# Machine Learning Project Report

## Fake Instagram ID Detection

Name: Naser Muhammad

Roll Number: 261936735

Name: Talal Akhtar

Roll Number; 261937182

## Objective

The objective of this project is to build a machine learning model that can identify fake Instagram profiles using key account attributes such as username structure, profile picture presence, follower and following counts, posting activity, and bio characteristics. The goal is to help in automating the detection of potentially malicious or bot-driven accounts on social platforms.

## Dataset

The dataset used in this project was sourced from Kaggle and contains 5000 Instagram user profiles. Each entry is labeled as either fake (1) or real (0). The dataset includes the following columns:

- profile pic  
- nums/length username  
- fullname words  
- nums/length fullname  
- name==username  
- description length  
- external URL  
- private  
- #posts  
- #followers  
- #follows  
- fake (label)

## Data Preprocessing

Before training the model, the dataset was cleaned and normalized. StandardScaler was used to bring all feature values to the same scale. The dataset was then split into 80% training data and 20% test data.

## Model Used

A Random Forest Classifier was selected due to its high accuracy and robustness in handling classification problems. The model was trained on the preprocessed data and tested on unseen data to evaluate its performance.

## Evaluation Results

The following metrics were recorded during evaluation on the test set:

Classification Report:

Class 0 (Real Accounts):  
- Precision: 0.98  
- Recall: 1.00  
- F1-Score: 0.99  
  
Class 1 (Fake Accounts):  
- Precision: 1.00  
- Recall: 0.98  
- F1-Score: 0.99  
  
Overall Accuracy: 99%  
This means the model is highly accurate in distinguishing fake accounts from real ones.

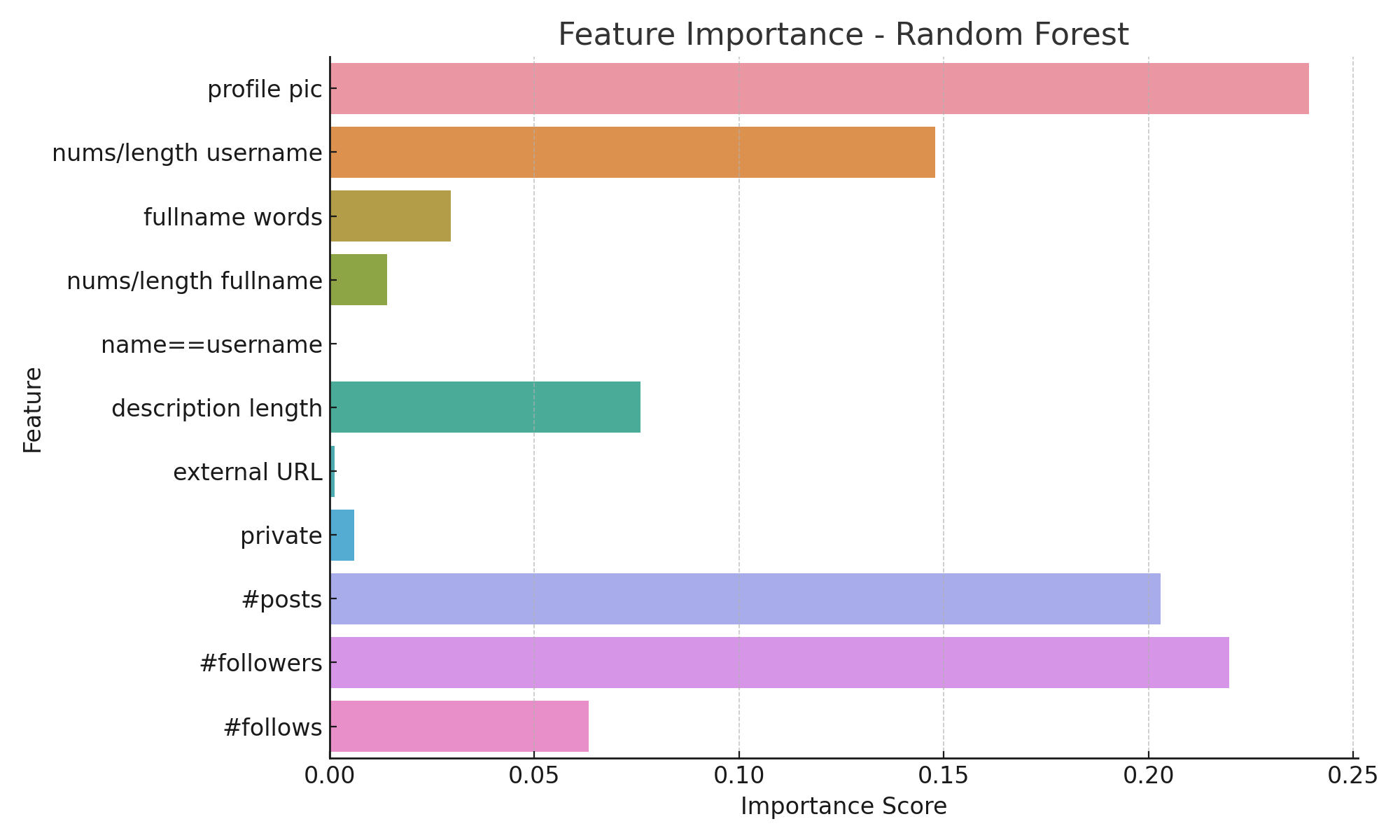
Confusion Matrix Interpretation:

The confusion matrix showed:  
- 496 Real accounts correctly predicted  
- 2 Real accounts misclassified as fake  
- 493 Fake accounts correctly predicted  
- 9 Fake accounts misclassified as real  
These results show very strong classification performance with only 11 misclassifications out of 1000 samples.

## Feature Importance

The Random Forest model allows us to visualize which features had the most influence in classification. The most important features were:  
- Profile picture presence  
- Number of followers  
- Number of posts  
- Username structure (e.g., numbers and length)  
- Description length  
These features help distinguish genuine user behavior from automated or suspicious accounts.

Feature Importance Visualization:



## CSV Output

The model outputs were saved to a CSV file located at `C:\Users\Naser\Downloads\Fake\_ID\_Detection\_Results.csv`. This file contains:  
- Scaled input features  
- Actual class labels  
- Predicted class labels  
- Probability that the account is fake  
This file helps in understanding which accounts were predicted correctly or incorrectly, along with the confidence score.

## Conclusion

This project successfully demonstrates how machine learning can be applied to detect fake social media accounts. Using a well-structured dataset and Random Forest Classifier, we achieved 99% accuracy. The project highlights the importance of features like profile pictures and engagement metrics in detecting fake profiles. In future work, deep learning models or more granular features (e.g., comments analysis, NLP on bios) could be explored for even more accuracy.