UNA APLICACIÓN DEL MODELO PROBIT SOBRE EL EFECTO VECINDARIO EN ENTORNOS EDUCATIVOS

An application of the probit model on the neighborhood effect in educational environments

CHARLI JHOAN BENAVIDES PARRA[1]

Recibido:01 de junio de 2023. Aceptado:07 de julio de 2023
DOI: http://dx.doi.org/10.21017/rimci.2023.v10.n20.a136

Resumen

El poder contar con métodos de análisis estadístico que permitan describir y predecir resultados en el ámbito educativo bajo un soporte tecnológico, se ha convertido en una necesidad en todos los niveles, desde la educación primaria hasta la superior, pues permite identificar falencias y oportunidades de forma local o general, tales como deserción, repetición y avance o fracaso estudiantil. Por ello, se pretende en este artículo, observar y estudiar las múltiples variables que podrían relacionarse con el desempeño de los estudiantes en los centros educativos, teniendo en cuenta la calidad del vecindario en primera instancia. Estimando un modelo espacial de precios de las viviendas, de modo que permita caracterizar a qué micro-vecindario están expuestos los hogares de los estudiantes. La calidad de vecindario se empleará como variable explicativa importante para analizar la probabilidad de influencia del desempeño en los estudiantes de los centros educativos, en el cual se pretende utilizar datos de ubicación geográfica para aplicar modelos de econometría espacial, el objetivo es que esta metodología pueda ser utilizada en los diferentes entornos académico, por ejemplo, en las aulas de educación superior donde se percibe más variabilidad en la distribución de las residencias de los estudiantes.

Palabras clave: Efectos vecindario; Desempeño académico; Continuidad; Deserción; Educación.

Abstract

Being able to rely on statistical analysis methods that allow describing and predicting results in the educational field under technological support has become a necessity at all levels, from primary to higher education, since it allows identifying shortcomings and opportunities in a local or general area, such as dropout, repetition, and student advancement or failure. That is why it is intended in this article to observe and study the multiple variables that could be related to the performance of students in educational centers, considering the quality of the neighborhood in the first instance. Estimating a spatial model of house prices to characterize which micro-neighborhoods the homes of the students are exposed to. The quality of the neighborhood will be used as an important explanatory variable to analyze the probability of influence on the performance of the students of the educational centers, in which it is intended to use geographic location data to apply spatial econometrics models. The objective is that this methodology can be used in different academic environments, for example, in higher education classrooms where more variability is perceived in the distribution of student residences.

Keywords: Neighborhood effects; Academic performance; Continuity; Dropout; Education.

I. Introducción

Es bastante la información empírica que data de los años 90 en la cual se establece la influencia del vecindario en el desempeño educativo de los estudiantes. Entiéndase el efecto vecindario como un elemento clave para evitar la exclusión social de las familias afectadas por las condiciones que

1 Matemático de la Universidad Distrital Francisco José de Caldas, de Bogotá, Colombia. Magister en Big Data & Data Science de la Universidad Internacional de Valencia, España. Especialista en Actuaría y Finanzas de la Universidad Antonio Nariño, de Bogotá, Colombia. Es estudiante de la especialización en Gerencia de Instituciones Educativas de la Corporación Universitaria Republicana, de Bogotá, Colombia. Actualmente se desempeña como profesor en educación media IB. ORCID: https://orcid.org/0009-0007-7643-7953 Correo electrónico: charly_benavides@hotmail.com.

presentan estos entornos vecinales, tales como infraestructura de los inmuebles, vías de acceso, seguridad, distancias etc.

El análisis teórico que se ha hecho alrededor de este tema no ha sido tan amplio, sin embargo, hay investigaciones que se dedican a analizar eventos relacionados a este fenómeno en busca de cuantificar la relación ya mencionada con el vecindario (específicamente la distancia entre la residencia y el lugar de estudio), este dato puede determinar el nivel de comodidad que les brindaría su vecindario para llegar al centro educativo.

De forma general y en estudios realizados en países como Inglaterra, se establece que existen factores por los cuales un estudiante deserta del centro educativo o continúa en el mismo independientemente de su desempeño, sin embargo, es escaso el número de estudios que expliquen de manera contundente a través de técnicas de ciencia de datos más especializadas, cuál es la incidencia sobre el tipo de vecindario, la calidad del mismo o distancia al colegio como algo determinante en el desempeño y permanencia de los estudiantes en el mismo.

En Bogotá, los diferentes vecindarios presentan variadas dinámicas socioeconómicas que están relacionadas con la conversión de los recursos económicos de quiénes habitan estos vecindarios, por ejemplo, alquiler, número de personas que conviven con el estudiante o el valor de los inmuebles en los que residen los estudiantes. Teniendo en cuenta esto último, se realizará la aplicación de un modelo espacial de precios de viviendas, con el fin de construir un indicador "Proxy" de la calidad del vecindario, teniendo en cuenta la estimación de un Spatial Error Model de los determinantes de los precios de las viviendas.

En este artículo se busca dar contribución a futuros trabajos sobre los determinantes del desempeño de los estudiantes en el ámbito educativo según las distancias de residencia, considerando un factor de contexto importante de acuerdo a la literatura teórica pero que hasta el momento ha sido omitido: La calidad del vecindario, ejecutando un análisis realizado con base en datos de elaboración propia que representen los resultados de pruebas académicas e información particular de los estudiantes junto a la información obtenida de precios comerciales de las viviendas, para luego aplicar técnicas de econometría espacial.

II. Marco teórico

A. Antecedentes

Algunos estudios observan un primer grupo en el cual se realizan estudios utilizando datos que fueron muestras experimentales para poder estimar los efectos que causa el vecindario en el desempeño educativo y otros ámbitos como la salud y criminalidad, en este programa se utilizó un esquema de distribución aleatoria de Boucher de subsidio a la vivienda en cinco ciudades de Estados Unidos entre 1994 y 1998, estos estudios fueron orientados hacia las familias con alta pobreza, en este programa se condicionaron las variables de forma experimental. En conclusión, no se encontraron evidencias de efectos significativos que ejercía el vecindario sobre las pruebas de matemáticas y lectura en los niños que participaron en el programa.

Se analiza un segundo grupo en el cual se utiliza un método cuasi-experimental para observar cuál es el factor de impacto del vecindario sobre los logros de los estudiantes, de acuerdo al Chetty y Hendren[1] este impacto se ve reflejado en la probabilidad de la asistencia a la universidad de jóvenes entre 18 y 23 años, Las observaciones muestran que este efecto es más notable cuando el individuo se ha visto influenciado desde temprana edad, específicamente antes de los 13 años.

En la literatura del ámbito educativo, se puede encontrar que la formación del capital humano se encuentra determinada por insumos que relacionan resultados obtenidos a partir de ciertos modelos matemáticos. Por ejemplo, la función de producción educativa de Checchi[2] se podría expresar como:

$$\Delta H_{it}=f(A_i,E_{it},H_{it},S_{it})$$

Donde, siendo **i** el individuo y **t** el período:

- ΔH_{it}: Formación de capital humano
- A_i: Habilidad individual
- E_{it}: Recursos educativos del centro
- H_{it}: Stock inicial del capital humano del individuo (entorno familiar y del vecindario)
- S_{it}: Fracción de tiempo para educarse

De acuerdo con esta función, la habilidad individual determina considerablemente a la formación de capital humano que podría concebirse como todo aquello que potencializa al niño en el momento en que toma su educación, y que no se puede conseguir en el mercado.

