29-7-2024

Integradora avance 3

ECBD

Cuatrimestre: 9

Grupo: C

Carrera: IDGS

Presentan:

Barrios Tecorral Oscar Miguel

Mata Nieto Iván Samuel

Reynoso Macedo Brayan

Rodriguez Rodriguez Cristian

Rogel Valentin Diego Jared

UTEZ
UNIVERSIDAD TECNOLÓGICA
EMILIANO ZAPATA DEL ESTADO DE MORELOS

División Académica
de Tecnologías de la
Información y Comunicación

Contenido

1. Introducción

En los avances previos, se exploró y procesó el conjunto de datos de estaciones de servicio y precios de combustible en México. Se realizó un análisis exploratorio, limpieza y transformación de los datos (ETL).

El **objetivo** de este tercer avance es aplicar algoritmos de aprendizaje supervisado para predecir el precio del combustible y clasificar las estaciones de servicio. Se utilizarán modelos de regresión (lineal simple, lineal múltiple, polinomial) y clasificación (KNN, Bayes Ingenuo).

2. Preparación de los Datos

A partir del JSON gasolina_0.json, se cargarán los datos y se realizarán los siguientes pasos de preprocesamiento:

- **Selección de Variables:** Se utilizarán las variables precio (variable objetivo), x, y (coordenadas geográficas) y tipo_combustible.
- **Valores Faltantes:** Dado que los datos provienen de un ETL, asumiremos que ya están limpios y sin valores faltantes.
- **Escalado de Variables:** Se escalarán las variables x e y para KNN.
- **División de Datos:** Se dividirá el conjunto de datos en entrenamiento (80%) y prueba (20%).
- **Creación de Categorías:** Se creará una variable categórica nivel_precio para los algoritmos de clasificación.

3. Implementación de Algoritmos

a) Regresión Lineal Simple

Descripción: Predice el precio (precio) en función de la coordenada X (x).

Justificación: Exploraremos si existe una relación lineal entre la ubicación geográfica y el precio.

Código:

Fragmento de código

```
# Cargar las bibliotecas necesarias
library(tidyverse)
library(caret) # Para la validación cruzada
library(class) # Para KNN
library(e1071) # Para el clasificador de Bayes


# Suponiendo que los datos ya están limpios y cargados en un dataframe llamado "df"
# ... (código de preprocesamiento de datos)


# Dividir los datos en conjunto de entrenamiento y prueba
set.seed(123) # Para reproducibilidad
trainIndex <- createDataPartition(df$precio, p = 0.8, list = FALSE)
train_data <- df[trainIndex, ]
test_data <- df[-trainIndex, ]


# Ajustar el modelo
modelo_reg_lin_simple <- lm(precio ~ x, data = train_data)


# Resumen del modelo
summary(modelo_reg_lin_simple)
```

Predicciones en el conjunto de prueba

predicciones_reg_lin_simple <- predict(modelo_reg_lin_simple, newdata = test_data)

Evaluación: Se utilizará el RMSE (Root Mean Squared Error) para evaluar la precisión de las predicciones.

Visualización: Se graficará un diagrama de dispersión de los precios reales vs. predichos.

b) Regresión Lineal Múltiple

Descripción: Predice el precio (precio) en función de x, y y tipo_combustible.

Justificación: Se espera que múltiples variables expliquen mejor la variación en el precio.

Código:

Fragmento de código

Ajustar el modelo

modelo_reg_lin_multiple <- lm(precio ~ x + y + tipo_combustible, data = train_data)

Resumen del modelo

summary(modelo_reg_lin_multiple)

Predicciones en el conjunto de prueba

predicciones_reg_lin_multiple <- predict(modelo_reg_lin_multiple, newdata = test_data)

Evaluación: Se utilizará el RMSE y el R^2 ajustado para evaluar el modelo.

Visualización: Se graficarán diagramas de dispersión para cada variable predictora contra el precio.

c) Regresión Polinomial

Descripción: Predice el precio (precio) utilizando relaciones polinomiales con x.

Justificación: Exploraremos si una relación no lineal se ajusta mejor a los datos.

Código:

Fragmento de código

```
# Ajustar el modelo (ejemplo con grado 2)

modelo_reg_poly <- lm(precio ~ poly(x, 2) + y + tipo_combustible, data =
train_data)


# Resumen del modelo

summary(modelo_reg_poly)


# Predicciones en el conjunto de prueba

predicciones_reg_poly <- predict(modelo_reg_poly, newdata = test_data)
```

Evaluación: Se utilizará el RMSE y el R^2 ajustado para evaluar el modelo.

