National College of Ireland

Project Submission Sheet – 2021/2022

Student Name:	Vishal Gajanan Patwardhan
Student ID:	X18190839@student.ncirl.ie
Programme:	MSc. In Data Analytics **Year:** 2021-2022
Module:	Data Mining & Machine Learning 1 (MSCDAD_B)
Lecturer:	Dr. Anu Sahni
Submission Due Date:	26/12/2021
Project Title:	A study into the performance of different Machine Learning techniques on three dataset – Germany Car Dataset, UK traffic Accidents & Heart Disease Indicator
Word Count:	5333

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
<u>ALL</u> internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	Vishal Gajanan Patwardhan
Date:	26/12/2021

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A study into the performance of different Machine Learning techniques on three datasets - Germany cars, UK traffic accidents and Heart Disease Indicator

Vishal Gajanan Patwardhan
School of Computing
National College of Ireland
Dublin, Ireland
x18190839@student.ncirl.ie

***Abstract*—This paper explains about six different machine learning methods applied on the three unrelated datasets naming Germany cars dataset, UK traffic accidents and Heart Diseases Indicator dataset. The six machine learning methods were linear regression, random forest (regression), K- nearest neighbor (also called as KNN), logistic regression, decision tree and random forest (classification). These models will predict the price of old cars, severity of accidents and detecting the heart disease using various factors which will be directly or indirectly affecting them. Further analyzing and comparing these models to find the best fit model for the dataset by evaluating them accuracy score and standard error associated with the model.**

Keywords — Car buying price, traffic accidents, Heart Disease, Regression, Classification, Algorithms.

I. INTRODUCTION

Machine Learning is a branch of Artificial Intelligence (AI) which deals with analytical model building and train the machine to adapt its environment independently. In the growing world of machine learning there is a need of answers for a critical question in the form of probability or in a percentage. Critical questions might come from any business organisation, health sector, and stock market handling industries. This report mainly focuses on how the same analytical machine learning models namely multiple linear regression, random forest (regression), K-nearest neighbour (KNN), logistic regression, decision tree and random forest (classification) can prove to be beneficial for the three un-related dataset of Germany cars dataset [1], UK traffic accidents dataset [2] and Heart disease indicator dataset [3] by predicting car prices based on the various independent factors, severity of the traffic accidents and identify whether or not the individual is having the heart disease. This prediction and analysis are done using some or many independent factors which are directly or indirectly associated with the predictor. Below is the short brief of all the three datasets used to answer the research questions.

A. Germany Cars Dataset

Everyone wants to own a personal car once in their lifetime to ease their comfort of travelling to places wherever and whenever they want. Some tend to buy a new car or a used car. There are many such online car markets which sells such (new or old) cars to fulfil the individual's dream of buying the cars. Autoscount24 [14] is one of such online car markets which is Europe based having presence in 18 countries with a traffic of more than 30 million users per month and more than 43,000 partnered dealers who provides humongous varieties of cars to feature in Autoscount24's online market. According to the study, there has been a rapid growth in demands of buying old cars in the recent years.

This dataset was chosen because there has been related work mentioned in one of the published journals by Pattabiraman Venkatasubbu, Mukkesh Ganesh [2] to predict the car prices using the random forest and multiple linear regression in their future work to find the more accurate model than their study using Multiple Regression, Lasso Regression and Regression Trees.

This dataset has 46,406 number of records or total number of old cars listed for re-sell in the online market. Being such a huge dataset will apply the appropriate machine learning models. Research question identified for this dataset is below:

1) How much would be the car re-sale price based upon the factors like car's mileage, horsepower (hp), fuel type, make year and gear type?

For the above question, the target variable identified is the price which is the continuous variable. It is important to predict the re-sale price of the car based on the certain factors which will help the Autoscout's [14] consultant to efficiently identify the approximate re-sale price of a car. Because it sometimes becomes difficult to define a price for re-sale cars.

B. UK Traffic Accidents Dataset

Traffic accidents are the incidents which occur when a motor vehicle collides with another vehicle, animals, people, or any other stationary objects in the road or in the road's debris. There is an increase in such accidents resulting into slight, serious, and fatal type of injuries which costs life of a person. In the year 2013, there has been around 54 million people across the world who got themselves injured in the traffic accidents [6]. Traffic accidents mainly occur due to the distractions while driving like using mobile phones, driver not having adequate amount of sleep, light conditions, and weather conditions. Accidents are classified with the three levels of accident severity ranging from slight, serious, and fatal.

This dataset was chosen because it had the accident records of UK from the year 2005 till 2015 which had around 17 Lakhs of the accidents recorded [2]. But to decrease the computation time for the analysis had to subset the dataset. Also, there were in-detailed factors and details about the surrounding of the accident which might or not be contributing to the actual accident cause or severity, but it can be useful by

finding some useful information based on them. Question based on the UK Traffic Accidents is below:

1) How severe (slight / serious) would be the accident based upon the several surrounding environmental factors?

For the above question, predictor variable would be accident severity which can be directly or indirectly linked to the various factors like road type, special conditions at site, speed limit, light conditions, weather conditions, area type which means whether accident occurred in urban area or rural area and day of week.

C. Heart Disease Indicator Dataset

Day by day there is a huge increase of people getting diagnosed with heart disease. As the heart being the most important functional part of the body, it needs to be taken care by following some best practices which can instructed by the physician or a dietician. Heart disease can be of different types such as Cardiovascular Disease, Coronary Artery Disease, Heart Failure, Heart Valve Disease and Congenital Heart Disease. In 2009, there were approximately 17.9 million people died from cardiovascular disease which roughly states that around 32% of all global deaths is due to the same disease [10].