El ejercicio en el cual se transmiten este tipo de habilidades se refuerza entre los padres y los hijos, creando así una correlación entre el capital humano adquirido por los padres y los hijos. Por otro lado, los niños se encuentran expuestos a ambientes residenciales específicos, por lo tanto, se encuentran condicionadas por preferencias tales como la ubicación más cercana a algún Centro Educativo en especial o las condiciones económicas de la familia del estudiante.

La fracción de tiempo que los individuos utilizan para educarse es un determinante en el capital humano, teniendo en cuenta que esta fracción de tiempo, se podría considerar como una inversión a futuro, que puede generar mayores retornos según sea el nivel de educación alcanzado, por lo tanto, se podría interpretar que los individuos enfrentarían un problema de optimización en el que se ha de maximizar la utilidad obtenida a partir del tiempo utilizado en la educación, dicha optimización sería del tipo:

$$V_i = w_{it}H_{it} - S_{it}W_{it}H_{it} - \gamma_t S_{it} + \frac{W_{it+1}H_{it+1} - S_{it+1}W_{it+1}H_{it+1} - \gamma_{t+1}S_{it+1}}{1+\rho}$$

con las restricciones de:

$$\Delta H_{it} = (A_i S_{it} E_{it} H_{it})^{\alpha},\ \alpha > 1$$

$$H_{it+1} = H_{it}(1-\delta) + \Delta H_{it}$$

donde:

- V_i: Valor actual de la utilidad obtenida
- W_{it}: Ingresos laborales
- γ_t: Costos de educación
- ρ: Tasa subjetiva de descuento intertemporal
- δ: Tasa de depreciación del capital humano

Por lo tanto, se puede obtener como un equilibrio en la demanda óptima de educación S^*_{it}, en el momento en que se igualan los costos y los beneficios al educarse, costos como aquellos directos γ_t qué se utilizan en matrícula, materiales, transporte, costo de vida y aquellos indirectos $S_{it}W_{it}$ como el costo de las oportunidades laborales, en cuanto a beneficio, hacemos referencia al aumento de los ingresos laborales producto del mayor stock de capital humano.

$$\beta_t H_t + \gamma_t = \frac{\beta_{t+1}}{(1+\rho)} \frac{\alpha \Delta H_{it}}{S_{it}}$$

Ese nivel óptimo depende de manera positiva en la expectativa de los retornos futuros β_{t+1} con su respectivo descuento de tasa ρ y su relativo retorno al valor presente β_t, también depende de las habilidades que son inobservables A_i, recursos del Centro Educativo E_{it} y el stock inicial del capital humano H_t; podría decirse que su dependencia es inversa antes Los costos directos de educarse γ_t.

$$S^*_{it} = \left(\frac{\beta_{t+1}}{\beta_t(1+\rho)} \frac{\alpha(A_t E_{it} H_{it})^{\alpha}}{H_{it} + \gamma_t/\beta_t}\right)^{\frac{1}{1-\alpha}}$$

Según esto, la asistencia del centro educativo o tiempo destinado a la educación es un insumo porque determina de forma directa al capital humano que acumula cada estudiante, también es un producto, pues su valor óptimo se encuentra determinado por los demás insumos de la función de producción educativa, por lo que estos factores ejercen una influencia directa en la formación del capital humano a través de los efectos que genera decidir si se permanece o no en la educación.

Antecedentes en Colombia

En Colombia la literatura de este tipo de temas en específico sobre el efecto vecindario y el desempeño educativo de los estudiantes no ha sido suficientemente estudiado, sin embargo, se pueden observar otras investigaciones que tienen en cuenta otro tipo de factores que podrían repercutir en el desempeño académico de los estudiantes. A continuación, se describen algunos de los estudios realizados en Colombia y relacionados con nuestro tema de interés:

Carlos Pérez[3] en su monografía «Efectos de vecindario como determinantes de la deserción estudiantil y el logro académico en la Universidad

del Valle», el problema de la identificación de los distintos factores que podrían determinar la deserción estudiantil y cierta probabilidad de graduación en la Universidad del valle particularmente, cuyo fin es describir las condiciones del lugar de residencia y su relación con el desempeño académico según el estudiante en comparación a sus pares académicos.

B. Factores de influencia en el efecto vecindario

En las taxonomías propuestas por Galster[4], se observan cuatro factores por los cuales es posible establecer una influencia de la calidad del vecindario del lugar de residencia sobre el desempeño académico educativo de los estudiantes: Interacción social, factores ambientales, geográficos e institucionales.

Factor de interacción social

Tiene mayor influencia sobre el efecto vecindario que se puede ejercer hacia el desempeño educativo de los individuos, dado que enfatiza en el rol que tienen las relaciones sociales e intergrupales desarrolladas en el vecindario y dirigidas hacia el bienestar de los residentes. Estos mecanismos pertenecientes a este grupo se refieren a los «procesos sociales endógenos a los vecindarios»[5].

Factores Ambientales

Hace referencia a las características que son particulares del espacio local, sean estos de forma natural o artificial, dichas características podrían afectar de alguna manera la salud mental o física de los residentes, la violencia o condiciones ambientales.

Factores geográficos e institucionales

Este tipo de factores hacen referencia a las condiciones de localización de un vecindario con respecto a otros centros económicos o políticos, existen vecindarios que tienen baja accesibilidad, hay alguna distancia considerable al compararlos con otros y esto genera un problema relacionado con los medios de transporte. Estas condiciones podrían llegar a limitar las oportunidades de los individuos residentes de un vecindario en particular, por ejemplo, cómo se limita la oferta educativa, enviando al individuo a ciertos costos unitarios directos o indirectos, tenemos el dilema del tiempo invertido en el transporte, este es un costo de oportunidad que afecta negativamente al individuo.

III. Metodología

Se hará uso de resultados de creación propia, que simulan una herramienta de evaluación denominada prueba PENSAR, esta prueba permite que los colegios tengan internamente evaluaciones que puedan suministrarles información de todo el proceso que se está llevando curricularmente dentro de la institución, esta prueba se construye dentro del marco de referencia que tiene el Ministerio de Educación nacional a través de lineamientos curriculares y estándares básicos de Educación por competencia.

Para este artículo, se trabajará con los resultados simulados en tres de estas pruebas aplicadas durante un periodo escolar para 189 estudiantes, en ellas se evalúan las áreas básicas que son: matemáticas, lenguaje, ciencias naturales, ciencias sociales e inglés, presentando los resultados a través de herramientas de visualización de datos. Se han asignado barrios de forma aleatoria suponiendo que allí residen los supuestos estudiantes (finalmente este es un plan piloto para aplicar en casos reales), junto a esta información se tendrá en cuenta la base de datos que facilita la secretaría de hacienda de Bogotá en cuanto a los impuestos prediales y valores comerciales de los hogares en la ciudad. Finalmente, por georreferenciación de los hogares en estudio y el análisis parcial de todas las variables asociadas se obtendrá conclusiones.

A. Variables

Se describen algunas características en términos del impuesto predial publicado por la Secretaría de Hacienda y atributos estructurales. La distribución de los precios comerciales de los inmuebles hace correspondencia al período del año 2022, para ello, se adjunta la siguiente tabla ilustrativa de los distintos valores según las localidades de la ciudad. Fig. 1.

