Logotipo, nombre de la empresa

Descripción generada automáticamente Logotipo, nombre de la empresa

Descripción generada automáticamente

Reporte 3

**Datos ENIGH**

**Classification**

**Materia: MAA(Aprendizaje Automático)**

**Profesor: José Anastacio Hernández Saldaña**

**Grupo: 03**

**Alumno: José de Jesús Almanza Trejo**

Contenido

[Introducción 3](#_Toc172480223)

[Definiciones 4](#_Toc172480224)

[Construcción de los datos 4](#_Toc172480225)

[Análisis descriptivo 5](#_Toc172480226)

[Clasificación 9](#_Toc172480227)

[Resultados 9](#_Toc172480228)

[Bibliografía 11](#_Toc172480229)

# Introducción

A continuación, se presenta un análisis descriptivo y aplicación de varios modelos de clasificación que pueda calcular el nivel socioeconómico (1=Bajo, 2=Medio Bajo, 3=Medio Alto y 4=Alto) reportado en la tabla de viviendas encuestadas en el ENIGH a partir de características conjuntas de hogares y personas que habitan la vivienda.

Se aplicarán los algoritmos de KNN Classification, Decision Tree Classification, SVM y regresión logística sobre una validación cruzada calculando el estadístico de accuracy sobre la base de pruebas para seleccionar el modelo con el mejor desempeño y mostrar sus características.

# Definiciones

## Construcción de los datos

La (ENIGH, s.f.) cuenta con las siguientes fuentes de información:

|  |  |  |
| --- | --- | --- |
| **Tabla** | **Descripción** | **Tipo de datos** |
| (viviendas, s.f.) | Se encuentran contenidas las características de las viviendas que habitan los integrantes de los hogares encuestados. | Materiales de construcción, amenidades, antigüedad, número de cuartos, con agua disponible, drenaje, etc |

De esta tabla se seleccionaron las siguientes características para realizar el ejercicio:

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Descripción** | **Clasificación** | **Tipo** |
| folioviv | Folio de la vivienda encuestada. | texto | llave |
| est\_socio | estrato socioeconómico | Discreta | cualitativa |
| tipo\_viv | tipo de vivienda | Discreta | cualitativa |
| mat\_pared | material de pared | Discreta | cualitativa |
| mat\_techos | material de techos | Discreta | cualitativa |
| mat\_pisos | material de pisos | Discreta | cualitativa |
| antiguedad | antigüedad de la vivienda | Discreta | cuantitativa |
| cocina | cuenta con cocina | Binaria | cuantitativa |
| cuart\_dorm | dormitorios | Discreta | cuantitativa |
| num\_cuarto | número de cuartos | Discreta | cuantitativa |
| disp\_agua | forma de abastecimiento de agua | Discreta | cualitativa |
| excusado | tiene excusado |  | cualitativa |
| bano\_comp | cuantos baños completos tiene | Discreta | cuantitativa |
| bano\_excus | cuantos baños con excusado tiene | Discreta | cuantitativa |
| bano\_regad | cuantos baños con regadera tiene | Discreta | cuantitativa |
| drenaje | destino del drenaje | Discreta | cualitativa |
| disp\_elect | fuente de donde se obtiene energía eléctrica | Discreta | cualitativa |
| focos\_inca | número de focos incandescentes | Discreta | cuantitativa |
| focos\_ahor | número de focos ahorradores | Discreta | cuantitativa |
| combustible | tipo de combustible usado en la cocina | Discreta | cualitativa |
| tipo\_adqui | tipo de adquisición de la vivienda | Discreta | cualitativa |
| tipo\_finan | tipo de financiamiento | Discreta | cualitativa |
| calent\_sol | cuenta con calentador solar | Binaria | cualitativa |
| calent\_gas | cuenta con calentador de gas | Binaria | cualitativa |
| medidor\_luz | cuenta con medidor de luz | Binaria | cualitativa |
| bomba\_agua | cuenta con bomba de agua | Binaria | cualitativa |
| tanque\_gas | cuenta con tanque de gas | Binaria | cualitativa |
| aire\_acond | cuenta con aire acondicionado | Binaria | cualitativa |
| calefacc | cuenta con calefaccion | Binaria | cualitativa |
| tot\_resid | número de residentes en la vivienda | Discreta | cuantitativa |
| tot\_hom | número de residentes hombres en la vivienda | Discreta | cuantitativa |
| tot\_muj | número de mujeres en la vivienda | Discreta | cuantitativa |
| tot\_hog | número de hogares en la vivienda | Discreta | cuantitativa |
| tam\_loc | tamaño de la localidad | Discreta | cualitativa |

