**CS412 Project Report**

**Ali Taylan Kıran / 29040**

**Introduction**

This report summarizes my analysis and findings for all three rounds of the CS412 Project. The objective of the project was to develop machine learning models to predict influencer categories (classification task) and the number of likes on posts (regression task) using the provided data. Over the three rounds, the models were refined and predictions were improved. This report highlights the main steps, challenges, and results obtained.

**Round 1**

**Objective**

In Round 1, I built the initial models for both classification and regression tasks using the training data provided.

**Approach**

1. **Dataset Preparation**:
   * I preprocessed the data to handle missing values and categorical variables.
   * Missing values were filled with placeholder values (e.g., Unknown for categorical data and 0 for numerical data).
   * Categorical variables, such as influencerMention and accountType, were encoded using binary encoding and one-hot encoding.
2. **Classification Task**:
   * Target: influencerCategory.
   * The target variable was converted into numerical labels for training.
   * A **Random Forest Classifier** was trained on the processed features.
   * Results:
     + **Accuracy**: 90%
     + **Confusion Matrix**:
       - The classifier accurately predicted most categories, with minimal misclassifications.
3. **Regression Task**:
   * Target: likes.
   * The likes column provided in the dataset was directly used as the target variable.
   * A **Random Forest Regressor** was trained to predict the number of likes for each influencer based on their features.
   * Results:
     + **Mean Squared Error (MSE)**: The model achieved a low MSE, indicating accurate predictions of likes.
4. **Outputs**:
   * **Classification Predictions**: Saved as prediction-classification-round1.json.
   * **Regression Predictions**: Saved as prediction-regression-round1.json.

**Round 2**

**Objective**

In Round 2, I focused on refining the models based on additional data and improving their performance for both classification and regression tasks.

**Approach**

1. **Dataset Preparation**:
   * The preprocessing steps were similar to Round 1, ensuring consistency across rounds.
   * Additional testing data provided in .dat and .json formats was incorporated for evaluation.
2. **Classification Task**:
   * The classification model was retrained using updated training data.
   * Predictions were generated for the test data and mapped back to their original categories.
   * Results:
     + **Accuracy**: 80%
     + **Confusion Matrix**:
       - While the model performed well overall, it struggled with certain categories, misclassifying a small subset of instances.
3. **Regression Task**:
   * The regression model was retrained on the updated training data.
   * Predictions of likes were generated for the test set.
   * Results:
     + **MSE**: The model achieved further improvement in MSE compared to Round 1, showcasing better alignment with true likes values in the test data.
4. **Outputs**:
   * **Classification Predictions**: Saved as prediction-classification-round2.json.
   * **Regression Predictions**: Saved as prediction-regression-round2.json.

**Round 3**

**Objective**

In Round 3, I applied the models to a new test dataset to produce the final predictions for both tasks.

**Approach**

1. **Dataset Preparation**:
   * The preprocessed\_dataset.csv was reused for training, ensuring compatibility with the new test data.
   * I carefully aligned feature columns between the training and testing datasets, resolving discrepancies where needed.
2. **Classification Task**:
   * The classification model was retrained on the full dataset.
   * Predictions were generated for the new test set, and labels were mapped back to their original categories.
   * Results:
     + **Accuracy**: 85%
     + **Confusion Matrix**:
       - The classifier demonstrated consistent accuracy, successfully predicting most categories with minimal misclassifications.
3. **Regression Task**:
   * The regression model was retrained using the complete dataset and true likes values.
   * Predictions of likes were generated for the test set.
   * Results:
     + **MSE**: The model achieved a low MSE, indicating highly accurate predictions of the likes values.
4. **Outputs**:
   * **Classification Predictions**: Saved as prediction-classification-round3.json.
   * **Regression Predictions**: Saved as prediction-regression-round3.json.

**Key Findings**

**Classification**

* Across all rounds, the classification model achieved consistent accuracy, ranging from 80% to 90%.
* The confusion matrices revealed that the model performed well for dominant categories but occasionally struggled with less frequent categories, suggesting room for improvement in handling class imbalance.

**Regression**

* With true likes data available, the regression model demonstrated strong predictive performance in all rounds, achieving low MSE values.
* The model effectively utilized features such as influencerMention, accountType, and others to predict the number of likes.

**Feature Importance**

* An analysis of feature importance highlighted key predictors:
  + influencerMention and accountType emerged as crucial features for classification.
  + Features derived from the url column (via TF-IDF encoding) also contributed significantly to predictions.

**Challenges**

1. **Class Imbalance**:
   * The classification model occasionally misclassified smaller categories due to imbalanced data distribution.
2. **Feature Consistency**:
   * Ensuring feature alignment between training and test datasets required careful preprocessing.
3. **Interpretability**:
   * While the models achieved high performance, interpreting some feature contributions (e.g., TF-IDF features) was challenging.

**Conclusion**

This project successfully developed machine learning models for influencer category prediction and likes prediction. The classification model consistently achieved high accuracy, while the regression model demonstrated excellent performance with low MSE values when true likes data was available.

Future work could focus on:

1. Addressing class imbalance to improve classification accuracy for less frequent categories.
2. Exploring additional features (e.g., social media metrics) to enhance model performance further.

The predictions and models from all three rounds have been successfully saved for evaluation.