Aquí se puede establecer que no hay simetría en cuanto al valor comercial de las viviendas discriminadas por localidad en Bogotá, de alguna

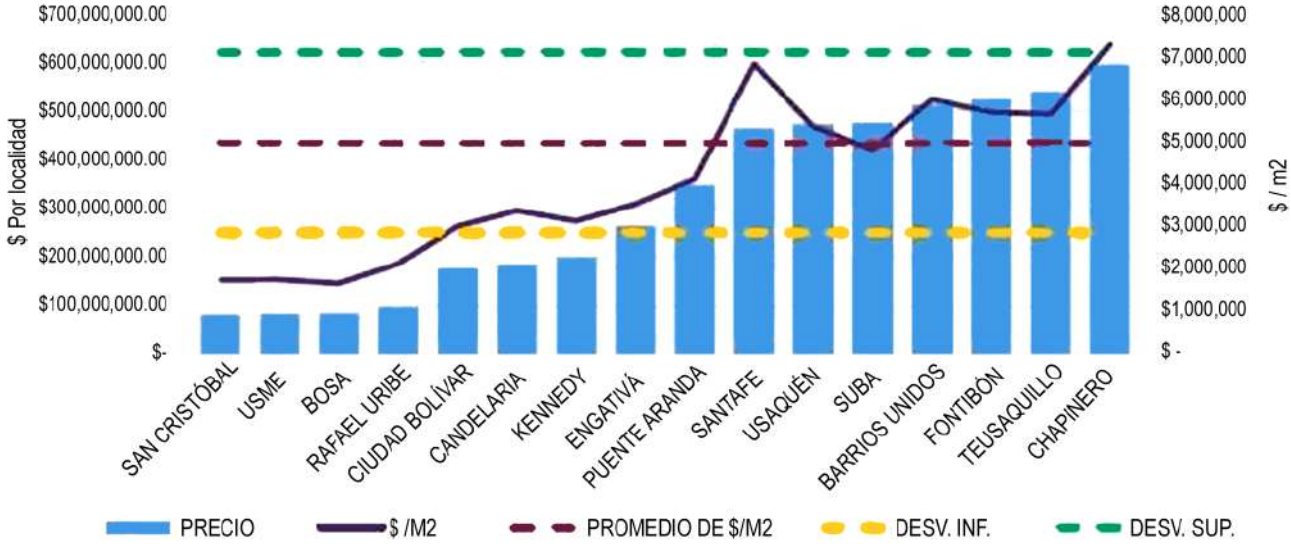

Fig. 1. El precio promedio de la vivienda nueva en Bogotá.

manera se puede establecer que reorganizando estas localidades desde el sur a norte es clara la diferencia que hay cuando nos acercamos a la parte norte de la ciudad donde se encuentran barrios como Chapinero, Teusaquillo y Fontibón, existe una gran diferencia entre estos barrios con valores obtenidos en localidades como San Cristóbal, es posible encontrar viviendas menores a $100.000.000 MCTE.

Por lo tanto, se puede desprender una relación elevada de dependencia espacial del precio de los inmuebles con su ubicación geográfica, de esta manera se puede justificar el uso de un modelo espacial para este estudio. En un análisis sobre el valor promedio del metro cuadrado, el más costoso de la ciudad se encuentra localizado en la localidad de Chapinero, con un promedio de $7’273.653[6].

Es probable encontrar que algunas de las familias de los estudiantes residan en proyectos de vivienda conjunta, es decir, en edificios con apartamentos más que en casas individuales, también se realiza un análisis sobre los costos de los proyectos de vivienda de Bogotá, el total de estos proyectos es de 369, de los 450 se encuentran ubicados en Usaquén, concentrándose allí el 49.8% del total de proyectos de vivienda activos.

Para realizar la aplicación del modelo probit[2], se debe convertir los datos relacionados con el desempeño educativo a variable binaria, por ello, la primera variable toma el valor de 1 en caso de que el estudiante haya obtenido una media por debajo de 50 puntos en alguno de los simulacros de las pruebas “Pensar” y 0 en caso contrario. Por otro lado, para la asistencia se tomará el valor de 1 si el estudiante presentó los tres simulacros en dicha institución y 0 en caso contrario. También se incluye la variable dummy «género» que toma el valor M si el estudiante corresponde a mujer y H en caso contrario. Tabla I de variables.

Tabla I. Descripción de Variables.

Variable	Descripción
Género	Género del estudiante
Código	Código del estudiante
Pérdida	1 si obtuvo menos de 50 puntos en al menos un simulacro, 0 de lo contrario
Dirección	Ubicación de la residencia del estudiante
Vecindario	Localidad del barrio del estudiante
Valor de inmuebles	Precia comercial promedio de las inmuebles
Edad	Edad del estudiante
Grado	Nivel educativo

2 El modelo probit se caracteriza por ser una regresión en la que la variable dependiente debe ser binaria. Su objetivo es estimar la probabilidad de que una situación con particularidades pueda comportarse de una manera específica, este es un modelo de clasificación binaria.

B. Etapas del proceso

Por medio de una aproximación y el uso de técnicas de econometría espacial a un modelo de precios hedónicos de las viviendas de Bogotá. Se aplicará el método mostrado por Dubin[7] con el objetivo de obtener un indicador sobre la calidad del vecindario por medio de los precios de las viviendas.

IV. Análisis descriptivo y modelos de clasificación

Empecemos observando la georeferenciación de estas residencias respecto a la ubicación geográfica de nuestro centro educativo al cual le asignamos una dirección aleatoria. Fig. 2.

Realizando un barrido sobre las localidades donde se registran las residencias de los estudiantes, se obtiene el diagrama de torta. Fig. 3.

Tiene sentido pensar que, en colegios o instituciones de educación media, la mayoría de los estudiantes residan en su misma localidad, situación que refleja este diagrama, pues entre Usaquén y Suba suman casi el 42% de las viviendas, esto en cambio no es tan común en centros de educación superior.

Podremos encontrar por variable, el conteo de los registros, la media, desviación estándar, valor mínimo, cuartiles y máximo valor, esto por supuesto se puede realizar estrictamente a nuestras variables de interés, sin embargo, se hace de manera global dado que estamos realizando un ejercicio exploratorio. Tabla II.

Se puede observar que los estudiantes pertenecen a la sección de primaria con los grados de segundo a quinto, cuyas edades oscilan entre los 7 y 12 años. Se observa el comportamiento en términos de distribución de cada una de las variables registradas en la base de datos. Fig. 4.

De allí se desprende que la mayoría de los estudiantes pertenecen a grado quinto, un porcentaje considerable ha reprobado las pruebas, la mayoría de los estudiantes tienen 10 años y casi hay un equilibrio entre niños y niñas.

