**Logotipo

Descripción generada automáticamente**

**Predicción de resultados de pruebas saber del 2024-1 en base a los resultados históricos de las pruebas, usando la información del estudiante, colegio y familia**

Juan Pablo Jaramillo Tobon

Daniel Alejandro Higuita Usuga

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor  
David Manuel Villanueva Valdes, Ingeniero de Datos

Universidad de Antioquia  
Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2024

| **Cita** | (Jaramillo Tobon & Higuita Usuga, 2024) |
| --- | --- |
| **Referencia**  **Estilo APA 7 (2020)** | Jaramillo Tobon, J. P. Higuita Usuga, D. A. (2024). *Predicción de resultados de pruebas saber del 2024-1 en base a los resultados históricos de las pruebas, usando la información del estudiante, colegio y familia*. Universidad de Antioquia, Medellín, Colombia. |

**** 

Especialización en Analítica y Ciencia de Datos, CohorteV.

Centro de Investigación Ambientales y de Ingeniería (CIA).

|  | Diagrama  Descripción generada automáticamente con confianza media |
| --- | --- |

Centro de Documentación Ingeniería (CENDOI)

**Repositorio Institucional:** http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.

Decano: Julio Cesar Saldarriaga Molina

Jefe departamento: Diego José Luis Botia Valderrama

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

**Dedicatoria**

A nuestras familias, sin su apoyo esto no sería posible

**Agradecimientos**

Al docente David Manuel Villanueva por su guía durante el desarrollo del trabajo

**Tabla de contenido**

[Resumen 8](#_heading=h.3rdcrjn)

[Abstract 9](#_heading=h.26in1rg)

[1. Descripción del problema 10](#_heading=h.35nkun2)

[1.1. Problema de negocio 10](#_heading=h.1ksv4uv)

[1.2. Aproximación desde la analítica de datos 10](#_heading=h.44sinio)

[1.3. Origen de los datos 10](#_heading=h.z337ya)

[1.4. Métricas de desempeño 11](#_heading=h.3j2qqm3)

[2. Objetivos 12](#_heading=h.4i7ojhp)

[2.1. Objetivo general 12](#_heading=h.2xcytpi)

[2.2. Objetivos específicos 12](#_heading=h.3whwml4)

[3. Datos 13](#_heading=h.2bn6wsx)

[3.1. Datos originales 13](#_heading=h.qsh70q)

[3.2. Datasets 18](#_heading=h.3as4poj)

[3.3. Analítica descriptiva 19](#_heading=h.1pxezwc)

[Referencias 23](#_heading=h.1jlao46)

**Lista de tablas**

[**Tabla 1**](#_heading=h.2u6wntf) [Campos de las pruebas Saber 11 2019-2 13](#_heading=h.9xzzu44ikn8y)

**Lista de figuras**

[**Figura 1** Codificación de la educación del padre y la madre **19**](#_heading=h.2r0uhxc)

[**Figura 2** Distribución de la educación del padre y la madre **19**](#_heading=h.2r0uhxc)

[**Figura 3** Codificación de los cuartos en el hogar **20**](#_heading=h.2r0uhxc)

[**Figura 4** Distribución de los cuartos en el hogar **20**](#_heading=h.2r0uhxc)

[**Figura 5** Codificación del género **21**](#_heading=h.2r0uhxc)

[**Figura 6** Distribución del género **21**](#_heading=h.2r0uhxc)

[**Figura 7** Codificación de si la familia tiene computador e internet **22**](#_heading=h.2r0uhxc)

[**Figura 8** Distribución de si la familia internet **22**](#_heading=h.2r0uhxc)

[**Figura 9** Distribución de si la familia computador **22**](#_heading=h.2r0uhxc)

