**Final Project**

You have to write a paper (6 to 8 pages) about your project.

**Choose a dataset from**

UC Irvine Machine Learning Repository (<https://archive.ics.uci.edu/ml/index.php>)

with at least 5000 instances and 20 attributes for classification or regression.

**Compare** how the different approaches seen in class perform on this dataset to

**Predict** accurately the classes or the values of the unlabeled data.

**Determine** what are the best hyper-parameters for each approach you are using.

-we need to split our dataset 75% of the data for training and 25% of the data for testing

**Linear Regression Single Variable**

-We want to minimize the error between our points and a line of best fit.

**Cross Validation**

-Rotate the 25% testing data through the data set, record the results for how well each method does and compare the results to each other. Since, we divided the data into 4 blocks, this is called Four-Fold Cross Validation.

**Python libraries**

-pip

-pandas

-Matplotlib Plotting

-sklearn

**Presentation**

You have to include a presentation of the research questions and the chosen methods to tackle them, Relevant Literature Review, a presentation of the results and discussion and a conclusion/future work.

**Table of Contents**

Abstract

General overview

Dataset Strategies

Motivation

Relevant Literature Review

Decision Tree

GaussianNB

Discussion of findings

Concluding remarks

References

Appendix

**Abstract**

Regression and classification are two fundamental tasks in machine learning. Regression intends to predict a continuous output variable, while classification intends to assign a categorical label to a data instance. Various algorithms will be discussed throughout including regression and classification algorithms, linear regression, logistic regression, decision trees, random forests, and support vector machines. Using these algorithms to compare the different approaches seen in class on this dataset to accurately predict the values of the unlabeled/training data.

Determine what are the best hyper-parameters for each approach you are using.

**General overview**

?

**Dataset Strategies**

The dataset that was used came from <https://archive.ics.uci.edu/ml/index.php>. The data is based on Diabetes in US hospitals between the years 1999 to 2008. The number of instances in this dataset is 100,000 and it has 55 different attributes. Inspection of the dataset revealed that the data required cleaning as well as data transformation. This is due to the dataset containing invalid data. This is in the form of question marks (?), string values (“Female”) and age ranges ([0-10)). This required dropping columns that were not relevant to the analysis or contained no usable data. {fig 1 – show dataset with unfiltered data} The next step was to transform the data. This involved

-add headers

-add parsing

-add formatting

-add pattern recognition (heatmap)

**Motivation**

-something to help benefit society

**Relevant Literature Review**

**Experimental setup and methodology**

**Discussion of findings**

**Concluding remarks**

**Future work**

-better data collection

-more varied

-graph {race}

**References**

**Appendix**