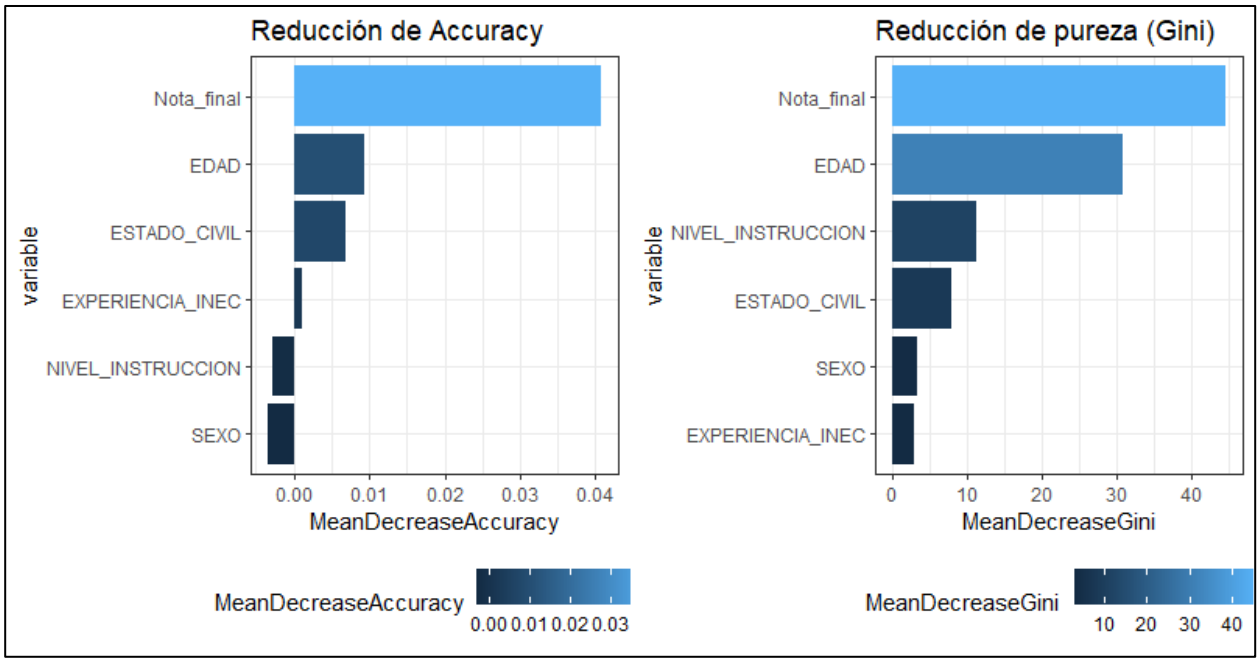


Gráfico 3.2: Importancia de los predictores.

Fuente: Elaboración propia

Tabla 3.2: Importancia de los predictores

Fuente: Elaboración propia

Predictor	Mean Decrease Accuracy	Mean Decrease Gini
Total	0,044	44,876
Edad	0,010	30,720
Sexo	-0,002	3,706
Estado Civil	0,006	7,951
Nivel Instrucción	-0,003	9,985
Experiencia Inec	0,001	2,806

3.3 Aplicación del método Random Forest

Para la aplicación del método Random Forest se usó la partición de toda la data dando un 80% para entrenar los datos y un 20% para testear los resultados.

Gracias al resultado de la selección de los predictores más importantes el conjunto de datos de entrenamiento contó con los atributos Nota final, Edad y Estado civil y cuya etiqueta de salida fue el Kpi Logístico, el cual ya se explicó su respectivo cálculo en el capítulo de Metodología en la sección de variables.

