# Machine Learning, Group Assignment 2022

Group A

**Maksim Gusev, Davin Mc Gowan, Rabin Sapkota, Senan Stanley,**

Atlantic Technological University

## Introduction:

Train two classifier methods and compare the results.

Kaggle data set - Traffic, Driving Style and Road Surface Condition.

## Methods:

Data preprocessing:

- Data concatenation;
- Feature assignment;
- NaN assignment;
- Standardization;
- Encoding Categorical Features;
- Training and test sets;

LR Model setting:

- Linear Regression – education;
- Linear Regression – prediction;
- Linear Regression – assessment;
- Linear Regression – metrics;

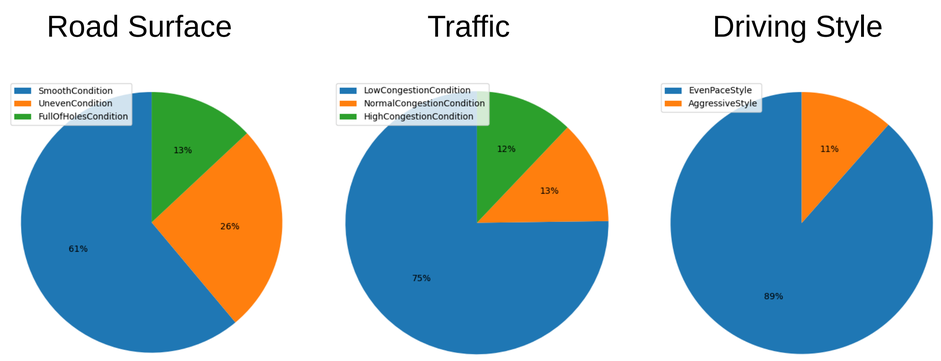
SVM Model setting:

- SVM – education;
- SVM – prediction;
- SVM – assessment;
- SVM – metrics;

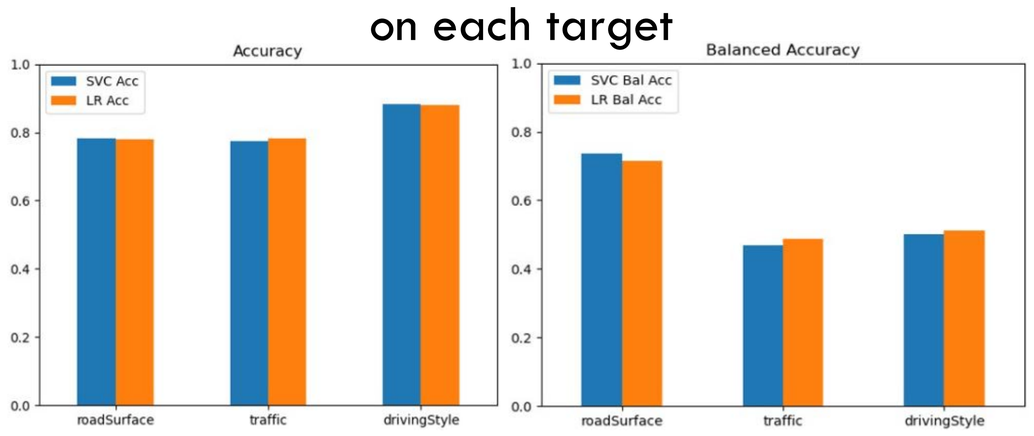
## Results:

Accuracy SVM/LR = 0.78/0.77
Balanced Accuracy SVM/LR = 0.73/0.71
F1-Score Breakdown SVM/LR = 0.86/0.85

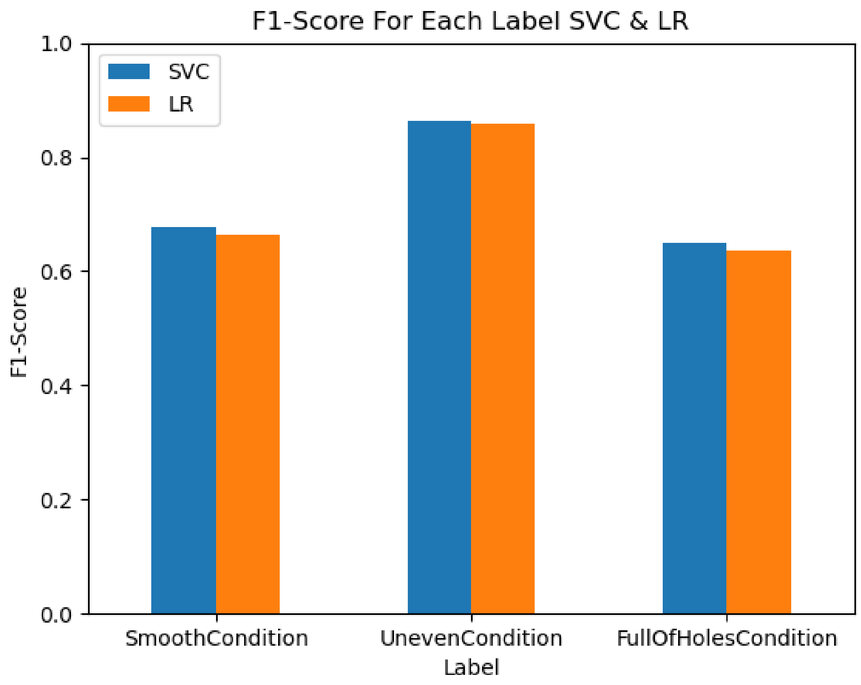
Label proportions for each variable




Accuracy & Balanced Accuracy for models trained on each target




F1-Score Breakdown for Road Surface




## Conclusions:

The imbalance of different labels in the target variables effected the classifiers accuracy and balanced accuracy. While the more unbalanced target variables showed higher accuracy, they also showed lower balanced accuracy indicating that the classifiers were doing a poor job of classifying less frequently encountered labels.F1 score more starkly reflects the poor precision and recall with the imbalanced data.

The SVM and Logistic Regression approaches are remarkably similar, and it's normal that the metrics gap is minimal. Low gap of metrics is a consequence of different approaches to loss function or that the SVM considers only points near the local boundary, while logistic regression considers global ones. In a situation with more balanced data, it could show bigger gaps in 2 methods metrics.

Improvements of models strongly depend on the quality and size of data. The results of the Linear Regression and Support Vector Machine models above, show a capacity for deployment in non-safety related functions. Further research into more complex models and hyperparameter optimization could yield a model more suited to compensate for imbalances in training data.

## Literature cited:

Jake VanderPlas, 2022. Python Data Science Handbook. Essential Tools for Working with Data.

## Acknowledgments:

Ollscoil Teicneolaíochta an Atlantaigh

Atlantic Technological University

## Further information:

https://github.com/DavinMcGowan/Machine-Learning-Group-Assignment-

https://www.kaggle.com/datasets/gloseto/traffic-driving-style-road-surface-condition