CS677 Final Project

Picking the Data Set

Look into the following sites as an example and select a time-series data set that interests you. The dataset should have numerical attributes and/or categorical attributes. One of the attributes should allow you to do classification tasks. You can also pick on dataset for time-series analysis and another dataset for other machine learning techniques.

1. https://www.kaggle.com/datasets
2. https://archive.ics.uci.edu/ml/index.php
3. Any other source of your choice

Preparing the data

- Import the data set into Pandas dataframe.
- Document the steps for the import process and any preprocessing had to be done prior to or after the import. Any python code used in the process should be included.

Analyzing the data

- Provide appropriate plots and interpretations for the attributes of the dataset. Analysis should include the standalone attributes as well relationships amongst the attributes.
- Do the time series analysis and forecasting predictions of the dataset. Provide the appropriate plots and interpretations.
- Do linear and logistic regression analysis on the data. Provide the appropriate plots and interpretations.
- Using Principal Component Analysis, determine which attributes are important for the analysis.
- Perform classification analysis using Naïve Bayes, Decision trees, and Support Vector machine algorithms. Provide the appropriate plots and interpretations.
- Do the clustering techniques on the dataset. Provide the appropriate plots and interpretations.
- Use pipelines where appropriate for the above techniques.

Presenting the Project

- **You will do your project presentation in the class on Dec 9th.**
- **The final code for the project will be due on Monday, Dec 9th, 5 PM EST.**

Submitting the Project

Upload a zip file (CS677Project_lastName.zip) containing all the source code (the notebook file), the presentation document if any (PDF or PPT), and a PDF of all the code and results.