Priyanka Kondapalli
CSC 9025 – Grand Challenges of Computing

Data Analytics Portal

Project Idea

The main idea of this project is to provide a platform for users to perform basic data analytics that help understand the data and relationship between various features in the dataset. Also the user can evaluate different models on their data sets and compare the accuracy scores.

Overview

Portal UI:

Below are the features planned to be implemented in the application

1. Descriptive Statistics – Describe the basic features of the data
2. Exploratory Data Analysis – Analyze data sets to summarize their main characteristics with visual methods.
3. Model Scores – View accuracy of various Statistical models that predict outcomes

The UI will allow the user to upload a dataset in .csv format to view the results for any of the above features

Input Dataset specifications

All the features in the dataset must be labeled, quantitative or categorical and any outcome feature should be ordered as the last column in the csv file.

1. Descriptive Statistics

After uploading the dataset, a python script is executed in the backend to display basic information about the data and the descriptive statistics

i. Data Types
ii. Mean
iii. Median
iv. Mode
v. Standard Deviation
vi. Variance
vii. Quartiles
viii. Missing Data
ix. Correlation

Once the descriptive statistics are displayed, we will deal with missing values.

Dealing with Missing Values

All the columns with more than 20% missing values will be dropped as part of the data cleaning process. We will use imputation method for filling the remaining missing values choosing mean of the column for imputation.

The cleaned data will be available for the user to download.

2. Exploratory Data Analysis

Exploratory data Analysis will use the cleaned dataset as input file and create plots useful for exploring the data and relationship between the features. We will generate the following plots

i. Histograms
ii. Box Plots
iii. Scatter Plots
iv. Correlation
v. Density Plot

Based on the plots the user may decide to drop any columns not highly relevant with the analysis.

3. Model Scores

Using the Model Scores the user will be able to view the accuracy scores of different models

i. Logistic Regression
ii. Decision Tree
iii. KNN
iv. Random Forest
v. Support Vector Machine

We may also include the precision and recall scores for each model

All the results will be presented to the user with a short description of what the method is actually used for and how does it work.

Tools & Technologies

Below are the tools and technologies we will be using to develop the application

Technologies	: ASP.NET, AJAX
Programming Language	: C#.NET
Scripting Language	: Python, jQuery
Tools	: Visual Studio 2017, AJAX Toolkit

Conclusion

These features will provide the user with a basic idea and knowledge to move to the next level of analyzing their datasets.