Machine Learning Algorithms used for prediction of Cardiovascular Disease, Road Accidents and Subscription of a term deposit

Terrance Thomas
School of Computing
National College of Ireland
Dublin, Irleand
lix18184928@student.ncirl.ie

***Abstract* — With the increase of deaths from Cardiovascular disease (CVD) increasing. It has become very important to predict if there is a chance that a person might have a CVD. This study focusses on having various inputs from a person about his/her habits and features (for e.g. Height, Weight, Smoking, Alcohol intake, Cholesterol etc.) to predict if the person might have a CVD. I would be applying Support Vector Machine (SVM) and Random Forest (RF) to predict if a person would have CVD.**

Every year millions of people die due to car accidents. Road accidents result up to 2.2% of global death (9th factor for death cause) [8]. Road accidents also results in fatal injuries. I will be using different Machine Learning algorithms to predict nature of the accident (i.e. Slight or Fatal). Decision Tree (DT) and K Nearest Neighbor (kNN) will be applied to find the best method.

Telemarketing is very important in the modern world as they can help the customers with a good product or service. It is the most efficient way of creating business contact as it allows you to reach your target audience quickly. This paper focuses on if a customer will subscribe to the marketing scheme by the bank by using Logistic Regression (LR).

Keywords—CVD, Road accidents, Telemarketing, prediction, DT, kNN, LR, SVM, RF, Data Mining, Machine Learning

I. INTRODUCTION

CVD is a general term used for a heart or a blood vessel disease. CVD's are divided into 3 types viz Coronary Heart Disease (CHD), Stroke and Peripheral Heart Disease (PHD). Modernization has resulted into increase of CVD. CVD are number 1 cause of death globally. About 17.9 million people died due to CVD in 2016 as per WHO. CHD occurs when heart's blood supply is blocked. Stroke occurs when the blood supply to the brain is stopped. PHD occurs when there is a blockage to the limbs.

Road traffic accidents are a huge threat that continues to cause casualties, injuries and fatalities on roadways daily. This results in huge losses both at the economic and social levels. According to WHO in 2013, 1.25 million deaths occurred due to road accidents.

Marketing campaigns are a technique of outsourcing by organizations with the goal of improving the financial posture of their businesses. Direct marketing targets fragments of clients by reaching them to meet a objective [1].

A. Motivation

CVD is treated by cardiologists, thoracic surgeons, vascular surgeons, neurologists, and interventional radiologists, depending on the organ system that is being treated. The heart is the organ that pumps blood to all tissues of the body. If the pumping action of the heart becomes inefficient, vital organs like the brain and kidney suffer and if the heart stops working, death occurs within minutes. The World Health Organization (WHO) has estimated that 12 million deaths occur worldwide, every year due to heart disease [2]. Due to the etiology and long duration, CVD needs lifetime treatment.

Medical diagnosis is an important yet complicated task that needs to be done accurately and efficiently. The automation of this system is very much needed to help the physicians to do better diagnosis and treatment. The representation of medical knowledge, decision making, choice and adaptation of a suitable model are some issues that a medical system should take into consideration. Medical progress is always supported by data analysis which improves the skill of medical experts and establishes the treatment technique for diseases. The purpose of medical diagnosis system is to assist physicians in defining the risk level of an individual patient [2]. At present, the classification algorithms with higher accuracy mostly depend on the complex model structure. In medical treatment, patients with different age and gender show great differences, so it is difficult to analyse with the same algorithm [3]. Machine Learning can be used to predict CVD. Prediction of CVD is regarded as one of the most important subjects of clinical data.

Road crash prediction models are very useful for highway safety. Crash frequency refers to the prediction of the number of crashes that would occur on a specific road segment or intersection in a time period, while crash severity models generally explore the relationship between crash severity injury and the contributing factors such as driver behaviour, vehicle characteristics, roadway geometry, and road-environment conditions. Road Accidents/Car accidents too result in many deaths very year. Even if a person survives an accident but the chances of him/her being severely injured is quite high.

