**Predicting Driving Behavior**

**Using** [**Kaggle data set:**](https://www.kaggle.com/datasets/outofskills/driving-behavior?select=train_motion_data.csv)

Keegan Henderson, [khenderson5@bellarmine.edu](mailto:khenderson5@bellarmine.edu)

**ABSTRACT**

Up to 150 word summary of your project.

1. **INTRODUCTION**

The data set I am working with was found on Kaggle. It focuses on driving behavior using numerous variables through accelerometers and gyroscopic sensors. This data set was created due to the AAA foundation for Traffic Safety reporting that 55.7% of deadly crashes were caused by aggressive driving.

The models I used were Logistic Regression, Random Forest, Decision Tree, and Support Vector Machine. I also used Voting classifier to give me the best combined result.

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

The data set was found on [Kaggle](https://www.kaggle.com/datasets/outofskills/driving-behavior?select=train_motion_data.csv), I found this set interesting as I hadn’t thought of driving categories and being able to predict them. The data was collected due to the number of fatal crashes (55.7%) being caused by aggressive driving. Therefore, they wanted to predict dangerous driving behavior quickly and accurately.

* 1. *Machine Learning Model*

Provide one or two paragraphs that describe the ML models (All four: MLR, SVR, DT, RF) for your data set and why you chose your final model. Explain in narrative how the model works. You do not have to provide a large amount of mathematical, but you can if want.

1. **EXPLORATORY ANALYSIS**

This section will be similar to your exploratory analysis project. First, provide a summary of the data set similar to your first exploratory analysis: *e.g. this data set contains 398 samples with 7 columns with various data types*. In this summary, provide the data types of your columns (in a table) and then rather than providing tabular statistics and plots for each variable, provide only statistics and plots that seem unusual. For example, if one or two variables have significant missing values or the distribution of the variable is skewed or looks unusual note that. Provide the unusual statistics or plots in this section. Provide any appropriate plots (e.g. correlation matrix, heatmaps, bar charts, etc.) that you deem necessary.

**Table 1: Data Types**

|  |  |
| --- | --- |
| *Variable Name* | *Data Type* |
| AccX | Float64 |
| AccY | Float64 |
| AccZ | Float64 |
| GyroX | Float64 |
| GyroY | Float64 |
| GyroZ | Float64 |
| Timestamp | Int64 |
| Class | Object |

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*

Not a whole lot of data prep, thought I needed to encode class however because it’s a classification problem there is no need, as it is the y variable.

* 1. *Experimental Design*

You will run your model several times with different parameters to see what different results you get. In a table, describe your experimental parameters. Three or four experiments are sufficient. This is where you will describe how you divided your data into train, validate and test data sets. For example:

Table X: Experiment Parameters

|  |  |
| --- | --- |
| **Experiment Number** | **Parameters** |
| 1 | 80/20 split for train, and test |
| 2 | 90/10 split for train, and test |

*Tools Used*

The following tools were used for this analysis: Python running the Google Collab environment for Dell Laptop was used for all analysis and implementation. In addition to base Python and Google Collab, 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.

I used these tools to aid my analysis. Python and google collab as a base platform. Then used mainly pandas and SKLearn for machine learning and manipulating the dataset.

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

Experiment 1:

Text

Description automatically generated with medium confidence

*Experiment 2:*

A screenshot of a computer

Description automatically generated with medium confidence

* 1. *Discussion of Results*

Discuss which of your models provided the best classification (or some other outcome if not classification). Explain why you think your best model was the best and why your worst model was the worst.

Logistic Regression was easily the worst at 0.37, possibly because it used all the data equally, and couldn’t find the best columns to utilize. All of the other ones were very close to one another. Random Forest at 0.997 and SVC, Decision Tree, and Voting Classifier at 1.0. I believe this is due to these algorithms being able to choose the best variable and put more weight within that one.

* 1. *Problems Encountered*

The largest problem I had was finding out the time, couldn’t figure out how to get it into hours 1-24. Would have liked to do some more implementation of that variable. I also felt like there was so much I could do with this set, and worried I am missing lots.

* 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?

I would like to see if there were hotspots of time where aggressive driving was rampant. There are also limitations as the data is from Romania, and therefore should only be used there.

* 1. *Improvements/Future Work*

What would you like to do to improve your model in future work? Do more experiments, use a different model, add/remove variables, find a different data set, etc?

I would like to run more experiments in the future and be able to use them on time.

1. **CONCLUSION**

Finish up with a paragraph or two of summarizing your problem, the results and your conclusions (good model, bad model, needs more work, etc.).

The goal was to be able to predict aggressive driving using cellular devices accelerometers and gyroscopes. The models that I found useful were very good and this is due to very good data collection, and picking the right models. I think this could be improved upon with more work, however with the limited timeframe I am very happy with these results.

**REFERENCES**

List any websites, books, articles, etc. that you found useful while you worked on this project. It is not necessary to cite the references in the paper unless you specifically mention it in the text.