Las variables con información vacía en las variables de estudio, se utilizará un algoritmo (SimpleImputer, s.f.) para rellenar estos datos con el valor que más se repite (moda) dado que todas las variables son discretas.

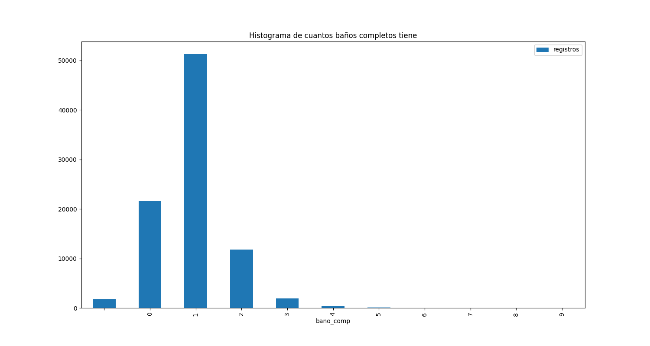
# Análisis descriptivo

Tenemos las siguientes características de las variables seleccionadas y construidas:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Conteo** | **Promedio** | **Min** | **Max** |
| est\_socio | 88,823 | 2.07 | 1 | 4 |
| tipo\_viv | 88,823 | 1.09 | 1 | 5 |
| mat\_pared | 88,823 | 7.78 | 1 | 8 |
| mat\_techos | 88,823 | 8.44 | 1 | 10 |
| mat\_pisos | 88,823 | 2.45 | 1 | 3 |
| antiguedad | 88,823 | 23.75 | 0 | 99 |
| cocina | 88,823 | 1.07 | 1 | 2 |
| cuart\_dorm | 88,823 | 2.07 | 1 | 10 |
| num\_cuarto | 88,823 | 3.75 | 1 | 19 |
| disp\_agua | 88,823 | 1.57 | 1 | 7 |
| excusado | 88,823 | 1.02 | 1 | 2 |
| bano\_comp | 88,823 | 0.95 | 0 | 9 |
| bano\_excus | 88,823 | 0.39 | 0 | 9 |
| bano\_regad | 88,823 | 0.04 | 0 | 9 |
| drenaje | 88,823 | 1.47 | 1 | 5 |
| disp\_elect | 88,823 | 1.05 | 1 | 5 |
| focos\_inca | 88,823 | 0.76 | 0 | 98 |
| focos\_ahor | 88,823 | 6.31 | 0 | 99 |
| combustible | 88,823 | 2.79 | 1 | 6 |
| tipo\_adqui | 88,823 | 2.16 | 1 | 4 |
| tipo\_finan | 88,823 | 4.02 | 1 | 5 |
| calent\_sol | 88,823 | 1.86 | 1 | 2 |
| calent\_gas | 88,823 | 1.66 | 1 | 2 |
| medidor\_luz | 88,823 | 1.07 | 1 | 2 |
| bomba\_agua | 88,823 | 1.74 | 1 | 2 |
| tanque\_gas | 88,823 | 1.90 | 1 | 2 |
| aire\_acond | 88,823 | 1.77 | 1 | 2 |
| calefacc | 88,823 | 1.96 | 1 | 2 |
| tot\_resid | 88,823 | 3.49 | 1 | 20 |
| tot\_hom | 88,823 | 1.68 | 0 | 13 |
| tot\_muj | 88,823 | 1.80 | 0 | 11 |
| tot\_hog | 88,823 | 1.01 | 1 | 5 |
| tam\_loc | 88,823 | 2.52 | 1 | 4 |

