Cardio-Vascular Disease prediction using Machine Learning Algorithms

Terrance Thomas
National College of Ireland
School of Computing
Dublin, Republic of Ireland
x18184928@student.ncirl.ie

***Abstract*— Cardiovascular diseases (CVD) are the number one cause of death globally: more people die each year from CVDs than from any other cause. An estimated 17.9 million people died of CVDs in 2016, accounting for 31 per cent of all deaths worldwide. 85 percent of these deaths are caused by a heart attack and stroke. The changing lifestyle, eating patterns and other unhealthy behaviors such as smoking, drinking has led people at a young age to a heart attack. ML algorithms can be used efficiently to forecast and make educated decisions using machine learning (ML), with the broad data available over the year. This prediction can be used to offer personalised treatments and care.**

Keywords—Cardiovascular disease, Prediction, Data mining, Heart Disease, Heart Attack, Machine Learning, Random Forest

I. RESEARCH AND INVESTIGATION INTO APPLICABLE TECHNIQUES

The goal of the study was to identify and analyse the factors of the human lifestyle in CVD prediction using data mining (DM) techniques. This data storage system has over thirty variables in the lifestyle, such as age, high salty diet, heartbeat rate, heart disease symptoms, and so on. This data storage system has over thirty variables in the lifestyle, such as age, high salty diet, heartbeat rate, heart disease symptoms, and so on. Considered input data is in a ratio of 75 percent for preparation and 25 percent for evaluating the early prediction method. The knowledge available in a repository is pre-processed and used in various categorization techniques such as the clustering of k-means and the algorithm of the decision tree (DT), which has the potential to provide further information relevant to the training and testing of the predicting system. Precision, recall, precision, and F-measures are calculated to illustrate the effects with the various root mean square error (RMSE), mean absolute error (MAE), and kappa statistic error measurements. K-means showed an accuracy of 94.56 percentage while DT had an accuracy of 93 percentage [1].

This research aims to use ML techniques to identify key features of CVD prediction. Through this work the causes that contribute to CVDs can be identified. Proposed excursion Boost the RF with an optimized linear model (RFRF-ILM) for predicting heart disease. An artificial neural network (ANN) for classifying cardiovascular disease, with the collection of apps and the methodology of studying backpropagation. The RFRF-ILM approach combines the linear model with the characteristics of a random forest (RF). RFRF-ILM achieves prediction of heart disease in high precision. The findings obtained by the suggested method of test case showed the best findings in contrast with the outcomes of other methods. A support vector system (SVM) is used to boost performance of the algorithms. The proposed algorithm reduces time and testing costs and increases the efficiency of treatment processes [2].

MIFH, a machine intelligence system for diagnosing heart disease, was created by the researchers. MIFH uses the Mixed Data Factor Analysis (MDFA) to extract and derive data set features and train predictive ML models. Results of the experiments show that MIFH performed well in terms of precision over many reference methods and comparable in terms of sensitivity and specificity MIFH returns the best possible solution of all input predictive models taking output parameters into consideration and increases device effectiveness, thereby allowing doctors and radiologists to better diagnose heart patients. The data set includes 76 features but was used for testing purposes with a limit of 14 features. Data imputation is performed with the new labels to fill in the missing values of the features to make the data set complete and appropriate for processing. Z-score normalization is performed for data standardization which exploits mean and standard deviation of the attributes to normalize the data. Extraction of the function (FE) is carried out using mixed data factor analysis (MDFA). The proposed method had an accuracy of 93.44 percent, the sensitivity of 89.28 percent and specificity of 96.96 percent [3].

The study proposes a method for determining rapidly the arterial stiffness of a patient and thus the risk of developing CVD without recourse to laborious blood tests. Simple measurement of the volume pulse of a patient measured at the fingertip ie digital volume pulse (DVP) using an infrared light absorption detector on the index finger is sufficient to predict the risk of their CVD. SVM approach yields a high degree of classification precision, with a substantially high percentage of true positive (i.e. sensitivity) achieved by 93 percent. Just 78 percent true negatives (i.e. the specificity) had a slightly lower score. The suggested approach is very useful for classifying patients into high pulse wave velocity (PWV) equivalent to high CVD risk and low PWV equivalent to low CVD risk due to the characteristics of their DVP waveform [4].