Tabla 3.3.1: Cantidad de observaciones resultantes por la división en datos de Train y de Test para el método Random Forest

Fuente: Elaboración propia

Train	Test
80%	20%
165	40

Se puede observar de la tabla 3.3.1 que al tomar el 80% de los datos para el conjunto de datos de entrenamiento, este tendrá un total de 165 observaciones con las cuales el modelo podrá aprender, por otro lado, con el 20% de los datos que se tomaron para que el modelo pueda poner a prueba su capacidad de predicción, es decir que el conjunto de datos de test se conformó por un total de 40 observaciones

La tabla 3.3.2 nos muestra los resultados de la hiperparametrización obtenida para el método de Random Forest, para la obtención de estos resultados se realizó el uso de hiperparámetros, los cuales fueron seleccionados mediante el resultado del método de K-Fold Cross Validation, los parámetros seleccionados fueron 10 particiones y 5 repeticiones, los

resultados se pueden observar en el gráfico 3.3.2, donde selecciona los mejores hiperparámetros.

Los hiperparámetros seleccionados donde obtuvo la mejor precisión mediante el método de K-folds Cross validation es 2 como números de predictores como ramificaciones y el tamaño mínimo de nodos 2.

Tabla 3.3.2: Resultados método Random Forest

Fuente: Elaboración propia

Predictores	Nodo Mínimo	Accuracy	Kappa
1	2	0,564	0,014
1	3	0,564	0,016
1	5	0,565	0,014
1	10	0,563	0,010
1	15	0,572	0,035
1	20	0,569	0,026
1	30	0,565	0,014
2	2	0,577	0,101
2	3	0,575	0,098
2	5	0,573	0,093
2	10	0,571	0,092
2	15	0,577	0,106
2	20	0,572	0,094
2	30	0,565	0,076
3	2	0,556	0,088
3	3	0,565	0,104
3	5	0,571	0,115
3	10	0,570	0,117
3	15	0,571	0,120
3	20	0,559	0,094
3	30	0,535	0,041