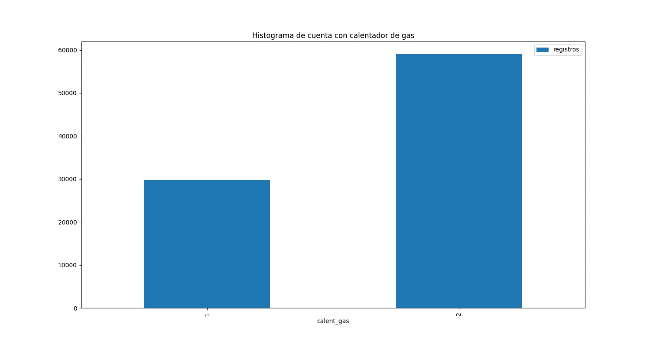
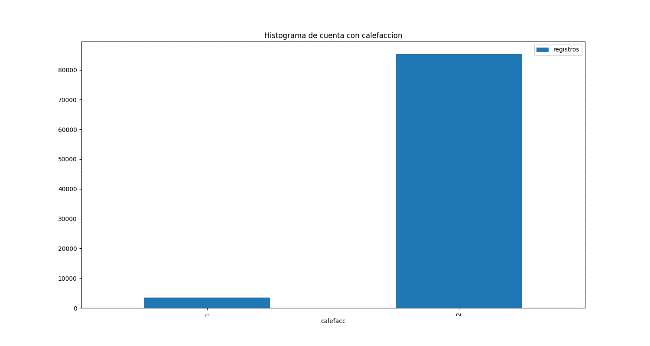
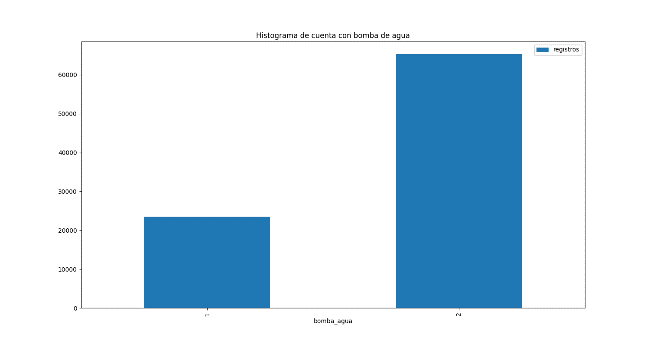
Se muestran algunas de las características de estas variables:

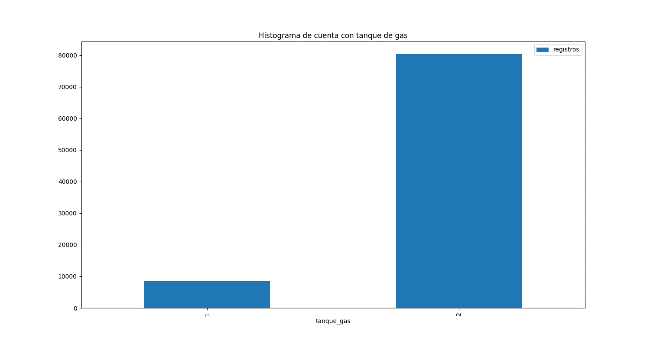
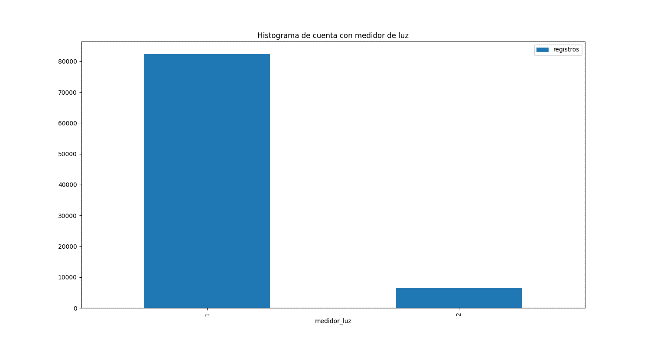
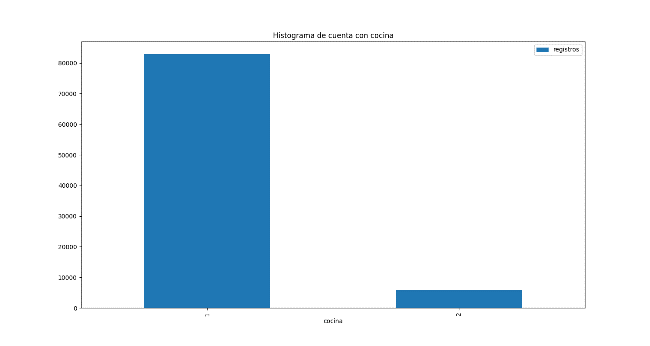
Gráfico, Histograma

