Jalisco Connection Points

Proyecto Programación para Análisis de Datos

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

Introducción

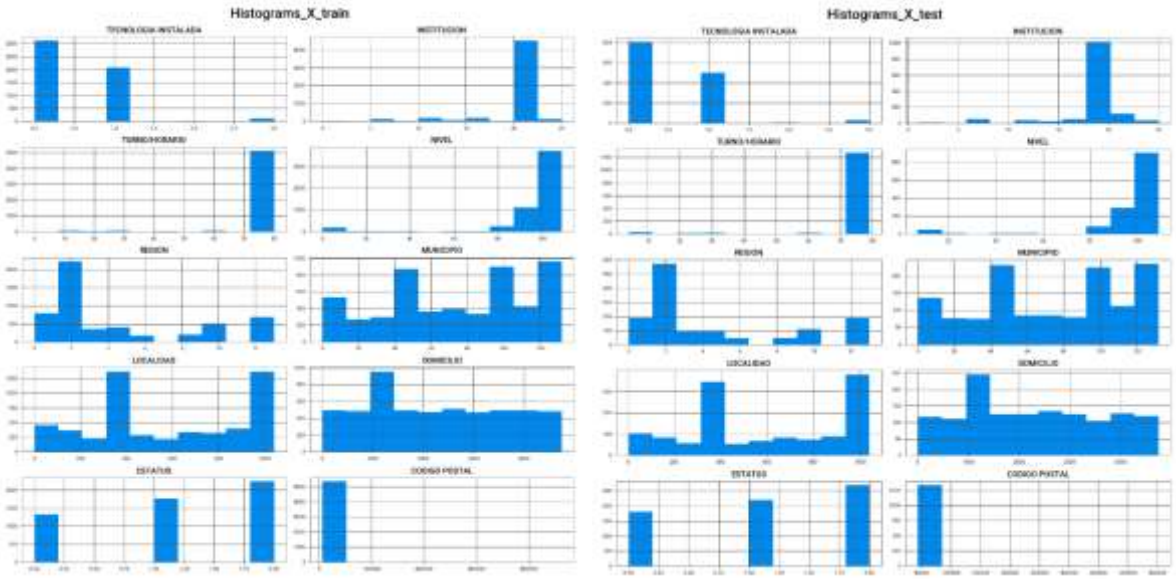
En este proyecto, se emplea el algoritmo KNN para predecir el ancho de banda contratado en 2014, por un listado de instituciones públicas de Jalisco en el año 2016. El objetivo principal es evaluar la precisión del modelo en función de las características de estas instituciones.

Los datos fueron obtenidos del sitio https://datos.jalisco.gob.mx/dataset/puntos-de-conexion-ejalisco en donde encontramos un dataset con las siguientes variables incluidas: **[clave de inmueble, tecnologia instalada, ancho de banda contratado 2014, institucion, nombre del centro, turno/horario, nivel, region, municipio, localidad, domicilio, codigo postal, longitud, latitud]** en donde encontramos 6716 registros.

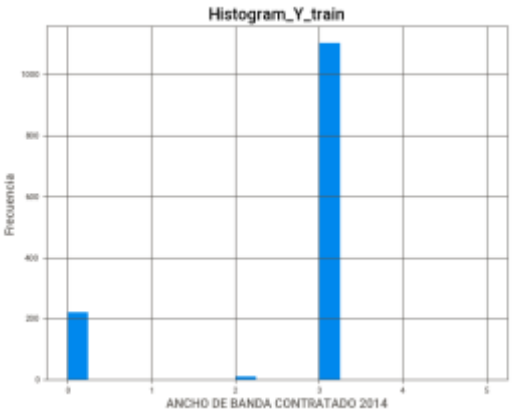
Desarrollo

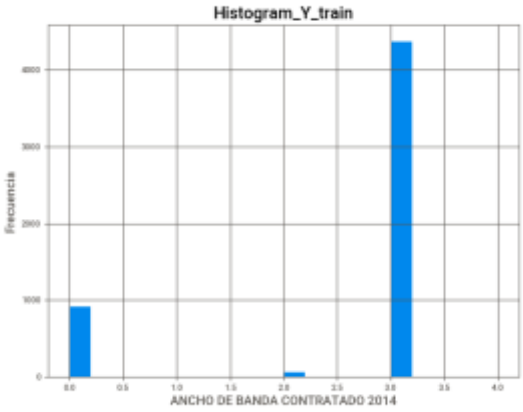
Dentro del preprocesamiento de datos se realizó una limpieza en donde todos los datos faltantes y N/D del dataset fueron remplazados por la primera moda de la columna correspondiente, esto ya que todas las columnas del dataset son categóricas, esto a su vez hiso que fuera necesario usar el *OrdinalEncoder*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