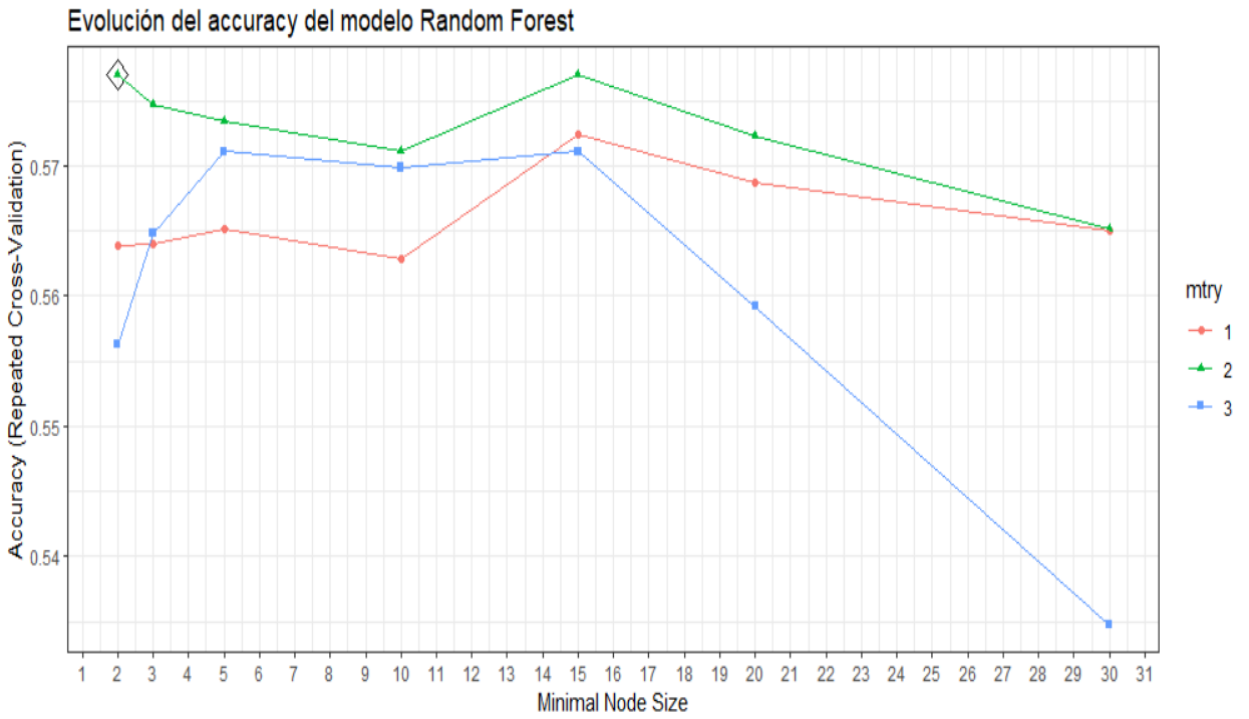


Gráfico 3.3.2: Resultado de los hiperparámetros seleccionados por el método K-Fold Cross Validation para el modelo Random Forest

3.4 Aplicación del método KNN

Para la aplicación del método KNN se usó la partición de toda la data dando un 80% para entrenar nuestros datos y un 20% para testear los resultados.

Tabla 3.4.1: Cantidad de observaciones resultantes por la división en datos de Train y de Test para el método KNN

Fuente: Elaboración propia

Train	Test
80%	20%
165	40

Al igual que para el método de Random Forest se tomó el mismo porcentaje de división tanto para los datos que formaron parte del conjunto de datos de entrenamiento, como para el conjunto de test, tal como se muestra en la tabla 3.4.1.

La tabla 3.4.2 nos muestra los resultados de la hiperparametrización obtenida para el método de KNN, para la obtención de estos resultados se realizó el uso de hiperparámetros, los cuales fueron seleccionados mediante el resultado del método de LOOCV.

El hiperparámetro donde se obtuvo la mejor precisión mediante el método de K-Folds Cross Validation es k=2, el cual indica que se obtiene mejor precisión utilizando dos como número de observaciones vecinas empleadas.

Tabla 3.4.2: Hiperparámetros considerados para el modelo de KNN, mediante la técnica K-Fold Cross Validation.