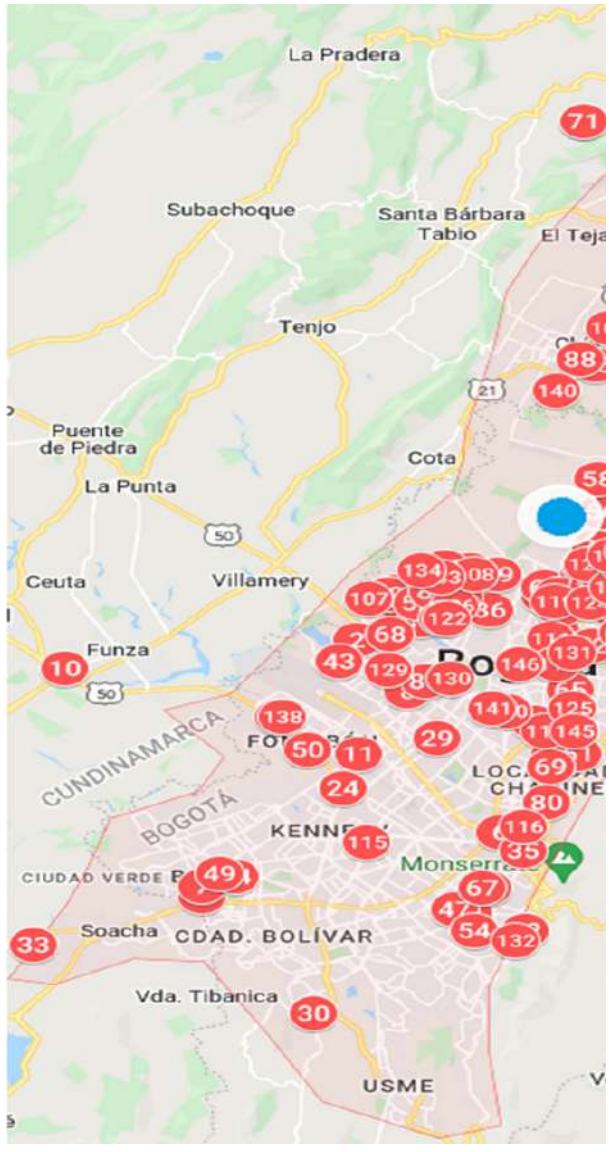

Fig. 2. Georeferenciación de las residencias de

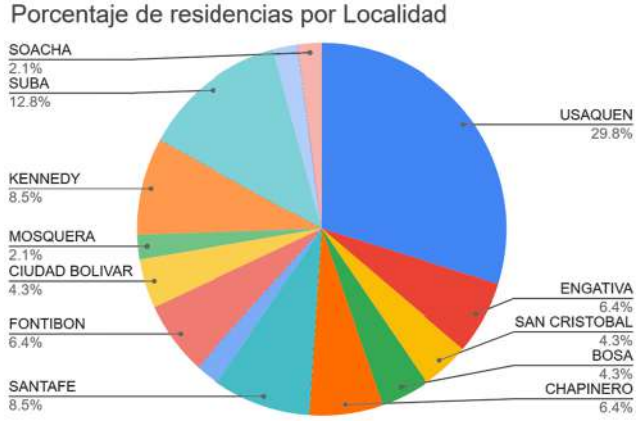

Fig. 3. Distribución de las viviendas de los estudiantes.

Tabla II. Descriptivo de variables.

	count	mean	std	min	25%	50%	75%	mx
G	189.0	3.671.958	1.161.578	2.0	3.0	4.0	5.0	5.0
Pérdidas	189.0	0.301587	0.460166	0.0	0.0	0.0	1.0	1.0
Edad	189.0	9.613.757	1.381.497	7.0	9.0	10.0	10.0	12.0
Género M	189.0	0.407407	0.492657	0.0	0.0	0.0	1.0	1.0

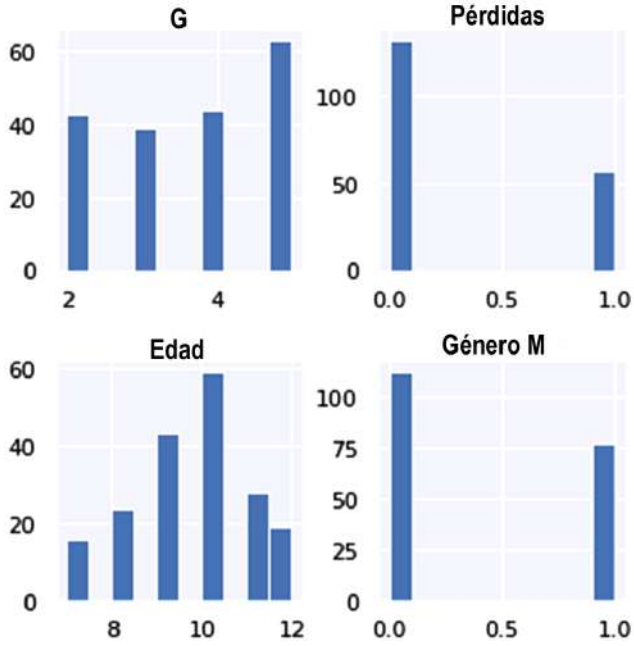

Fig. 4. Visualización de variables.

Se observa ahora la correlación entre las variables en una matriz de correlación. Tabla III.

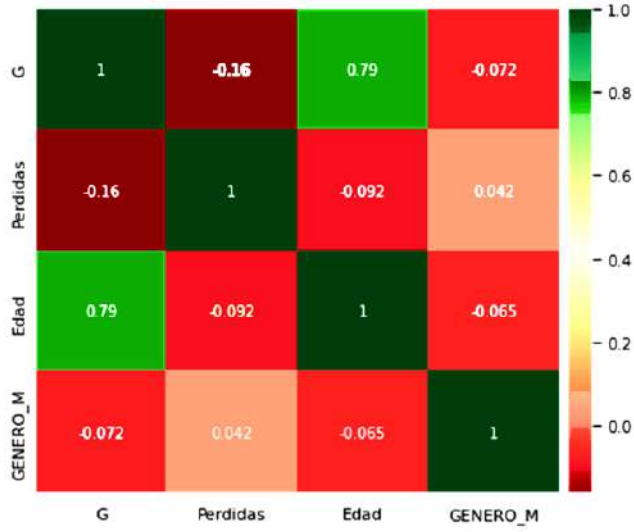

Tabla III. Correlación de las variables.

Como es de suponer, la edad guarda relación directa con el grado. Con el objetivo de reconocer una posible relación interesante entre algunas variables de la base de datos, aplicaremos algunas técnicas de aprendizaje supervisado. Ver Fig. 5.

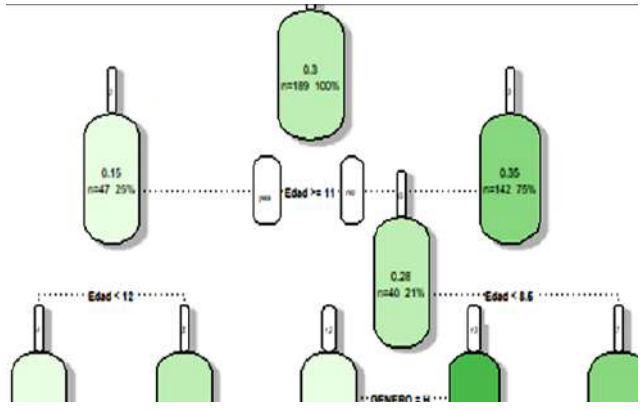

Fig. 5. Árbol de decisión.

A. Árbol de decisión

Vamos a emplear clasificación utilizando el modelo de árboles de decisión, esto para clasificar si un estudiante reprueba uno o más simulacros, en este caso, asignaremos una clase o categoría a la observación basada en variables independientes también llamadas «predictores».

