Data Analytics I

CS5525

Final Term Project (FTP)

The goal of the final term project is to gain practical experience with machine learning classifiers. The objective of the FTP is to apply various machine learning classifiers to a real dataset. The expectation is to apply the course learning objectives to a real dataset and recommend a classifier that classifies the selected dataset with the highest performance. The final term project mainly consists of three phases:

1- Exploratory Data Analysis (EDA)
2- Regression Analysis
3- Classification Analysis
4- Clustering analysis and association mining

The first step in FTP is the data selection. The dataset must satisfy the following criteria:

- Pick and an interesting, applied real world dataset from industry.
- It must be a multivariate dataset with at least 50K observations.
- It must contain numerical & categorical data with at least 2 for each category.
- It could be a time series or non-time series.
- It must come from non-classified (public) database.
- Update the provided google excel sheet with the selected dataset as soon as possible. The deadline to select term project dataset is by 3/13/2023.
- https://docs.google.com/spreadsheets/d/1JoHc7Xh2FnC7LQqxGm4wSenRIAtiEOEnfjMQ5lOp7dE/edit#gid=0
- If two students select the same dataset, the dataset will be assigned to the first student and the second student needs to pick another dataset. First come first serve.
- The dataset must be split into 80-20% train-test.
- You need to use python to accomplish the objectives.
- You are free to use any package in python.

There are several resources available to acquire dataset i.e.

- https://www.kaggle.com/
- https://archive.ics.uci.edu/ml/index.php
- https://datasetsearch.research.google.com/
- https://analyticsindiamag.com/top-10-popular-publicly-available-datasets-deep-learning-research/

I. Phase I: EDA

You need to define which feature is the target and which variables are considered attributes. The explanatory data analysis could consist of the followings:

- Data preprocessing
    - Data cleaning: pick a method to fix the missing data or nan's objects
    - Aggregation [if applicable]

- Down sampling [if applicable]
- Dimensionality reduction/feature selection:
  - Random Forest Analysis
  - Principal Component Analysis
  - Singular Value Decomposition Analysis
  - Write down your observations about above methods.
- Discretization & Binarization: Label Encoding/one hot encoding. Write down your observations.
- Variable Transformation: Normalization, standardization, differencing.
- Anomaly detection/Outlier Analysis and removal [i.e., distance-based/density-based or clustering-based]. Write down your observations.
- Sample Covariance Matrix display through heatmap graph. Write down your observations.
- Sample Pearson Correlation coefficients Matrix display through heatmap graph. Write down your observations.

II. Phase II: Regression Analysis

- T-test analysis
- Association Analysis
- F-test analysis
- Final regression model
- Confidence interval analysis
- Stepwise regression and adjusted R-square analysis.
- Collinearity analysis i.e VIF method.

III. Phase III: Classification Analysis:

In this phase you need to apply various machine learning classifiers to the selected dataset and pick the best technique and recommend a classifier with the highest performance. Some of the main classifiers are:

- Decision tree
- Logistic regression
- KNN
- SVM
- Naïve Bayes
- Random Forest
- Neural Network

Phase IV: Clustering and Association [independent study]:

This phase of the project is the independent research. The following algorithms needs to be applied to your dataset. Write down your observations about the clustering and association mining analysis of the selected dataset.

- K-mean algorithm
- DBSCAN algorithm
- Apriori algorithm

To compare the performance of different classifiers, it is expected that you have the followings:

- Display Confusion matrix
- Display Precision
- Display Sensitivity or Recall
- Display Specificity
- Display F-score
- Display ROC curve

In this phase you need to have a graphical representation of the classification result. You need to show how different classes are classified and misclassified using the selected machine learning classifier graphically.

A formal report and presentation are required by the deadline.

SPECIFIES

The final formal report (pdf format) must be typed and should contain the following sections: [APA format]

1- **Cover page.**
2- **Table of content.**
3- **Table of figures and tables.**
4- **Abstract.**
5- **Introduction**. An overview of the procedures to accomplish the FTP objectives and an outline of the report.
6- **Description of the dataset**: You need to provide a description on the selected dataset and how the dataset satisfies the dataset criteria. You need to specify which variable in the selected dataset will serve as dependent variable and which ones serve as independent variables. You will need to explain the importance of the selected dataset in industry.
7- **Phase I**: see above
8- **Phase II**: see above
9- **Phase III**: see above
10- **Recommendations:** This section of your FTP report provide a summary and recommendations after classifying the dataset. Recommendation is an important section of your final report which could include the followings:
    a. What did you learn from this project?
    b. Which classifiers perform the best for the selected dataset?
    c. How do you think you can improve the performance of the classification? This could be in the future work section.
11- A **separate appendix** should contain supporting python codes that is developed for this project.
12- **References**
13- The **soft copy of your python programs** needs to be submitted separately as a .py to verify the results in the report. Make sure to include the dataset in your submission. Make sure to run your

code before submission. If the python code generates an error message, 50% of the term project points will be forfeited.

14- Include a **readme.txt** file that explains how to run your python code. All the results in your report must be regenerated to grant the grade.
15- The FTP is defined to be individual unless an approval is granted for collaboration. All the coding must be done individually, and it must be genuine. Copying a code from internet without proper citation will be considered as a **plagiarism** and FTP grade will be disregarded. Make sure to write your own code to avoid future complications.
16- All figures in your report must include a proper x-label, y-label, title, and a legend [if applicable]. Pick an appropriate theme or style for the plotted graphs. If you have a table inside your report, then make sure to include a proper title. Including grid is optional.
17- **Final presentation:** You will be given 20 minutes to present your term project to the class. The presentation weighs 20% of the term project grade. You need to create a power point for your presentation and submit the power point presentation. Due to the number of students in class and time constraint, some students will be asked to record their presentation and submit the recorded video. Due date by **Monday May 1st**
18- **Final formal report submission** weighs 80% of the FTP and is due by **Friday, May 5th**.

Upload the **final report (as a single pdf**) plus **the .py file(s)** through course shell canvas by the due date.