



SAUDI DIGITAL ACADEMY
AI Bootcamp

Artificial Intelligence

Natural Language Processing (NLP)

2025

Represent By

Team Members:

- GHALA ALOTAIBI
- HANAN ALNBHANI
- SARAH ALQAHTANI
- SHATHA KAMAL

Group Number: 2

Project Number: 3



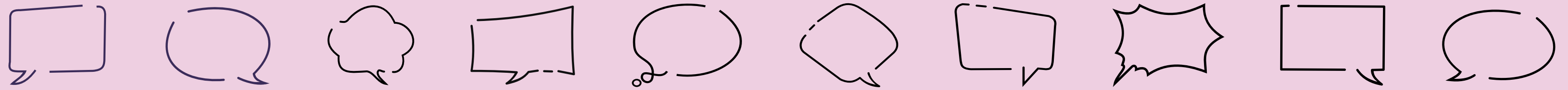


Table Of Contents

- 1.Introduction
- 2.Data Understanding & Pre-processing
- 3.Text Vectorization using TF-IDF
- 4.Model Training
- 5.Model Evaluation & Performance Metrics
- 6.Model Comparison & Best Model Selection
- 7.Predictions & Interpretation with LIME
- 8.Challenges & Solutions
- 9.Labour Division & Team Organization



quickly



Introduction

Natural Language Processing (NLP) is a branch of AI that helps computers understand and analyze human language. It is used in applications like chatbots, sentiment analysis, and text classification.



Main Objective of the Project

This project involves detecting fake news articles by classifying them into "real" or "fake" categories using machine learning.

below

Understand Dataset



Data



data

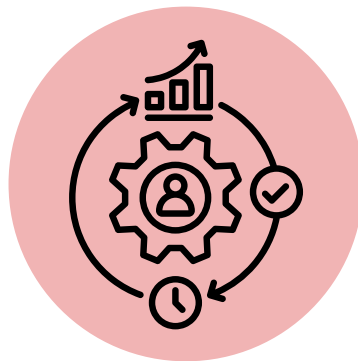


validation data



validation_data

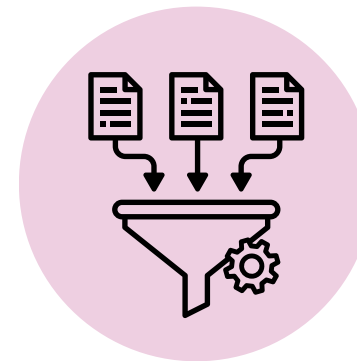
Data Preparation & Pre-processing



1

Drop unnecessary
columns

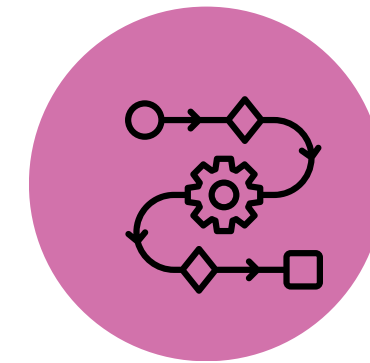
```
# Drop unnecessary columns
data = data[['title', 'text', 'label']]
data.dropna(inplace=True)
```



2

Text cleaning
function

```
# Text cleaning function
def preprocess_text(text):
    #Function to clean and preprocess text
    text = text.lower() # Convert to lowercase
    text = re.sub(r'\d+', '', text) # Remove numbers
    # Remove punctuation
    text = text.translate(
        (str.maketrans('', '', string.punctuation)))
    # Remove extra spaces
    text = re.sub(r'\s+', ' ', text).strip()
    return text
```



3

Features and
target variable

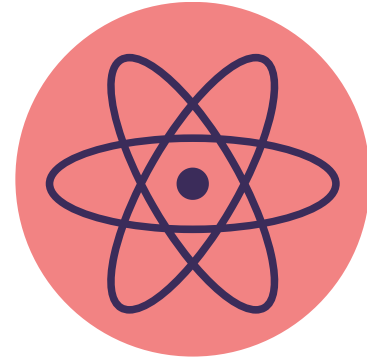
```
# Features and target variable
X = data['title'] + " " + data['text']
y = data['label']
```

Vectorization Using TF-IDF

Convert text to numerical features

```
vectorizer = TfidfVectorizer(max_features=5000)  
X_train_tfidf = vectorizer.fit_transform(X_train)  
X_test_tfidf = vectorizer.transform(X_test)
```





Model Training

Split The Dataset Into Training & Test Sets

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
print(f"Training Data: {len(X_train)} samples")
print(f"Test Data: {len(X_test)} samples")
```

F or

A nd

N or

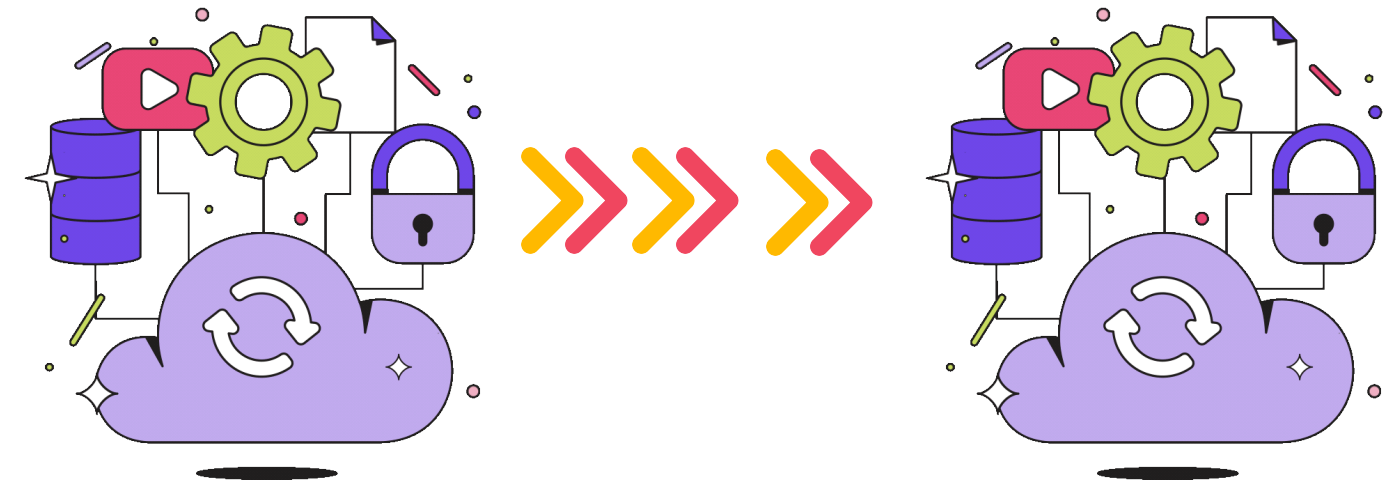
B ut

O r

Y et

S o

Models Training



1

KNN

