# 

# **Capstone Project for**

# **IBM Data Science Professional Certificate**

# **Comparative Study on Classic Machine learning Algorithms**

Mohsen Selseleh

2020,september

1. **A description of the problem and a discussion of the background.** (**15 marks**)

Collision severity (Accidents) is under-researched in all over the world, but the probable factors and correlation between these factors and collision severity outcomes is not known. In this study, machine learning based algorithms were employed to predict and classify collision severity.  The main aim of this research is to evaluate and compare different approaches to modeling collision severity. Four machine learning based models were developed: KNN, TREES, LOGISTIC REGRESSION AND SVM. These machine learning algorithms were validated using 10-fold cross-validation technique. The three-machine learning based algorithms were compared with one another. The results of the study reveal that the predictions of machine learning algorithms are like each other.I will be doing a comparative study over different machine learning supervised techniques like  Logistic Regression, K nearest neighbors and Decision Trees.. I will investigate their basic logic, advantages, disadvantages, assumptions.

1. B **description of the data and how it will be used to solve the problem. (15 marks)**

I found the data from Kaggle data sets:

<https://www.kaggle.com/mohitpote/collision-severity-accidents>

1. **Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.**

To begin this exploratory analysis, first import libraries and define functions for plotting the data using matplotlib. Depending on the data, not all plots will be made.I have used KNN, TREE, LOGISTIC REGRESSION AND SVM

Just like linear regression, Logistic regression is the right algorithm to start with classification algorithms. Even though, the name ‘Regression’ comes up, it is not a regression model, but a classification model. It uses a logistic function to frame binary output model. The output of the logistic regression will be a probability (0≤x≤1), and can be used to predict the binary 0 or 1 as the output ( if x<0.5, output= 0, else output=1).

K-nearest neighbors is a non-parametric method used for classification and regression. It is one of the easiest ML technique used. It is a lazy learning model, with local approximation.

Decision tree is a tree-based algorithm used to solve regression and classification problems. An inverted tree is framed which is branched off from a homogeneous probability distributed root node, to highly heterogeneous leaf nodes, for deriving the output. Regression trees are used for dependent variable with continuous values and classification trees are used for dependent variable with discrete values.

SVM is a supervised machine learning algorithm which can be used for classification or regression problems. It uses a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs.

**LR vs Decision Tree**:

* Decision trees supports nonlinearity, where LR supports only linear solutions.
* When there are large number of features with less data-sets (with low noise), linear regressions may outperform Decision trees/random forests. In general cases, Decision trees will be having better average accuracy.
* For categorical independent variables, decision trees are better than linear regression.
* Decision trees handles collinearity better than LR.

**LR vs SVM:**

* SVM supports both linear and non-linear solutions using kernel trick.
* SVM handles outliers better than LR.
* Both perform well when the training data is less, and there are large number of features.

**LR vs KNN:**

* KNN is a non -parametric model, whereas LR is a parametric model.
* KNN is slow in real time as it has to keep track of all training data and find the neighbor nodes, whereas LR can easily extract output from the tuned θ coefficients.

1. **Results section where you discuss the results.**

Accuracy coefficients are like each other and are strong, therefore all of these methods could be used for prediction collision severity.

|  |  |
| --- | --- |
| Method | ACCURACY |
| KNN | 0.895 |
| TREE | 0.896 |
| LOGISTIC REGRESSION | 0.896 |
| SVM | 0.84 |

1. **Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.**

I recommend that another method as Forest, Naïve method, are be considered and their accuracy compared with these methods.

1. **Conclusion section where you conclude the report.**

I could conclude that ML classical algorithms have their strong position in the field.