SDAIA
الهيئة السعودية للبيانات والذكاء الاصطناعي
Saudi Data & AI Authority

T5 BootCamp Data Science Projec

Quora Question Pairs

Submit to:
Mr.Ali El-kassas

By:
Arwa Essa
Nadia hajrasi
Alaa Alghamdi
Bushra alzhrani

Introduction :

Quora is a platform for Q&A, just like StackOverflow. But quora is more of a general-purpose Q&A platform that means there is not much code like in StackOverflow.

One of the many problems that quora face is the duplication of questions. Duplication of question ruins the experience for both the questioner and the answerer. Since the questioner is asking a duplicate question, we can just show him/her the answers to the previous question. And the answerer doesn’t have to repeat his/he

Objectives :

1. Handling text data
2. Convert text data into numbers
3. Processing data by entering it into modeling
4. predict which of the provided pairs of questions contain two questions with the same meaning.

Tools :

- Python and Jupyter Notebook
- Matplotlib, pyLDAvis and wordcloud for visuializations
- Numpy and Pandas for data manipulation
- NLTK, tashaphyne, and Farasa for NLP preprocessing
- Sklearn for ML algorithms
- Matplotlib, pyLDAvis and wordcloud for visuialization

Algorithm:

1.Read data
From kaggle website, which contains more than 400,000 rows and 6 random randomized plots consisting of 10,000 points.

2.We looked at the data
 Visualization data
 Delete rows that contain null values

Algorithm:

Text preprocessing :

Removing multiple spaces ,single characters from the start,Punctuations, capital letters,stopwords And apply lemmatize for text.

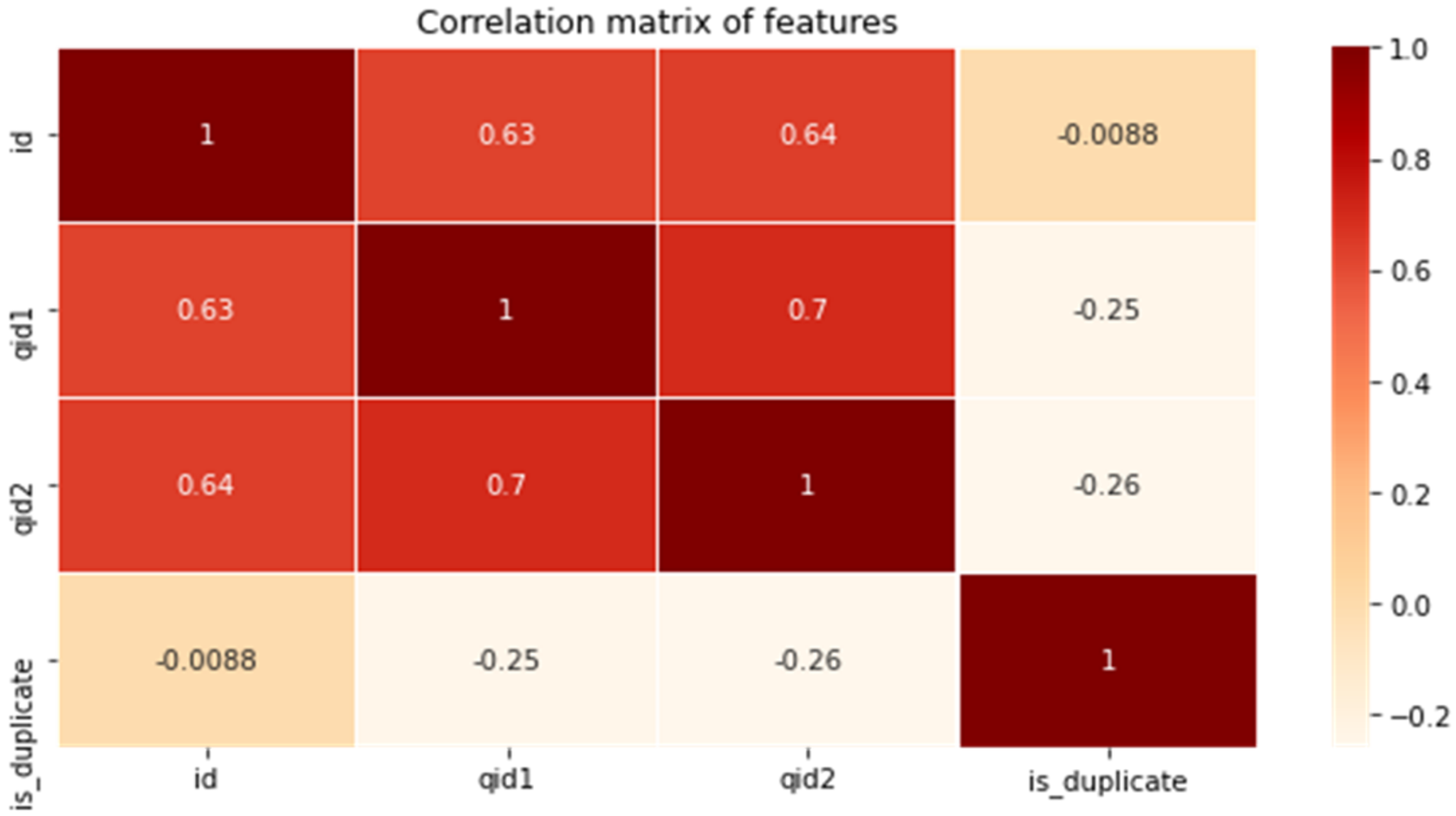
Algorithm:

3. Features Engineering
- new columns that contain the length of the questions,
- new columns that contain the number of words and
-new columns that contain the percentage of convergence between each question and another by use jaccard simalrity and cosine simalrity

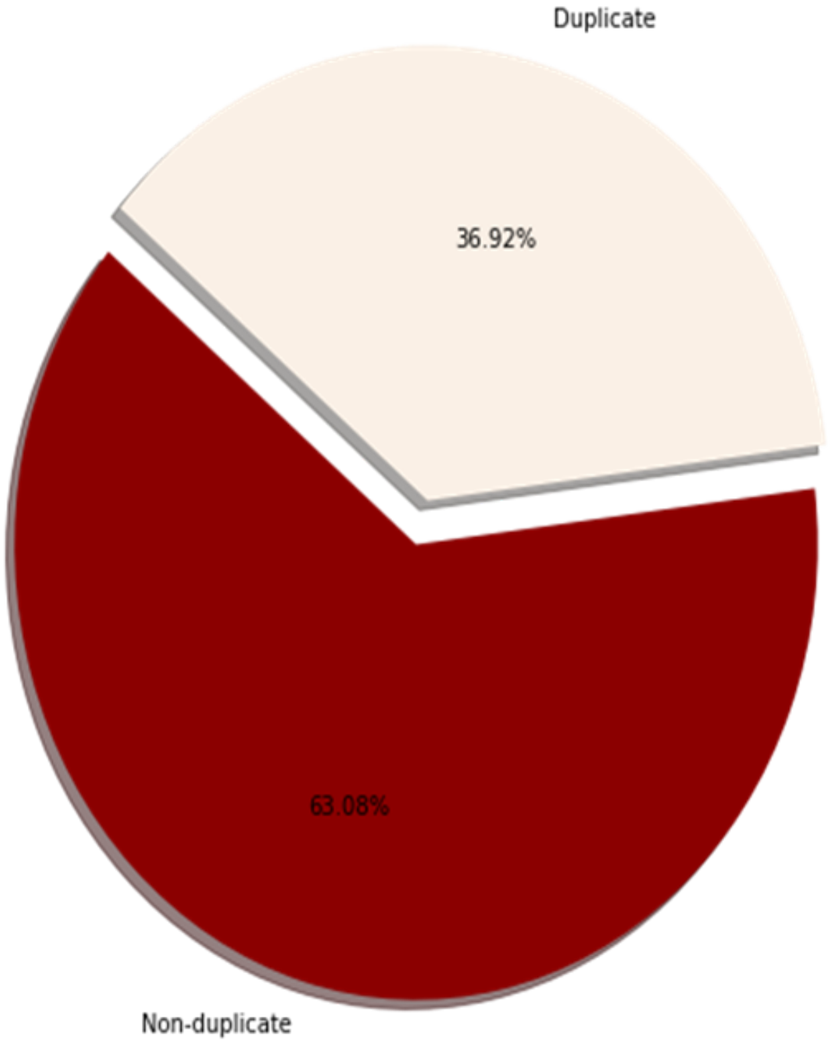
Algorithm:

4. Modeling
 - Convert text to numbers using TFIDF
 - Four types of classification models were used for machine learning
- Choosing the best scoring in validation

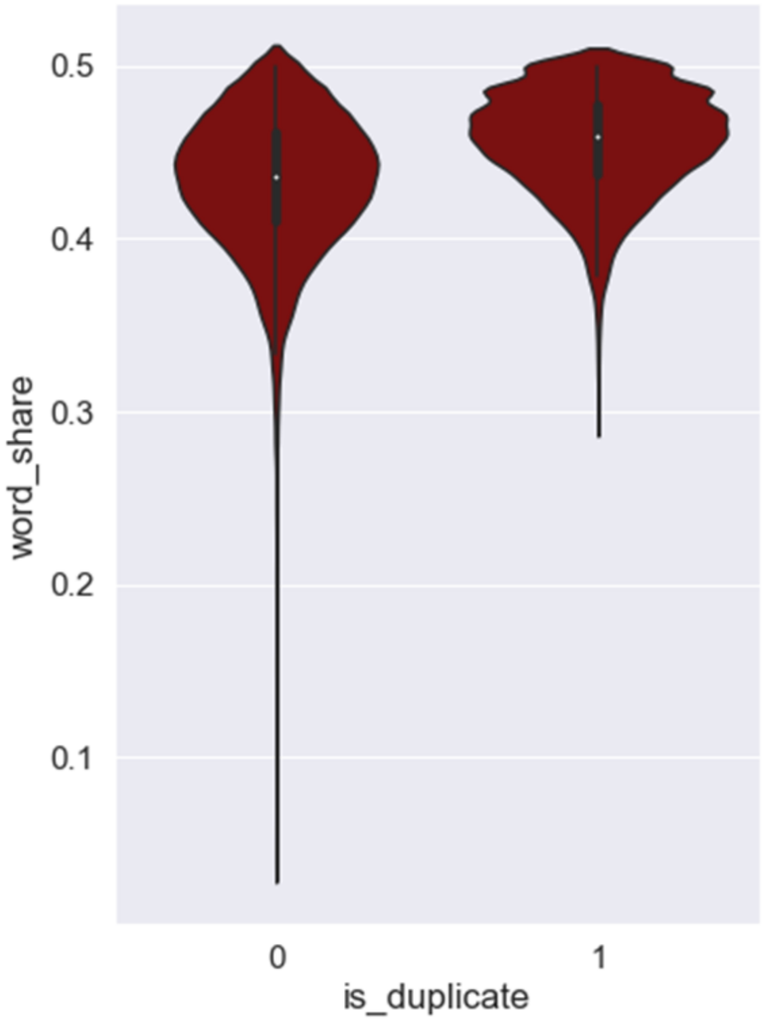
Data visualization :