The work proposes a predictive model for predicting whether a person has heart disease or not. This is accomplished by comparing the accuracy of applying rules to the results of the individual SVM, gradient boosting (GB), RF, naive bayes (NB) classifier, and logistic regression (LR) on the dataset taken in a region to present a precise model of CVD. The documents are divided into 2 datasets: dataset preparation and validation of datasets. A total of 920 records is present along with 76 medical-related attributes. Only 14 attributes are used for the proposed system. Best accuracy was of LR which was 86.51% followed by 84.26 % of NB and 84.26% of GB. RF had an accuracy of 80.89 % while SVM had the lowest accuracy of 79.77%. Using better and better

data sets and improved analytical methods, these accuracies can also be increased [5].

The paper recommends a method based on a smart approach focused on the principal component analysis (PCA) and adaptive neuro-fuzzy inference Method (ANFIS). This system has two stages: Dimensions of the heart disease dataset with 13 attributes are reduced to 7 attributes using PCA at the first stage. During the second stage, heart disease diagnosis is performed using ANFIS. In ANFIS, neural network learning capabilities and Fuzzy logic reasoning capabilities are combined to offer a better prediction. Pre-processing data involves data cleaning, integration of data, data transformation and data reduction. Routines for data cleaning aim to fill in missing values, smooth out noise while finding outliers and correct data incoherence. The most probable value for filling the missing values is used in the data. Routines for data transformation turn the data into suitable types of mining. For classification purposes standardization is useful. The dimensions of a large dataset can be minimized by using PCA as one of the commonly used statistical methods. This method transforms the original data into new dimensions. The proposed method gave an accuracy of 93.2 percent and had a 90.91 percentage true positive value [6].

The authors review the findings of neural network (NN) studies that examine the key markers for heart disease electrocardiogram (ECG) diagnosis. NN is qualified to identify the ECG signals by different possible states. The heart rate variability (HRV) parameters are derived from ECG signals and used as functions for NN inputs. There was a lot of noise at the dataset. Cleaning of the data was done to remove them. Following data cleaning, extractions of the features were performed on the dataset. The NN Learning Vector Quantification (LVQ) is used for the identification and classification of ECG features. Vector Quantization Neural Learning Network (LVQNN) is a competitive network which employs supervised learning. The key feature is to run a NN from the input with an ECG signal or derived functions to produce an inference as the source of a cardiovascular disorder or a possible disease. A qualified NN can be used to identify ECG signals and thus help to diagnose the disease and to assess the right symptoms or kinds of disease [7].

The paper aims to boost cardiovascular risk prediction with data on ML and DNA methylation. The dataset was split into 20 subsets to perform a nest-cross validation process. Of these, the system was circularly trained in 18 subsets, and the remaining two are used as validation and test sets. Two models were applied to the dataset. First, bagging of LR with lasso penalty (BLR). RF classifier is second. In all tests, the RF outperformed the BLR. Area under the receiver operating characteristics (AUC) of RF was 90 percent [8].

The paper proposes a flexible tree boosting architecture, called XGBoost, resulting in substantial improvements in performance. The experimental testing of the proposed algorithm resulted in promising results compared to traditional shallow ML techniques, in terms of performance of various performance metrics including prediction accuracy and model building time. Boosting refers to a learning category of algorithms by adding many simpler models to suit the models. XGBoost uses the subtle art of penalizing the trees. Subsequently, different terminal nodes are allowed on the trees. Using XGBoost penalty shortens the vine. The downside is that not all the weights of the leaf are reduced by the same factor, whereas the estimated weights of the leaf that use less evidence in the data will be reduced more severely. The dataset contains data on activity and heart health gathered through a questionnaire on physical activity, cardiovascular fitness check and physical activity tracking. Using Grubb's method, outlier identification and exclusion are completed. This approach is useful because it considers outer identification and exclusion of full data set values. The modes and mean are replaced with missing data. Usage of mean and standard deviation to achieve data scaling. F-Score process of choosing the most appropriate and important functions from the dataset. F-Score is a simple method of classifying the two groups with the actual values. XGBoost results were compared with other ML algorithms. XGBoost took only 15 seconds to give the results however the accuracy was 89.9%. Other ML techniques like k-nearest neighbours (KNN), RF, SVM, NB, DT, LR and ensemble learning gave an accuracy of 90.1%, 90.1 percent, 90.2 percent, 89.9 percent, 90 percent, 90 percent and 89.3 percent respectively [9].