```
KNeighborsClassifi  
eighborsClassifier(
```

2

Naive Bayes

```
▼ MultinomialNB  
MultinomialNB()
```

3

Logistic
Regression

```
LogisticRegressio  
gisticRegression()
```

4

XGBoost

```
XGBClassifier  
booster=None, call  
=None, colsample b
```

5

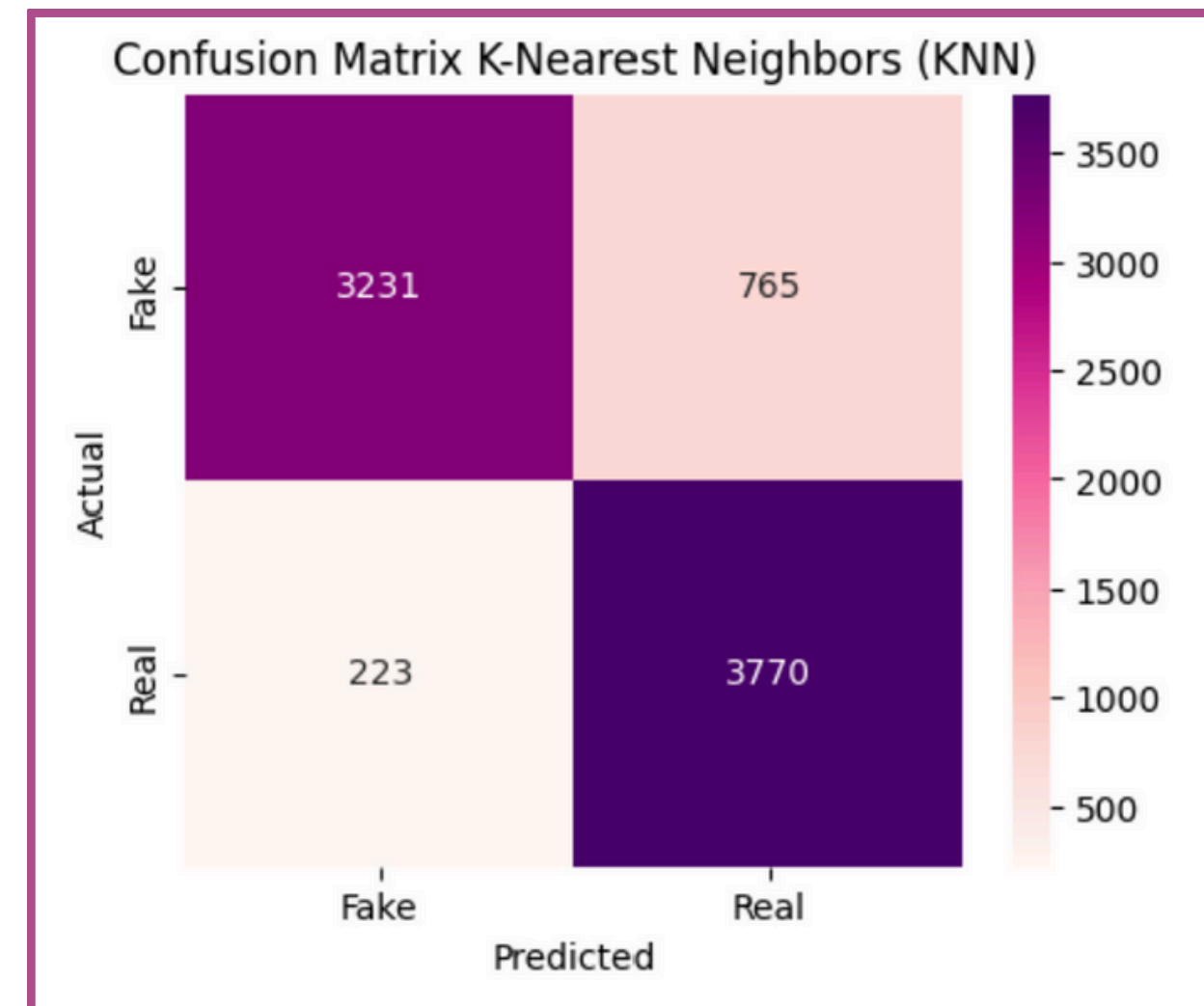
Random Forest