Descripción generada automáticamenteGráfico, Histograma

Descripción generada automáticamenteGráfico

Descripción generada automáticamenteGráfico, Gráfico de barras

Descripción generada automáticamenteGráfico

Descripción generada automáticamenteGráfico

Descripción generada automáticamenteGráfico, Histograma

Descripción generada automáticamenteGráfico, Gráfico de barras

Descripción generada automáticamenteGráfico, Histograma

Descripción generada automáticamenteGráfico

Descripción generada automáticamente con confianza bajaGráfico

Descripción generada automáticamenteGráfico

Descripción generada automáticamenteGráfico, Histograma

Descripción generada automáticamenteGráfico, Histograma

Descripción generada automáticamenteImagen que contiene Gráfico

Descripción generada automáticamenteGráfico

Descripción generada automáticamenteGráfico, Histograma

Descripción generada automáticamenteGráfico, Histograma

Descripción generada automáticamenteGráfico, Histograma

Descripción generada automáticamenteGráfico, Gráfico de barras

Descripción generada automáticamenteGráfico

Descripción generada automáticamenteGráfico, Gráfico de barras

Descripción generada automáticamenteGráfico

Descripción generada automáticamenteGráfico

Descripción generada automáticamente

# Clasificación

Se realizan varios ejercicios de clasificación con los diversos algoritmos vistos en la clase, para poder seleccionar el mejor modelo aplicaremos estos algoritmos sobre una base de entrenamiento que consta del 60% de la población y el 40% restante será para medir el desempeño de este en una población no observada anteriormente.

## Resultados

Se corren iterativamente algunos de los algoritmos modificando los parámetros que estos reciben para encontrar la mejor solución, encontrando los siguientes mejores modelos por algoritmo con sus características:

|  |  |  |  |
| --- | --- | --- | --- |
| Algoritmo | parámetros | Accuracy Train | Accuracy Test |
| KNN Classifier | n\_neighbors=6 | 0.7394 | 0.6408 |
| Decision Tree Classifier | depth=11, min\_samples\_split=50, min\_samples\_leaf=50 | 0.7243 | 0.7030 |
| Support Vector Machines | C=1.0 | 0.6752 | 0.6695 |
| Logistic Regression | C=0.9 | 0.6822 | 0.6775 |

Del cuadro anterior podemos seleccionar el modelo de Decision Tree con parámetros de depth=11, min\_samples\_split = 50 y min\_samples\_leaf = 50 como el mejor modelo generado comparando su accuracy en Test.

Confusion matrix del segmento test de los mejores modelos por algoritmo:

KNN con n\_neighbors = 6:

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

Decision Tree con depth=11, min\_samples\_split = 50 y min\_samples\_leaf = 50:

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

Support Vector Machines con C=1.0:

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

Logistic Regression con C=0.9:

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

# Bibliografía

ENIGH. (s.f.). Obtenido de https://www.inegi.org.mx/rnm/index.php/catalog/901/study-description

SimpleImputer. (s.f.). Obtenido de https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html

viviendas. (s.f.). Obtenido de https://www.inegi.org.mx/rnm/index.php/catalog/901/data-dictionary/F68?file\_name=viviendas