Marketing campaigns are widely applied in all business. It includes promoting an object or a service via mass communication. For instance, radio, television, telephone, print, social media and other ways to contact with the targeted audience. When organization contacts with customers through these communication methods, it is called direct marketing [4]. On the other hand, when the target audience is very wide, calling all clients takes a huge amount of time. The possible solution, which is described in this work, is to construct classification models, compare them between each other for identifying the most accurate one. The classification models aim to recognize a person with an increased probability to make a purchase. It allows to focus on potential buyers and not to spend time on inactive population [4].

B. *Project Objectives*

The importance of treating CVD on an early stage and considering if a person gets CVD it has to be treated lifetime; For dataset 1; the primary objective is

- To predict if the user has CVD

How weather and negligence can lead to death or fatal injuries and how using the data the response team can aid the required person has led to the primary objective of dataset 2:

- To predict the severity of an accident

Since a lot of resource (time and money) is invested in a marketing campaign and every organization wants the ROI for the investment has led to the primary objective of dataset 3:

- To predict if the customer subscribes for the marketing scheme

II. Related Work

The paper [5] focuses on assessing the arterial stiffness which is directly related to CVD. The research takes into consideration the digital volume pulse. SVM is used which gives an accuracy of more than 85%. The Pulse Wave Velocity (PWV) is classified as low and high. High PWV means high chances of CVD [5]. The short coming of the research conducted is that there were only 134 test subjects. The subject size is comparatively very small considering that the research paper wants to assess CVD [5]. The study [6] suggests a risk assessment model than consist of risk prediction, clustering and regression analysis. The model gave an accuracy of 90.62%. the paper analyses the relationship between patient's risk level and dangerous factors. Artificial Neural Network (ANN) was used in the Risk prediction [6]. This research [2]suggests a system that is based on Principal Component Analysis (PCA) and Adaptive Neuro Fuzzy Inference System (ANFIS). PCA reduces the attributes to 7 from 13 and then the ANFIS conducts the diagnosis. Accuracy obtained by this method is 93.2%. CVD was classified into 5 classes as the final output ranging from absent (0) to Serious (4). Here 303 instances of data were used of which only 200 instances were used for training and the remaining 103 was used for test [2]. The number of instances is low. The research should be done with a higher number of data and then use the model to predict the accuracy. The research [7] shows classification of decision trees was used. The predictor variable is Low Density Lipoprotien (LDL) while 27 variables (age, cholesterol, smoking etc.) were used as input. The model was applied on 1800 people of different cities. Time related variables and clustering and other data mining methods could be used for prediction and to compare with the current method [7]. The paper [3] shows application of a new 2-layer neural network method called Softmax Regression model is used. The paper compares results with kNN and Back Propagation Neural Network (BPNN). Softmax regression gives an accuracy of 94.44% which outperforms kNN (77.78%) and BPNN (72.27%) [3]. The paper [8] tries to predict CVD using ANN. 1500 data were used for the analysis. 4 variables were considered for the prediction Blood Pressure, Fasting Blood Sugar, Thalach and Cholesterol. 70% of data was used for training and 15% was used for test validation and remaining data was used for ANN's performance. Post validation the accuracy was 66.7%. The accuracy of the model is less than other existing methods [8]. The research paper [9] uses pattern recognition and data mining for risk prediction in a CVD domain. Naive Bayes (NB), DT, kNN and Neural Network (NN) have been applied to the data and NB has the best performance of all the other methods [9].