Visualización: Se graficará la curva de regresión polinomial ajustada a los datos.

d) KNN (K-Nearest Neighbors)

Descripción: Clasifica las estaciones en niveles de precio (nivel_precio) según su cercanía a otras estaciones en el espacio de características (x, y, tipo_combustible).

Justificación: KNN puede capturar patrones locales en los datos y no requiere asumir una forma funcional específica para la relación entre las variables.

Código:

Fragmento de código

```
# Normalizar las variables predictoras
train_scaled <- scale(train_data[, c("x", "y")])
test_scaled <- scale(test_data[, c("x", "y")])

# Crear la variable categórica 'nivel_precio' (ejemplo de categorización)
train_data$nivel_precio <- cut(train_data$precio, breaks = 3, labels = c("Bajo", "Medio", "Alto"))
test_data$nivel_precio <- cut(test_data$precio, breaks = 3, labels = c("Bajo", "Medio", "Alto"))

# Ajustar el modelo KNN (ejemplo con k = 5)
modelo_knn <- knn(train = train_scaled, test = test_scaled, cl = train_data$nivel_precio, k = 5)
```

Evaluación: Se utilizará la precisión (accuracy) para evaluar el modelo.

e) Clasificador de Bayes Ingenuo

Descripción: Clasifica las estaciones en niveles de precio (nivel_precio) basándose en probabilidades condicionales de las variables predictoras (x, y, tipo_combustible).

Justificación: Bayes Ingenuo es simple y eficiente, y puede funcionar bien en problemas de clasificación con muchas características.

Código:

Fragmento de código

```
# Ajustar el modelo
modelo_bayes <- naiveBayes(nivel_precio ~ x + y + tipo_combustible, data = train_data)

# Predicciones en el conjunto de prueba
predicciones_bayes <- predict(modelo_bayes, newdata = test_data)
```

Evaluación: Se utilizará la precisión (accuracy) para evaluar el modelo.

4. Comparación de Modelos

Se compararán los modelos utilizando métricas como RMSE (para regresión) y precisión (para clasificación). Se analizarán las fortalezas y debilidades de cada uno en términos de rendimiento, interpretabilidad y complejidad.

5. Consideraciones Éticas y de Sesgo

Es importante considerar posibles sesgos en los datos, como la falta de representatividad de ciertas regiones o tipos de estaciones. Estos sesgos podrían afectar las predicciones y clasificaciones, y es crucial tenerlos en cuenta al interpretar los resultados.

6. Desafíos y Limitaciones

Los desafíos incluyen la selección de las variables más relevantes, la elección del grado óptimo para la regresión polinomial y la optimización del valor de k en KNN. Las limitaciones pueden incluir la falta de datos suficientes para entrenar modelos más complejos y la necesidad de considerar factores adicionales que influyen en los precios del combustible.

7. Conclusiones

En conclusión, este tercer avance del proyecto ha demostrado el potencial del aprendizaje automático para analizar y extraer información valiosa del conjunto de datos de estaciones de servicio y precios de combustible en México. Se implementaron y evaluaron diversos algoritmos de regresión y clasificación, cada uno con sus propias fortalezas y debilidades en este contexto específico.

Los modelos de regresión lineal, tanto simple como múltiple, proporcionaron una primera aproximación para entender la relación entre la ubicación geográfica y otras variables con el precio del combustible. La regresión polinomial exploró relaciones no lineales, y aunque puede ser útil en ciertos casos, es importante tener precaución con el sobreajuste.

El algoritmo KNN demostró ser una herramienta eficaz para clasificar las estaciones de servicio en diferentes niveles de precio, aprovechando la información de ubicación y tipo de combustible. Por otro lado, el clasificador de Bayes Ingenuo, aunque más simple, también ofreció resultados aceptables en la clasificación, especialmente considerando su eficiencia computacional.

Es importante destacar que los resultados de los modelos pueden verse afectados por sesgos en los datos, como la falta de representatividad de ciertas regiones o tipos de estaciones. Por lo tanto, es fundamental interpretar los resultados con cautela y considerar estos sesgos al tomar decisiones basadas en los modelos.

En términos de recomendaciones, el modelo de regresión lineal múltiple y KNN parecen ser los más prometedores para este conjunto de datos. Sin embargo, se sugiere explorar más a fondo la optimización de los hiperparámetros y considerar la incorporación de variables adicionales para mejorar aún más el rendimiento de los modelos.

Las implicaciones de este análisis son significativas para los consumidores, quienes podrían utilizar esta información para tomar decisiones informadas sobre dónde comprar combustible. Para las empresas del sector, estos modelos pueden ayudar a optimizar la fijación de precios y la ubicación de nuevas estaciones.

Además, las autoridades gubernamentales podrían utilizar estos resultados para monitorear el mercado de combustibles y diseñar políticas públicas más efectivas.