**Paper Title (Something Interesting)**

**Find a relevant dataset from Kaggle**

Hayden Horn, hhorn@bellarmine.edu

Jacob Schelonka, Jschelonkat@bellarmine.edu

**Our project uses SVR, DT, and RF in order to predict the likelihood of an individual being accepted to a particular university. We will use three experiments using SVR, all differing in their splits and one with slightly different data. For the DT and RF experiments, both splits will be equivalent to each other. These were originally meant for comparison purposes, however both experiments preformed equal or better than the three SVR experiments. The splits made slight differences. Overall, all 5 experiments worked well and predicted accurately. On all five experiments we did see that some experiments would be extremely close to one or two predictions, but then miss the other predictions by full percentage points. Due to this we changed the splits and data to make them more accurate for all 4 validation sets. Decision Tree was almost 100% accurate, but it was off by a few decimal places on two of the validation sets.**

1. **INTRODUCTION**

We are using “Graduate Admissions 2” from Kaggle.com. This data set is designed for people to use the independent variables such as “University Rating” and “CGPA” to predict the admission of certain students. We will be using support vector regression, because from previous assignments, it has shown to be a great tool for predicting an output based on independent variables.

1. **BACKGROUND**
   1. *Data Set Description*

I chose to work on this dataset, because for linear regression it is easier to be more accurate when you have multiple independent variables that all go into the prediction. There are 7 independent variables being used to predict the administration of students. Every category is numerical, which will make predicting admission easier.

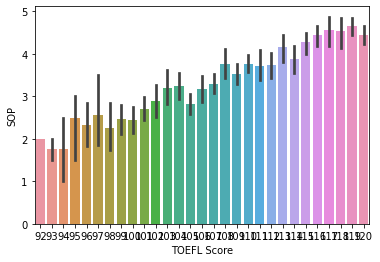
* 1. *Machine Learning Model*

We are using support vector regression because it is used for classification, regression, and outliers’ detection. you can use support vector regression methods not only for classification but also are really good for regression. Since it is a regression algorithm instead of using the curve as a decision boundary it uses the curve to find the match between the vector and position of the curve. It will be perfect for this test score dataset. We will also use Random Forest and Decision Tree as a comparison to our experiments.

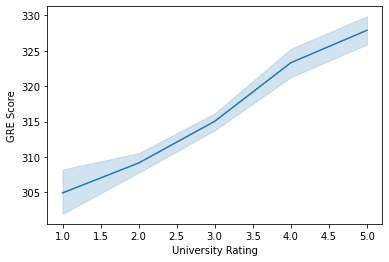
1. **EXPLORATORY ANALYSIS**

This dataset is comprised of 9 categories, all either integers or float. There are no null values in this dataset. All these categories seem to accurately correlate to what would be expected in a real-life scenario.Graphical user interface, Teams

Description automatically generated with medium confidence

****

**This shows a pretty accurate correlation between SOP and TOEFL Score.**

****

**This shows that university rating and GRE Score also correlate accurately.**

**Overall, this dataset seems well put together and realistic. There are a few outliers, but they seem to have minimal effect on the data.**

**Table 1: Data Types**

|  |  |
| --- | --- |
| *Variable Name* | *Data Type* |
| V1: GRE SCORE | Int64 |
| V2: TOEFL SCORE | Int64 |
| V3: UNIVERSITY RATING | Int64 |
| V4: SOP | Float64 |
| V5: LOR | Float64 |
| V6: CGPA | Int64 |
| V7: Research | Float64 |

1. **METHODS**

In this section, describe how you prepared the data for your model and performed multiple experiments using different parameters for the model.

* 1. *Data Preparation*

Since all 7 variables show an obvious correlation to the chance of admission, and none are null, we had to do very little to prep the data. The first experiment split the data 8/10. The second experiment had the data split ¼. The last two split the data by 1/3.

* 1. *Experimental Design*

Table X: Experiment Parameters

|  |  |
| --- | --- |
| **Experiment Number** | **Parameters** |
| 1 | SVR: Split = 8/10 |
| 2 | SVR: Removed SOP and split = 1/4 |
| 3 | SVR: Normal, Split = 1/3 |
| 4 | DT: Normal, Split = 1/3 |
| 5 | RF: Normal, Split = 1/3 |

* 1. *Tools Used*

The following tools were used for this analysis: Python running the Anaconda environment for Microsoft computer was used for all analysis and implementation. In addition to base Python, the following libraries were also used: Pandas 0.18.1, Numpy 1.11.3, Matplotlib 1.5.3, Seaborn 0.7.1, SKLearn 0.18.1. Pandas was needed to read the file and give us information on the data set. NumPy and mat plot didn’t play a large role but are still nice to have for plotting and looking at data. Seaborn and SKLearn was needed for graph, finding correlations, and making predictions. SVR, DT, and RF were all used for these experiments, but what was supposed to be the main three experiments used SVR.

1. **RESULTS**
   1. *Mean square Error and R-Square calculation*

Experiment 1: R2 = 0.775 MSE = 0.063

Experiment 2: R2 = 0.797 MSE = 0.061

Experiment 3: R2 = 0.610 MSE = 0.083

Experiment 4: R2 = 0.613 MSE = 0.083

* 1. *Discussion of Results*

Since Decision Tree got 2 of the 4 validation predictions exactly right, Decision tree performed the best. All 3 predictions were accurate, but none of them were perfect at guessing all for 4 of the chance of administration predictions. Even though SOP didn’t have the best correlation, taking it out of the X value barely changed the original prediction. SVR without ‘SOP’ was the most inaccurate experiment.

* 1. *Problems Encountered*

Going through the data, it’s obvious this data can be very randoms at time, and although there is enough to be consistent, outliers also are in play. Due to this, the predictions will be slightly more inaccurate than we wished. However, all four experiments ended up still being accurate.

* 1. *Limitations of Implementation*

Discuss the limitations of your model. Is there is reason it might not be the best way to model the data? What other models might work better?

* 1. *Improvements/Future Work*

To improve our work, we would have worked all experiments with Decision Tree. This is because how accurate it was with a 1/3 split. If we played around with the numbers and categories used, we could have gotten it even closer to perfect. Also, maybe a different dataset, however it did work well despite outliers.

1. **CONCLUSION**

**Decision Tree made two perfect predictions and was only 0.09 values away from being a perfect validation prediction. The other three are not perfect, but still are only inaccurate by small amounts. All together all of the experiments worked well.**