This dataset was chosen to study the large amount of dataset comprising of 2,53,681 people's data with their various physical and other underlying disease along with their habits. Together which they can bring more information. And it is much more important to identify such disease beforehand so that humans take appropriate medications and consultations from the doctors. With the help of the machine learning models, it allows to predict various factor which leads to heart diseases. Below is the research question compounded to the heart disease.

1) Is the patient diagnosed with heart diseases related to their other health problems and their levels?

As per above research question, heart disease (yes or no) is the predictor with having independent factors like Body Mass Index (BMI), whether a patient smokes, having high Blood Pressure (BP), high cholesterol level, and diagnosed with diabetes, sex, and age.

II. Related work

These three datasets were chosen based on the earlier published academic papers and journals by the authors. Going through the published paper and journals it helped to grow more insights of these three datasets and put me into the right direction of the model building technique approach.

A. Germany Cars Dataset

In [3] authors have tried to build price predication model for the used cars in Bosnia and Herzegovina by applying three machine learning techniques which were Artificial Neural Networks, Support Vector Machine, and Random Forest. Dataset was collected though web scraping from autopijaca.ba web portal using the PHP programming language. There was the process of data cleaning which was mentioned to acquire best accuracy out of the model's outcome. Upon applying the model and evaluating them their results for Random Forest was 41.18% accuracy, Artificial Neural Network gave 42.35%, and Support Vector Machine gave 48.23%. This evaluation results led to ensemble of multiple machine learning algorithm which helped them to gain more accuracy of 92.38%. After building the model they compared their performance and found the best model that fit into their available dataset. Final model had the accuracy of about 87.38 % when tested on the test data which is quite good in terms of car price prediction as there are many factors which needs to be taken care of while building the model. Further they integrated the model into the Java application. Future work mentioned was to test their model on different datasets in which they desired to apply it on eBay [17] and OLX [16].

In [4], authors predicted the price of used cars in Mauritius. They performed the machine learning technique on the historical data found in the daily newspapers. They applied total of four machine learning models namely multiple linear regression analysis, KNN, naïve bayes, and decision trees. During the evaluation of all the models they found the mean error of Rs. 51,000 was found in the multiple linear regression whereas Rs. 27,000 for Nissan card and for Toyota's car it was Rs. 45,000 mean error using kNN model. And for Naïve bayes, accuracy was between 60-70% using different combinations performed on the independent variables. During the analysis they found that decision trees and naïve bayes were not contributed well to handle the numeric values. So, the reason the converted price to ranges of values but that too led to in-accuracies in their models. As their analysis was performed on very less amount of data, they would like to collect more and more data to apply the same model to test the accuracies of the models used in the paper.

In [5], researchers predicted the car prices using the data which was available on www.pakwheels.com[18]. Supervised learning Machine learning technique applied was the Multiple Linear Regression which offered them about 98% percent of accuracy which is excellent. This accuracy states that researchers took many efforts to clean the data well and selected relevant set of predictor variables. Main predictor variable which gave highest accuracy were car's model, make year, city, version, colour, mileage, rims, and power steering. Their future work was to apply the same dataset to fuzzy logic, KNN and genetic algorithm.

In [6], authors performed the supervised machine learning method using linear regression and K Nearest Neighbour (kNN). Authors collected the dataset from the Kaggle website which is provides large number of various datasets. During the model building technique, they splitted the dataset into two datasets naming them as test and train dataset. They performed regression model on the train and further on the test dataset to achieve the accuracy of about 85% for kNN whereas 71% accuracy for linear regression model which fitted best into the optimized model. Final model was validated using 5 and 10 k-fold methods. Future work mentioned was to apply more advanced machine learning method and validating with different optimization techniques.

B. UK Traffic Accidents Dataset

In [7], researchers termed the traffic accidents as one of most important concerns of the world as it leads to serious and fatal injuries. They researchers aimed to develop a model which will be capable of predicting the severity of the traffic accidents using the accidents dataset from Setúbal, Portugal for the period 2016 to 2019. In addition, they also proposed to build a machine learning model based on the historical data of

the traffic accidents. Authors used supervised machine learning methods such as decision trees, random forest, logistic regression, and naïve bayes. And used DBSCAN and hierarchical clustering for unsupervised machine learning models. During evaluation of the models performed there was 73% accuracy for both random forest and logistic regression and 65% for decision trees. Researchers highlighted that in majority, motorcycles accounted more accidents than the other vehicles. In their future work they would like to run the same model in the recent dataset of 2020/21to check how well their model perform. Moving forward they would like to perform Artificial Neural Networks and deep learning machine learning method as they perform well and are successful. As per analysis they would also like to focus more on building machine learning algorithms solely on motorcycles accidents.

In [8], authors showed the great concern of the traffic accidents fatalities in the world which also affects the economic loss every year. In the study, they collected the data of traffic accidents from Regional Traffic Division and collected weather data from Regional Directorate of Meteorology for the period from 2005 to 2016. They tried to build the link between the accidents and the weather to find out if there are any factors of weather affecting the traffic accidents. To analyse the same, the applied five major machine learning techniques namely k-Nearest Neighbour, Naïve Bayes, Multilayer Perceptron, Decision Trees, Support Vector Machine, one statistical method, Logistic Regression. During evaluation they found that Decision Trees, k-Nearest Neighbour, and Multilayer Perceptron gave the highest accuracies of 90%, 90% and 92% respectively than other models. And further analyses proved that cloudiness, traffic control existence and Ground surface temperature showed very positive effects on the traffic accidents whereas maximum temperature and weather parameters showed insignificant to contribute to the models.

