Prueba de conocimiento VIVI: Predicción precio vivienda

Lina María Ortiz Parra

Contenidos

Limpieza y manipulación de datos

- De los 222 características que se tenían de cada vivienda escogi las 20 más relevantes para la predicción del precio de una nueva vivienda.

tipo_avaluo	object
sector	object
alcantarillado_en_el_sector	object
acueducto_en_el_sector	object
gas_en_el_sector	object
energia_en_el_sector	object
estrato	object
demanda_interes	object
nivel_equipamiento_comercial	object
habitaciones	object
sala	object
bano_privado	object
total_cupos_parquedaro	object
numero_total_depositos	object
valor_total_avaluo	object
cocina	int64
zona_de_ropas	int64
Longitud	float64
Latitud	float64
area_privada	object
valor_area_privada	object

Limpieza y manipulación de datos

- Se eliminan los datos que son NaN.
- Se corrigen los datos que tienen caracteres especiales.
- Elimino los datos que no corresponden a la categoria en que se encuentran.
- Datos que son valores como valor total avaluo, entre otros son de tipo object, los cambio a float.
- Codificación de variables categóricas mediante dummies, la cual una variable artificial creada para representar un atributo con dos o más categorías.
  tipo avaluo, sector, alcantarillado en el sector, acueducto en el sector, gas en el sector, energia en el sector, demanda interes, nivel equipamiento comercial

Técnicas usadas

Se creo tres modelos diferentes para la predicción de datos

1. **Regresión lineal:** predice el valor del avaluo en función de las características independientes dadas.
2. **Random Forest (Bosque Aleatorio):** Un Random Forest es un conjunto de árboles de decisión combinados con bagging. Al usar bagging, lo que en realidad está pasando, es que distintos árboles ven distintas porciones de los datos. Ningún árbol ve todos los datos de entrenamiento. Esto hace que cada árbol se entrene con distintas muestras de datos para un mismo problema.
3. **Máquina de vectores de soporte (SVM):** Es un algoritmo de aprendizaje supervisado que se utiliza en problemas de clasificación y regresión

Resultados y análisis

- Para realizar los modelos se divide la información, una parte para entrenamiento y otra para pruebas. En este caso se utilizo 80% de la información para entrenar y 20% para prueba.
- Para cada modelo se calculo la raíz del error cuadrático medio (RMSD)

Modelo	RMSD
Regresión lineal	848371835.5960336
Bosque Aleatorio	1079204.668367289
SVM	37961456.071260445

- En el modelo que se obtuvo mejor resultado fue en Bosque aleatorio.

Dificultades encontradas

- Las características de la vivienda que eran palabras tenían caracteres especiales, los cuales habían que limpiar antes de realizar cualquier clasificación
- No se podían realizar un análisi con todas las variables ya que tomaba mucho tiempo en procesarlo.
- El documetno tenía muchos datos con NaN por lo que se tuvieron que eliminar para que no generara errores.