**Siglas, acrónimos y abreviaturas**

**ML** Machine Learning

**ICFES** Instituto Colombiano para la Evaluación de la Educación

# Resumen

Las Pruebas Saber 11 son un examen estandarizado que realizan los estudiantes de último año de secundaria en Colombia, la prueba muestra el nivel académico de los colegios del país y mide lo que cada estudiante aprendió durante once años de educación. Debido a los malos resultados en las pruebas en todo el país, el Ministerio de Educación de Colombia necesita identificar a los estudiantes que podrían tener un mal desempeño en el examen para poder brindarles apoyo. Este proyecto utilizará modelos de aprendizaje automático para predecir los puntajes que un estudiante podría obtener en la prueba, antes de realizarla, utilizando datos recopilados por el icfes y publicados por el gobierno colombiano en su página de datos abiertos, el modelo podría ayudar a identificar y dar información sobre quién podría tener un mal desempeño en la prueba, lo cual puede ayudar al ministerio de educación a dirigir sus programas de apoyo a esos estudiantes y adaptarlos a sus necesidades.

*Palabras clave*: Machine Learning,Examen,Colegio, Educación, Prueba

**Repositorio Github:** <https://github.com/juanjaramillo98/Seminario/tree/main>

# Abstract

The Knowledge Exam 11 is a standardized exam which students in their last year of high school take in Colombia, the test shows the academic level of the country's schools and measures what each student learned during eleven years of education. Due to poor performances in the tests around the country, the ministry of education in Colombia needs to identify students who might perform poorly in the exam in order to give them support. This project is going to use machine learning models to predict the scores a student could get in the test, before they take it, using data collected by the icfes and published by the colombian government on their open data page, the model could help to identify and give insights about who might perform badly on the test, this could help the ministry of education to direct their support programs to those students and adapt them to the student’s needs.

*Keywords***:** Machine Learning,Exam, High School, Education, Test

**Github Repository:** <https://github.com/juanjaramillo98/Seminario/tree/main>

# Descripción del problema

## Problema de negocio

La media de los resultados de las pruebas saber 11 han empeorado y hay una preocupación a nivel nacional sobre la calidad de la educación que reciben los estudiantes [1]. El ministerio de educación quiere mejorar resultados de las pruebas saber 11 mediante la toma de iniciativa con apoyos escolares. Se utilizaran datos de resultados anteriores de las pruebas para entrenar modelos que permitan predecir el resultado de la prueba. El ministerio de educación cuenta con los resultados históricos de las pruebas saber 11 de años anteriores con la información de cada estudiante, su familia y colegio. Las métricas podrían ser la precisión del modelo a la hora de predecir la puntuación de estudiantes que no fueron beneficiados con los apoyos escolares y el aumento en los puntajes reales vs las predicciones en los grupos de estudiantes los cuales se vieron beneficiados por los apoyos escolares.

## Aproximación desde la analítica de datos

Para identificar estudiantes que puedan tener resultados deficientes se implementará un modelo de regresión que permita predecir el resultado de las pruebas saber 11 de cada estudiante. El modelo se entrenará con los datos del estudiante, colegio y su familia de previas ediciones de las pruebas saber 11. El modelo tendrá la capacidad de predecir los resultados de las próximas pruebas, y con dicha información el ministerio podrá agrupar a todos los estudiantes que puedan llegar a tener un resultado deficiente y dentro de este grupo identificar similitudes entre ellos para mejorar dichas problemáticas. El ministerio también podría utilizar las predicciones para identificar conjuntos de estudiantes dentro de la misma geografía y así poder optimizar los apoyos escolares dándole mejor efectividad y eficiencia.