On average about 3287 people die by Road accidents. The paper [10] suggests using Fuzzy Logic (FL) model for the dataset. FL uses Fuzzy Tree (FT) to divide data into Total and Fatal accidents. The paper shows which area of Greece is more prone to accidents [10]. The paper [11]divides the dataset into 2 parts State Highway (SH) and Ordinary District Roads (OHD) and estimates the severity of the accident. The paper shows which areas are prone to accidents (Cross-intersection, R-intersection etc.) and determines that Cross intersection is where most accidents occur. However, the paper does not consider other parameters for the accident [11]. The study [12] uses Seattle data to analyse the accidents. Factors such as weather and road conditions are considered. NB, RF, Multilayer perceptron (MLP) and AdaBoost was used. NB had the worst accuracy while the other models predicted almost the same. But this study did not consider human behaviour for analysis [12]. In order to predict the fatal and non-fatal injuries in an accident, the paper [13] uses WEKA (Waikato Environment for Knowledge Analysis) data mining decision tree (J48, ID3 and CART) and NB classifiers are built to model the severity of injury. J48 had the highest accuracy but NB has better AUC and ROC results. The study revealed that fatal events occurred during raining condition and midnight also, Serious accidents occurred in Foggy weather [13]. This study [14] considers various dataset while predicting the accidents. NB, RF, Instance-Bases learning with parameter k (IBK) uses kNN and ZeroR Classifier were used. 2549 instances were used. IBK performed the best [14]. Traffic accidents were predicted using a Decision-making system. PCA was used to remove inconsistent data. Big Data technology was used for data integration and storage in a data warehouse. The model [15] suggests that less severity accidents occur in urban areas. Risk of accident increase as per 2nd and 3rd quarter of the year. Also, most accidents tend to occur between noon and 5 pm [15]. In this research [16], Self Organization Map (SOM) is used. SOM is based on NN and is an unsupervised learning method. K means is used to compare with SOM. The paper concludes stating SOM has better results than K means [16].

Bank Direct Marketing offers customers with various products and service. It is important to predict what are the number of people that will the product or service. The paper [17] suggests a method to analyse asymmetric information using SMOTE algorithm and Rotation Forest (PCA)-J48. The performance of the proposed system is evaluated and compared with DT, RF, Rotation Forest, NB etc. The proposed method solves the data imbalance and inequality [17]. The research [18] shows how machine learning can be applied to Digital marketing. Digital Marketing plays a vital role in 21st century for higher conversion rates. The study [19] tries to predict if Deep Neural Network if a given customer is proper for bank telemarketing or not. The model uses 45211 phone calls over a 30-day period as dataset. The model achieves a 76.70% accuracy. Of the other models which were used for comparison only LR was the closest with 75.19% accuracy [19]. The paper [1] is the most relevant for the objective as the main objective directly aligns with the objective of my project. The paper tries to Predict customer response to bank direct telemarketing campaign. Different Machine Learning techniques was

applied to the model viz Multilayer Perceptron Neural Network (MLPNN), DT, LR and RF. RF gave the highest accuracy [1]. The paper [4] shows the as evaluated various classification models for prediction of bank telemarketing campaign results about the probability of the subscription of the customer to the deposit. RF gave the best accuracy followed by ANN. Boosting algorithms can be used to increase the accuracy of the models applied [4]. The research [20] focuses on optimizing the predicting done on a telemarketing by classification. It introduces to a new technique that implicitly fosters most significant features and predicts the classes of clients according to the types of these features. It compares with different optimizing. NB, SVM, DT and ANN were applied on the dataset [20].

III. METHODOLOGY

This study follows the Knowledge Discovery in Databases (KDD) process to achieve its results. KDD follows a non-trivial extraction of useful information from the Database. The KDD process is shown in Figure 1. KDD starts with selection of data followed by Pre-processing and cleaning of data. Now this data is transformed as per requirement. Now on this data, Data mining techniques/algorithms are applied for visualization or pattern finding. From the visualizations and patterns the final knowledge is obtained .

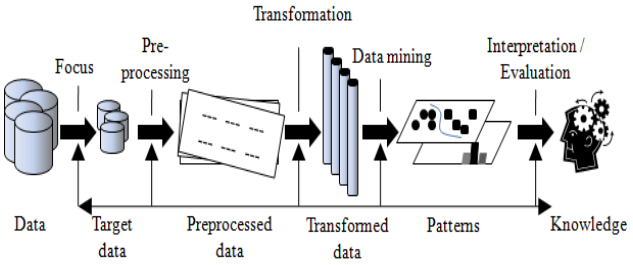

Figure 1: KDD flow

A. *Data Collection*

All the dataset used in the project are from Kaggle [21].

Dataset 1: This dataset is about CVD. It has attributes about the patient like age, height, gender etc. and patient habits like smoking, alcohol etc. which are used for testing if there is a chance that the user has CVD. The dataset has 13 columns and 70,000 rows [22].

Dataset 2: This dataset is about road accidents in UK. The dataset has 19 columns and 251382 rows. The dataset consists of information vehicle details such as make, engine etc., Accident area details and weather details [23].