```
RandomForestClassif  
domForestClassifier
```

Models Evaluation

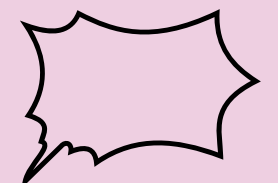
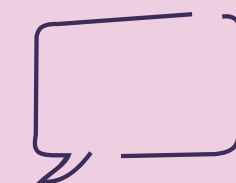
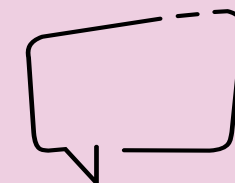
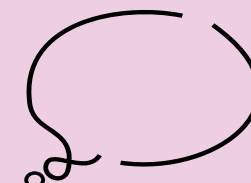
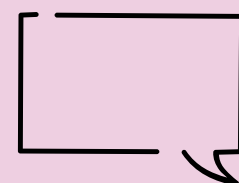
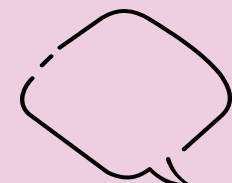
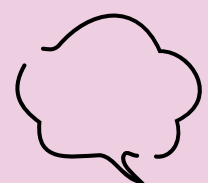
1
KNN

Confusion matrix



Accuracy Calculation

Accuracy K-Nearest Neighbors (KNN): 87.63%



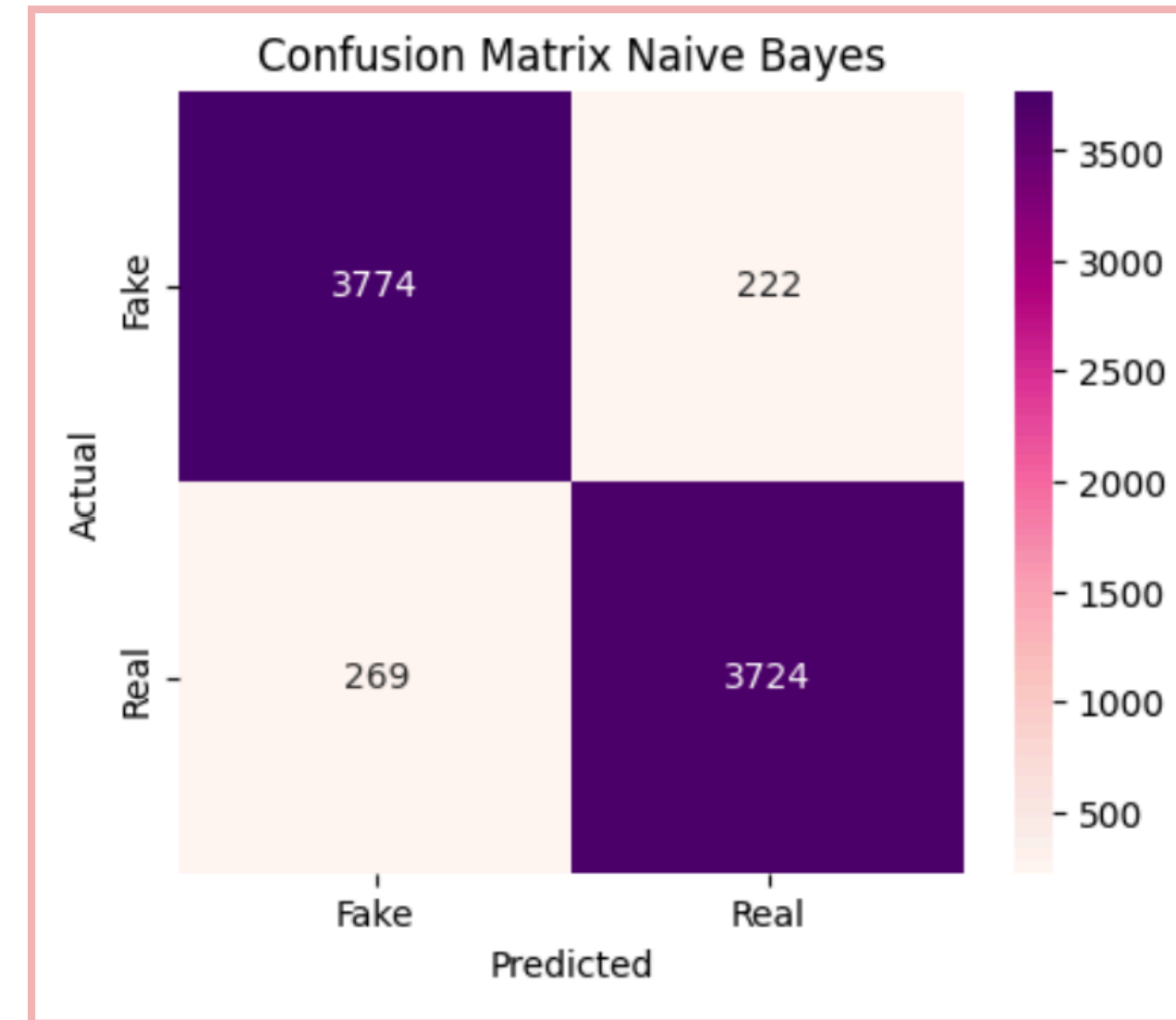
Models Evaluation

2

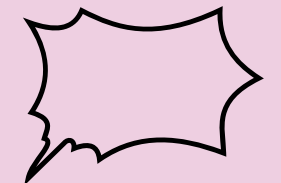
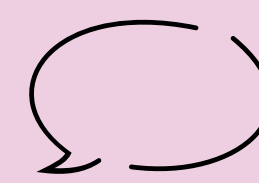
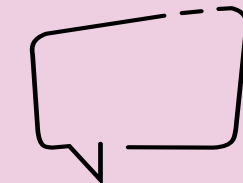
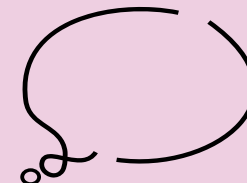
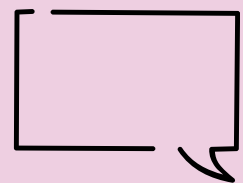
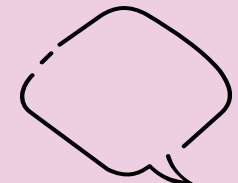
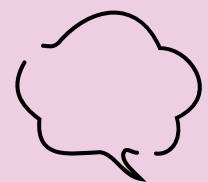
Naive Bayes

