Jalisco Connection Points

Esteban Javier Berumen Nieto
ITESO (Instituto Tecnológico y de Estudios Superiores de Occidente)

10 de abril del 2024

Introducción

En este proyecto, se emplea el algoritmo KNN para predecir el ancho de banda contratado en 2014, por un listado de instituciones públicas de Jalisco en el año 2016. El objetivo principal es evaluar la precisión del modelo en función de las características de estas instituciones.

Dado que una predicción precisa del ancho de banda contratado para las instituciones públicas no solo es importante para la gestión eficiente de la infraestructura de TI y la asignación de recursos financieros, sino que también tiene un impacto significativo en la calidad y disponibilidad de los servicios públicos digitales ofrecidos a los ciudadanos.

Se tomó la decisión de utilizar el algoritmo de KNN dado que como veremos más adelante la distrubución de los datos no es normal. Ademas de que con KNN no es nesesario encontrar una correlacion entre las variables y como tambien veremos precticamente no la hay

Los datos fueron obtenidos del sitio Datos Abiertos Jalisco en donde encontramos un dataset con las siguientes variables incluidas: [**clave de inmueble, tecnologia instalada, ancho de banda contratado 2014, institución, nombre del centro, turno/horario, nivel, región, municipio, localidad, domicilio, código postal, longitud, latitud**] en donde encontramos 6716 registros.

CLAVE DE INMUEBLE	TECNOLOGIA INSTALADA	ANCHO DE BANDA CONTRATADO 2014	INSTITUCION	NOMBRE DI
1401423H	iMAX	2 MB	SEJ	VENUSTIAN
1401423H	iMAX	2 MB	SEJ	USAER
1401423H	iMAX	2 MB	SEJ	ADOLFO LO
1401425E	iMAX	2 MB	SEJ	MA GUADAL
1401426X	iMAX	2 MB	SEJ	VALENTIN C
1401427B	iMAX	2 MB	SEJ	20 DE OVIEI
1401427B	iMAX	2 MB	SEJ	SOR JUANA
1401428F	iMAX	2 MB	SEJ	PREPARATC
1401428F	iMAX	2 MB	SEJ	ESCUELA P
1401428F	iMAX	2 MB	SEJ	PREPARATC

Desarrollo

Dentro del preprocesamiento de datos se realizó una limpieza en donde todos los datos faltantes y N/D del dataset fueron remplazados por la primera moda de la columna correspondiente, esto ya que todas las columnas del dataset son categóricas, esto a su vez hizo que fuera necesario usar el ***OrdinalEncoder*** de la librería ***sklearn***, para poder realizar la codificación de las variables, además se dividió el dataset en un train y un test en donde el train es el **20%** del dataset original y el test es el **80%**.

Train data 📊

CLAVE DE INMUEBLE	TECNOLOGIA INSTALADA	ANCHO DE BANDA CONTRATADO 2014	INSTITUCION	NOMBRE DI
1409096A	iMAX	2 MB	SEJ	FERNANDO
SC_189	ADSL	2 MB	SC	SAMARTIN
JCSSA007643	ADSL	10 MB	SSJ	ZAPOTLANI
1409248A	ADSL	100 MB	SEJ	LAZARO CA
DIF_MUN_066	ADSL	10 MB	DIF	N/D
177IJJ	ADSL	2 MB	GOB	ESPACIO P
1404097H	iMAX	2 MB	SEJ	VICENTE SU
1409435K	iMAX	2 MB	SEJ	TENAMAZT
1409079X	iMAX	2 MB	SEJ	FRANCISCO
1404431X	iMAX	2 MB	SEJ	NETZAHUA