Dataset 3: This dataset is about Bank Marketing. The dataset has 21 columns and 41188 rows. The dataset has information about different attributes of the customer [24].

B. *Data Prepocessing*

1) Missing, and Null Values Handling: The code ".na" was used to find na values in all the datasets. The same function is used to replace the missing values by mead or median.

2) Data Exploration and Transformation: Majority of the time taken is in this stage. This stage is very crucial as I had to sekect te features and transform the data in a way to achieve desirable results.

- Feature Selection: In this process, we select the columns which are required for the model. This process can be done manually as well as using libraries like Baruto.

a) Dataset 1: All the columns and rows were used for the model.

b) Dataset 2: All the columns except 'Accident_Index', 'Latitude','Longitude', 'Datetime' were used. Since the row count was more than 2,00,000; I used "sample_n" function to select 20,000 rows randomly and used this data for the processing.

c) Dataset 3: The columns default and 'contact', 'month', 'day_of_week', 'duration', 'campaign', 'pdays', 'poutcome', 'emp.var.rate', 'cons.price.idx', 'euribor3' were used.

- Oversampling and Under sampling: Oversampling and Under sampling is used to adjust the class distribution of a data set. It is used to overcome imbalance data.

- Feature Selection: Columns data have been converted into factors to suit the requirement of the model.

- Holdout Sample: All the dataset has been split into test and train. The data have been either split into 75:25 or 70:30 manner

C. *Data Mining Algorithms:* The Machine Learning Classificatons and regressions used for the dataset are as follows:

1) Dataset 1:

a) Random Forest Classification: It is a superivsed learning method. It consist of multiple DT. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction is better than any individual tree

b) Support Vector Machines: It is a supervised learning method. In this algorithm, we plot each data as a point in n-dimensional space (n = number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes. Support Vectors are simply the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes (hyper-plane/ line).

2) Dataset 2:

a) Decision Tree: It is a superivsed learning method. A DT is a decision support tool that uses a tree like model of decisions and their possible consequences. A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node.

b) K Nearest Neighbours: It is a supervised learning method. It classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If , K = 1, then the case is simply assigned to the class of its nearest neighbor.

3) Dataset 3:

a) Logistic Regression: It is a superivsed learning method. It is used for analysing a dataset in which there are one or more independent. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables.

IV. Model Evaluation

1) For Classification:

a) Sensitivity: It is propotion of positive examples correctly classified. It is the ratio of True Postive to the sum of True Positive and False Negative. In other words, it is the ratio of True Positive to Positive

b) Specificity: It is propotion of negative examples correctly classified. It is the ratio of True Negative to the sum of True Negative and False Postive. In other words, it is the ratio of True Negative to Negative

c) Accuracy: It is propotion of correct predictions to total number of predictions

d) Kappa: It is used to control only those instances that have correctly classified by chance. It is calculated by difference between Total Accuracy and Random Accuracy to 1 – Random Accuracy

e) ROC: ROC is Receiver Operating Characteristic curve. ROC plots the true positive rate against false positive rate

f) AUC: AUC stands for Area Under the Curve. It is a measure of seprability. It tells how much model is capable of distinguishing between classes.

2) For Regression:

a) R Squared: It is a ration between how good the model is and how good is the navive mean model. It is Scale free

*b) MSE:*MSE stands for Mean Square Error. It measures average square error of the predictions. Lower the score the better is the model

c) RSE: It stands for Residual Standard Error. It is average amount that the response will deviate from the true regression line

V. Results

1) Dataset1:

a) Random Forest Classification: RF classification was applied on dataset 1 and the result is shown in figure 2. The accuracy of the model is 94.37% while the kappa score is 88.74%

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 23933  2449
         1   279 21801

               Accuracy : 0.9437
                 95% CI : (0.9416, 0.9457)
    No Information Rate : 0.5004
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8874

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9885
            Specificity : 0.8990
         Pos Pred Value : 0.9072
         Neg Pred Value : 0.9874
             Prevalence : 0.4996
         Detection Rate : 0.4939
   Detection Prevalence : 0.5444
      Balanced Accuracy : 0.9437