Confusion matrix

Confusion matrix



Accuracy Naive Bayes: 93.85%



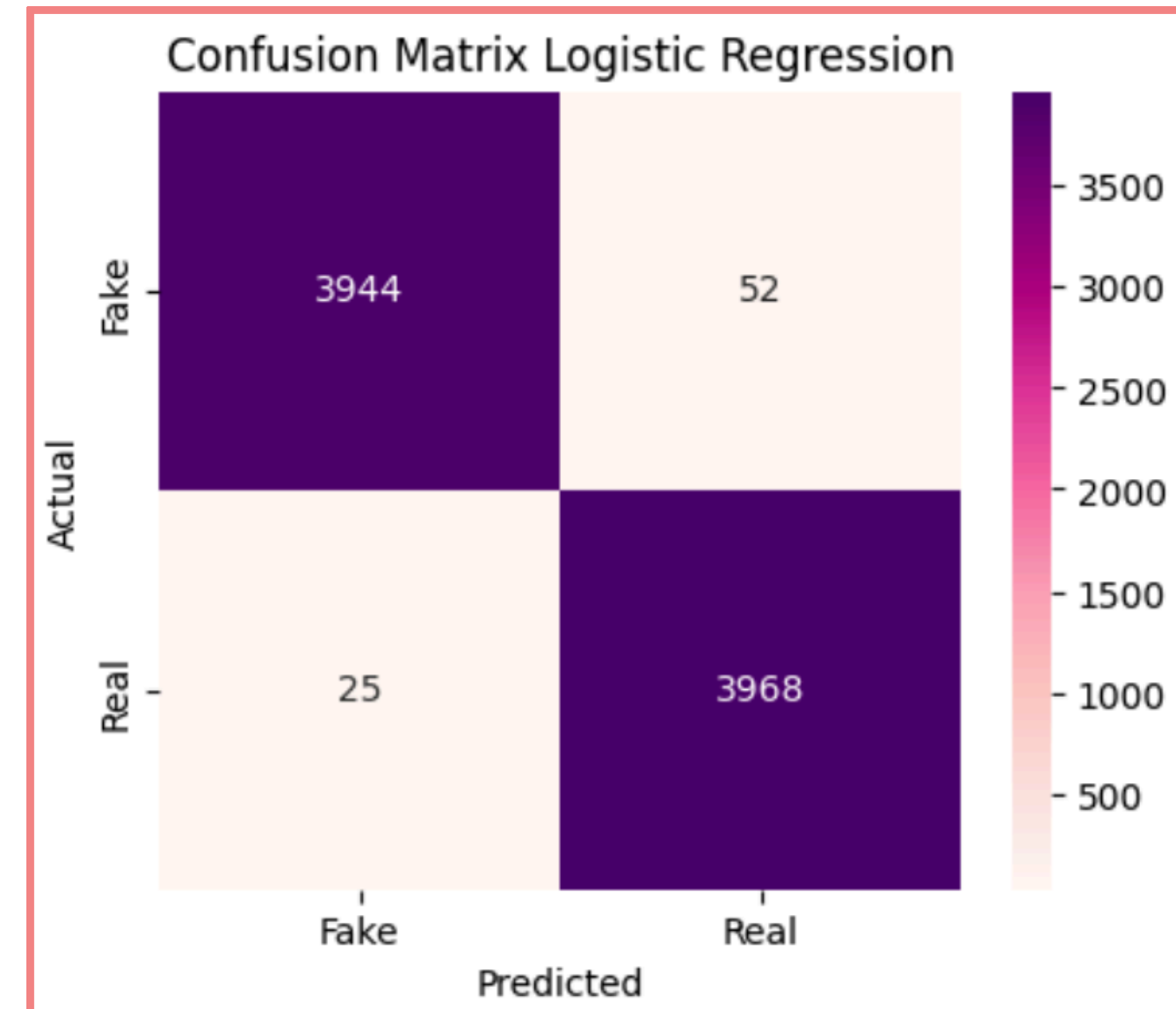
Models Evaluation

3

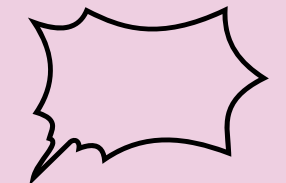
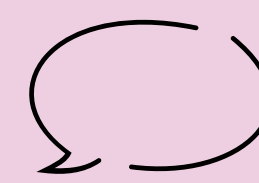
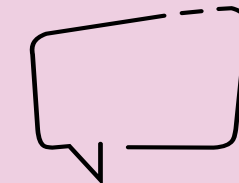
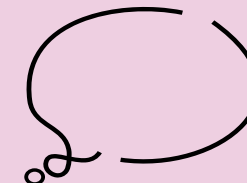
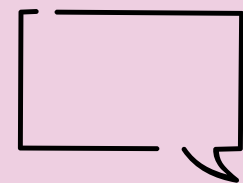
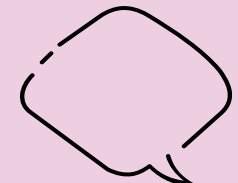
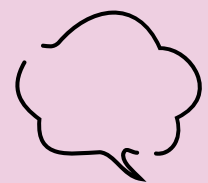
Logistic Regression

Confusion matrix

Accuracy Calculation



Accuracy Logistic Regression: 99.0362%



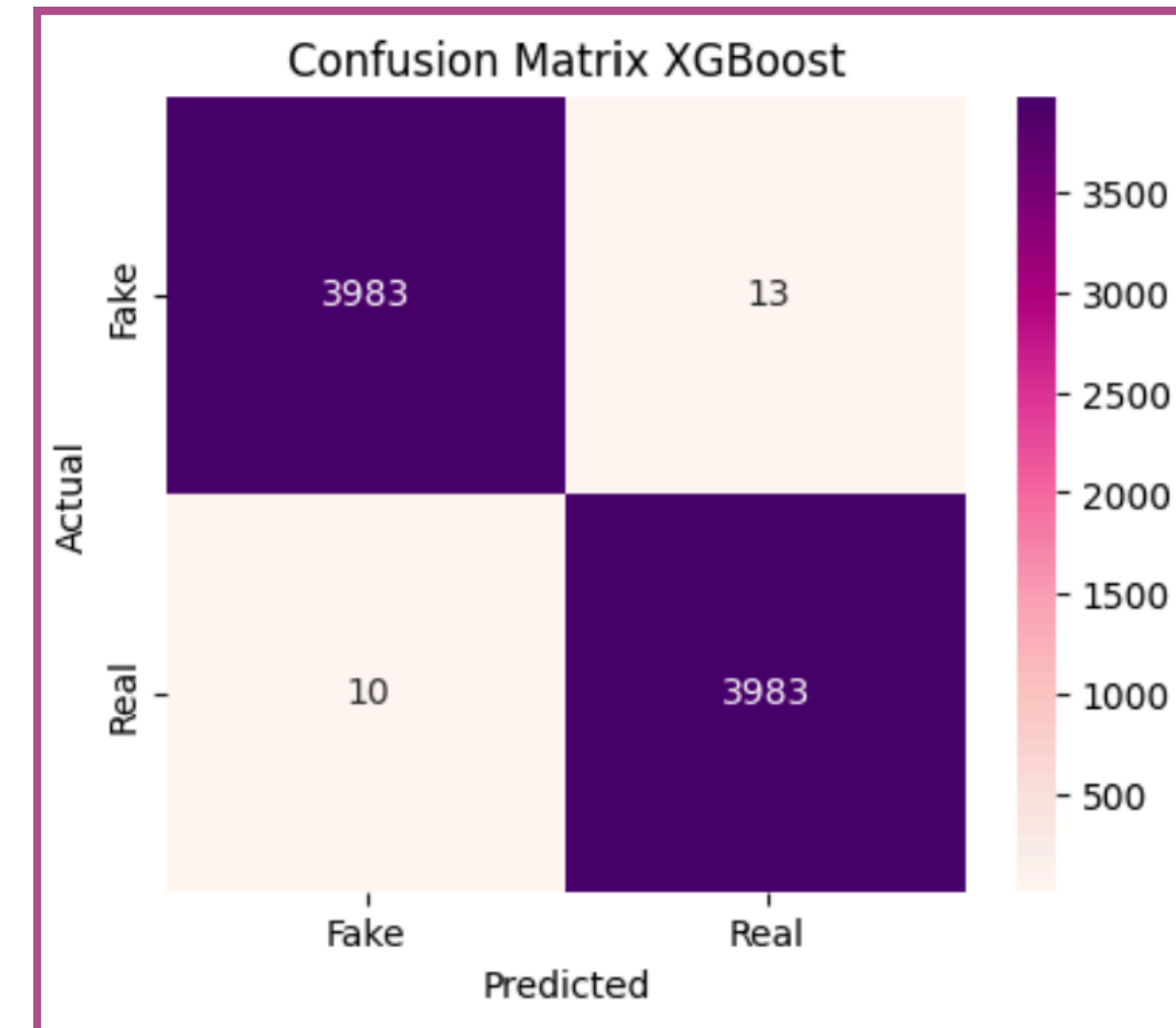
Models Evaluation

4

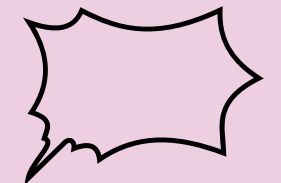
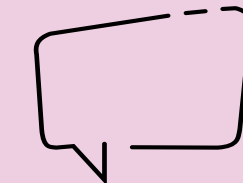
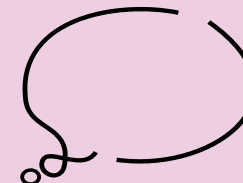
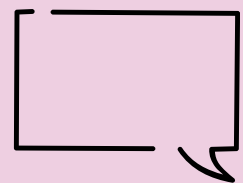
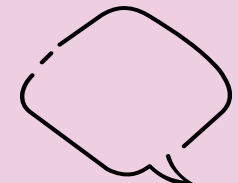
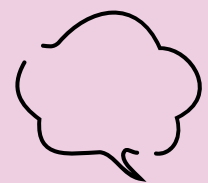