## Origen de los datos

Los datos provienen de los resultados históricos de las pruebas saber 11 que el Icfes ha recolectado en cada edición de las pruebas, estos datos comprenden información específica del estudiante, su colegio y la situación socioeconómica de su hogar. Los datos se encuentran consolidados anonimizados y disponibles en la plataforma de datos abiertos del gobierno [2]

## Métricas de desempeño

Para calificar el modelo se tendrán en cuenta 2 métricas las cuales son R2 y R2 ajustado las cuales nos permitirán identificar que tanto describe nuestro modelo la variabilidad en los datos, además de ser métricas simples de fácil interpretación. Estas métricas se realizarán bajo el método de cross validation dado que nuestro modelo es bastante complejo y podríamos presentar sobre ajuste, esta también es la razón por la cual no solo tendremos R2 dado que el R2 ajustado tiene en cuenta la cantidad de características a la hora de evaluar el modelo, se pretende esperar valores de R2 y R2 ajustado mayores a 0.6, dado que se realizarán predicciones en un ámbito social y humano el cual no puede pretender precisiones perfectas, pero al menos valores que nos permitan tener confianza a la hora de tomar decisiones. Para poder evaluar los resultados de las tomas de decisiones para mejorar los resultados, se pretende comparar los resultados ponderados de los grupos de estudiantes que recibieron los apoyos y las predicciones ponderadas de esos mismos grupos de estudiantes, así podemos minimizar el impacto de estudiantes atípicos y analizar si los apoyos tuvieron efectos significativos.

# Objetivos

## Objetivo general

Desarrollar e implementar un modelo de regresión con la capacidad de predecir los futuros resultados de las pruebas saber 11 de cada estudiante, que permita identificar estudiantes que puedan tener bajo rendimiento en las pruebas para ayudar a la toma de decisiones acerca de la implementación de planes de apoyo escolar

## Objetivos específicos

* Realizar un análisis exploratorio de los datos para reducir la dimensionalidad del dataset y dejar caracteriticas que se relacionen con el puntaje del estudiante al menos en un 70%
* Diseñar un método de codificación de variables categóricas ordinales
* Codificar los datos categóricos de la fuente de tal forma que se puedan implementar métodos que evalúan la correlación con la variable dependiente
* Identificar el modelo de regresión que entregue las mejores métricas para nuestro conjunto de datos
* Identificar los hiperparametros adecuados que optimicen las métricas del modelo
* Evaluar el rendimiento de la arquitectura final del modelo mediante el método de cross validation y con las métricas de R2 y R2 ajustado
* Entrenar el modelo final obteniendo valores superiores a 0.6 en las métricas de R2 y R2 ajustado

# Datos

## Datos originales

Los datos se recolectaron de la página de datos abiertos del gobierno Colombiano, los datos incluyen: 546212 registros y 82 campos, con los resultados anonimizados de las pruebas Saber 11 del segundo semestre del año 2019, además tienen la información del estudiante, su familia y colegio. Los datos se encuentran en formato CSV y con un tamaño de 396 MB.

Las columnas se encuentran nombradas de tal manera que es fácil de identificar 6 grupos de columnas:

* ESTU: Datos del estudiante
* FAMI: Datos de la familia
* COLE: Datos de colegio
* PUNT: Resultados de la prueba general y por asignatura
* PERCENTIL: Percentiles de los resultados de la prueba general y por asignatura
* DESEMP: Desempeño general y por asignatura

La siguiente tabla muestra los campos, sus descripciones y tipo de dato:

#### **Tabla 1**

#### *Campos de las pruebas Saber 11 2019-2*

| **Nombre del Campo** | **Descripción** | **Tipo del Dato** |
| --- | --- | --- |
| ESTU\_TIPODOCUMENTO | Tipo de documento del estudiante | Texto |
| ESTU\_NACIONALIDAD | Nacionalidad del estudiante | Texto |
| ESTU\_GENERO | Género del estudiante | Texto |
| ESTU\_FECHANACIMIENTO | Fecha de nacimiento del estudiante | Timestamp |
| PERIODO | Periodo de las pruebas | Texto |
| ESTU\_CONSECUTIVO | Código único para identificar al estudiante | Texto |
| ESTU\_ESTUDIANTE | Campo que identifica el registro como perteneciente a un estudiante | Texto |
| ESTU\_TIENEETNIA | Campo que muestra si el estudiante pertenece a alguna etnia | Texto |
| ESTU\_PAIS\_RESIDE | País donde vive el estudiante | Texto |
| ESTU\_ETNIA | Etnia del estudiante | Texto |
| ESTU\_DEPTO\_RESIDE | Departamento donde vive el estudiante | Texto |
| ESTU\_COD\_RESIDE\_DEPTO | Código del departamento donde vive el estudiante | Texto |
| ESTU\_MCPIO\_RESIDE | Municipio donde vive el estudiante | Texto |
| ESTU\_COD\_RESIDE\_MCPIO | Código del municipio donde vive el estudiante | Texto |
| FAMI\_ESTRATOVIVIENDA | Estrato de la vivienda de la familia | Texto |
| FAMI\_PERSONASHOGAR | Personas que viven en el hogar | Texto |
| FAMI\_CUARTOSHOGAR | Habitaciones en el hogar | Texto |
| FAMI\_EDUCACIONPADRE | Nivel de educación del padre | Texto |
| FAMI\_EDUCACIONMADRE | Nivel de educación de la madre | Texto |
| FAMI\_TRABAJOLABORPADRE | Trabajo del padre | Texto |
| FAMI\_TRABAJOLABORMADRE | Trabajo de la madre | Texto |
| FAMI\_TIENEINTERNET | La familia tiene internet | Texto |
| FAMI\_TIENESERVICIOTV | La familia tiene televisor | Texto |
| FAMI\_TIENECOMPUTADOR | La familia tiene computador | Texto |
| FAMI\_TIENELAVADORA | La familia tiene lavadora | Texto |
| FAMI\_TIENEHORNOMICROOGAS | La familia tiene horno microondas o gas | Texto |
| FAMI\_TIENEAUTOMOVIL | La familia tiene automóvil | Texto |
| FAMI\_TIENEMOTOCICLETA | La familia tiene motocicleta | Texto |
| FAMI\_TIENECONSOLAVIDEOJUEGOS | La familia tiene consola de videojuegos | Texto |
| FAMI\_NUMLIBROS | Cuantos libros tiene la familia | Texto |
| FAMI\_COMELECHEDERIVADOS | En la familia se comen alimentos con derivados de leche | Texto |
| FAMI\_COMECARNEPESCADOHUEVO | En la familia se consume pescado o huevo | Texto |
| FAMI\_COMECEREALFRUTOSLEGUMBRE | En la familia se consumen cereal, frutas o legumbres | Texto |
| FAMI\_SITUACIONECONOMICA | Situación económica de la familia | Texto |
| ESTU\_DEDICACIONLECTURADIARIA | Tiempo que le dedica el estudiante a leer | Texto |
| ESTU\_DEDICACIONINTERNET | Tiempo que le dedica el estudiante al internet | Texto |
| ESTU\_HORASSEMANATRABAJA | Horas que trabaja el estudiante a la semana | Texto |
| ESTU\_TIPOREMUNERACION | Tipo de pago que recibe el estudiante | Texto |
| COLE\_CODIGO\_ICFES | Identificación que le da el Icfes al colegio | Texto |