       'Positive' Class : 0
```

Figure 2 RF Classification Output

*b) Support Vector Machines:*SVM classification was applied on dataset 1 and the output is shown in figure 3. The accuracy is 73.26% while the kappa value is 46.51%

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 18723  7387
         1  5572 16780

               Accuracy : 0.7326
                 95% CI : (0.7286, 0.7365)
    No Information Rate : 0.5013
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4651

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7707
            Specificity : 0.6943
         Pos Pred Value : 0.7171
         Neg Pred Value : 0.7507
             Prevalence : 0.5013
         Detection Rate : 0.3863
   Detection Prevalence : 0.5388
      Balanced Accuracy : 0.7325

       'Positive' Class : 0
```

Figure 3 SVM Output

2) Dataset 2:

a) Decision Tree: DT was appllied on dataset 2. The accuracy of DT is shown in figure 4 where the accuracy changes as per the number of neighbours. The ROC curve for the DT is shown in figure 5. The overall output of the model is shown in figure 6.

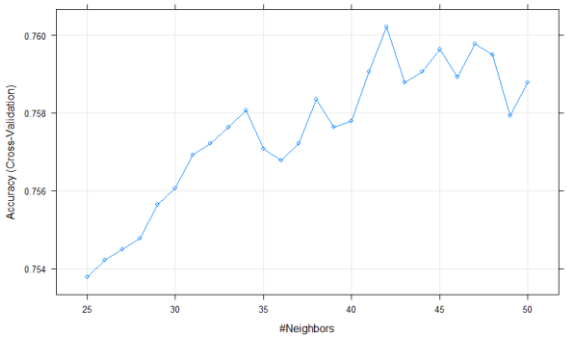

Figure 4 DT Accuracy

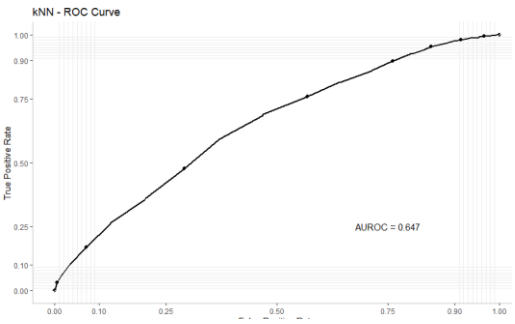

Figure 5 ROC for DT

```
> accuracy_baseline
[1] 0.753
> knn_accuracy <- round(knn_confusion_matrix$overall['Accuracy'], 3)
> knn_precision <- round(knn_confusion_matrix$byClass['Pos Pred Value'], 3)
> knn_recall <- round(knn_confusion_matrix$byClass['Sensitivity'], 3)
> knn_f1_score <- round(2*((knn_precision * knn_recall) / (knn_precision + knn_recall)), 3)
> knn_roc <- round(calc_auc(roc)$AUC, 3)
> test_df_normalized <- test_df_normalized[ , !(names(test_df_normalized) %in% c('pred', 'pred_prob', 'error'))]
> data.frame(accuracy_baseline, knn_accuracy, knn_precision, knn_recall, knn_f1_score, knn_roc)
          accuracy_baseline knn_accuracy knn_precision knn_recall knn_f1_score knn_roc
Accuracy             0.753        0.761         0.602      0.092         0.16   0.647
>
```

Figure 6 DT Output

b) kNN: kNN was applied on dataset 2. The DT as root, node and leaves is shown in figure 7. The ROC curve for DT is shown in figure 8. Finally, the DT output is shown in figure 9.