XGBoost

Confusion matrix

Accuracy Calculation



Accuracy XGBoost: 99.71%



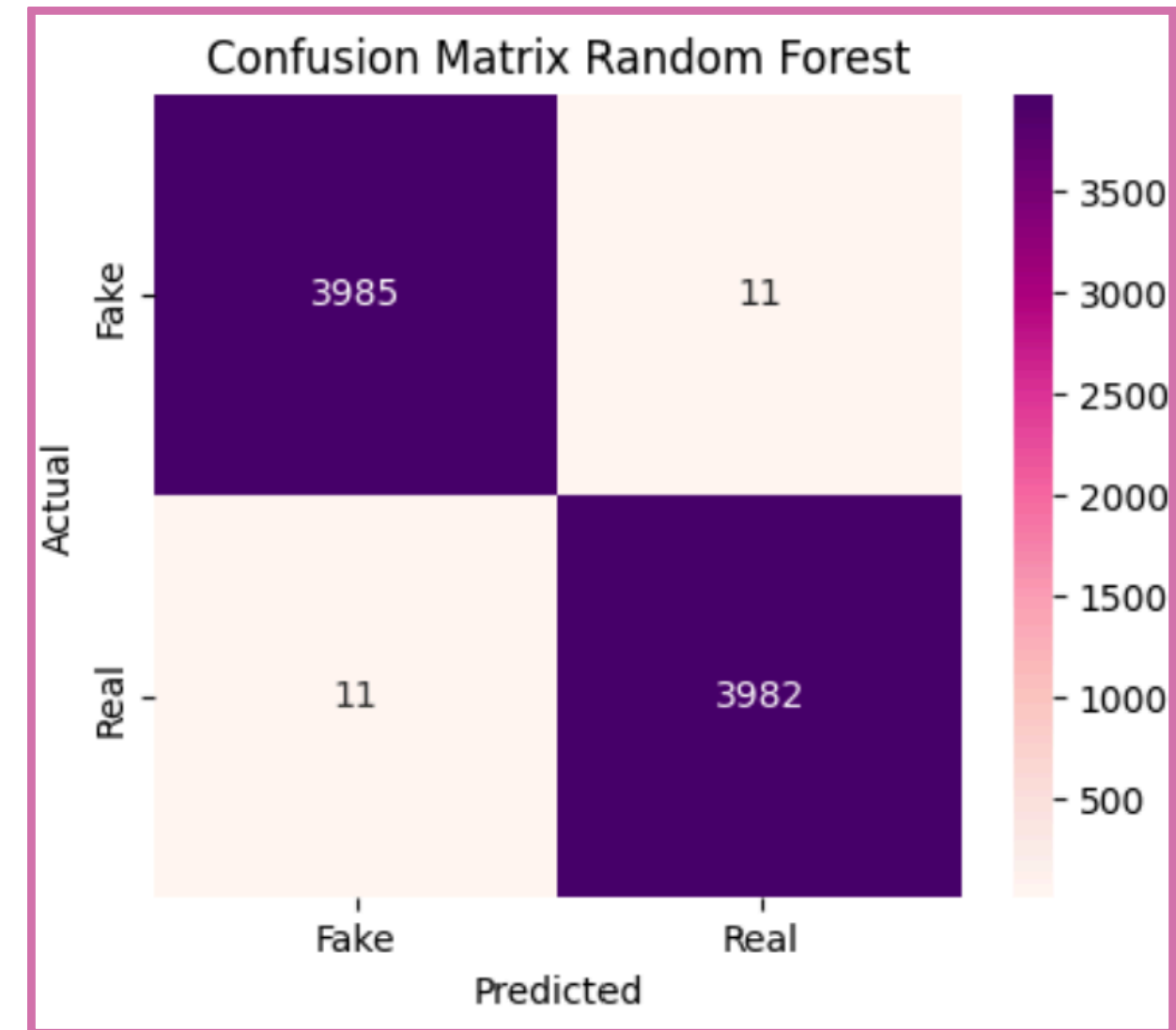
Models Evaluation

5

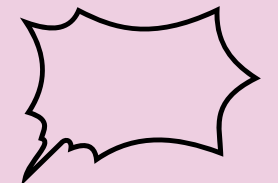
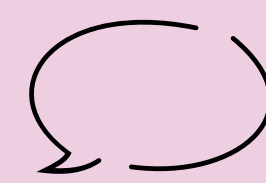
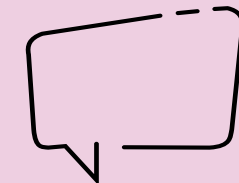
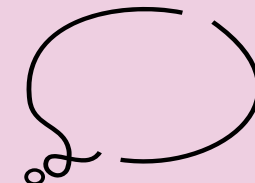
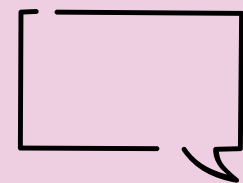
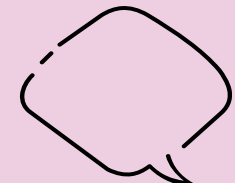
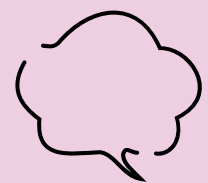
Random Forest

Confusion matrix

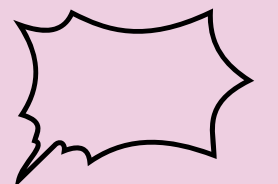
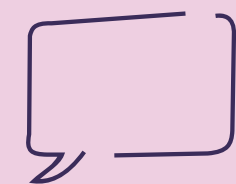
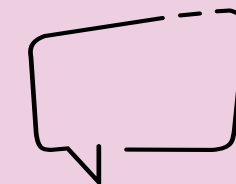
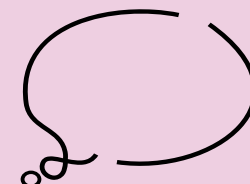
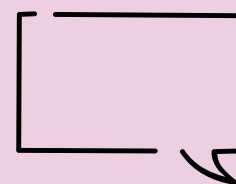
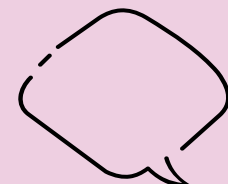
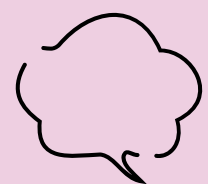
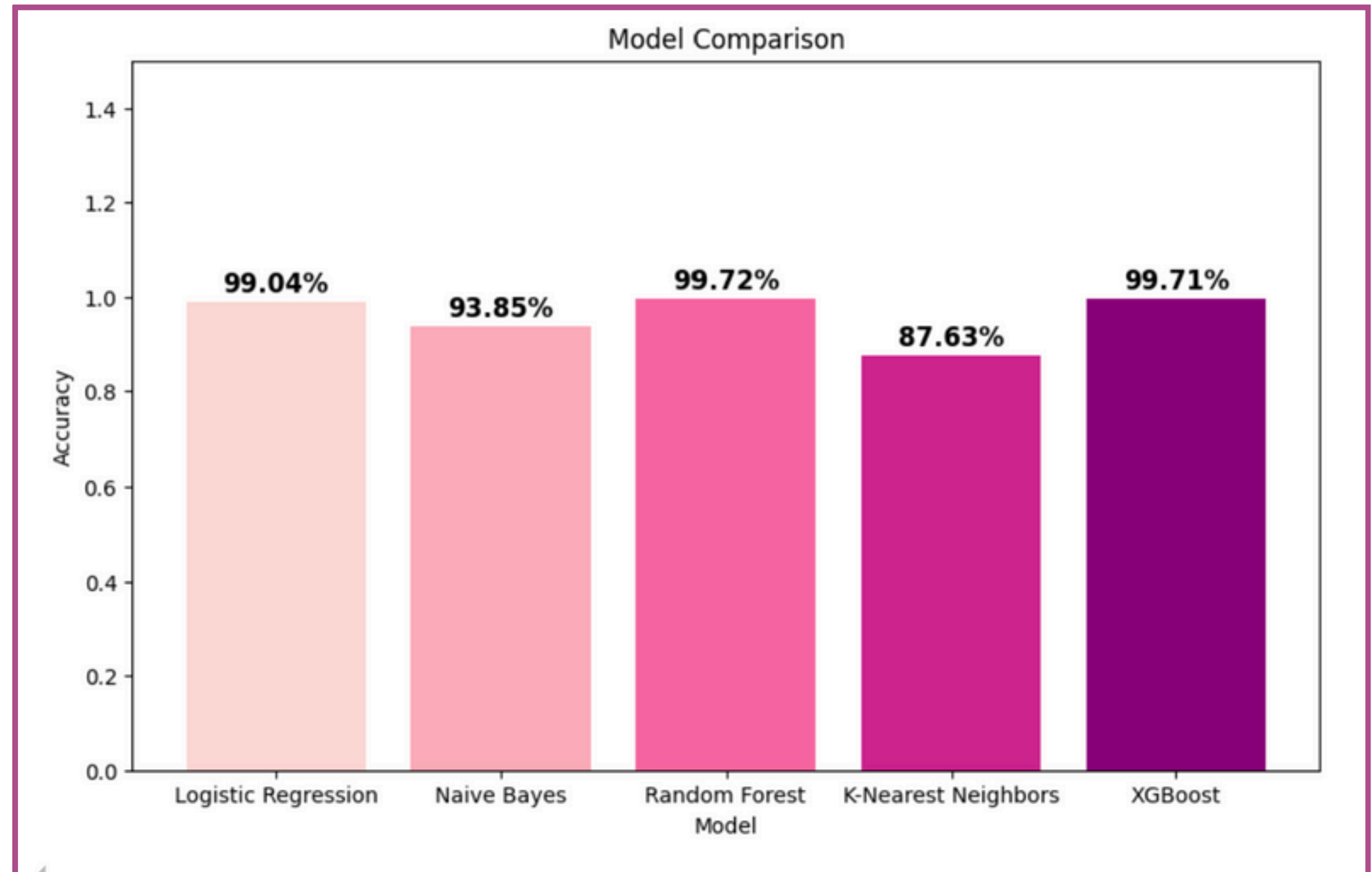
Accuracy Calculation