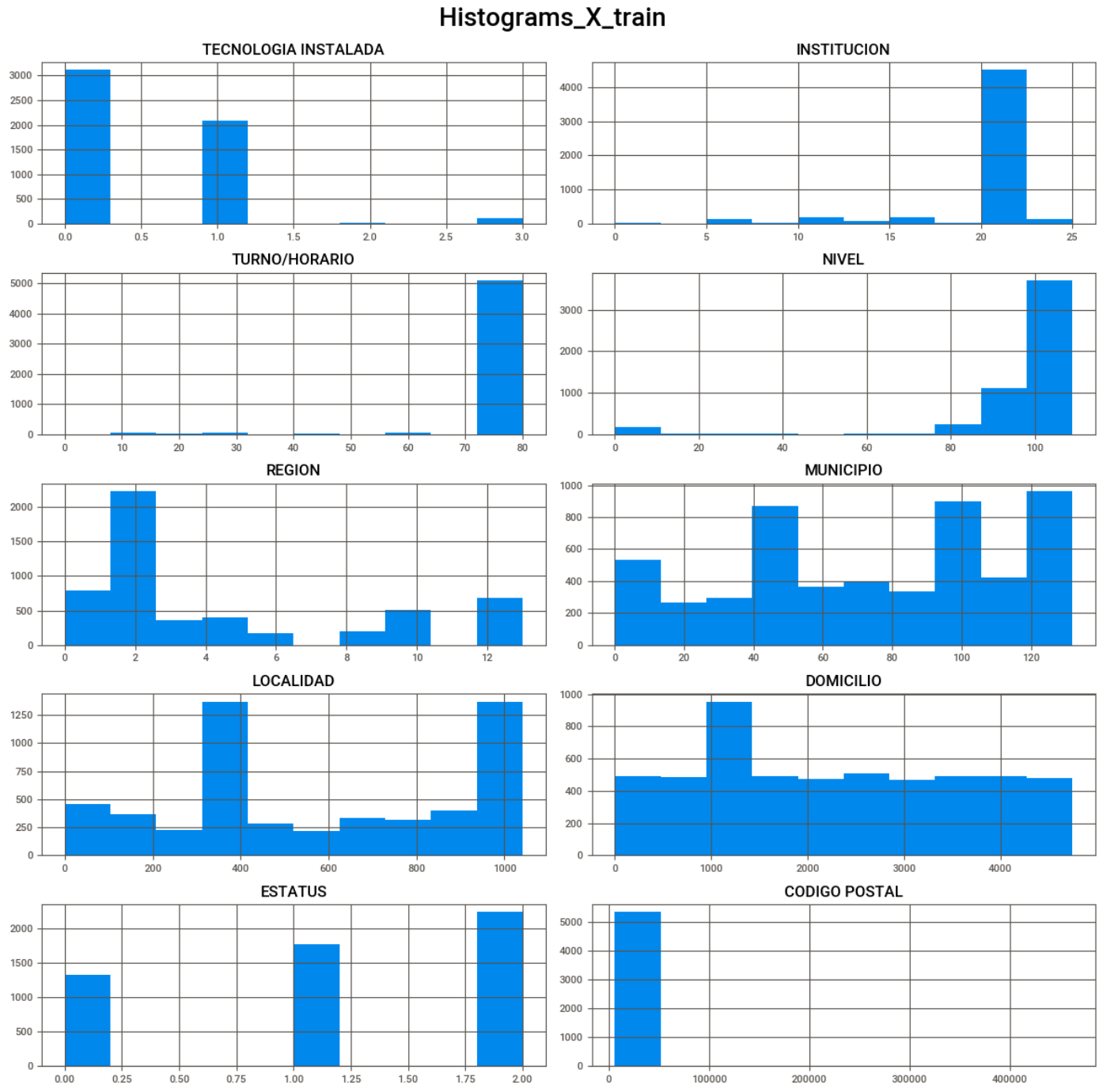
Test data 📊

CLAVE DE INMUEBLE	TECNOLOGIA INSTALADA	ANCHO DE BANDA CONTRATADO 2014	INSTITUCION	NOMBRE DI
GOB_AHUALULCO	iMAX	2 MB	GOB	PRESIDENC
SSJ_400	ADSL	2 MB	SSJ	TEPEHUAJI
1404880H	ADSL	10 MB	SEJ	FRAY JUAN
1404259J	iMAX	2 MB	SEJ	NIÑOS HER
1410827C	iMAX	2 MB	SEJ	RAFAEL JIM
1409144F	ADSL	2 MB	SEJ	BENITO JU/
1410694C	iMAX	2 MB	SEJ	LAZARO CA
1410974B	iMAX	2 MB	SEJ	CUAUHTEM
1405880C	iMAX	2 MB	SEJ	PATRIA
SSJ_239	iMAX	2 MB	SSJ	C.S.R. BUEN