Como entrada se tendrá la edad y el género de los estudiantes y nos va a interesar su desempeño en los simulacros, lo interesante es detectar a través de la información del estudiante si él llega a reprobar o no un simulacro o más, por lo tanto, estamos interesados en saber cuáles son esas variables que influyen o determinan ese desempeño para establecer un modelo.

Es importante recordar que en este caso «GÉNERO» es una variable no numérica a diferencia

de «Edad», el campo «Perdidas» tendrá las salidas 0 y 1, dónde 0 es que no perdió ningún simulacro (No obtuvo menos de 50 puntos), en cambio 1 representa la pérdida de al menos uno de los simulacros.

Haciendo uso del lenguaje de programación R, se obtiene el siguiente árbol de decisión:

Del cual se interpreta:

- El 30% de los estudiantes reprobaron al menos uno de los simulacros.
- Los estudiantes menores a 11 años tendrían un porcentaje de reprobación del 35%.
- El 38% de los estudiantes de 9 y 10 aprobarían.
- La mitad de los estudiantes menores a 8 años de género femenino reprobaron al menos un simulacro.

B. Random Forest

Utilizando Python y el modelo de clasificación "Random forest" para generar datos de entrenamiento para variables independientes como variable dependiente, al evaluar el modelo con los datos de pruebas tenemos 71% de accuracy promedio. Fig. 6.

```
[7]  1  # Seleccionamos las características para el modelo
     2  data = df[['GENERO_M', 'Perdidas', 'Edad', 'G']]
     3  data.head()
```

	GENERO_M	Perdidas	Edad	G
0	1	0	9	2
1	1	1	8	2

```
[34] 1  BA_model.fit(X_train, y_train)

     RandomForestClassifier(min_samples_leaf=18, n_estimators=50, random_state=50)

[35] 1  # Accuracy promedio
     2  BA_model.score(X_test, y_test)

     0.7083333333333334
```

Fig. 6. Random forest.

Al construir la matriz de confusión se puede encontrar todas las predicciones a partir del conjunto de datos de test, nos dará en la diagonal principal, los elementos que fueron correctamente clasificados, en este caso de los 48 datos de prueba 34 correctamente fueron identificados como casos de no pérdida de simulacros, sin embargo ninguno de los casos en los que se perdía al menos un simulacro, fue reconocido correctamente, generando falta de seguridad al establecer que las variables edad, género y grado sean suficientes para determinar una relación con la posibilidad de perder al menos un simulacro. Fig. 7.

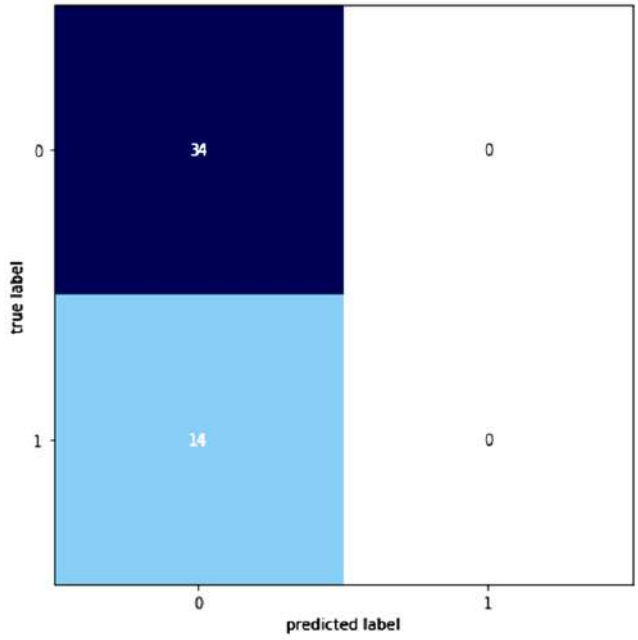

Fig. 7. Matriz de confusión.

Teniendo en cuenta los resultados obtenidos, nace la necesidad de incluir una nueva variable a nuestro estudio, en este caso, aquella que relaciona al vecindario.

C. Justificación de las localidades como vecindario

Se intentará encontrar un soporte técnico que permita rechazar o aceptar la toma de localidades como vecindario para nuestro estudio. Esto a través de distintas técnicas de Clustering. Utilizando Python, se representará en el plano, la distribución de las residencias de los estudiantes que hacen parte de la muestra. Fig. 8.

Dado que se ha elegido la variable «Dirección» es útil clasificar estas direcciones en posibles grupos, para ello, podremos realizar un estudio empleando técnicas de clustering, para tener una visión aproximada de la pertinencia al elegir las localidades de Bogotá como nuestra base para definir los vecindarios.

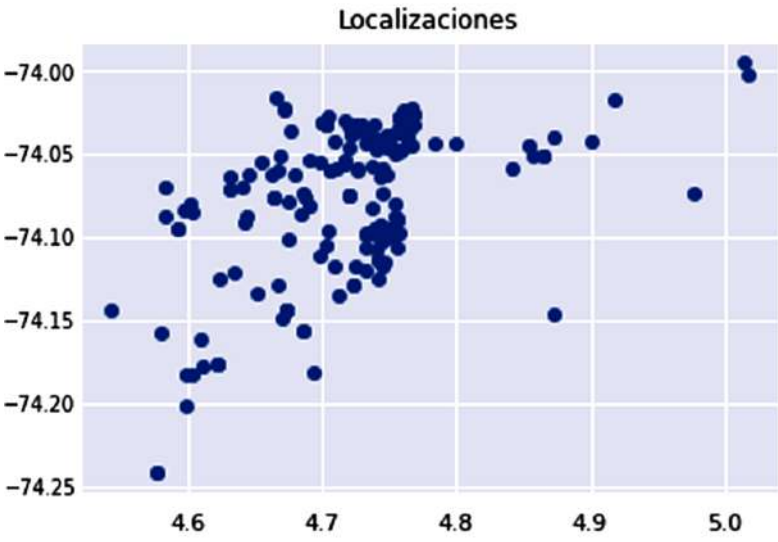

Fig. 8. Dispersión de residencias.

Con el objetivo de utilizar clustering jerárquico, se ejecuta el método "Ward", tenemos una correlación correspondiente a 0.59, valor que no es suficiente y por ello requerimos replantear las características del clustering.

Eligiendo un método diferente (single) y especificando la métrica (euclidiana), obtenemos una correlación de 0.82, valor aceptable con el que se representa el respectivo dendograma. Fig. 9.

Si se condiciona la distancia a 0,2 se tiene 9 posibles grupos, dando como resultado el dendograma truncado. Fig. 10.