Accuracy Random Forest: 99.72%

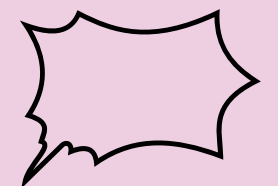
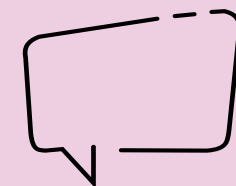
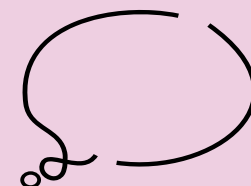
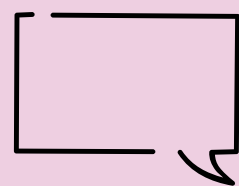
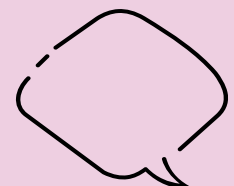
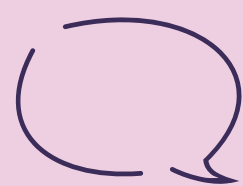
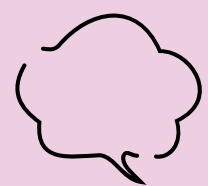
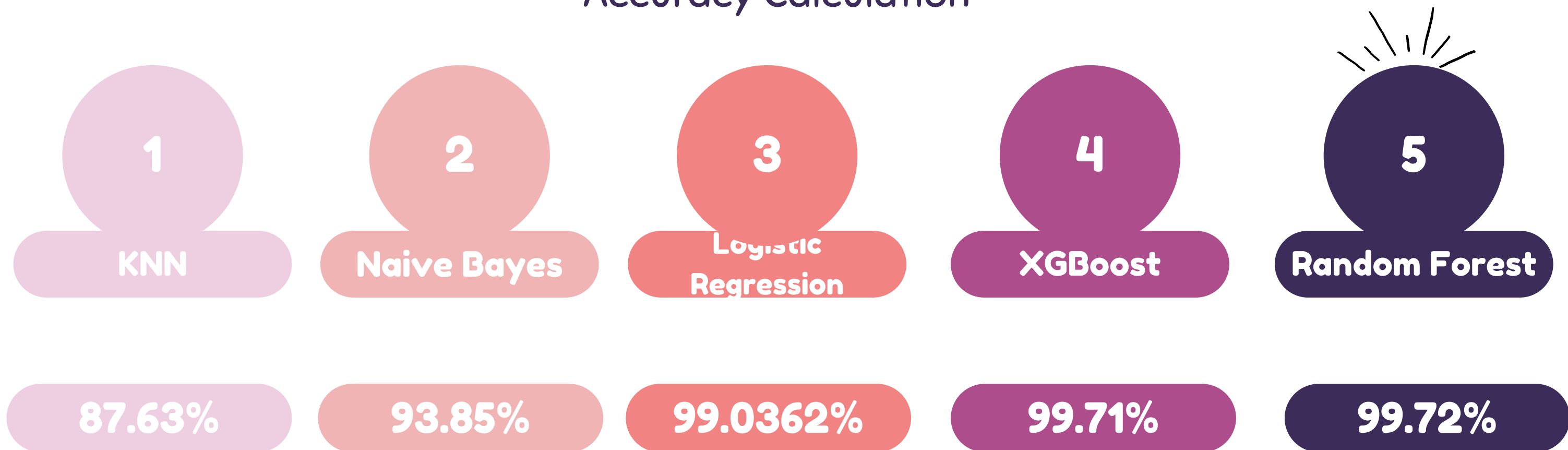


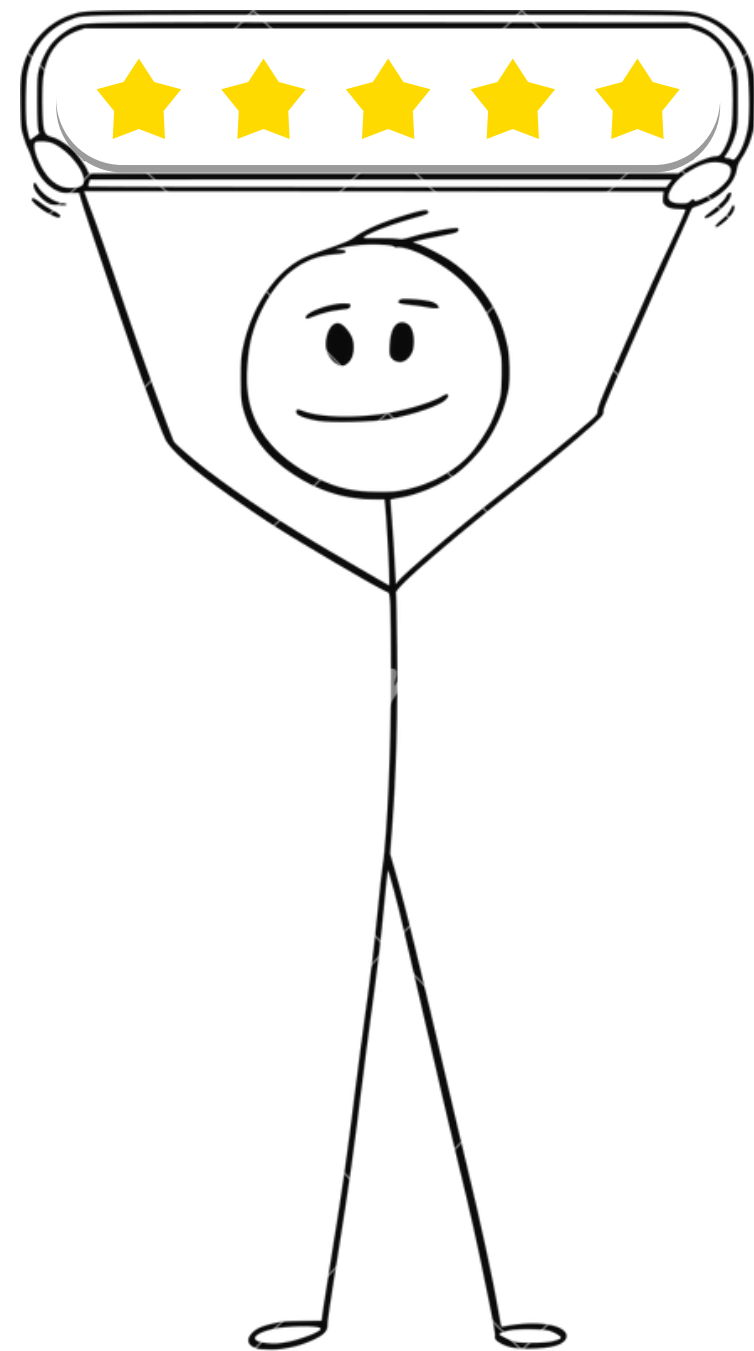
Models Comparison



Models Evaluation

Accuracy Calculation

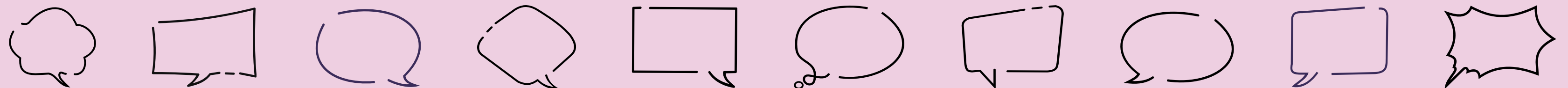




Best Model Selection

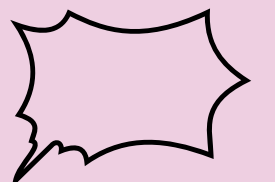
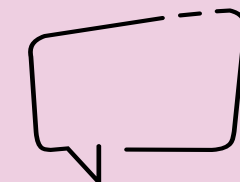
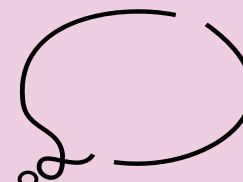
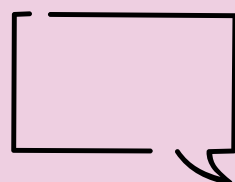
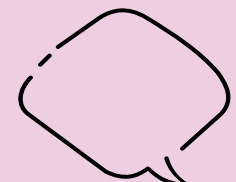


The best-performing model is: “Random Forest”



predictions on validation data

