COLLEGE OF ENGINEERING
COMPUTER SCIENCE
VIRGINIA TECH

Machine Learning I

CS 5805

Final Term Project (FTP)

The goal of the final term project is to obtain practical experience with feature engineering, classification, clustering, and association rule mining. The objective of FTP is to apply various machine learning algorithms 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 consists of three phases:

1- Feature engineering & exploratory data analysis (EDA)
2- Regression Analysis [on a selected continuous numerical feature]
3- Classification Analysis [on a selected categorical feature binary classification or multi label classification]. For multi label classification perform:
    a. One-vs-all
    b. One-vs-one
4- Clustering analysis and association rule mining.

I. Phase I: Feature Engineering & EDA [20 pts]

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

- Data preprocessing
    - Data cleaning: pick a method to fix the missing data or nan's objects.
    - Check for data duplications and removal.
    - Aggregation [if applicable]
    - Down sampling [if applicable]
    - Dimensionality reduction/feature selection:
        - Random Forest Analysis
        - Principal Component Analysis and condition number
        - Singular Value Decomposition Analysis
        - VIF
        - You need to make sure that the collinearity does not exist in the data matrix.
        - Write down your observations about the above methods and the dimensionality reduction method that you picked for this project.
    - Discretization & Binarization: Label Encoding/one hot encoding (avoid the dummy variable trap). 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.
- Balanced or imbalanced data. If the target is imbalanced, what method did you use to make it balanced?

II. Phase II: Regression Analysis [15 pts]

For this phase you will need to make a prediction on a continuous numerical feature using a multiple linear regression model. You need to incorporate the following operations in your analysis. You need to plot the train, test and predict variable in one plot. Show the final model and develop table showing the R-squared, adjusted R-square, AIC, BIC and MSE.

- T-test analysis
- F-test analysis
- Final regression model and prediction of dependent variable.
- Confidence interval analysis
- Stepwise regression and adjusted R-square analysis.

III. Phase III: Classification Analysis: [24pts]

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. The purpose of this phase is to improve the performance of classification and make sure that the classifier is not overfitted or underfitted. You need to perform grid search for hyper parameter search for each classifier. The following classifiers are needed for this project:

- Decision tree
  - Pre-pruning, post pruning, grid search for optimum parameters: 'criterion', 'splitter', 'max_depth', 'min_samples_split', 'max_features', 'ccp_alpha'
  - Optimize Cost Complexity function.
- Logistic regression
- KNN
  - Find optimum K using elbow method.
- SVM
  - With linear kernel, polynomial kernel, radial base kernel.
- Naïve Bayes
- Random Forest
  - Bagging, Stacking, Boosting
- Neural Network
  - Multi-layered perceptron.

For each classifier and evaluate its performance, it is expected that you have the followings: **[5 pts]**

- Display Confusion matrix
- Display Precision
- Display Sensitivity or Recall
- Display Specificity

- Display F-score
- Display ROC and AUC curve
- Stratified K-fold cross validation

You need to display/perform above to each classifier. Create a table that compares the performance of different classifiers and **recommends the best classifier for the dataset**. You need to have a graphical representation of the classification result and show how the different classes are classified/misclassified.

Phase IV: Clustering and Association [independent study]: [10pts]

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

- K-mean or K-mean++ algorithm
  - Silhouette analysis for the k selection, within-cluster variation plot,
- DBSCAN algorithm
- Apriori algorithm

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]. You can use latex or other editors to create the final report. **[6 pts]**

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 of 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 variable s. 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- **Phase IV**: see above.
11- **Recommendations:** This section of your FTP report provides 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.
    d. What features are associated with the target variable?

e. Number of clusters in this feature space.

12- A **separate appendix** should contain supporting python codes that are developed for this project.
13- **References**

Submission Notes

- 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.
- 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.
- 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 code from the internet without proper citation will be considered **plagiarism** and FTP grade will be disregarded. Make sure to write your own code to avoid future complications.
- 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.
- **Final presentation:** You will be given 20 minutes to present your term project. 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. You need to record your presentation and submit the recorded video. Due date by **December 7th .**
- **The final formal report submission** weighs 80% of the FTP and is due by **December** 8th.
- Upload the **final report (as a single pdf**) plus **the .py file(s)** through canvas by the due date.