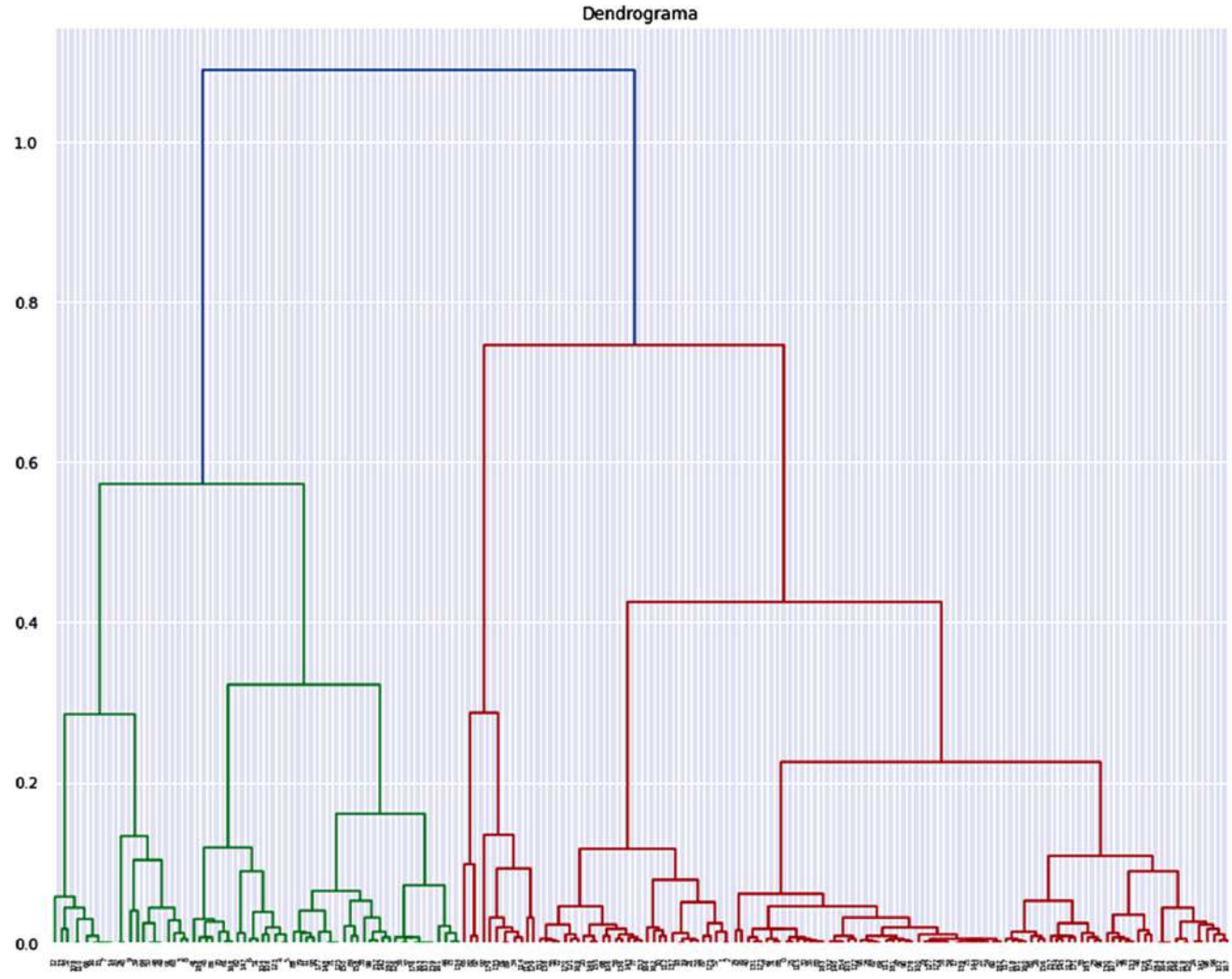

Fig. 9. Dendograma.

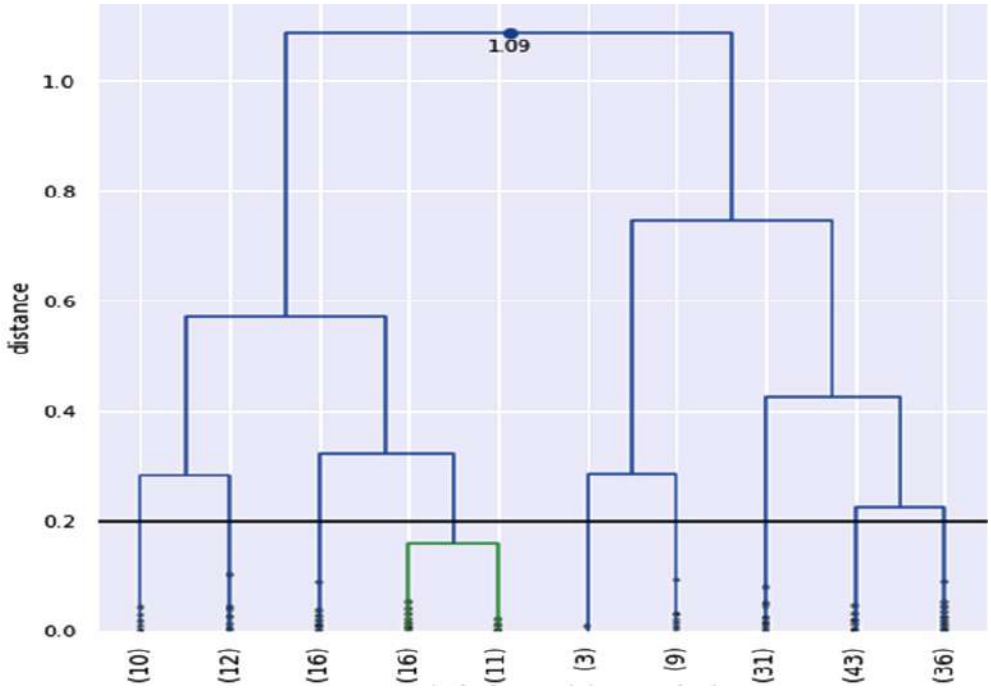

Fig. 10. Dendograma truncado.

Con esto se tiene un primer candidato para un número de clusters (nueve), se puede interpretar que los dos vecindarios con mayor número de muestras se relacionan con las localidades de Suba, Usaquén y Engativá, que son zonas más concentradas y más cercanas a la ubicación del colegio.

Representando las últimas 10 uniones por método "Single" se tiene. Fig. 11.