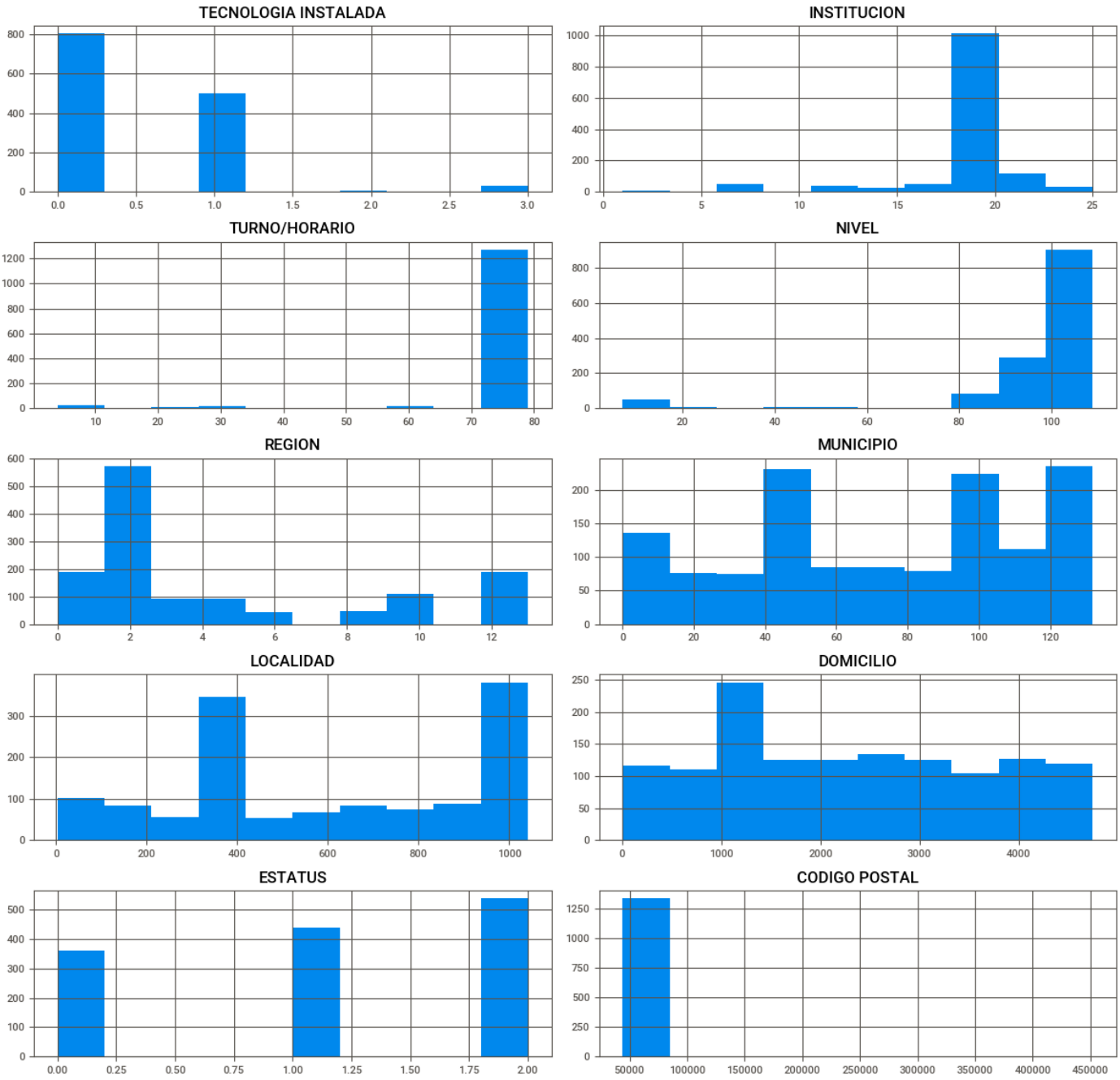
A continuación, se eliminaron algunas columnas debido a diferentes problemas u objetivos; tales son las columnas de: **clave de inmueble, ancho de banda contratado 2014, nombre del centro, longitud,**