| COLE\_COD\_DANE\_ESTABLECIMIENTO | Identificación que le da el Dane al colegio | Texto |
| COLE\_NOMBRE\_ESTABLECIMIENTO | Nombre del establecimiento | Texto |
| COLE\_GENERO | Género de los estudiantes en el colegio | Texto |
| COLE\_NATURALEZA | Naturaleza del colegio | Texto |
| COLE\_CALENDARIO | Tipo de calendario del colegio | Texto |
| COLE\_BILINGUE | El colegio es bilingüe | Texto |
| COLE\_CARACTER | Carácter del colegio | Texto |
| COLE\_COD\_DANE\_SEDE | Identificación que le da el Dane a la sede del colegio | Texto |
| COLE\_NOMBRE\_SEDE | Nombre de la sede del colegio | Texto |
| COLE\_SEDE\_PRINCIPAL | Es la sede principal | Texto |
| COLE\_AREA\_UBICACION | Ubicación del colegio | Texto |
| COLE\_JORNADA | Jornada del colegio | Texto |
| COLE\_COD\_MCPIO\_UBICACION | Código del municipio donde está el colegio | Texto |
| COLE\_MCPIO\_UBICACION | Nombre del municipio donde está el colegio | Texto |
| COLE\_COD\_DEPTO\_UBICACION | Código del departamento donde está el colegio | Texto |
| COLE\_DEPTO\_UBICACION | Nombre del departamento donde está el colegio | Texto |
| ESTU\_PRIVADO\_LIBERTAD | Está privado de la libertad | Texto |
| ESTU\_COD\_MCPIO\_PRESENTACION | Código del municipio donde presentó el icfes | Texto |
| ESTU\_MCPIO\_PRESENTACION | Nombre del municipio donde presentó el icfes | Texto |
| ESTU\_DEPTO\_PRESENTACION | Nombre del departamento donde presentó el icfes | Texto |
| ESTU\_COD\_DEPTO\_PRESENTACION | Código del departamento donde presentó el icfes | Texto |
| PUNT\_LECTURA\_CRITICA | Puntaje de lectura crítica | Numérico |
| PERCENTIL\_LECTURA\_CRITICA | Percentil en lectura crítica | Numérico |
| DESEMP\_LECTURA\_CRITICA | Desempeño en lectura crítica | Numérico |
| PUNT\_MATEMATICAS | Puntaje en matemáticas | Numérico |
| PERCENTIL\_MATEMATICAS | Percentil en matemáticas | Numérico |
| DESEMP\_MATEMATICAS | Desempeño en matemáticas | Numérico |
| PUNT\_C\_NATURALES | Puntaje de ciencias naturales | Numérico |
| PERCENTIL\_C\_NATURALES | Percentil en ciencias naturales | Numérico |
| DESEMP\_C\_NATURALES | Desempeño en ciencias naturales | Numérico |
| PUNT\_SOCIALES\_CIUDADANAS | Puntaje en sociales | Numérico |
| PERCENTIL\_SOCIALES\_CIUDADANAS | Percentil en sociales | Numérico |
| DESEMP\_SOCIALES\_CIUDADANAS | Desempeño en sociales | Numérico |
| PUNT\_INGLES | Puntaje en inglés | Numérico |
| PERCENTIL\_INGLES | Percentil en inglés | Numérico |
| DESEMP\_INGLES | Desempeño en inglés | Texto |
| PUNT\_GLOBAL | Puntaje global | Numérico |
| PERCENTIL\_GLOBAL | Percentil global | Texto |
| ESTU\_INSE\_INDIVIDUAL | Índice Socioeconómico Individual del estudiante | Texto |
| ESTU\_NSE\_INDIVIDUAL | Nivel Socioeconómico del estudiante | Texto |
| ESTU\_NSE\_ESTABLECIMIENTO | Nivel Socioeconómico del establecimiento | Texto |
| ESTU\_ESTADOINVESTIGACION | Estado de investigación | Texto |
| ESTU\_GENERACION-E | Está en algún programa del gobierno | Texto |