In paper [9], researchers aimed to reduce the accident effects by identifying the possibilities of various factors using the limited resources of the data. Data was collected from the GES vehicle accidents data from 1995 to 2000. They presented three machine learning models such as Neural Networks, hybrid decision trees and neural network, and decision trees using concurrent hybrid models. Evaluating and comparing all the models it proved that hybrid decision tree with neural network gave the best performance with respect to individual approaches. And the future work mentioned was to test the same model on the large dataset with recent accidents data.

In [10], authors collected the accidents dataset of Abu Dhabi for the period from 2008 to 2013 which had 5,973 numbers of accidents recorded. Each accident event had 48 different factors which were recorded at the time of accidents. These factors were reduced to 16 after pre-processing of the data. In this analysis authors used WEKA (Waikato Environment for Knowledge Analysis) which is the tool for data-mining and to build the ANN classifiers. They built the two classifiers using two different ways. By dividing the whole dataset into 90% to train the second classifier whereas 10% to test it. After experimenting the ANN, it was observed that clustering significantly increases the accuracy if the model. ANN resulted accuracy of about 81.6% for train dataset and 74.6 % for test dataset. In their future work, they plan to implement this model as decision support tool for Abu Dhabi Emirate Traffic Office.

C. *Heart Disease Indicator*

In paper [11], researchers showed how the heart disease is a big concern which needs to be predict beforehand so that the respective medications and consultation might provide by the doctors. They created the prediction system which can predict that if the patient is likely to get diagnosed by heart disease or not based on their medical history and factors. Supervised machine learning method such as Naïve Bayes, Logistic regression, and KNN were performed to predict the same. Wherein Naïve Bayes could not prove to be the best model for the data but Logistic regression and KNN were able to predict the heart disease with good accuracy. Evaluating the model, it was observed that after the data cleaning KNN and logistic regression gave the accuracy of 87.5%.

In [12], researchers applied the supervised machine learning methods such as Naïve Bayes, decision tree, K-nearest neighbour, and random forest algorithm. To apply these model they used the datasets of heart disease patients provided by UCI repository from the Cleveland database. This dataset had around 300 records of the patients with having 76 attributes. Out of these 76 attributes only 14 number of attributes proved to be useful to predict the heart disease. During evaluation it was found that on training dataset KNN, Decision tree, and random forest had the accuracy of 78.9%, 73.6%, and 84.21% respectively whereas testing dataset had 90.7%, 80.2%, and 82.8% respectively. By summarising the accuracy percentages of all the models, researchers found that KNN shows the highest accuracy score with respective other models. They mentioned future work as to build more complex and combination of the model to achieve high accuracy for the prediction of the heart diseases.

III. METHODOLOGY

Data mining is the process of finding the relations between the variables and patterns present in the huge dataset to predict the outcomes. This is achieved by combining the statistical learning, machine learning and database system.

In this paper, Knowledge Discovery in Database (KDD) methodology was performed. KDD is the process of exploring the important and useful knowledge from the large dataset which can prove to be beneficial in the business and health sectors to predict some critical outcomes.

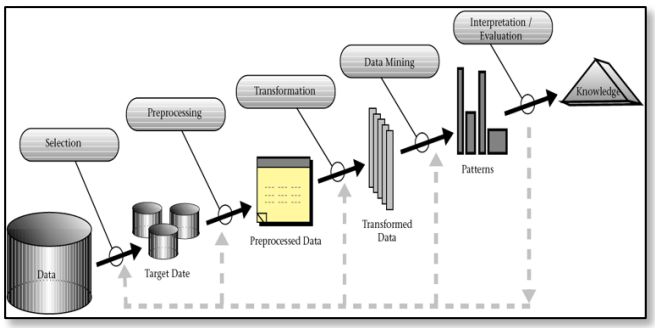

Fig. 1. Knowledge Discovery in Database (KDD) methodology [19]

Fig. 1. shows that KDD is a iterative process which involves progressive levels or a steps to discover knowledge from the data.

- Data selection is a process in which a subset or target dataset is selected from the large dataset to discover the knowledge.
- Data pre-processing is the process in which data is cleansed and pre-processed by eliminating missing values or by handling them by replacing into relevant data.
- Data transformation is the process in which insignificant or unwanted independent variables are removed from the dataset.
- Data mining in the context of machine learning is the process to build the models such as regression and classification and finding the patterns.
- Interpretation/Evaluation is the process in which evaluation is done based on cross validation methods, checking the accuracies of the model to conclude.

IV. Implementation, Results & Evaluation

A. *Germany Cars Dataset*

1) *Dataset Description:*

Germany car dataset contained total of 46,405 rows which are nothing but total number of cars listed for resale along with 10 columns.

2) *Model Selection:*

Multiple Regression:

As per the reference papers, there was multiple regression model were performed on the car price prediction [5] [6] which showed the best accuracy of 98% and 85% respectively, so out of curiosity decided to apply multiple regression model to check the accuracy of car re-sale price for the Germany's car dataset.

Random Forest (Regression):

In paper [3], researcher applied the Random Forest which could not perform well and generated accuracy of only 41.18%, so to analyse the same on the Germany's dataset chose to apply Random Forest Regression model.