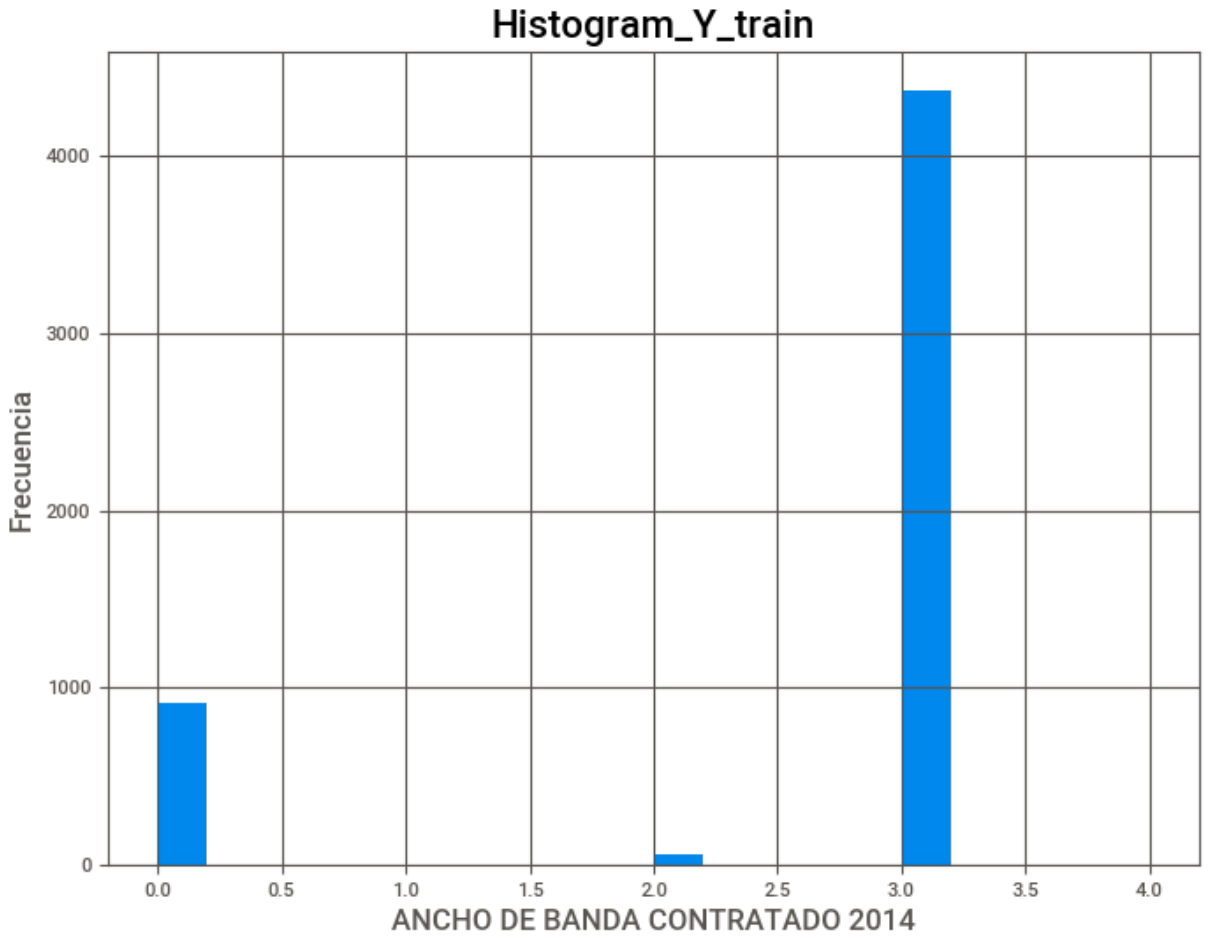
latitud. En el caso del ancho de banda, esta columna se convierte en el target. También se realizaron los histogramas de las diferentes columnas, así como un mapa de correlación de las variables.