A continuación, se eliminaron algunas columnas debido de diferentes problemas u objetivos; como lo 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.

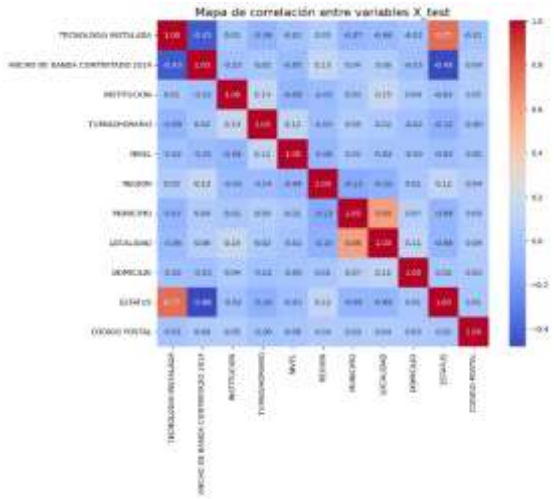


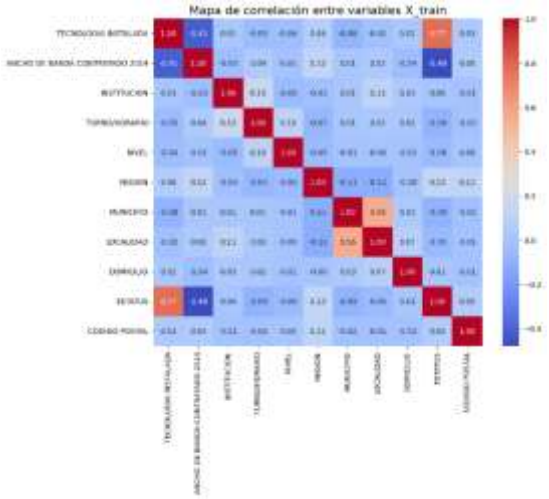
También se realizaron histogramas para el traget tanto en la parte del train como en la parte del test.





Mapa de correlación de variables.





Se implemento 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.

```
k_values = [x for x in range(1,50,2)]
scores = []
for k in k_values:
    clf = KNeighborsClassifier(n_neighbors= k )
    clf.fit(X_train,y_train)
    acc = clf.score(X_test,y_test)
    scores.append(acc)
```

```
plt.figure(figsize=(10, 6))
plt.plot(k_values, scores, marker='o')
plt.title('Scores de KNN para diferentes valores de K')
plt.xlabel('Número de vecinos (K)')
plt.ylabel('Accuracy')
plt.xticks(k_values)
plt.grid(True)
plt.savefig('../reports/figures/knn_scores.png')
plt.show()
```

Resultados

Visto lo anterior podemos decir que el algoritmo llega su mejor precisión con 21 vecinos en donde nos un 84.5% de precisión

```
clf = KNeighborsClassifier(n_neighbors= 21 )
clf.fit(X_train,y_train)
clf.score(X_test,y_test)
0.846441947565543
```

Conclusiones

Al intentar predecir el ancho de banda que debería de tener una institución pública en Jalisco basándonos en diferentes características, el algoritmo knn nos da una precisión de hasta el 84.5% utilizando 21 vecinos, además de esto como vimos en los mapas de correlación la mayoría de las variables no tiene no gran correlación entre si, sin embargo vemos que las que tiene más correlación con el ancho de banda son la tecnología instalada y el estatus, por lo que podríamos decir que ancho de banda depende u poco mas de estos que de las demás características.

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