3) *Methodology:*

As described, KDD methodology was followed throughout the model building process technique. Supervised machine learning model applied on this dataset were Multiple Linear Regression and Random Forest (Regression). Below is the step by step application of KDD methodology performed on this dataset.

i. Data Selection & Cleaning:

Germany car dataset was obtained from Kaggle web portal which is basically the web scraping of Autoscount24 online car market. Data was glanced and found some missing values which were removed from the dataset.

ii. Data Pre-processing & Transformation:

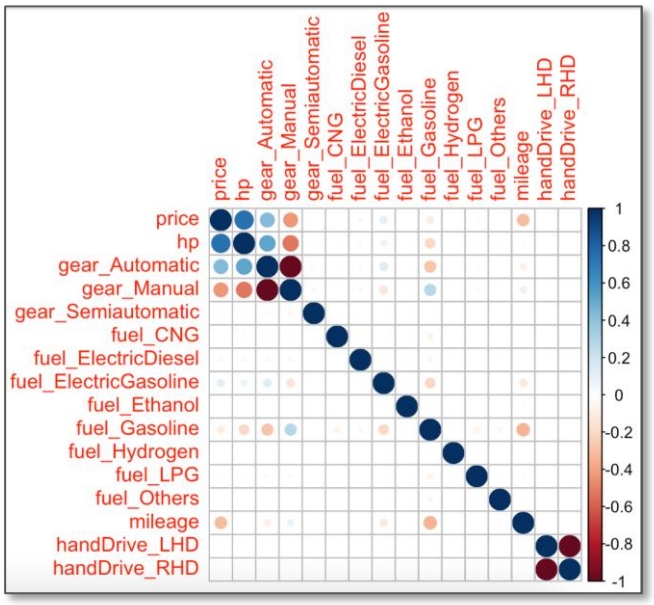

Fig. 2. Correlation matrix for variables of Germany Car Dataset

Based on the Fig. 2. Correlation map generated shows the correlation between all the variables in which the colour of the cube gradually getting bright represents the strong correlation between the variables. It helps to identify the independent variable(s) which is strongly correlated with the dependent variable. In the dataset "gear" variable had blank values which led to eliminate them from the dataset. Further "gear", "fuel", and "handdrive" being the categorical variable was converted into factors and then created dummy variables of the same.

Boxplot was plotted to identify if there were outliers present in the continuous variables "hp", "mileage", and "price". And as per Fig. 3. found that there were outliers present in these variables. All the outliers for three variables were removed by increasing the Interquartile range (IQR) and further excluding some more outliers which were not removed post increasing the IQR by excluding them from the dataset. Boxplot in the left shows when the outliers were present and the boxplot in the right shows the outliers after the IQR treatment.

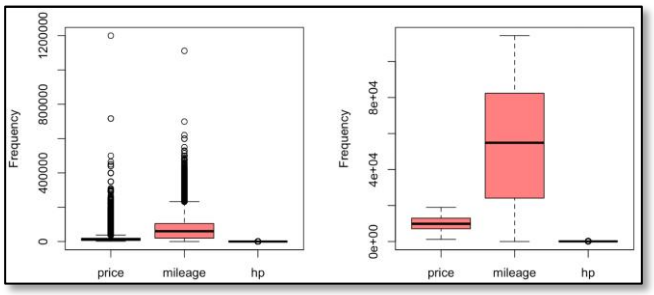

Fig. 3. Boxplot for price, mileage and hp

Upon plotting the histogram to check whether continuous variables "hp" and "price" were found that they were skewed so by applying log10 and sqrt transformations to them led to minimize the skewness in them.

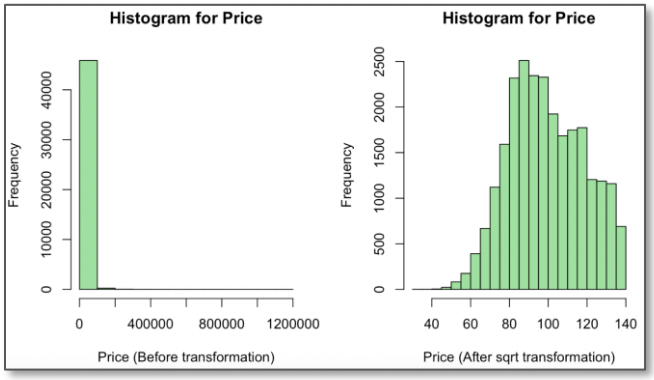

Fig. 4. Histogram for Price

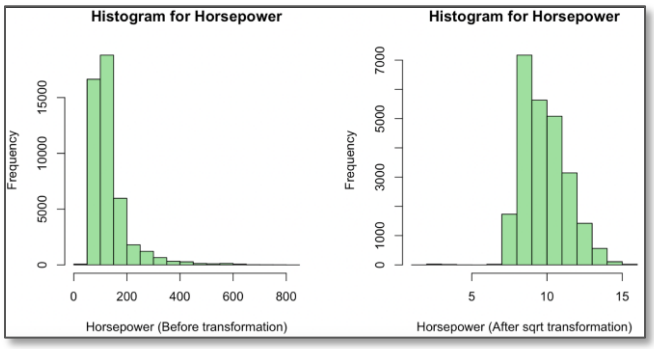

Fig. 5. Histogram for HP

iii. Data Mining:

For Regression model data was split into the ratio of 80% - 20% for training and testing data respectively with the seed (123). And as mentioned earlier Linear Multiple Regression and Random Forest Classification model was implemented to predict the car re-sale price.

