Proyecto 1 Etapa1

Ministerio de Comercio, Industria y Turismo de Colombia, la Asociación Hotelera y Turística de Colombia – COTELCO

Maria Alejandra Estrada Garcia -202021060
Santiago Martínez Novoa - 202112020
Marilyn Stephany Joven Fonseca - 202021346

Entendimiento y enfoque analítico

Part 01

El problema

El negocio

Mejorar la promoción y gestión turística en Colombia mediante el análisis de las características de los sitios turísticos para identificar qué los hace atractivos o poco recomendados. Esto permitirá tomar decisiones informadas para aumentar la popularidad de los destinos y fomentar el turismo en el país.

Objetivos

1. Identificar las características que hacen atractivo un hotel sobre el resto y como estas afectan la opinión y la probabilidad de recomendación de los usuarios.
2. Diseñar un modelo de aprendizaje automático que, dado una reseña de un usuario, permita a las cadenas hotelera y a COTELCO clasificarlo como relacionado a una calificación del hotel en una escala entera de 1 a 5.

Enfoque analítico

Para este proyecto de procesamiento de lenguaje natural (NLP), nos centraremos en el enfoque basado en inteligencia artificial (IA) para analizar las características de los sitios turísticos y comparar su atractivo para los turistas. El objetivo principal del negocio es identificar oportunidades de mejora en el turismo mediante técnicas de aprendizaje automático.

SVM

Las SVM son algoritmos de aprendizaje supervisado utilizados principalmente para la clasificación. Son eficaces en la clasificación de textos, como las reseñas de hoteles, debido a su capacidad para manejar eficazmente conjuntos de datos de alta dimensionalidad

Random Forest

Son un conjunto de árboles de decisión que se entrenan con diferentes subconjuntos del conjunto de datos y luego combinan sus resultados para realizar predicciones.

Logistic Regression

La regresión logística se utiliza comúnmente para problemas de clasificación binaria, como determinar si una reseña es positiva o negativa. Es un modelo simple pero efectivo que estima la probabilidad de que una instancia pertenezca a una clase particular.

Entendimiento y preparación de los datos

Part 02

Decisiones

de transformación de los datos

1. Eliminación de duplicados:

Se eliminaron los datos duplicados.

2. Identificación del idioma:

Se examinaron los diferentes idiomas utilizados y se dejaron reseñas unicamente en español

3. Limpieza de caracteres

Se boraron puntos, comas, tildes, otros simbolos y se dejó todo el texto en minuscula.

4. Quitar palabras insignificantes:

Se eliminaron las palabras que no traen ningun significado, también conocidas como palabras vacías.

Decisiones

de transformación de los datos

5. Verificación del vocabulario

Se obtuvo un conteo de las palabras y se eliminaron aquellas que solo aparecen una vez.

6. Normalización

Se realizó Stemming y Lematización para reducir el vocabulario a una raíz de la palaba

7. Vectorización de texto

Se convierte el texto en vectores númericos para el respectivo análisis

8. Balanceo de clases

Se tomó una cantidad equitativa para cada valor de la columna Class.

Modelado y evaluación

Part 03

Métricas de evaluación

Evaluación cuantitativa

Accuracy:

Esta métrica indica la proporción o porcentaje de registros que fueron clasificados correctamente por el modelo.

Precisión:

Este enfoque calcula un promedio de la precisión del modelo para cada categoría de clasificación.

Recall o sensibilidad:

Se refiere a la tasa de "verdaderos positivos", es decir, qué proporción de los textos fueron correctamente asignados a su categoría correspondiente de calificación obtenido.

F1 Score:

Teniendo en cuenta que la precisión y la sensibilidad son métricas complementarias, el puntaje F1 es un promedio ponderado de ambas métricas.