Fuente: Elaboración propia

K	Accuracy	Kappa
1	0,552	0,091
2	0,624	0,224
3	0,600	0,179

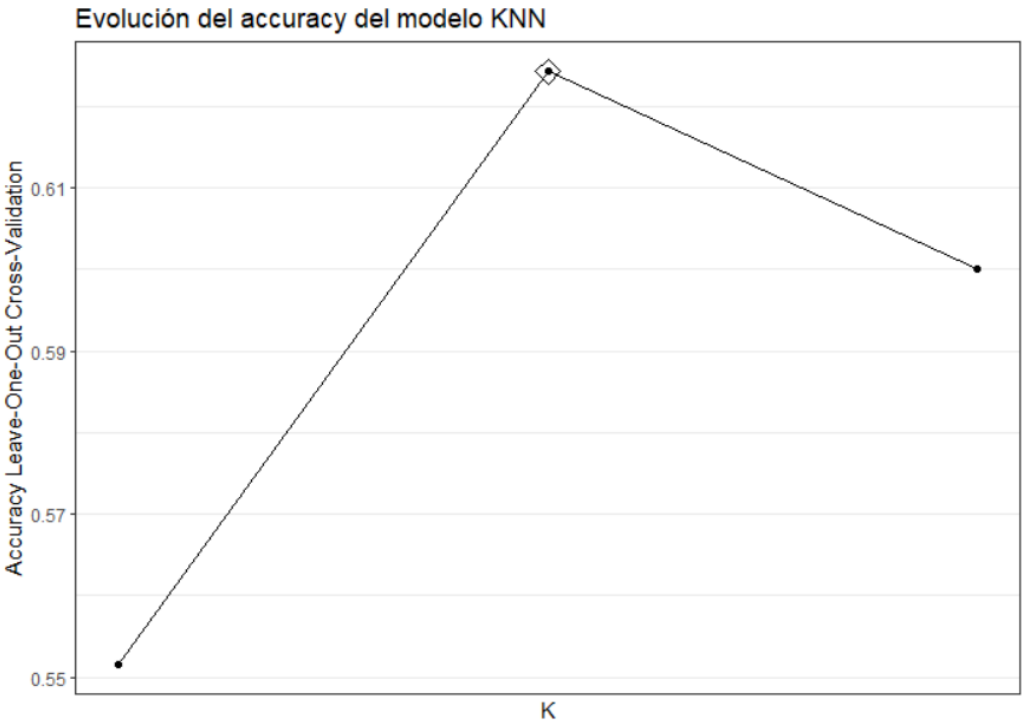


Gráfico 3.4.1: Resultado de los hiperparámetros seleccionados por el método Leave One-Out Cross Validation para el método KNN

3.5 Comparación método Random Forest y KNN

Las tablas siguientes muestran los resultados obtenidos en ambos modelos (Random Forest – KNN) con sus debidas medidas que hemos tomado a consideración para la selección del mejor modelo clasificador para nuestra data, estos resultados son directamente en nuestra data de prueba.

Tabla 3.5.1: Matriz de confusión para el método Random Forest

Fuente: Elaboración Propia

Random Forest

	Predicción	
Real	Aceptable	No Aceptable
Aceptable	9	8
No Aceptable	6	17

Tabla 3.5.2: Matriz de confusión para el método KNN

Fuente: Elaboración Propia

KNN		
	Predicción	
Real	Aceptable	No Aceptable
Aceptable	15	2
No Aceptable	5	18

Tabla 3.5.3: Matriz comparativa de Precisión (Accuracy) entre los métodos Random Forest y KNN

Fuente: Elaboración propia

	General	
	Random Forest	**KNN**
Accuracy	0,65	0,83
Accuracy Lower	0,48	0,67
Accuracy Upper	0,79	0,93

Tabla 3.5.4: Matriz comparación de métricas de evaluación de los modelos Random Forest y KNN

Fuente: Elaboración propia

	Por Clase	
	Random Forest	**KNN**
Sensitivity	0,53	0,88
Specificity	0,74	0,78
False Negative	0,57	0,12
False Positive	0,26	0,22

3.6 Indicadores para seleccionar el mejor modelo

3.6.1 Resultados Generales

Como pudimos observar en las tablas anteriores tenemos distintas técnicas para medir cual modelo está clasificando mejor, la Accuracy (precisión) que tuvo nuestro modelo de Random Forest fue 65% mientras que el modelo KNN fue de 83%.

A su vez si los analizamos por intervalos de confianza podemos denotar que con el método de Random Forest se obtuvo un intervalo de precisión entre (0.48,0.79) y para el método de KNN se obtuvo un intervalo de confianza de precisión de (0.67,0.93).

3.6.2 Resultados Random Forest

Sin embargo, para tomar una mejor decisión analizaremos la clasificación que obtuvo el modelo de Random Forest en primer lugar. En donde tuvo una sensibilidad del 53% eso significa que clasificó correctamente a los individuos con Kpi Aceptable como Aceptable, a su vez tuvo una especificidad del 74% lo que indica que clasificó correctamente a los individuos con Kpi No aceptable como No aceptable.

A su vez podemos observar que clasificó de formar errónea a individuos que eran aceptables como no aceptables en un 57% conocido como falsos negativos. Y se obtuvo falsos positivos en un 26% lo que significa que siendo individuos con un Kpi No aceptable los clasifico como aceptable.