```
# Make predictions using the best model
predictions = model.predict(X_validation_tfidf)
# Add the predictions to the validation dataframe
validation_data['label'] = predictions
```

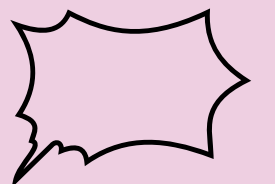
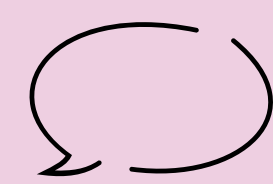
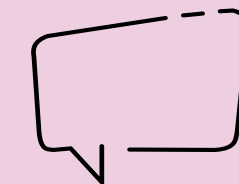
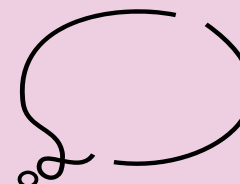
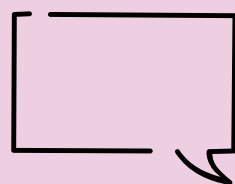
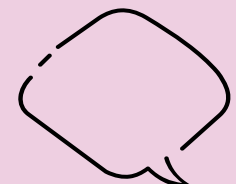
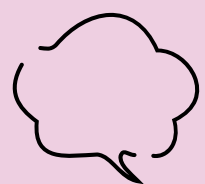


Save the result to a new CSV file