4) Model Results & Evaluation:

Model result and evaluation being the last step of KDD will evaluate both the models based upon their outcomes.

Multiple Linear Regression:

Fig. 6. Represent the final linear regression plot which clearly explains how the model performed throughout. Normal Q-Q plot can be observed where most of the points are normally distributed. F - statistics and P -value proved to be statistically significant in the model. Residual vs Fitted plot are normally distributed but not bad. Model passed the VIF and Durbin Watson test. RMSE accounted for the test model is 11.43585.

After performing trial and error method, Fig. 6. Shows the final output of the multiple regression model, with the significant variables selected 'hp', 'fuel_CNG', 'fuel_Diesel', 'fuel_ElectricGasoline', and 'mileage' gave the accuracy of about 65.74% when trained on the training data whereas 65.22% accuracy on testing data.

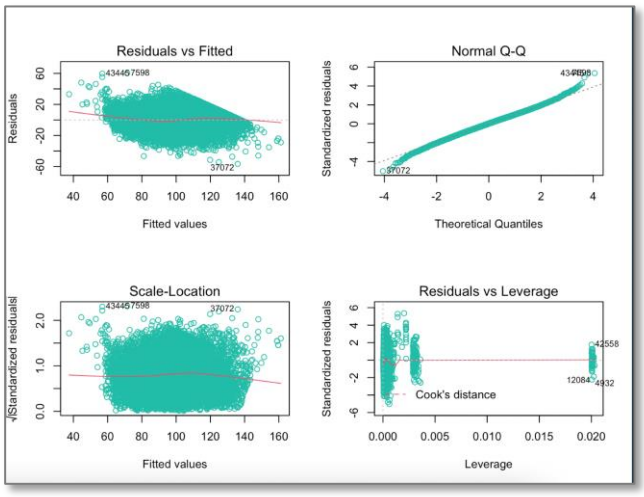

Fig. 6. Linear Regression Plot

Random Forest (Regression):

Random Forest was applied on the same dataset with significant variable selected. And cross-validation (10 fold) was After executing the model, it was observed that model gives best accuracy when mtry is 3. Running variable Importance for the random forest model showed that 'mileage' variables remains the best predictor for the model at 100% whereas 'fuel_CNG' is not at all significant variable for predicting 'price'. RMSE observed after testing on test data was 10.37668. Comparing both the model where RMSE for Multiple Linear Regression and Random Forest is 11.44 and 10.38 respectively which states that Random Forest fits best into this dataset to predict the car re-sale price.

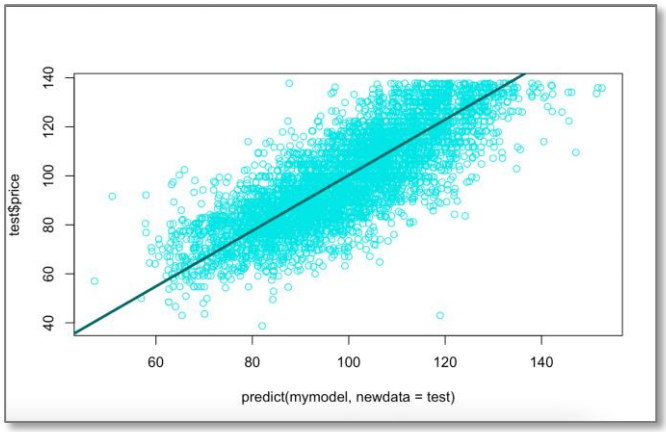

Fig. 7. Actual vs Prediction plot for Linear Regression Model

Fig. 7. Show that after plotting actual and predicted model for linear regression, most of the points are normally distributed to the regression line but some points are away from the

regression line which tells that there is a standard error of 11.44.

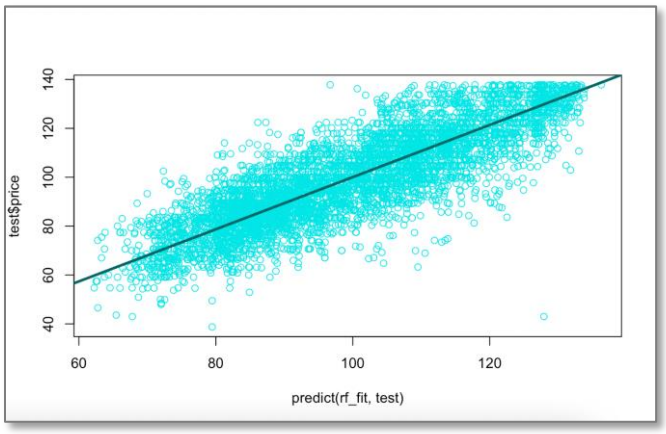

Fig. 8. Actual vs Prediction plot for Random Forest Model

Fig. 8. shows that after plotting actual and predicted model for Random Forest, most of the points are normally distributed to the regression line but some points are away from the regression line which tells that there is a standard error of 10.38.

B. UK Traffic Accidents Dataset

1) Dataset Description:

UK traffic accident dataset contained total of 17,80,653 rows which are nothing but total number of accidents occurred between the period 2005 to 2015 along with 32 columns. Machine learning technique was applied on only the sample of the period Nov 2014 to Dec 2014.

2) Model Selection:

Glancing through the previous published papers, in [7], [8], [9], and [10] it is observed that researchers received great accuracy when KNN and Random Forest Classifier was applied on the accident dataset. Which states that these two models performs best on the accidents dataset. So being the curiosity to perform the same models on the UK accidents data.