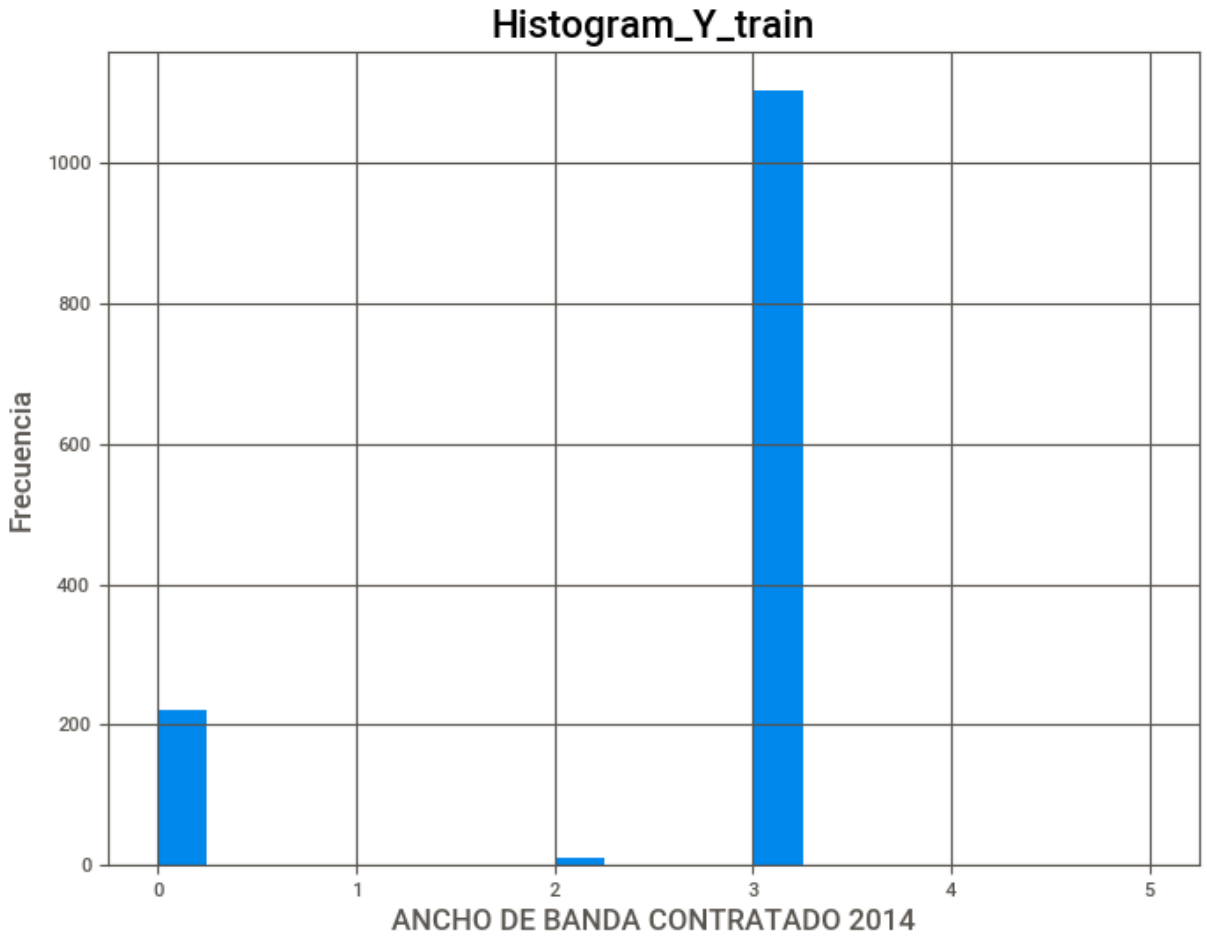
En los histogramas veremos el eje y como la frecuciencia y el eje x como la variable ya codificada



