Summer 2023 CIberCATTS

Machine Learning Project Outline

1. Dataset Selection:

    The IMDB dataset is obtained from HuggingFace datasets. This dataset comprises movie reviews from IMDB and consists of 25,000 highly polar movie reviews for training and 25,000 for testing. Additionally, there are extra unlabeled data available for use.

    Alternatively, you can select a suitable dataset from Kaggle that aligns with your project goals and interests. Ensure that the dataset includes both input features and a target variable for classification or regression tasks.

2. Data Exploration and Visualization:

    Load the dataset and perform exploratory data analysis (EDA) to understand the structure, patterns, and distributions of the data. Use visualizations such as histograms, scatter plots, and heatmaps to gain insights into the relationships between variables.

3. Data Preprocessing:

    Preprocess the dataset to handle missing values, outliers, and categorical variables. Perform text cleaning tasks, including removing punctuation and unwanted characters, converting to lowercase, removing stopwords, tokenization, normalization (stemming and lemmatization), and feature extraction using bag-of-words or other chosen methods. Split the data into training and testing sets.

4. Model Selection and Training:

    Choose three machine learning algorithms suitable for your dataset and problem (e.g., decision tree, logistic regression, random forest). Train each model using the training set. Use appropriate techniques, such as cross-validation, to optimize model hyperparameters.

5. Model Evaluation:

    Evaluate the trained models using the testing set. Calculate performance metrics such as accuracy, precision, recall, and F1 score for classification tasks, or mean squared error or R-squared for regression tasks. Compare the performance of the three models and identify the best-performing one.

Tip: On the course’s GitHub, there is a notebook provided for you that covers different model evaluation techniques. You can seek help from it.

6. Interpretation and Analysis:

Analyze the results and interpret the findings. Identify the important features contributing to the model's predictions using feature importance techniques.

Tip: On the course’s GitHub, there is a notebook provided for you that covers identifying feature importance in fandom forest. You can get help from it.

7. Conclusion and Reporting:

Summarize the project findings, including the performance of the models, insights gained from data visualization, and interpretations. Communicate the results effectively through visualizations, reports, or presentations.

Tip: Remember to document your code, explain your thought process, and provide clear interpretations of the results.