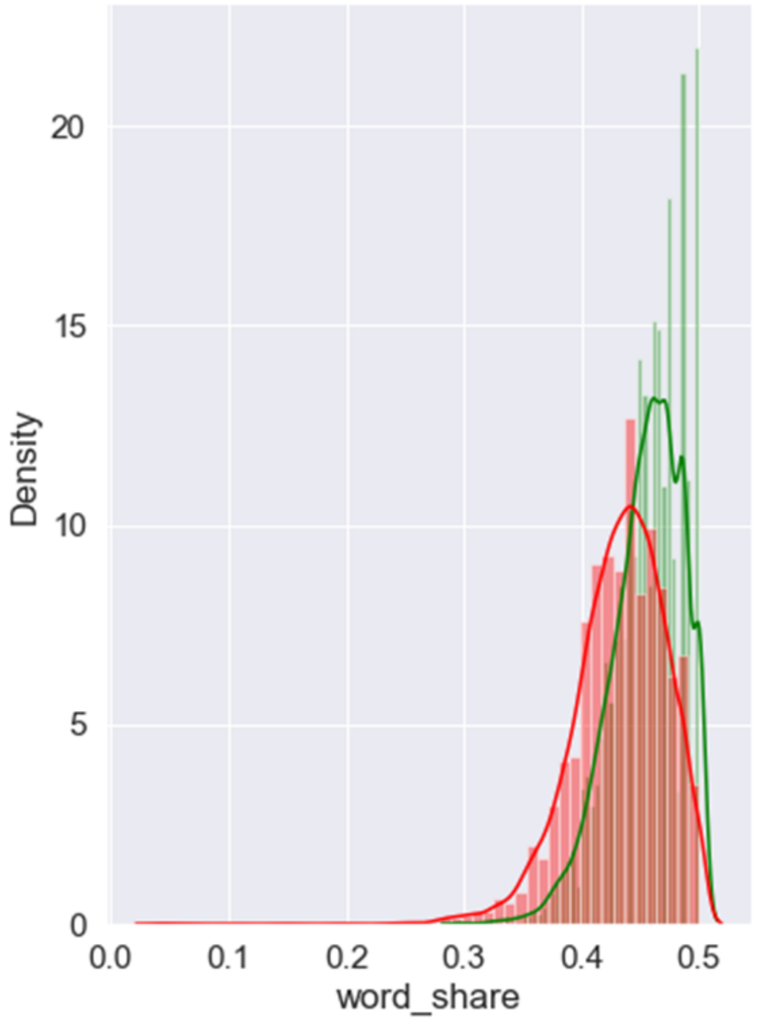


Data visualization :



Data visualization :