Histograms_X_test

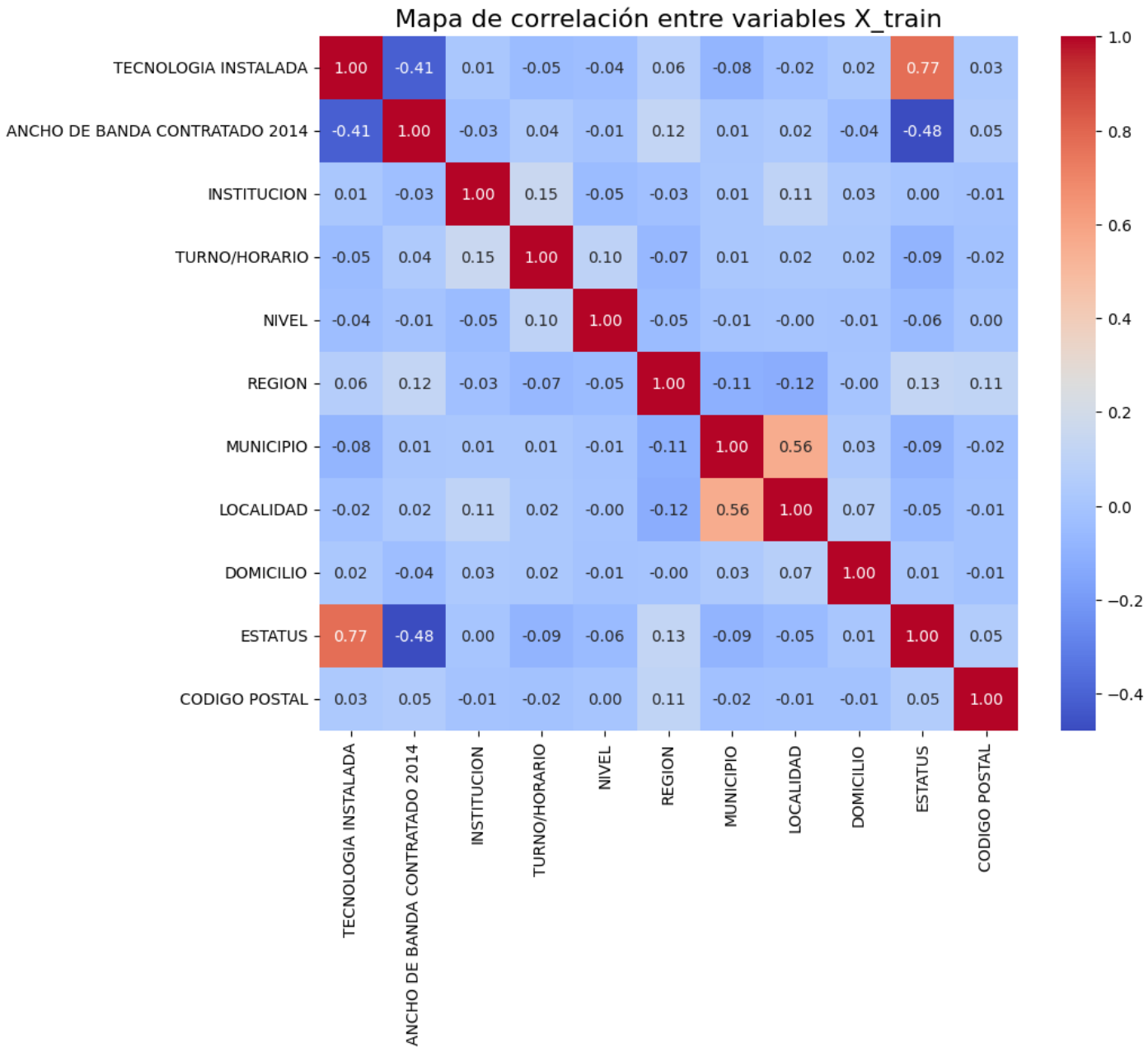


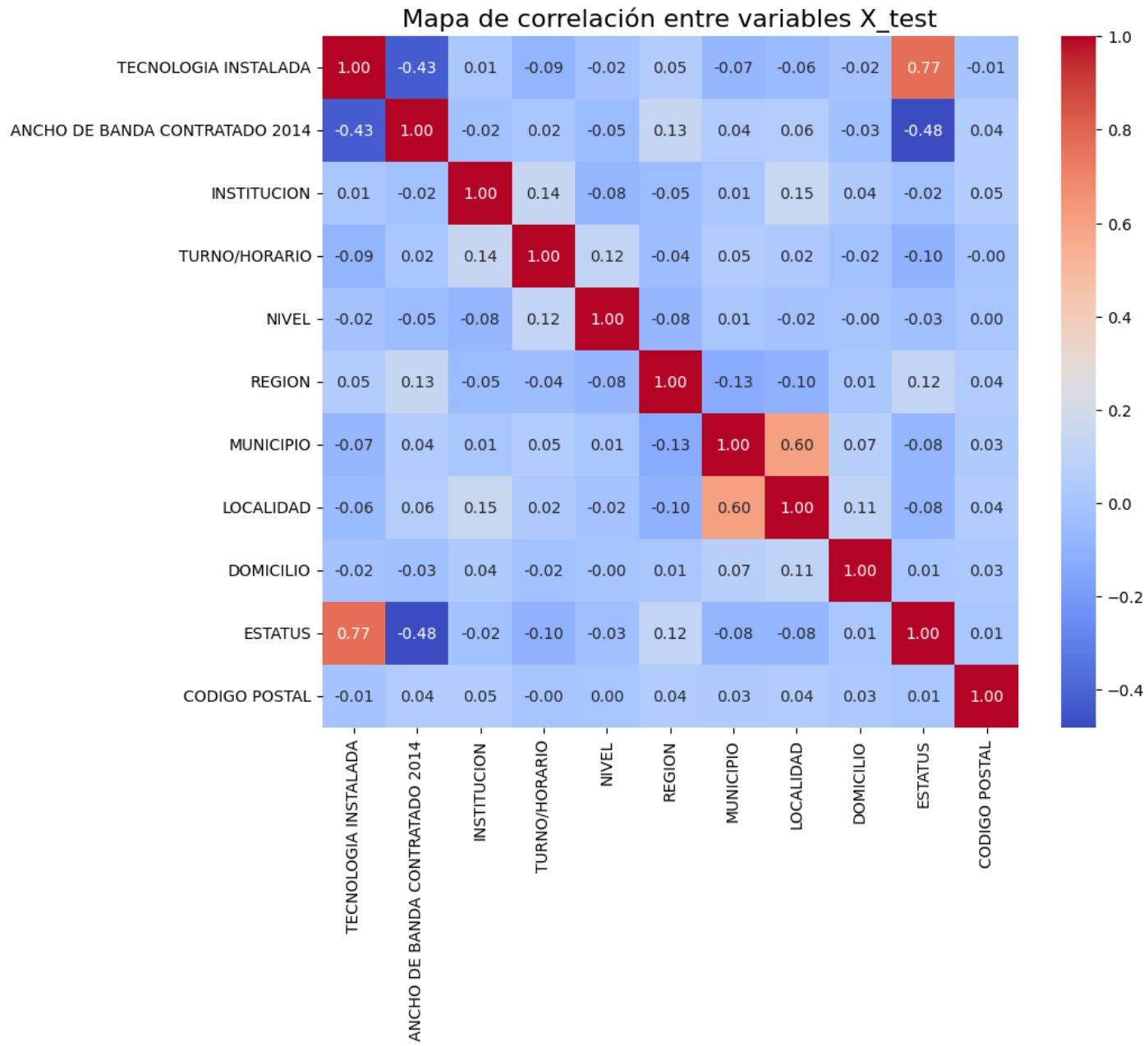

Histogram_Y_train
Frecuencia
4000
3000
2000
1000
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
ANCHO DE BANDA CONTRATADO 2014



Como ya mencionó anteriormente en ninguno de los features encontramos una distribución normal, gracias a esto, tambien podemos ver que en entre el train y el test hay una gran simetría.