3.6.3 Resultados de KNN

Al realizar el mismo análisis en el modelo KNN se obtuvo una sensibilidad del 88% eso significa que clasificó correctamente ese porcentaje a los individuos con Kpi Aceptable como Aceptable, a su vez tuvo una especificidad del 78% lo que indica que clasificó correctamente a los individuos con Kpi No aceptable como No aceptable.

A su vez podemos observar que clasificó de formar errónea a individuos que eran aceptables como no aceptables en un 12% conocido como falsos negativos. Y se obtuvo falsos

positivos en un 22% lo que significa que siendo individuos con un Kpi No aceptable los clasificó como aceptable.

CAPÍTULO 4

4. CONCLUSIONES Y RECOMENDACIONES

En este capítulo se presentan las conclusiones y recomendaciones respecto a los resultados obtenidos en este trabajo, en donde se indicará el mejor modelo seleccionado y el porqué de la selección, el cual tiene el fin de ser un parámetro más a tomar a consideración por el área de recursos humanos de la institución para la selección de nuevo personal de encuestadores.

4.1 Conclusiones

- El modelo de Random Forest para nuestro estudio nos dio un resultado muy bajo en precisión general y en métricas específicas, obtuvimos resultados poco favorables para la elección ya que clasificaba de forma errónea, dando muchos falsos positivos como falsos negativos, por lo cual quedo descartado.

- El modelo de KNN posee una excelente sensibilidad como especificidad puesto que se obtuvieron porcentajes de 88% y 78% respectivamente, demostrando que el modelo clasifica mejor cuando un individuo tiene características que probablemente lo clasifican con un Kpi Logístico aceptable como encuestador.

- El modelo de KNN obtuvo un excelente porcentaje en una métrica a nuestra consideración muy importante puesto que al tener un porcentaje del 12% de falsos

negativos se considera que probablemente el modelo no clasificará de forma errónea a un candidato con características asociadas a tener un buen Kpi como encuestador.

- El modelo de KNN se considera mejor clasificador cuando existen mayor cantidad de variables cuantitativas como predictores, puesto que al tratarse de hallar la mayor similitud que existe en las características de los individuos tanto como edad y su calificación previa al ingreso se obtiene una precisión aceptable.

- El modelo de KNN es el modelo seleccionado puesto que cumple con lo deseado ya que tanto en términos generales al obtener una precisión del 82% y en específicos tuvo excelentes porcentajes.

- Mediante el análisis de correspondencia múltiple se pudo hallar una fuerte similitud en el comportamiento de factores como Estado Civil casado/a, soltero/a e individuos con experiencia en el INEC, con el factor de respuesta Kpi Logístico como Aceptable, y de forma adversa existe una fuerte atracción entre factores sexo masculino, estado civil Unión Libre y no experiencia en el INEC, con un factor de Kpi Logístico como No Aceptable.
- Mediante el análisis de correspondencia simple se pudo denotar que existe similitud en la nota final de preselección entre individuos con los factores estado civil soltero/a, y tercer nivel de instrucción, a su vez también se pudo notar que existe relación en la nota obtenida de preselección en individuos con factores de estado civil casado/a y con un nivel de instrucción técnico superior o egresado.
- Al analizar la relación mediante análisis de correspondencia simple de la variable edad, pudimos notar que existe relación de edades entre individuos con nivel de instrucción

de egresado o técnico superior con el factor casado/a. Como dato curioso se pudo notar que al realizar el análisis de correspondencia existe similitud en los datos, entre las edades de estado civil divorciado con estudiantes universitarios.

- La implementación de un dashboard generó de manera más interactiva y eficiente la forma de transmitir los estudios realizados en el presente proyecto, consiguiendo una mejor interacción por parte del usuario y un mejor detalle de los descubrimientos realizados.