Los datos pueden descargarse de la página de datos libres del gobierno colombiano[3], hay dos formas de descargas: descargar el csv completo, no hay restricciones en la descarga, o extraerlos a través de la API, la cúal es limitada a 1000 registros por request.

## Datasets

Los datos deben ser imputados primero para reemplazar el valor “-”, ese valor significa

que el campo no fue llenado, los registros que tienen ese carácter no pueden descartarse, ya que se perdería aproximadamente el 96% de los datos, por esto se emplea el método de la categoría más frecuente[4], para reemplazar esos valores.

Los datos de tipo texto deben ser codificados para que puedan ser usados por modelos de ML, para este proyecto se utilizan tresa tipos de codificaciones:

* **Ordinary Encoding:** para valores categóricos que tienen algún orden establecido.
* **One Hot encoding:** para valores categóricos no ordinales y aquellos que tengan 15 o menos valores únicos[5]
* **One Label encoding:** para valores categóricos que no tienen jerarquía.
* **Cyclic encoding**: para día y mes de nacimiento, esto con el fin de reducir la distancia que hay entre los valores de estos campos.

Además los datos deben estandarizarse ya que se encuentran en unidades diferentes y esto puede llevar a que los modelos de regresión le den más peso a unos campos que a otros [6], para el escalado se utiliza el StandardScaler de Skelarn

Finalmente se crean dos datasets:

* **Dataset de Entrenamiento**: 80% de los datos transformados para entrenar el modelo
* **Dataset de Prueba:** 20% de los datos transformados para evaluar el modelo.

## Analítica descriptiva

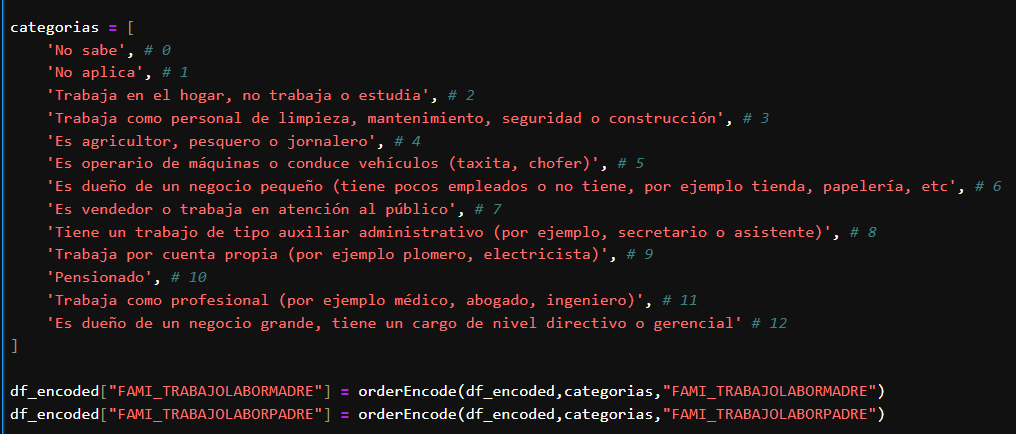
Se realizará un análisis descriptivo preliminar de los datos para comprender mejor su

distribución y características. Este análisis incluirá:

* La educación del padre y la madre
* Los cuartos del hogar
* El género del estudiante
* Si la familia tiene internet y computador

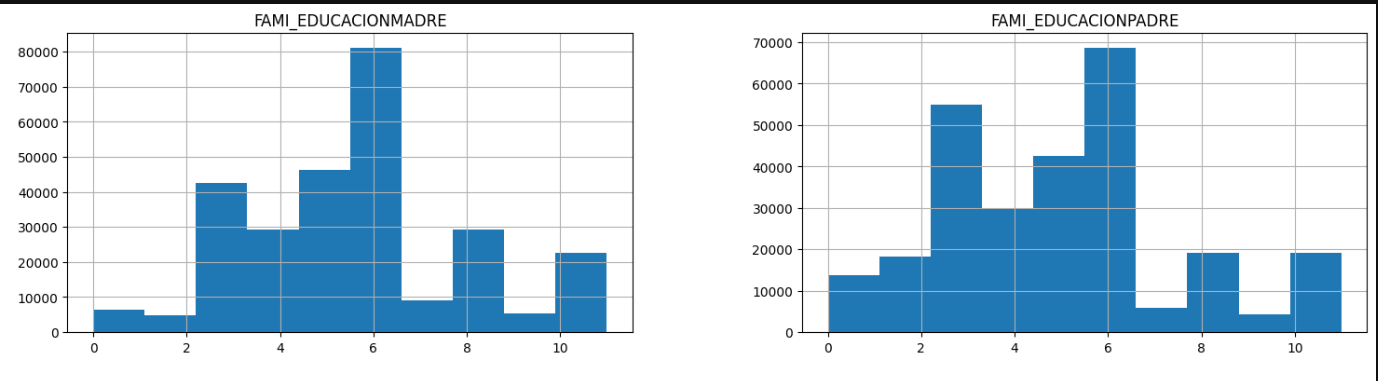
La codificación de la educación del padre y la madre se hizo por order encode y la codificación es la siguiente:

### **Figura 1** *Codificación de la educación del padre y la madre*



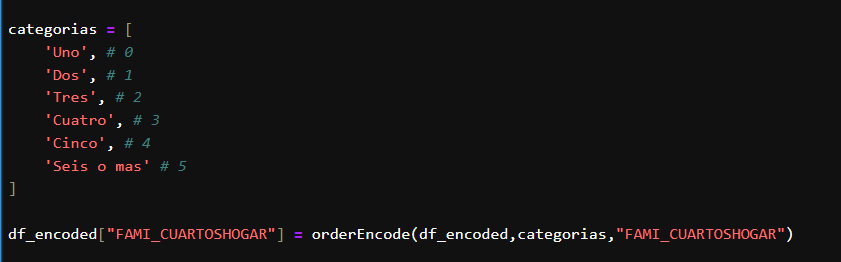
Las distribuciones de los datos se ven a continuación:

### **Figura 2** *Distribución de la educación del padre y la madre*



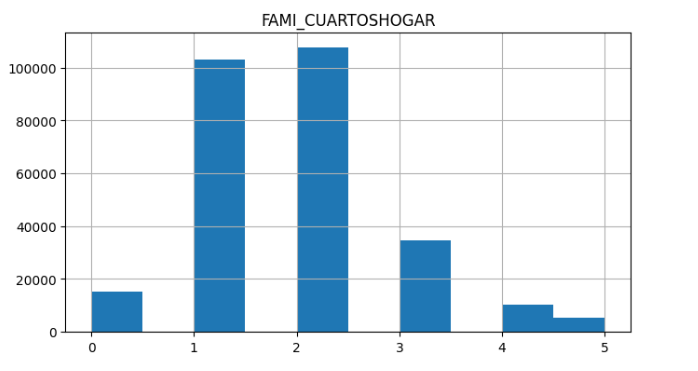
La codificación de los cuartos del hogar se hizo por order encode y la codificación es la siguiente:

### **Figura 3** *Codificación de los cuartos en el hogar*



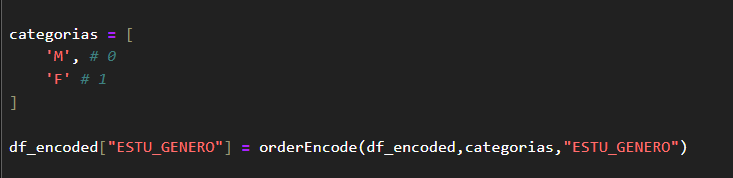
La distribución de los datos se ve a continuación:

### **Figura 4** *Distribución de los cuartos en el hogar*



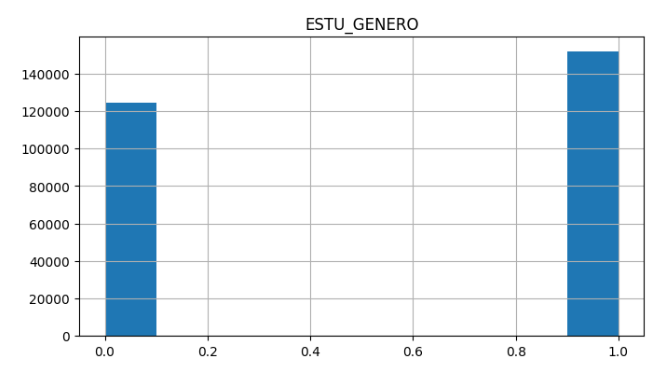
La codificación del género es la siguiente:

### **Figura 5** *Codificación del género*



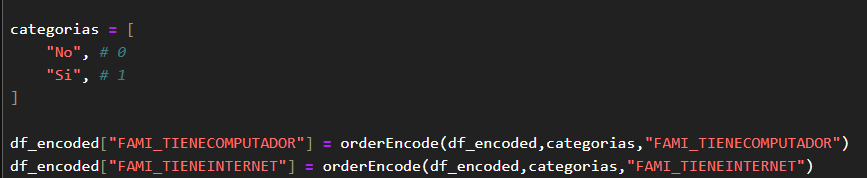
La distribución de los datos se ve a continuación:

### **Figura 6** *Distribución del género*



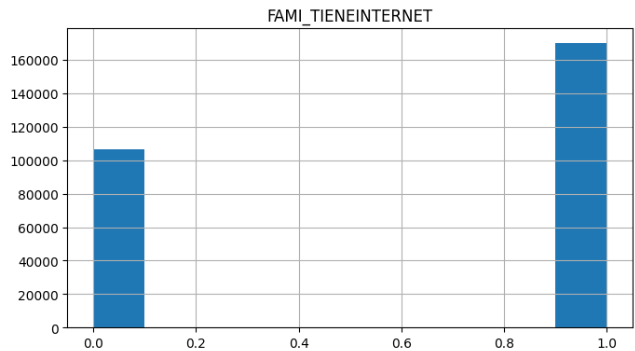
Finalmente la codificación del si se tiene computador e internet es la siguiente:

### **Figura 7** *Codificación de si la familia tiene computador e internet*



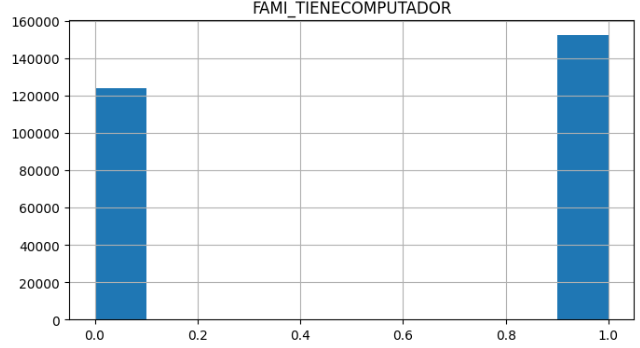
Las distribuciones de los datos se ven a continuación:

### **Figura 8** *Distribución de si la familia internet*



### 

### **Figura** 9 *Distribución de si la familia computador*



# Referencias

[1]*Acerca del Examen saber 11°*.<https://www.icfes.gov.co/acerca-del-examen-saber-11%C2%B0>

[2]Orduz, M. C. (2022, February 17). Preocupantes resultados de la educación en el país.<https://www.eltiempo.com/vida/educacion/pruebas-saber-11-resultados-son-peores-que-hace-siete-anos-652099>

[3]Instituto Colombiano para la Evaluación de la Educación - ICFES. (2019, December 24). Saber 11° 2019-2: Datos Abiertos Colombia. Saber 11° 2019-2 | Datos Abiertos Colombia. <https://www.datos.gov.co/Educaci-n/Saber-11-2019-2/ynam-yc42/about_data>

[4]Singhal, S. (2023, August 29). Defining, analysing, and implementing imputation techniques. Analytics.<https://www.analyticsvidhya.com/blog/2021/06/defining-analysing-and-implementing-imputation-techniques/>

[5]Dansbecker. (2018, January 22). Using categorical data with one hot encoding. Kaggle. <https://www.kaggle.com/code/dansbecker/using-categorical-data-with-one-hot-encoding>

[6]*When and why to standardize your data*. Built In. (n.d.). <https://builtin.com/data-science/when-and-why-standardize-your-data>.