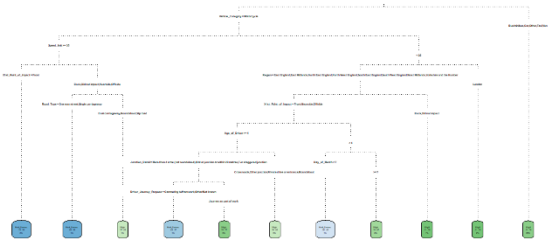

Figure 7 DT breakdown

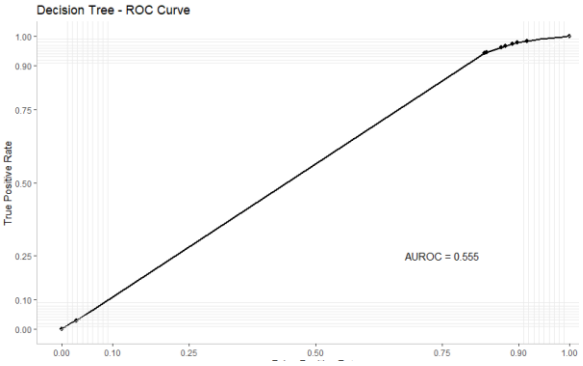

Figure 8 ROC curve for DT

```
> dt_f1_score <- round(2*((dt_precision * dt_recall) / (dt_precision + dt_recall)), 3)
> dt_roc <- round(calc_auc(roc)$AUC, 3)
> test_df_categorical <- test_df_categorical[ , !(names(test_df_categorical) %in% c('pred', 'pred_prob', 'err
> data.frame(accuracy_baseline, dt_accuracy, dt_precision, dt_recall, dt_f1_score, dt_roc)
          accuracy_baseline dt_accuracy dt_precision dt_recall dt_f1_score dt_roc
Accuracy             0.753       0.76        0.578     0.113       0.189  0.555
```

Figure 9 DT Output

3) Dataset 3:

a) Logistic Regression: LR was applied on dataset 3. The output of LR is shown in fugre 10. The accuracy of the mdoel is 91.09% while the kappa score is 46.6%.

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 11865   917
         1   306   641

               Accuracy : 0.9109
                 95% CI : (0.906, 0.9156)
    No Information Rate : 0.8865
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.466

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9749
            Specificity : 0.4114
         Pos Pred Value : 0.9283
         Neg Pred Value : 0.6769
             Prevalence : 0.8865
         Detection Rate : 0.8642
   Detection Prevalence : 0.9310
      Balanced Accuracy : 0.6931