```
# Save predictions to CSV
validation_data.to_csv('validation_predictions.csv', index=False)
print("Predictions saved to 'validation_predictions.csv'")
# Display the first few rows of the validation data with predictions
validation_data.head()
```

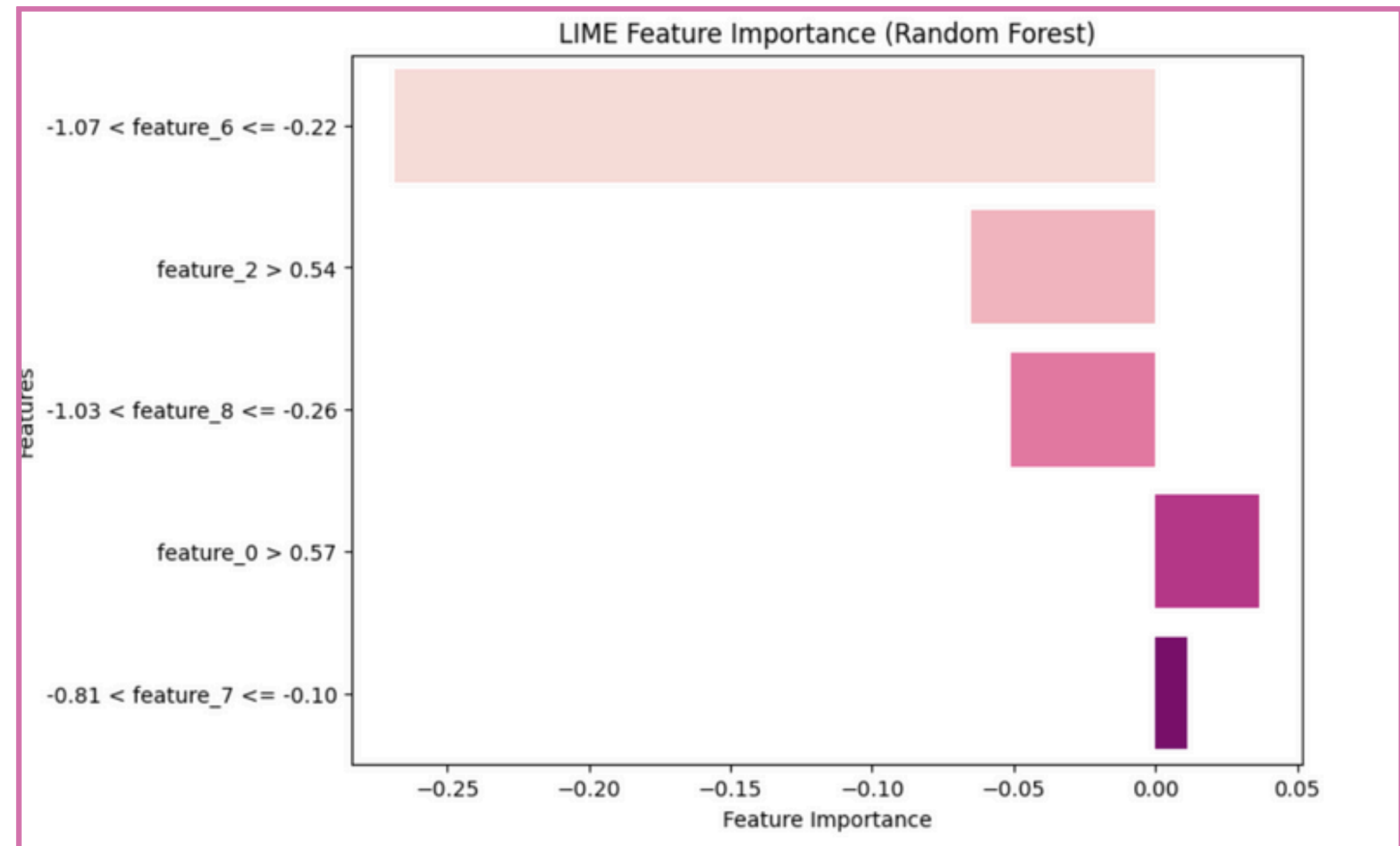
Predictions saved to 'validation_predictions.csv'

	label	title	text
0	1	UK's May 'receiving regular updates' on London...	LONDON (Reuters) - British Prime Minister Ther...
1	1	UK transport police leading investigation of L...	LONDON (Reuters) - British counter-terrorism p...
2	1	Pacific nations crack down on North Korean shi...	WELLINGTON (Reuters) - South Pacific island na...
3	1	Three suspected al Qaeda militants killed in Y...	ADEN, Yemen (Reuters) - Three suspected al Qae...
4	1	Chinese academics prod Beijing to consider Nor...	BEIJING (Reuters) - Chinese academics are publ...



Explain Predictions with LIME

Local Interpretable Model-agnostic Explanations



Challenges

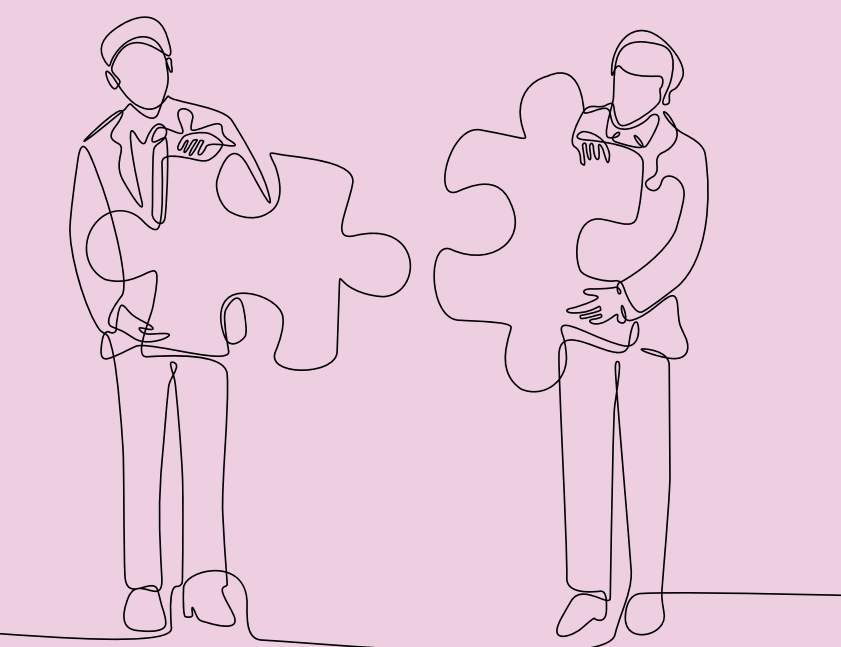
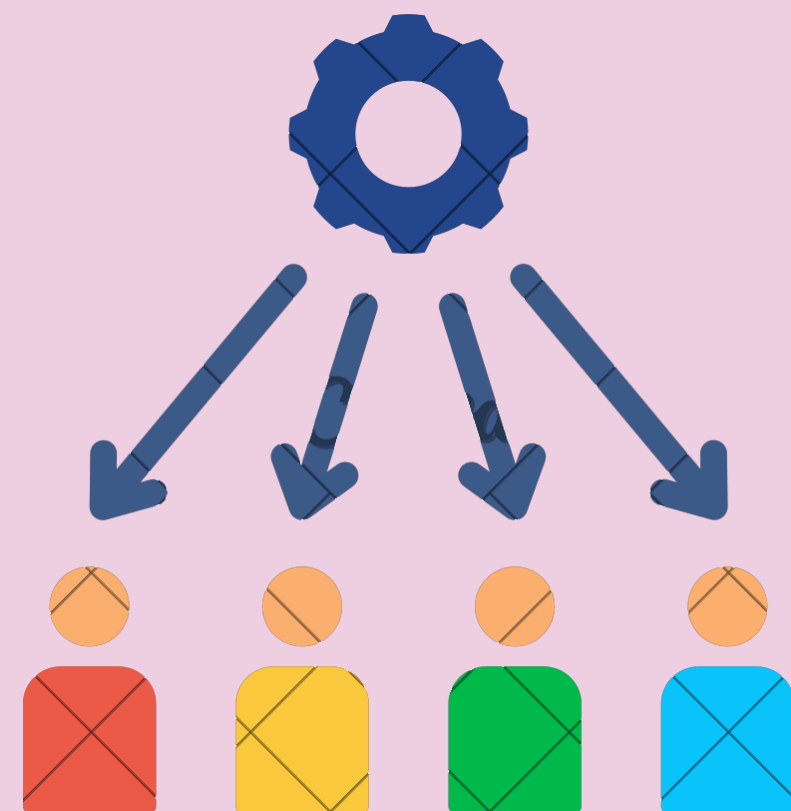
&

Solutions

Challenges

Solutions

1. Difficulty in selecting the best model	We tested multiple models and compared their accuracy to find the best-performing one.
2. Understanding model predictions	We used LIME (Local Interpretable Model-agnostic Explanations) to interpret model decisions.
3. Balancing workload among team members	We assigned tasks based on each member's expertise and skill set.
4. Time management and meeting deadlines	We utilized Slack to track progress and ensure efficient collaboration.



Organizing Labour Division in Our Team Strategies & Implementation

Ghala AlOtaibi

KNN - Logistic Regression - XGBoost

Hanan Alnbhani

Presentation -LIME

Sarah Alqahtani

Random Forest

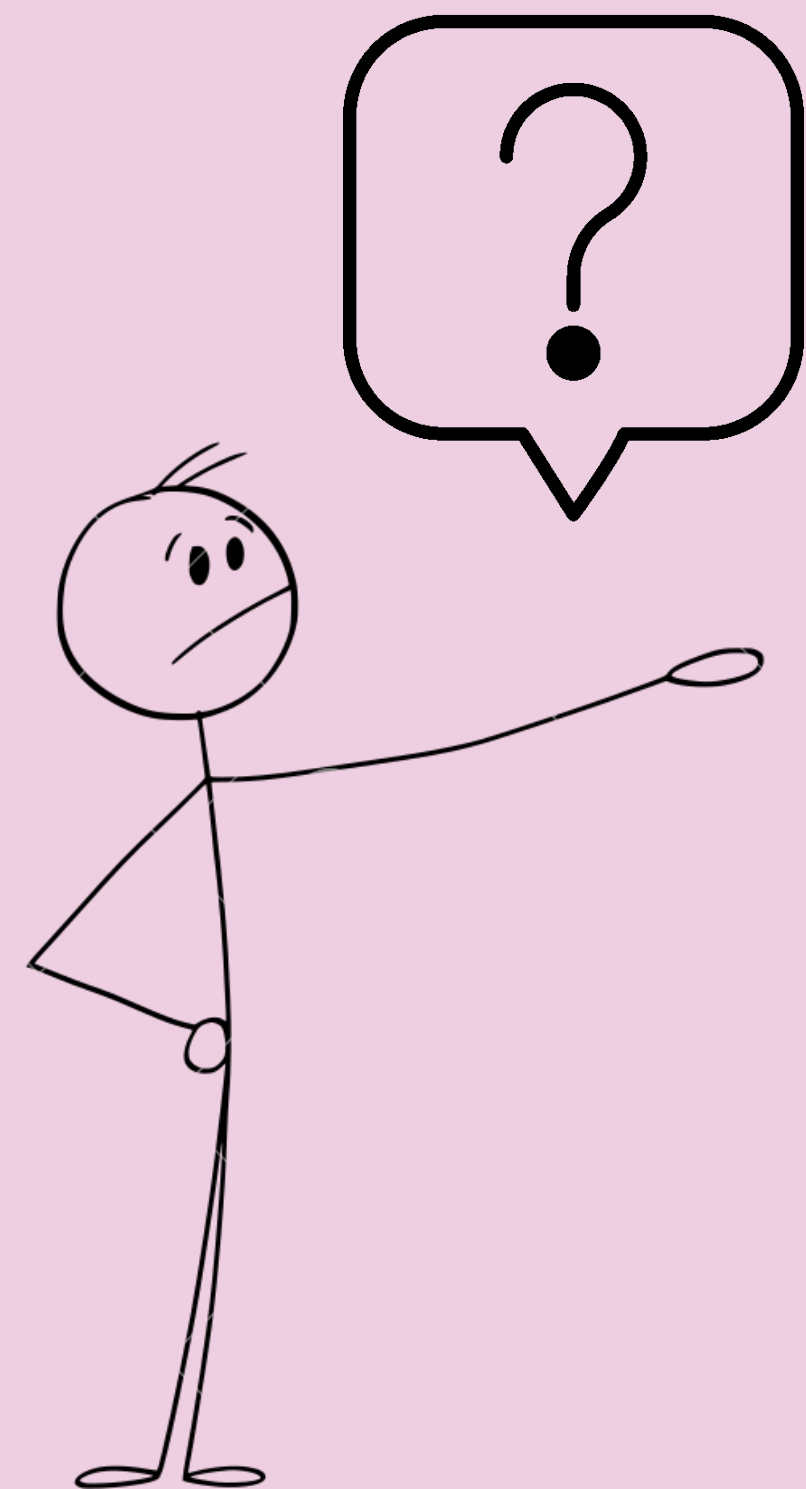
SHATHA KAMAL

Naive Bayes - Read Me



Thank you[♡]

For your kind listening



ANY
Questions?