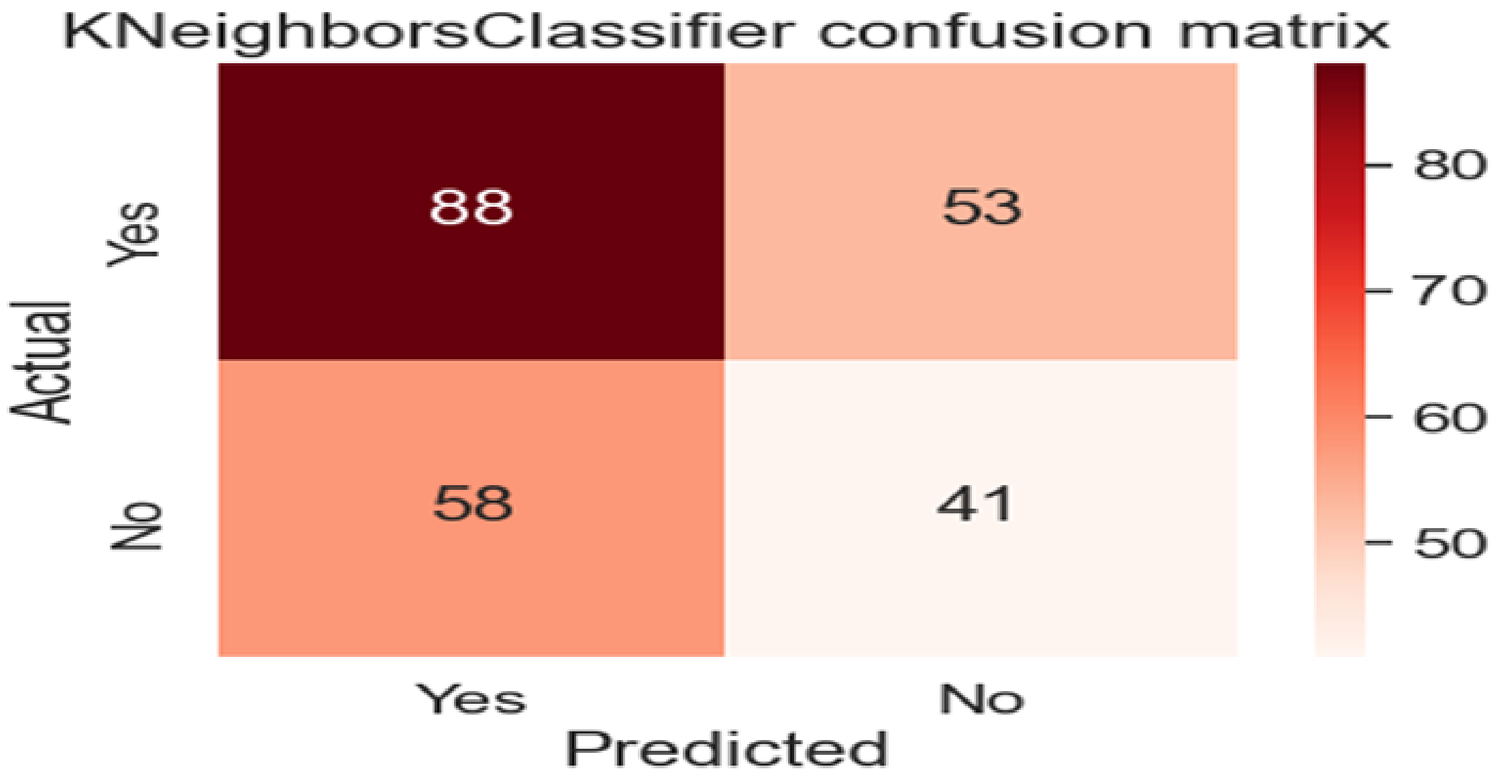




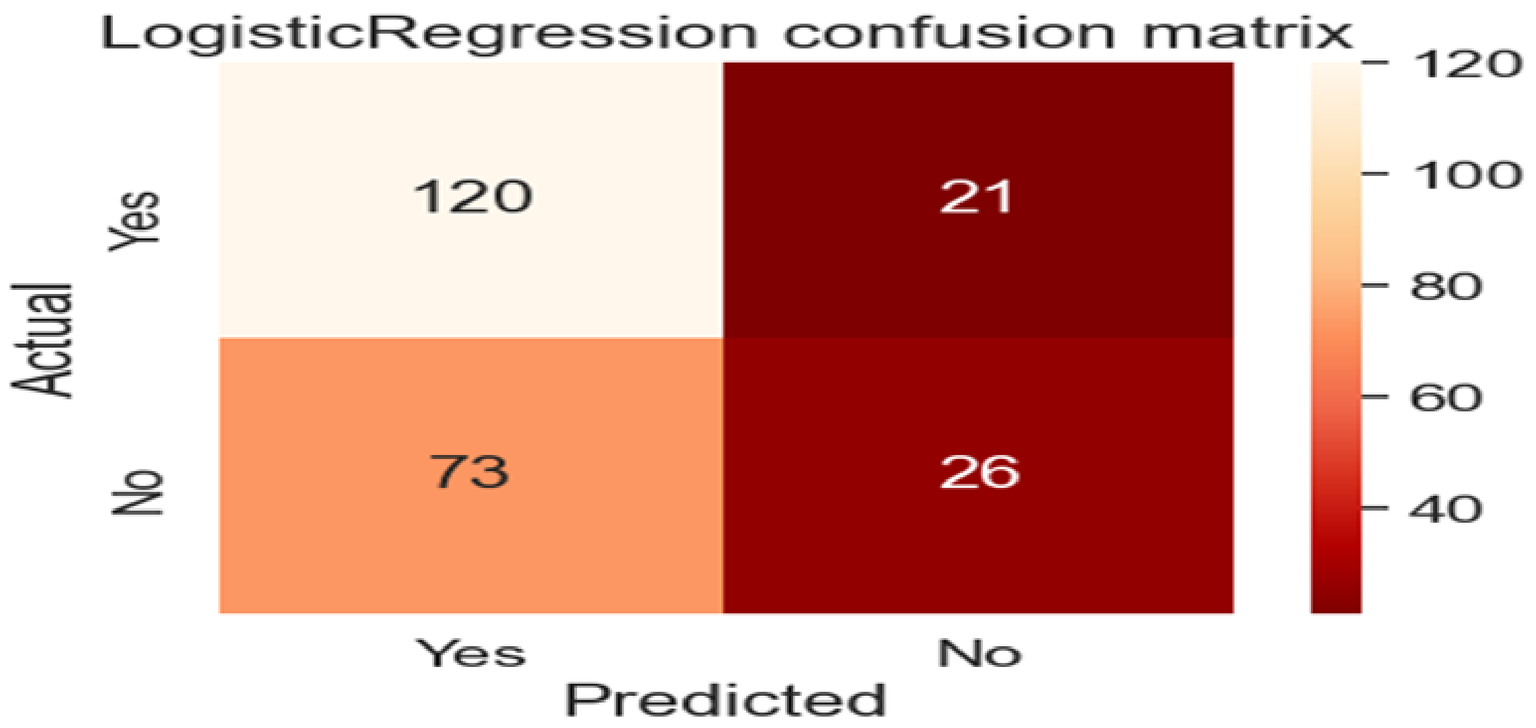
Data visualization :