3) Methodology:

As described, KDD methodology was followed throughout the model building process technique. Supervised machine learning model applied on this dataset were K- Nearest Neighbour and Random Forest Classifier. Below is the step by step application of KDD methodology performed on this UK accidents dataset.

i. Data Selection & Cleaning

UK car accidents data was obtained from Kaggle web portal which was collected from the UK's Government Road and Safety website. Data was well maintained and there were many less number of missing values as well as noise. As mentioned earlier number of accidents records being around 17,80,653 between 2005 and 2015, took the sample for the period Nov 2014 to Dec 2014. Removed the irrelevant variables from the dataset. There were missing values in 'Urban_or_Rural_Area' variable which were removed from the dataset. After cleaning the data number of records came down to 25,282, then the machine learning model was performed on the same dataset.

ii. Data Pre-processing & Transformation

Dataset had a categorical variables which were converted to factors which needs to be done for applying classification models. Most of the variables were numeric in nature which had to convert into factors. Transformation was applied to 'Accident_Severity' variable which had 3 levels of categorical variable (Slight , Serious and Fatal) where in Serious and Fatal were combined to one to convert 'Accident_Severity' to binary variable. Likewise same transformation was applied to 'Urban_or_Rural_Area' variable where it being the 2 level categorical variable with factor value as 1 and 2 it was converted to binary value (0's and 1's).

iii. Data Mining

For Classification model data was split into the ratio of 70% - 30% for training and testing data respectively with the seed (123). And as mentioned earlier K – Nearest Neighbour and Random Forest model was implemented to predict the severity of the accident.

4) Model Results & Evaluation:

Model result and evaluation being the last step of KDD will evaluate both the models based upon their outcomes.

K- Nearest Neighbour (KNN) Classification:
KNN model was implemented on the dataset where to select significant value of k. where k = 9 was tested for the best k value to be fitted in to the model. Fig. 9. shows the confusion matrix for k = 9 tested on test data.

```
> confusionMatrix(table(knn.9,test.acc_labels))
Confusion Matrix and Statistics

     test.acc_labels
knn.9    0    1
    0 6463  735
    1    3  384
```

Fig. 9. Confusion Matrix for KNN on test data

Accuracy for KNN classifier was 90% with Sensitivity of 99% and specificity was recorded as 34.32 % and Kappa statistics was 0.4690. By observing these parameter it states that k = 9 is the best k value for this dataset.

Random Forest Classifier:

Random forest classifier was implemented on the same training data which showed the accuracy of 86% where Sensitivity, Specificity, and Kappa statistics were noticed to be 99%, 9 %, and 0.15 respectively which shows that Random forest is not the best model for this dataset.

```
> rf$confusion
      0  1 class.error
0 15069 41 0.002713435
1  2547 40 0.984538075
```

Fig. 10. Confusion Matrix for Random Forest on test data

Fig. 10. Explains the confusion matrix between predicted and actual data. Where is shows that 15069 times random forest predict that its severity of the accident would be not serious and 40 times it predicted that it would be severe accident. Which states that random forest doesn't fit into this model.

	K - Nearest Neighbour (k=9)	Random Forest Classifier
Accuracy	0.9027	0.8676
Sensitivity	0.9995	0.9994
Specificity	0.3432	0.0978
Kappa	0.4698	0.1549

Fig. 11. Performance comparison of KNN and Random Forest Classifier

Fig. 11. Shows the performance of both KNN and Random Forest Classifier for the accidents data where it concludes that KNN neighbour classification at k =9 fits best into the UK accidents dataset.

C. *Heart Disease Indicator*

1) *Dataset Description:*

Heart Disease dataset contained total of 23,893 rows which are nothing but total number of patients along with 22 columns or variables which contains medical history factors data.

2) *Model Selection:*

In [11], [12] researchers have implemented KNN, Logistic Regression, Decision Tree, and Random Forest. And looking at their model selection it shows that Logistic Regression and Decision Tree performs best showing best accuracy on the heart disease prediction.

3) *Methodology:*

As described, KDD methodology was followed throughout the model building process technique. Supervised machine learning model applied on this dataset were Logistic Regression and Decision Tree. Below is the step by step application of KDD methodology performed on this Heart Disease Indicator dataset.

i. Data Selection & Cleaning

Heart Disease Indicator dataset was obtained from Kaggle web portal. Data was glanced and found some missing values which were removed from the dataset.

ii. Data Pre-processing & Transformation

Dataset had 20 variables which needed to convert from numeric to factors. Also there were some special characters in the variable name which had to be handle by replacing them by general format.

iii. Data Mining

For implementing the classification model, data was split into the ratio of 80% - 20% for training and testing data respectively with the seed (123). And as mentioned earlier Logistic Regression and Decision tree classification model was implemented to predict the heart disease for a patient with knowing their medical history.

4) *Model Results & Evaluation:*

Model result and evaluation being the last step of KDD will evaluate both the models based upon their outcomes.

Logistic Regression:

Logistic Regression model implemented in the heart disease indicator dataset independent variables selected were 'HighBP', 'HighChol', 'CholCheck', 'BMI', 'Smoker', 'Stroke', 'Diabetes', 'PhysActivity', 'Fruits', 'Veggies', 'HvyAlcoholConsump', 'AnyHealthcare', and 'NoDocbcCost'.

```
> table(ActualValue=testing$HeartDiseaseorAttack, PredictedValue=res>0.5)
           PredictedValue
ActualValue FALSE  TRUE
          0 52069   300
          1  4949   337
```

Fig. 12. Confusion Matrix for Logistic Regression (res > 0.5)

Fig. 12. shows the logistic regression model predicts 52069 times correct prediction of the patient not diagnosed with heart disease whereas it also predicts pretty good when the patient is diagnosed with heart disease with 4949 times of the total patients. To increase the performance and find correct response value plotted the ROC curve.