Algunas otras cosas que podemos ver en algunos features específicos son: La mayoría de los regitros de horario corresponde a solo un horario (Matutino)


Mapa de correlación entre variables X_train
TECNOLOGIA INSTALADA
ANCHO DE BANDA CONTRATADO 2014
INSTITUCION
TURNO/HORARIO
NIVEL
REGION
MUNICIPIO
LOCALIDAD
DOMICILIO
ESTATUS
CODIGO POSTAL
1.00 -0.41 0.01 -0.05 -0.04 0.06 -0.08 -0.02 0.02 0.77 0.03
-0.41 1.00 -0.03 0.04 -0.01 0.12 0.01 0.02 -0.04 -0.48 0.05
0.01 -0.03 1.00 0.15 -0.05 -0.03 0.01 0.11 0.03 0.00 -0.01
-0.05 0.04 0.15 1.00 0.10 -0.07 0.01 0.02 0.02 -0.09 -0.02
-0.04 -0.01 -0.05 0.10 1.00 -0.05 -0.01 -0.00 -0.01 -0.06 0.00
0.06 0.12 -0.03 -0.07 -0.05 1.00 -0.11 -0.12 -0.00 0.13 0.11
-0.08 0.01 0.01 0.01 -0.01 -0.11 1.00 0.56 0.03 -0.09 -0.02
-0.02 0.02 0.11 0.02 -0.00 -0.12 0.56 1.00 0.07 -0.05 -0.01
0.02 -0.04 0.03 0.02 -0.01 -0.00 0.03 0.07 1.00 0.01 -0.01
0.77 -0.48 0.00 -0.09 -0.06 0.13 -0.09 -0.05 0.01 1.00 0.05
0.03 0.05 -0.01 -0.02 0.00 0.11 -0.02 -0.01 -0.01 0.05 1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4



Se implementó el algoritmo KNN en el cual se probó con diferentes k para poder establecer cuál sería el número de vecinos mediante la prueba del codo.

Número de vecinos (k)

23

1 30

Precisión del modelo KNN con **23** vecinos: **0.84869**

Generar gráfica

Resultados

Visto lo anterior, podemos decir que el algoritmo llega a su mejor precisión con 23 vecinos en donde nos un ofrece 84.8% de precisión

```
from sklearn.neighbors  import KNeighborsClassifier
clf = KNeighborsClassifier(n_neighbors= 23 )
clf.fit(X_train,y_train)
clf.score(X_test,y_test)
```

0.846441947565543

Conclusiones:

Nuestro análisis de predicción del ancho de banda para instituciones públicas en Jalisco utilizando el algoritmo KNN ha arrojado resultados prometedores. Logramos obtener una precisión del 84.8% al utilizar 23 vecinos en el modelo. Esta precisión nos brinda confianza en la capacidad del modelo para predecir de manera efectiva el ancho de banda necesario en base a las características proporcionadas.

Al explorar las relaciones entre las diferentes variables, observamos que la mayoría de ellas tienen una correlación baja entre sí. Sin embargo, identificamos que las variables de tecnología instalada y estatus muestran una correlación significativa con el ancho de banda contratado. Esto sugiere que estas dos características pueden ser factores determinantes en la cantidad de ancho de banda requerido por una institución.

En resumen, aunque nuestras características no muestran una correlación fuerte entre sí, hemos encontrado que la tecnología instalada y el estatus son variables importantes a considerar al predecir el ancho de banda necesario. Esto destaca la importancia de tener en cuenta no solo la cantidad de ancho de banda disponible, sino también el tipo de tecnología utilizada y el estatus de la institución al planificar la infraestructura de red.

Referencias

Secretaría de Educación Pública de Jalisco. (s/f). Puntos de Conexión eJalisco [Conjunto de datos]. Recuperado de https://datos.jalisco.gob.mx/dataset/puntos-de-conexion-ejalisco
scikit-learn. (s/f). Nearest Neighbors. Recuperado de https://scikit-learn.org/stable/modules/neighbors.html