Question classification Modeling

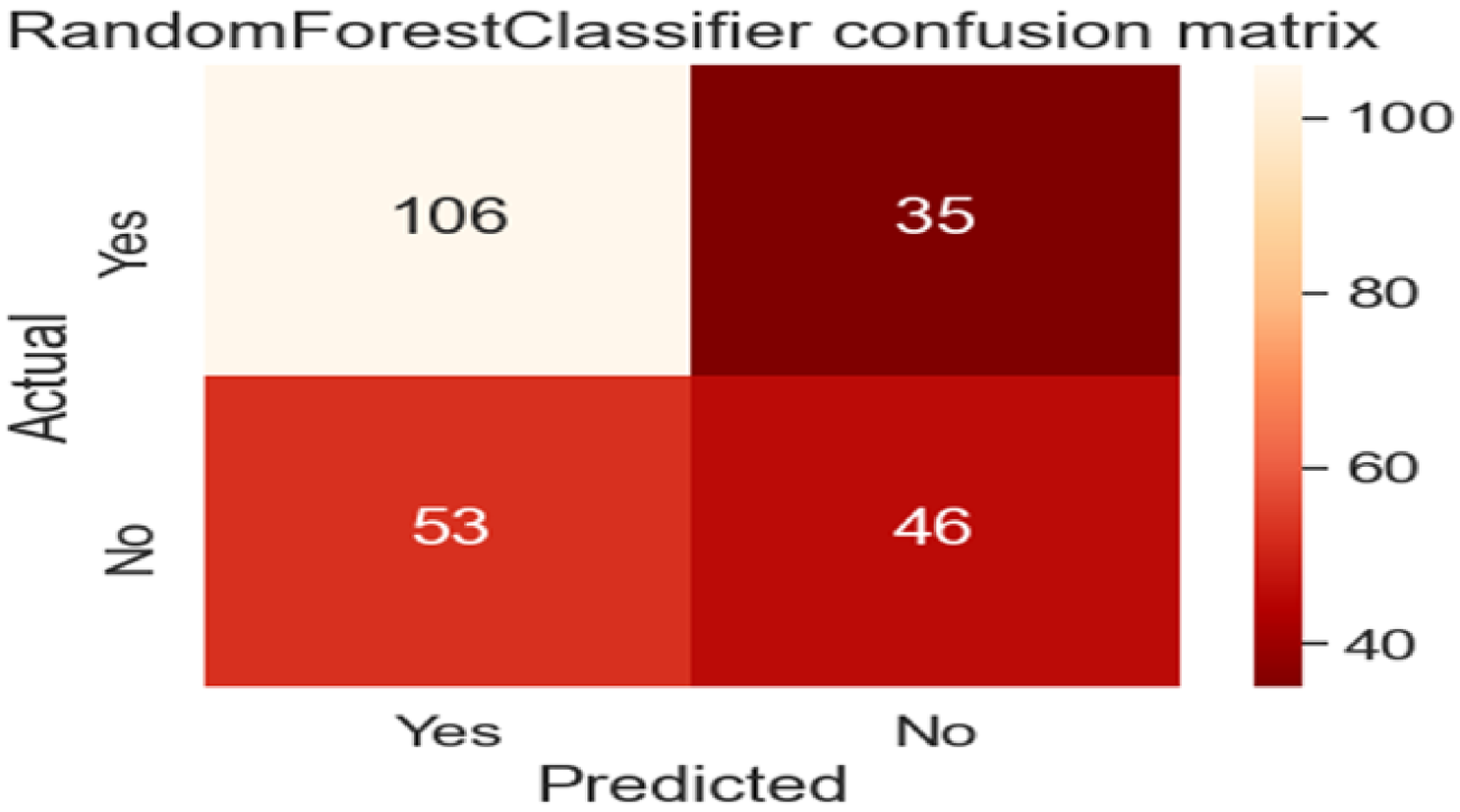
Data visualization :