4.2 Recomendaciones

- Comparar los resultados con estudios previamente hechos, da una noción más amplia de los modelos que se pueden utilizar y como llevar una correcta directriz para un estudio.
- Cuando se desea realizar un análisis de calidad se debe considerar realizar un previo e intenso conocimiento a la base de datos para así poder realizar una correcta manipulación y análisis de dichos datos.
- Tomar en cuenta y consultar a expertos sobre métricas que se manejen para así poder realizar una correcta creación y uso de nuevas variables que logran ampliar nuestro análisis.
- Siempre buscar y preguntar sobre cualquier duda que se presente con la base de datos puesto que dichas respuestas pueden cambiar de forma positiva la directriz de un estudio y hallar soluciones más beneficiosas.

BIBLIOGRAFÍA

Aggarwal, S. (15 de junio de 2020). *What is Cross-Validation?* Obtenido de towards data science: https://towardsdatascience.com/what-is-cross-validation-622d5a962231

Amat Rodrigo, J. (Octubre de 2020). *Árboles de decisión, random forest, gradient boosting y C5;0*. Obtenido de cienciadedatos.net: https://www.cienciadedatos.net/documentos/33_arboles_de_prediccion_bagging_random_forest_boosting#Introducci%C3%B3n

Castillo, C., & Gustavo, E. (2021). Machine learning en la mejora del proceso de selección del personal administrativode la Corte Superior de Justicia de Lima, 2020. *(Tesis de maestría).* Universidad César Vallejo, Lima.

CEPAL. (18 de Diciembre de 2020). *Qué es la anonimización?* Obtenido de Comisión Económica para América Latina y el Caribe: https://biblioguias.cepal.org/c.php?g=495473&p=4961125

Chiavenato, I. (2011). *Administración de recursos humanos.El capital humano de las organizaciones.* México: Mc Graw Hill.

Cortés Cortés, M., & Iglesias León, M. (2004). *Generalidades sobre la Metodología de la Investigación.* México: Universidad Autónoma del Carmen.

DATOS, A. E. (2016). *Orientaciones y garantías en los procedilientos de ANONIMIZACIÓN de datos personales.*

González Ortiz , M. (28 de Abril de 2021). *Qué son los KPI y por qué son importantes en una empresa IT*. Obtenido de OpenWebinars: https://openwebinars.net/blog/que-son-los-kpi-y-por-que-son-importantes-en-una-empresa-it/

INEC. (2013). *Ecuador - Encuesta Nacional de Empleo, Desempleo y Subempleo - Diciembre 2011, RONDA XXXIV-12-2011*. Obtenido de Instituto nacional de estadísticas y censos: https://anda.inec.gob.ec/anda/index.php/catalog/269/datafile/F2/V190

INEC. (2021). *ACTUALIZACIÓN CARTOGRÁFICA Y PRECENSO.*

Johnson, K., & Kuhn, M. (2013). *Applied Predictive Modeling* (Vol. 26). New York: Springer.

Martínez Noriegas, H. A., Medrano Broche, B. E., Fernández Capestany, L., & Tejeda Rodríguez, Y. E. (2013). Análisis Multivariado de datos como soporte a la decisíon en la selección de estudiantes en prpoyectos de software. *Ingeniería Industrial*, 130-142.

Morales, F. C. (16 de Julio de 2020). *Productividad laboral*. Obtenido de Economipedia.com: https://economipedia.com/definiciones/productividad-laboral.html

Penguin, W. (17 de Octubre de 2018). *Qué es un KPI: Significado y ejemplos de Key Performance Indicators*. Obtenido de yoseoMarketing: https://www.yoseomarketing.com/blog/kpi-significado-ejemplos/#kpi,-%C2%BFque-es

Raschka, S., & Mirjalili, V. (2020). *Python machine learning. Marcombo.*

VanderPlas, J. (2016). *Python data science handbook: Essential tools for working with data. .* O'Reilly Media, Inc.

APÉNDICES

APÉNDICE A

Gráfico 4.1: Imagen 1 Dashboard

Fuente: Elaboración propia

Gráfico 4.2: Imagen 2 Dashboard