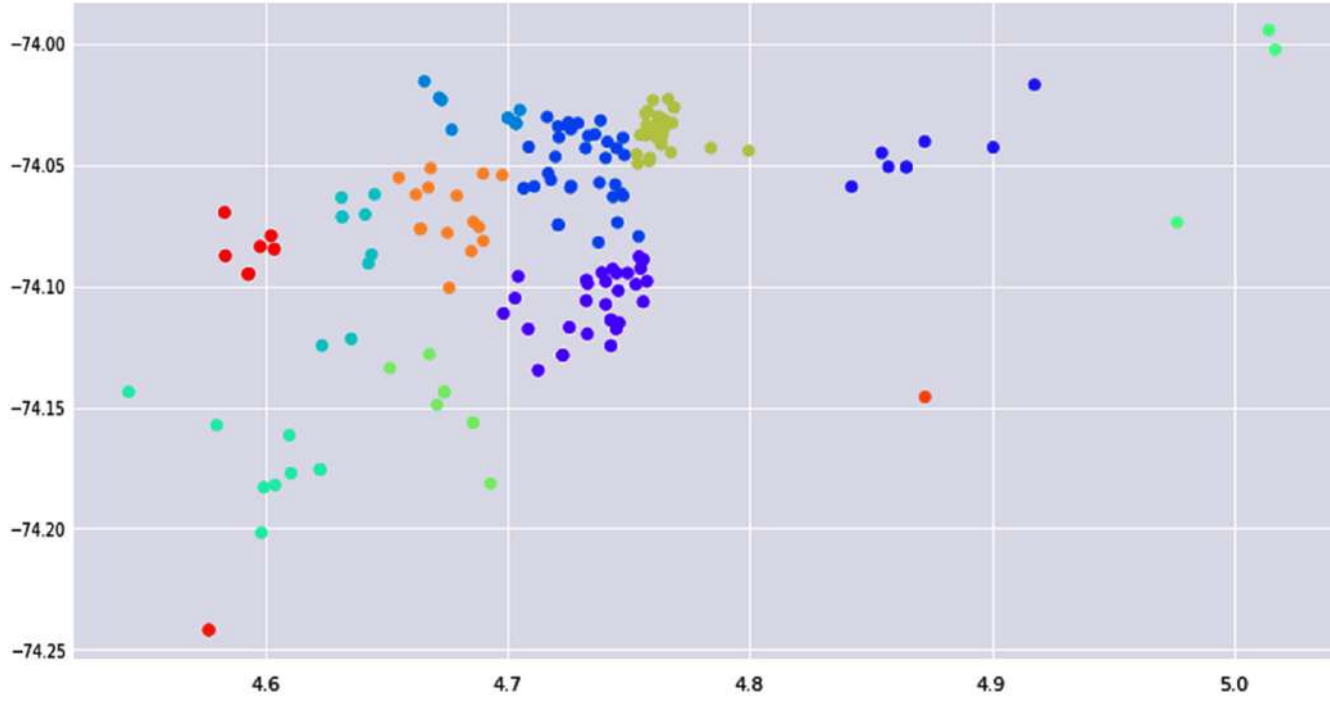

Fig. 11. Método Single.

Para seguir ampliando la visión, se hará uso de clustering basado en densidad con 12 clusters y su representación en el plano es, fig. 12.

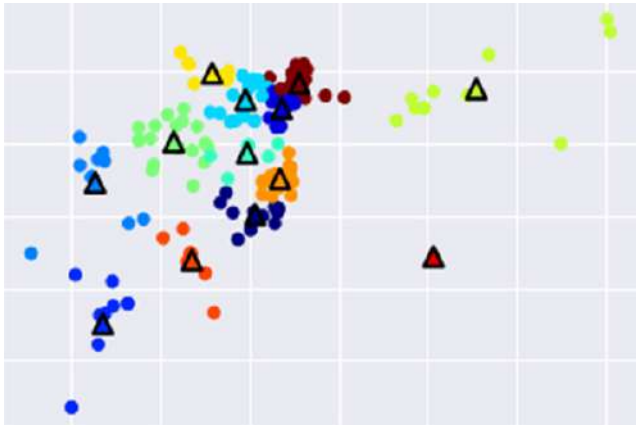

Fig. 12. Clustering basado en densidad.

Se puede observar una similitud en la distribución de las agrupaciones aun cuando la cantidad de grupos es mayor.

Ahora, basados en particiones, por medio del método del codo, se establece el número apropiado de clusters según la técnica k-means. Fig 13.

Se puede observar que la gráfica presenta un decrecimiento monótono a partir del valor k=13, por ende, este es el número más apto de agrupamientos a realizar con estos puntos de residencia, lo que relaciona 13 tipos de vecindario adecuados para este análisis.

Lo anterior, permite pensar que es oportuno tomar las agrupaciones de acuerdo a la cantidad de localidades en Bogotá y sus alrededores, por ello se podría continuar tomando a los vecindarios como localidades.

V. Aplicación del modelo

A continuación, se realizará la ejecución del modelo Probit, no sin antes utilizar regresión múltiple con la ayuda del lenguaje de programación R, luego

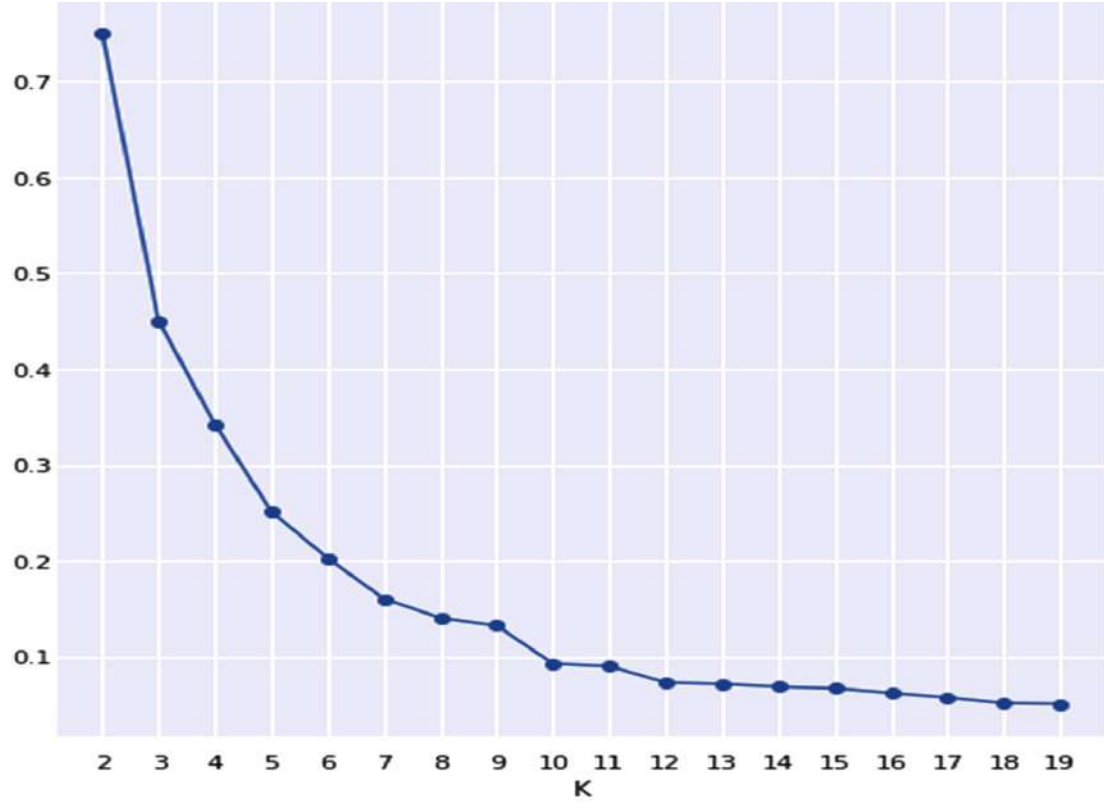

Fig. 13. Método del codo.

de analizar los resultados, sabremos si es necesario o no justificar el uso del modelo Probit, ejecutándolo con el lenguaje nombrado anteriormente.

```
install.packages("Rcpp",
repos="https://rcppcore.github.io/drat")
install.packages('readxl')
library(readxl)
file.choose()
ruta_excel <-
"C:\\Users\\STUDENT\\Downloads\\ESTUDIANTES.xlsx"
datos <- read_excel(ruta_excel)
```

A. Base de datos:

- Pérdida: toma el valor de 1 si el estudiante perdió al menos una prueba, 0 si no perdió ninguna.
- Preciovivienda: Hace referencia al valor promedio de las casas de la localidad donde reside el estudiante.
- G: Grado del estudiante.
- Género: 0 si es masculino 1 si es femenino
- Edad: Edad del estudiante, como es de esperarse, muy correlacionada con la variable grado.