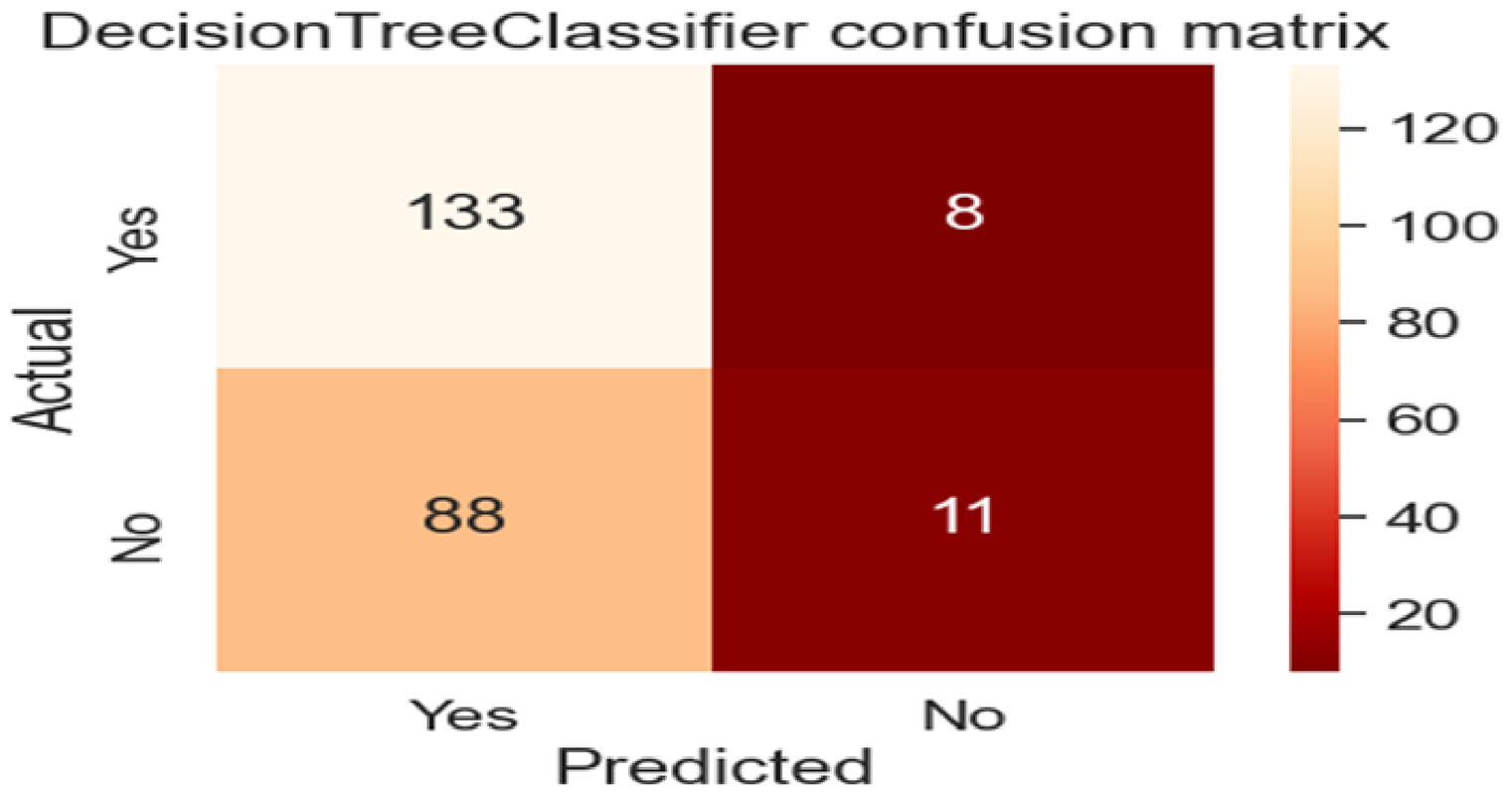
Data visualization :



Data visualization :



Data visualization :



Data visualization :

Model	Score Training	Score Validate
Decision Tree	66.40%	66.04%
Random forest	99.75%	72.02%
Neighbors	99.61%	54.22%
Logistic Regression	91.46%	69.10%

Topic Modelling :

Four types of topic modeling have been applied:
LSA , LDA , NMF ,K- means.


Any question ?

Thank you for listening