       'Positive' Class : 0
```

Figure 10 LR Output

VI. Conclusion and Future Work

All the 3 datasets were cleaned, transformed and then various data mining techniques were applied like DT, RF, SVM, kNN and LR.

In Dataset 1 both RF and SVM were applied. RF has clearly better accuracy and kappa score than SVM. In Dataset 2 both DT and kNN were applied. Looking the ROC curve and output, it is visible that kNN has the better accuracy, precision and ROC. Finally, on Dataset 3 LR was applied which gave an accuracy of 91.09%.

Future work will include adding more regression and classification models and comparing them with the output of the already applied algorithms. Also, various method such as XGBoost and AdaBoost can be used to increase the accuracy of the model.

Based on the results, a model can be suggested to predict accidents, check if patient has a CVD or chances of a customer subscribing to a scheme. Additional parameters can also be considered to check if they do impact the model with a larger set of data.

An improved model on CVD prediction can help a lot of people and help to reduce the risk at an early stage. Predicting accidents can help the response team to judge better and help to take measure to avoid such scenarios. While predicting the outcome of a direct marketing would help to get the best ROI.

VII. References

[1] J. A.-F. a. M. Jayabalan, "Predicting customer response to bank direct telemarketing campaign," in *2017 International Conference on Engineering*

Technology and Technopreneurship (ICE2T), Kuala Lumpur, Malaysia, 2017.

[2] B. A. N. G, "An intelligent approach based on Principal Component Analysis and Adaptive Neuro Fuzzy Inference System for predicting the risk of cardiovascular diseases," in *2013 Fifth International Conference on Advanced Computing (ICoAC)*, Chennai, India, 2013.

[3] S. H. L. a. F. J. L. Hao, "Classification of Cardiovascular Disease via A New SoftMax Model," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Honolulu, HI, USA, 2018.

[4] K. B. a. A. Y. E. Zeinulla, "Comparative study of the classification models for prediction of bank telemarketing," in *2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT)*, Almaty, Kazakhstan, Kazakhstan, 2018.

[5] S. C. M. P. J. C. a. A. J. S. R. Alty, "Cardiovascular disease prediction using support vector machines," in *2003 46th Midwest Symposium on Circuits and Systems*, Cairo, Egypt, 2003.

[6] S. C. R. L. D. T. Y. T. a. J. L. C. Zhu, "Design and Development of a Readmission Risk Assessment System for Patients with Cardiovascular Disease," in *2016 8th International Conference on Information Technology in Medicine and Education (ITME)*, Fuzhou, China, 2016.

[7] M. H. S. a. S. A. A. Kajabadi, "Data mining cardiovascular risk factors," in *2009 International Conference on Application of Information and Communication Technologies*, Baku, Azerbaijan, 2009.

[8] J. T. a. B. K. Dewangan, "Analysis of cardiovascular diseases using artificial neural network," in *2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, Solan Himachal Pradesh, India, India, 2018.

[9] T. J. P. a. K. Somasundaram, "An empirical study on prediction of heart disease using classification data mining techniques," in *IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM -2012)*, Nagapattinam, Tamil Nadu, India, 2012.

[10] L. I. A. K. a. B. P. A. Katsoukis, "Classification Of Road Accidents Using Fuzzy Techniques," in *2018 Innovations in Intelligent Systems and Applications (INISTA)*, Thessaloniki, Greece, 2018.

[11] G. K. a. E. H. Kaur, "Prediction of the cause of accident and accident prone location on roads using data mining techniques," in *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Delhi, India, 2017.

[12] B. N. a. C. X. X. Xia, "Real-Time Traffic Accident Severity Prediction Using Data Mining Technologies," in *2017 International Conference on Network and Information Systems for Computers (ICNISC)*, Shanghai, China, China, 2017.

[13] D. K. S. a. E. A. T. T. K. Bahiru, "Comparative Study on Data Mining Classification Algorithms for Predicting Road Traffic Accident Severity," in *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, Coimbatore, India, 2018.

[14] L. K. E. P. a. P. K. S. C. Sugetha, "Performance Evaluation Of Classifiers For Analysis Of Road Accidents," in *2017 Ninth International Conference on Advanced Computing (ICoAC)*, Chennai, India, 2017.

[15] A. E. F. F. Z. E. a. M. S. H. El Alaoui El Abdallaoui, "Decision Support System for the Analysis of Traffic Accident Big Dat," in *2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, Las Palmas de Gran Canaria, Spain, Spain, 2018.

[16] A. V. S. a. P. S. Kasbe, "A review on road accident data analysis using data mining techniques," in *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, Coimbatore, India, 2017.

[17] P. R. a. S. Jaiyen, "Bank direct marketing analysis of asymmetric information based on machine learning," in *2015 12th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, Songkhla, Thailand, 2015.

[18] M. K. N. E. a. S. Z. A. Miklosik, "Towards the Adoption of Machine Learning-Based Analytical Tools in Digital Marketing," *IEEE Access* , vol. 7, pp. 85705 - 85718, 2019.

[19] C. L. S. J. a. S. C. K. Kim, "Predicting the success of bank telemarketing using deep convolutional neural network," in *2015 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, Fukuoka, Japan, 2015.

[20] W. C. a. S. H. C. S. T. Koumétio, "Optimizing the prediction of telemarketing target calls by a classification technique," in *2018 6th International Conference on Wireless Networks and Mobile Communications (WINCOM)*, Marrakesh, Morocco, Morocco, 2018.

[21] kaggle, "kaggle," kaggle, [Online]. Available: https://www.kaggle.com/. [Accessed 15 November 2019].

[22] kaggle, "cardiovascular disease dataset," [Online]. Available: https://www.kaggle.com/sulianova/cardiovascular-disease-dataset. [Accessed 20 November 2019].

[23] S. Leone, "adm project road accidents in uk," kaggle, [Online]. Available: https://www.kaggle.com/stefanoleone992/adm-project-road-accidents-in-uk. [Accessed 21 November 2019].

[24] H. Yamahata, "Bank Marketing," kaggle, [Online]. Available:

https://www.kaggle.com/henriqueyamahata/bank-marketing. [Accessed 22 November 2019].

[25] Kaggle, "Kaggle," Kaggle, [Online]. Available: https://www.kaggle.com/. [Accessed 25 November 2019].