**Project Report**

**AI Lab(D1)**

****

**Instructor:**

**Usama Amjad**

**Submitted By:**

**Anas Iftikhar(F2020266208)**

**Muqaddas Iftikhar(F2022332052)**

**Talha Ahmad(F2019266302)**

**Session**

**Spring 2024**

**University of Management and Technology**

**C-II Johar Town Lahore Pakistan**

**INSTAGRAM FAKE ACCOUNT DETECTION**

1. **Introduction**

The research entails applying several machine learning algorithms to analyze a dataset of Instagram accounts in order to identify phony accounts. The processes from data cleansing and investigation to model construction and assessment are described in depth in this study.

**2. Data Overview**

The dataset includes a number of Instagram profile-related characteristics, such as:  
  **nums/length username:** Ratio of numerical characters in the username.

 **fullname words:** Number of words in the full name.

 **nums/length fullname:** Ratio of numerical characters in the full name.

 **name==username:** Binary feature indicating if the name is the same as the username.

 **description length:** Length of the profile description.

 **external URL:** Presence of an external URL.

 **private:** Account privacy status.

 **#posts:** Number of posts.

 **#followers:** Number of followers.

 **#follows:** Number of follows.

 **fake:** Target variable indicating if the account is fake.

**3. Data Cleaning and Exploration**

**3.1 Initial Analysis**

* **Shape:** The dataset contains multiple rows and columns.
* **Info:** The dataset includes various data types, mostly integers and floats.
* **Description:** Statistical summary of the dataset was generated.
* **Null Values:** No missing values were identified.
* **Duplicates:** Duplicate rows were found and removed.

**3.2 Data Visualization**

* **Numerical Characters in Username:** Distribution plotted using a histogram.
* **Number of Words in Fullname:** Distribution plotted using a histogram.
* **Name Equals Username:** Bar plot to visualize the count.
* **Description Length:** Distribution plotted using a histogram.
* **External URL Presence:** Bar plot to visualize the count.
* **Private Account Status:** Bar plot to visualize the count.
* **Number of Posts, Followers, and Follows:** Distribution plotted using histograms.
* **Fake Account Status:** Bar plot to visualize the count.

**3.3 Feature Relationships**

* **Boxplots:** Relationship between target variable (fake) and various features like numerical characters in username, number of words in fullname, description length, number of posts, followers, and follows.
* **Countplots:** Relationship between the target variable and binary features like external URL presence and private account status.

### 4. Data Preprocessing

* **Scaling**: Min Max Scaler was used to normalize the data.
* **Data Type Conversion**: Certain features were converted to integers.
* **Dropping Columns**: nums/length fullname was dropped as part of the cleaning process.

### 5. Model Building

#### **5.1 Splitting Data**

* **Train-Test Split**: Data was split into training (80%) and testing (20%) sets.

#### **5.2 Model Training**

* **Random Forest Classifier**: A Random Forest Classifier was trained with default parameters.

### 6. Model Evaluation

#### **6.1 Predictions**

* **Accuracy**: The model achieved an accuracy of 0.91 on the test set.
* **Classification Report**: Detailed report showing precision, recall, f1-score for each class.

#### **6.2 Confusion Matrix**

* **Plot**: The confusion matrix was plotted to visualize the true positives, true negatives, false positives, and false negatives.

### 7. Conclusion

The Random Forest Classifier successfully used the features provided to distinguish between fake and real accounts, demonstrating strong performance with an accuracy of 91%.