B. Regresión múltiple

Aplicamos regresión múltiple, relacionando pérdida con el valor de las viviendas, grado, género y edad. Fig. 14.

```
reg = lm(Perdidas ~ 0 + Preciovivienda + G + GENERO +
Edad, datos)
plot(datos$Preciovivienda, datos$Perdidas,
xlab='Localidad', ylab='Perdida')
abline(reg)
```

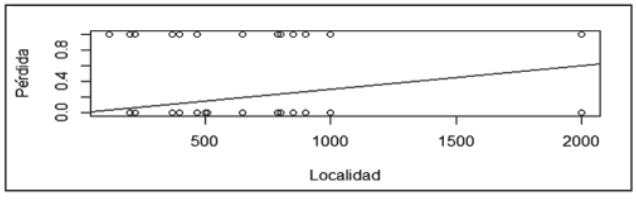

Fig. 14. Relación entre el valor de la vivienda (en millones) y el desempeño en las pruebas.

Con la ayuda de "summary(reg)" obtenemos:

```
Call:
lm(formula = Perdidas ~ 0 + Preciovivienda + G + GENERO + Edad,
    data = datos)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5106 -0.3106 -0.2343  0.5848  0.8508

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
Preciovivienda  3.187e-11  9.587e-11   0.332 0.739970
G              -1.189e-01  4.158e-02  -2.859 0.004738 **
GENERO          3.809e-02  6.776e-02   0.562 0.574690
Edad            7.184e-02  1.926e-02   3.730 0.000255 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4584 on 185 degrees of freedom
Multiple R-squared:  0.3181,    Adjusted R-squared:  0.3034
F-statistic: 21.58 on 4 and 185 DF,  p-value: 1.264e-14
```

Al preguntar si este modelo tiene alguna validez, se debe remitir a la prueba F, se observa un p-valor menor a 0.05, por lo tanto se puede rechazar la hipótesis de que este modelo no es válido, así que guarda validez el uso de este método, si se quiere saber qué tan explicativo es, se debe tener en cuenta R cuadrado ajustado, este dice que tiene una predicción del 30%, por supuesto, no es suficientemente significativo, esto invita a probar otros métodos como Probit, en cuanto a los coeficientes, se puede observar que el grado y la edad son aquellos únicos que guardan un nivel de significancia al ser menores al 0.05, por lo tanto, este modelo no rechaza la hipótesis, es decir, el precio de la vivienda y el género no tienen significancia en este modelo.

C. Modelo Probit

Es momento de ejecutar el modelo Probit en R, aplicando el código:

```
PROBIT = glm(Perdidas ~ 0 + Preciovivienda, datos, family
= binomial(link=probit))
summary(PROBIT)
```

Se obtiene:

```
Call:
glm(formula = Perdidas ~ 0 + Preciovivienda + G + GENERO + Edad,
    family = binomial(link = probit), data = datos)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0954  -0.8903  -0.6988   1.3002   1.7994

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
Preciovivienda -1.001e-11  2.736e-10  -0.037   0.9708
G              -2.327e-01  1.202e-01  -1.937   0.0528 .
GENERO          5.990e-02  1.961e-01   0.305   0.7600
Edad            3.236e-02  5.487e-02   0.590   0.5553
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 262.01  on 189  degrees of freedom
Residual deviance: 225.94  on 185  degrees of freedom
AIC: 233.94

Number of Fisher Scoring iterations: 4
```

En términos de interpretación se entiende que la variable que realmente es significativa en este estudio es G (Grado), este coeficiente es negativo, lo que indica que a mayor grado es menor la probabilidad de obtener un bajo rendimiento en las pruebas.

VI. Conclusiones

Aunque en el marco teórico se incluyen variables sociodemográficas más profundas como la dedicación que los papás ofrecen a sus hijos en la orientación de su proceso escolar, hay posibilidad de obtener resultados precisos dirigidos a este tipo de preguntas, sin tener un detalle profundo sobre estas condiciones, ya que esta es una primera etapa de inspección sobre las variables establecidas en la metodología, de esta manera en futuros estudios, si se podrá utilizar este artículo como base exploratoria e incluir los datos que hacen falta.

Frente a la distribución de las residencias de los estudiantes a lo largo de la ciudad y sus alrededores, la mayor concentración se presenta en las localidades cercanas a la institución, este se encuentra ubicado en la localidad de Suba, localidad que limita con Usaquén Engativá y Santa Fe, hay más del 55% de concentración, pues solo en la localidad de Suba se presenta el 13%, mientras que en Usaquén el 30% aproximadamente (localidad con mayor valor comercial promedio por vivienda).

Al observar la correlación entre las variables se puede establecer que la variable grado y edad son las que más se relacionan con un casi 80%, estas variables sin tener en cuenta la ubicación de la residencia, no presentan una relación considerable, por ejemplo, el género apenas presenta un 4% de correlación con el desempeño académico de los estudiantes.

Llevando a cabo análisis de dispersión de las residencias de los estudiantes, gracias a la variable "dirección", se obtiene la latitud y longitud. Información tratada en la técnica de Machine learning, arrojando que el valor más apropiado para el número de vecindades es 13, lo cual juega a favor en la elección de las localidades como vecindarios, ya que, al obtenerse 16 particiones entre localidades de la ciudad y municipios aledaños, tres de ellos hacen parte de las municipalidades que limitan con Bogotá, por lo tanto, observamos prudente trabajar con el rango 9 a 16.

Al aplicar el modelo de Probit, incluimos nuestra variable clave (vecindario), por medio de una regresión múltiple observamos que no hay un efecto significativo entre un variable vecindario y el desempeño académico de los estudiantes, a partir en su en la ejecución de las pruebas pensar, efecto contrario relacionado a la variable Grado.

Teniendo en cuenta las oportunidades que presenta este artículo, se manifiesta la necesidad de una intervención en políticas públicas con el objetivo de mejorar los resultados de las pruebas nacionales, teniendo en cuenta las condiciones propias de los centros educativos, de esta manera reducir la reproducción intergeneracional de desigualdad y los efectos negativos que puedan ejercer el vecindario sobre el desempeño de un estudiantil.

Referencias

[1] Chetty, R. and Hendren, N. «The Impacts of Neighborhoods on Intergenerational Mobility: Childhood Exposure Effects and County-Level Estimate», Harvard University. 2015.

[2] Checchi, «D. The economics of education: Human capital, family background and inequality», Cambridge University Press. 2006.

[3] Pérez Ramírez, C. «Efectos de vecindario como determinantes de la deserción estudiantil y el logro académico», Universidad del Valle. 2012.

[4] Galster y Santiago, A. M, «What's the'hood got to do with it? Parental perceptions about how neighborhood mechanisms affect their children», Journal of urban affairs, pp. 201 – 226. 2006.

[5] Bracco L. «Efecto vecindario en el desempeño educativo», Montevideo-Uruguay. 2019.

[6] Precios m2. Metrocuadrado. https://www.metrocuadrado.com/noticias/herramientas/precios-m2/.

[7] Dubin, R. A. «Spatial autocorrelation and neigh borhood quality», Regional science and urban economics, pp. 433 – 452. 1992.

Códigos:

Random Forest y Clustering:
https://github.com/charlybenavides/TFM
Árbol de decisión y modelo Probit:
https://github.com/charlybenavides/TFM-R