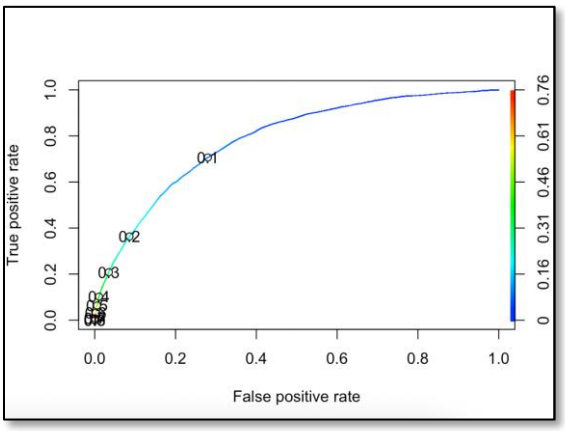

Fig. 13. ROC plotted for Actual vs Predicted Value

Referring Fig. 13, ROC curves show the True positive vs False positive rate of the Actual and predicted values of the train and testing dataset. Based on the figure tried selecting another response value to check if the prediction accuracy increases or not. But found that it degrades the performance of the model.

```
> table(ActualValue=testing$HeartDiseaseorAttack, PredictedValue=res>0.4)
           PredictedValue
ActualValue FALSE  TRUE
          0 51811   558
          1  4746   540
```

Fig. 14. Confusion Matrix for Predicted value > 0.4

Decision Tree:

Decision Tree was implemented on this dataset to predict if patient is diagnosed with the heart disease or not based on medical history and their routine.

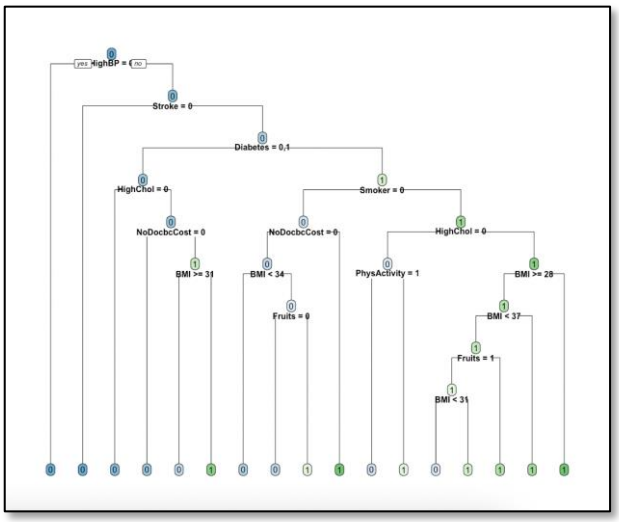

Fig. 15. Decision Tree of Heart Disease Indicator

```
> confusionMatrix(predict_pd,testing$HeartDiseaseorAttack)
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 52174  5049
         1   195   237

               Accuracy : 0.909
                 95% CI : (0.9067, 0.9114)
    No Information Rate : 0.9083
    P-Value [Acc > NIR] : 0.275

                  Kappa : 0.07

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.99628
            Specificity : 0.04484
         Pos Pred Value : 0.91177
         Neg Pred Value : 0.54861
             Prevalence : 0.90832
         Detection Rate : 0.90493
   Detection Prevalence : 0.99251
      Balanced Accuracy : 0.52056

       'Positive' Class : 0
```

Fig. 16. Decision Tree of Heart Disease Indicator

As we can see the Fig. 15 shows the decision tree formation of the model which looks quite good with having significant prediction. But in Fig. 16. Showing the confusion matrix, it shows that it has accuracy of about 90%, Sensitivity of about 99% but the specificity is quite low at 0.044 which needs to prune to optimize the decision tree outcome. So, applying the pruning.

After the cross-validation and pruning was applied to the model, Fig. 17. Shows the output of the same. Further choosing the Complexity Parameter value as 0.00053743 and pruned again the model.

```
          CP nsplit rel error  xerror      xstd
1 0.00155855      0   1.00000 1.00000 0.0069744
2 0.00053743      5   0.99092 0.99113 0.0069466
3 0.00042995      7   0.98984 0.99194 0.0069491
4 0.00020000      8   0.98941 0.99317 0.0069530
```

Fig. 17. Complexity parameter (CP) table of Decision tree

Complexity parameter plays an important role in optimizing the decision tree model. But unfortunately, it could not optimize the model at greater extent.