Evaluación Cuantitativa

Precisión: 49.87%
Accuracy: 50.77%
Recall: 50.77%
F1 Score: 49.66%

Support Vector Machine SVM

Precisión: 35.76%
Accuracy: 37.99%
Recall: 37.99%
F1 Score: 34.46%

Random Forest

Precisión: 48.85%
Accuracy: 49.31%
Recall: 49.31%
F1 Score: 49.03%

Interpretation of findings

Resultados

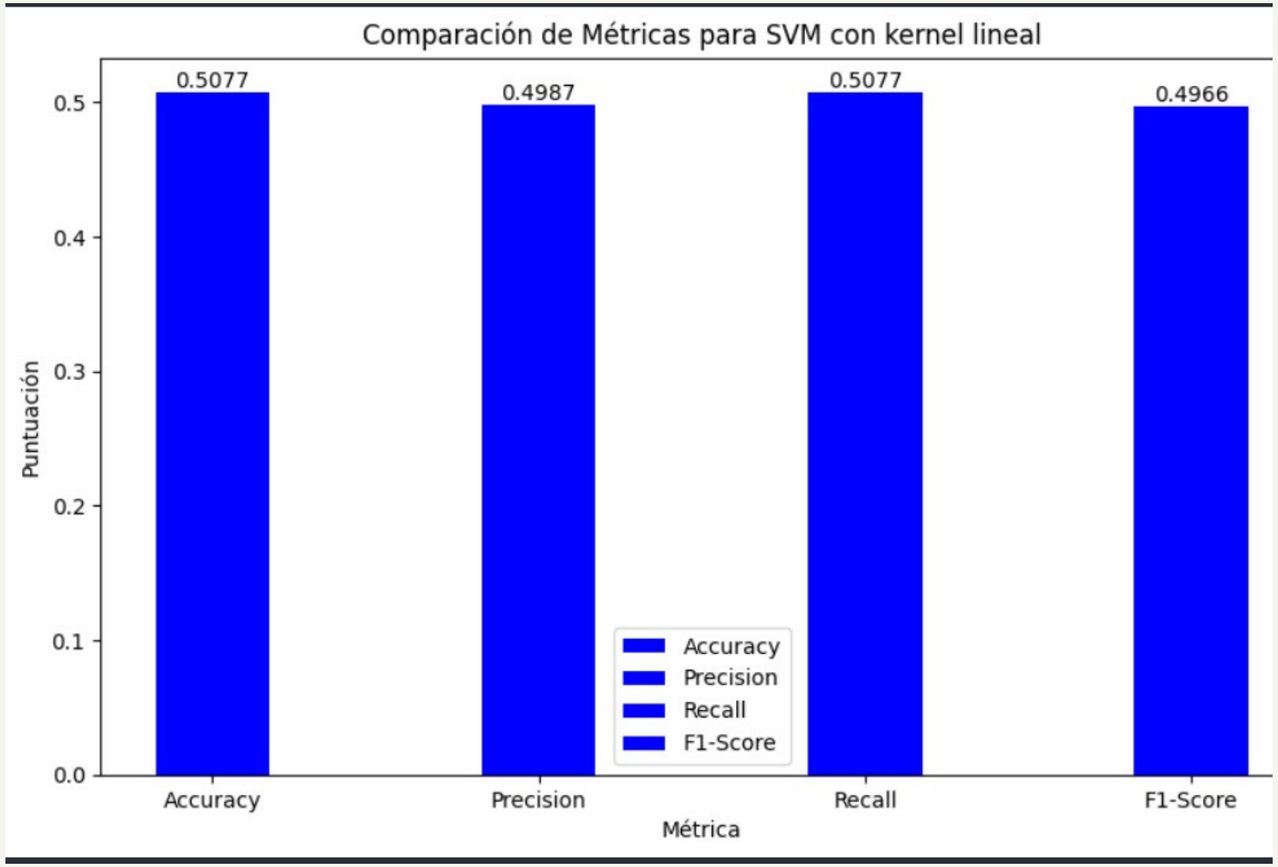
Part 04

Evaluación cualitativa

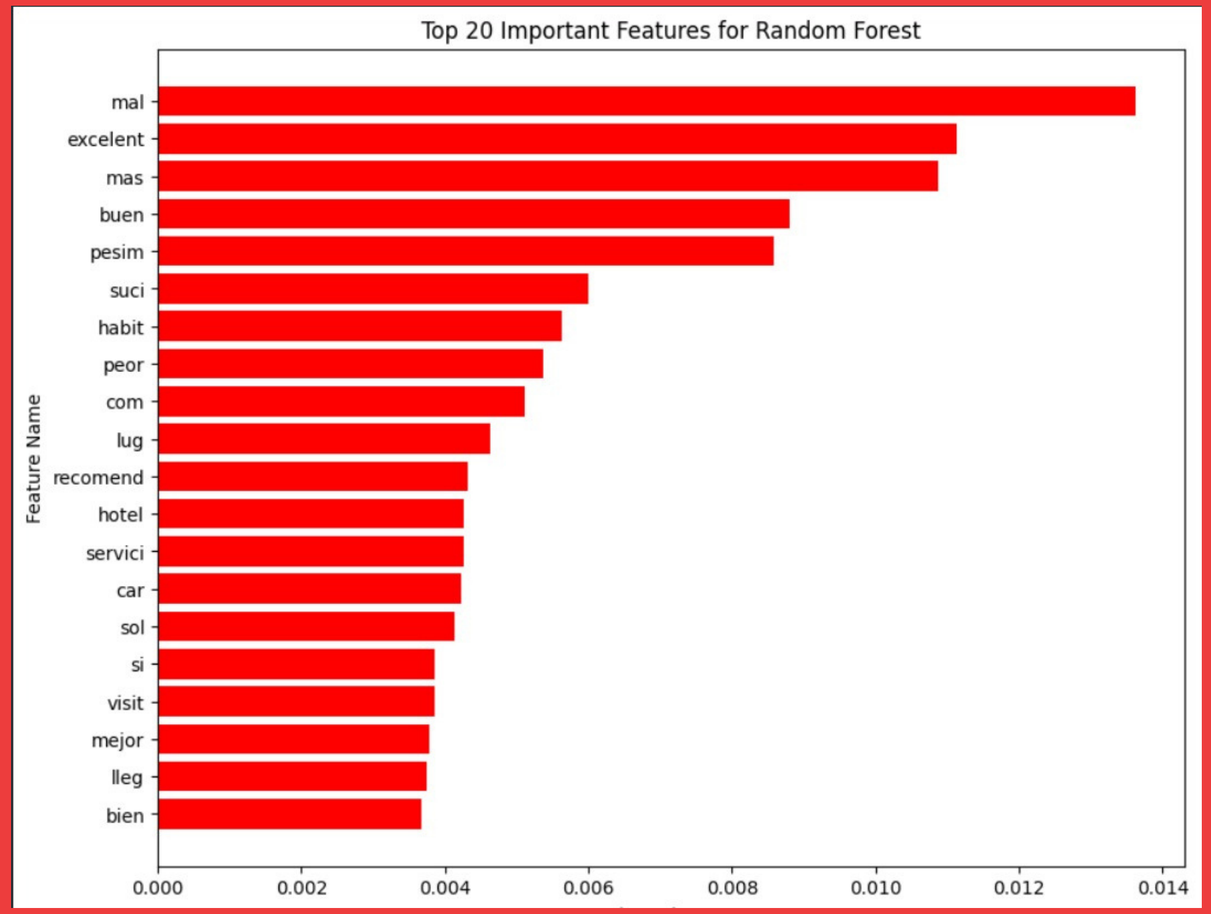
Análisis de las métricas de calidad

De los modelos implementados, se recomienda usar el modelo de Support Vector Machine (SVM) esto se debe a que presenta los mejores indicadores en la evaluación.

Las tres raíces de palabras más importantes, según el modelo Random Forest, fueron "mal", "excelent" y "mas". Estas palabras sugieren que en las reseñas analizadas, se mencionan aspectos como la calidad del servicio o producto ("excelent"), la comparación con otras opciones ("mas"), y la experiencia general ("al"). Por otro lado, se observan otras raízes de importancia como "suci", "com" y "servici", estas nos indican que la suciedad, la comida y el servicio son de gran importancia



Gráfica


Top 20 Important Features for Random Forest
Feature Name
mal
excelent
mas
buen
pesim
suci
habit
peor
com
lug
recomend
hotel
servici
car
sol
si
visit
mejor
lleg
bien
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.014