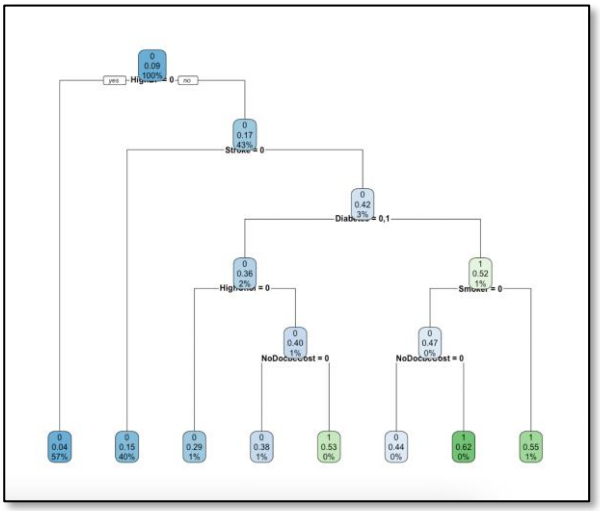

Fig. 18. Final Decision Tree of Heart Disease Indicator

V. Conclusions & Future Work

This study mainly focuses on how the same analytical machine learning models namely multiple linear regression, random forest (regression), K-nearest neighbour (KNN), logistic regression, decision tree and random forest (classification) can prove to be beneficial for the three un-related dataset of Germany cars dataset [1], UK traffic accidents dataset [2] and Heart disease indicator dataset [2] by predicting car prices based on the various independent factors, severity of the traffic accidents and identify whether or not the individual is having the heart disease. This prediction and analysis are done using some or many independent factors which are directly or indirectly associated with the predictor.

Prediction of car's re-sale price was implemented using the supervised machine learning technique such as Multiple Linear Regression and Random Forest Regression. In future, would like to apply more model machine learning algorithms.

Prediction of severity of accidents was performed using machine learning techniques such as K – Nearest Neighbour (KNN) and Random Forest Classifier. In which KNN outperformed the Random Forest Classifier. In future, would like to apply hybrid models of two or more models to find the highest accuracy of the model.

Heart Disease was implemented using machine learning modelling technique such as logistic regression and decision tree. Where both the models did not perform great so, in the future would like to apply Artificial Neural Network on the same dataset.

REFERENCES

[1] Kaggle.com. 2021. *Germany Cars Dataset*. [online] Available at: <https://www.kaggle.com/ander289386/cars-germany> [Accessed 26 December 2021].

[2] Kaggle.com. 2021. *UK Car Accidents 2005-2015*. [online] Available at: <https://www.kaggle.com/silicon99/dft-accident-data> [Accessed 26 December 2021].

[3] Kaggle.com. 2021. *Heart Disease Health Indicators Dataset*. [online] Available at: <https://www.kaggle.com/alexteboul/heart-disease-health-indicators-dataset> [Accessed 26 December 2021].

[4] Ripublication.com. 2021. [online] Available at: <http://ripublication.com/irph/ijict_spl/ijictv4n7spl_17.pdf> [Accessed 26 December 2021].

[5] Ijcaonline.org. 2021. [online] Available at: <https://www.ijcaonline.org/archives/volume167/number9/noor-2017-ijca-914373.pdf> [Accessed 26 December 2021].

[6] Ijirase.com. 2021. [online] Available at: <https://www.ijirase.com/assets/paper/issue_1/volume_4/V4-Issue-3-686-689.pdf> [Accessed 26 December 2021].

[7] Mdpi.com. 2021. [online] Available at: <https://www.mdpi.com/2073-431X/10/12/157/pdf> [Accessed 26 December 2021].

[8] Ijisae.org. 2021. View of Predicting the Severity of Motor Vehicle Accident Injuries in Adana-Turkey Using Machine Learning Methods and Detailed Meteorological Data. [online] Available at: <https://www.ijisae.org/IJISAE/article/view/709/pdf> [Accessed 26 December 2021].

[9] Ajith.softcomputing.net. 2021. [online] Available at: <http://ajith.softcomputing.net/isda-mam.pdf> [Accessed 26 December 2021].

[10] 2021. [online] Available at: <https://www.researchgate.net/publication/301708789_Severity_Prediction_of_Traffic_Accident_Using_an_Artificial_Neural_Network_Traffic_Accident_Severity_Prediction_Using_Artificial_Neural_Network> [Accessed 26 December 2021].

[11] Iopscience.iop.org. 2021. [online] Available at: <https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012072/pdf> [Accessed 26 December 2021].

[12] Link.springer.com. 2021. [online] Available at: <https://link.springer.com/content/pdf/10.1007/s42979-020-00365-y.pdf> [Accessed 26 December 2021].

[13] Mdpi.com. 2021. [online] Available at: <https://www.mdpi.com/2076-3417/11/18/8352/pdf> [Accessed 26 December 2021].

[14] Autoscout24.com. 2021. *AutoScout24 Europe's car market for new and used cars*. [online] Available at: <https://www.autoscout24.com/> [Accessed 26 December 2021].

[15] Who.int. 2021. *Cardiovascular diseases (CVDs)*. [online] Available at: <https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)#:~:text=Cardiovascular%20diseases%20(CVDs)%20are%20the,%2D%20and%20middle%2Dincome%20countries> [Accessed 26 December 2021].

[16] Olx.ba. 2021. *OLX.ba - Svijet Kupoprodaje*. [online] Available at: <https://www.olx.ba/> [Accessed 26 December 2021].

[17] eBay. 2021. *Cars for sale | eBay*. [online] Available at: <https://www.ebay.ie/b/Cars/9801/bn_1839037> [Accessed 26 December 2021].

[18] eVentures, P., 2021. *Used Cars - Buy Used Cars in Pakistan | PakWheels*. [online] Pakwheels. Available at: <https://www.pakwheels.com/used-cars/> [Accessed 26 December 2021].

[19] 2021. [online] Available at: <https://www.researchgate.net/publication/333622164_Integrated_Data_Mining_and_Knowledge_Based_System_to_Predict_and_Advice_of_Diabetes/figures?lo=1> [Accessed 26 December 2021].