The research suggests a novel method aimed at identifying significant features through the application of machine learning techniques that will enhance the accuracy of cardiovascular disease prediction. The predictive model is applied with various combinations of functions and different known classification techniques. The suggested approach is a hybrid solution that uses the combination of the properties of the Linear Approach (LM) and RF to be called HRFLM. Data pre-processing phase followed by selection of DT entropy dependent features. The apps' range and simulation appear to be repeated for various combinations. The performance of each generated model is reported on the basis of 13 functions and the ML techniques used for each iteration and output. There are a total of 303 medical records in the dataset, where 6 records are with some missing values. Such 6 records were excluded from the list and the remaining 297 records of patients are used in pretreatment. Pre-processing data for 297 medical records revealed that 137 records showed a value of 1 for cardiac disease, while the remaining 160 showed a value of 0 for cardiac disease and the remaining 160 showed a value of 0 for cardiac disease absence. Many ML techniques, including DL, NB, RF, Linear Generalized Model (GLM), DTGB, LR, and SVM, are also used. The proposed HRFLM model reached an 88.7 per cent degree of accuracy. HRFLM has been very effective in predicting heart disease [10].

A comparison of ML models on the prediction of CVDs using cardiovascular risk factors of patients is the research. The list contains records of 70,000 patients. Total variables in the dataset are 12. 49% of data has CVDs while 51% doesn't have CVD. Data were cleaned by eliminating the value of diastolic blood pressure that is greater than systolic blood pressure. The value of height and weight and tends to be out of the range of height and weight importance of the grown person was removed. Lastly, a new body mass index (BMI) variable was developed from known height and weight. The models which were running are RF, NB, KNN, and LR. RF had the best accuracy of 73 percent, sensitivity of 65 percent and specificity of 80 percent. However, the model accuracy can be improved to get better results [11].

The work focuses on the study of CVDs using data mining methods for patients of all age groups. The pre-processed dataset consists of all the 16 attributes of 299 available patients. The dataset has features like age, gender, fast grume

sugar, dig, prop etc. The pre-processed data is divided into two groups (present, absent) using data mining approaches based on CVD's severity. 3 different models are applied on the dataset viz NB, random tree (RT) and reduced error pruning (REPTree). RT had the highest accuracy of 88.6% when 10 attributes were only considered [12].

The paper proposes an improved prediction of CVD quality using a better pre-processing method. This helps recognize a patient's heart disease and directs a doctor to better determine whether a person has or does not have cardiovascular disease. The dataset consists of 303 instances consisting of 13 attributes. Missing values for each column are replaced by mean values. Algorithms SVM, KNN, and NB are applied to the dataset. SVM had the best result with 86.8 percent accuracy [13].

For the prediction of heart disease, the researchers initially used local heart binary pattern (LBP) and PCA were used in conjunction with artificial neural network (NN). LBP and PCA are intended to extract the feature and reduce the set size and NN for the classification and verification of the accuracy of the unit. Input is patient image data in the proposed system and then stored as a dataset followed by multiple preprocessing techniques such as greyscale conversion, noise reduction and enhancement followed by extraction of the feature using LBP and PCA method, which extracts the feature data and is compressed with PCA. Then the NN applied which gives the final outcome of heart disease predictions. The proposed method of prediction was developed using the Matlab framework and predicts about 95 per cent of heart problems [14].

From the above papers, it is seen that RF and SVM have been applied to most of the datasets. The performance of RF is consistent among different dataset and the same will be applied on the dataset for this research.

II. THE TECHNIQUE EMPLOYED

There are several DM approaches that can be extended to individual projects. Data are collected and processed at a dramatic pace through a wide number of fields. A new generation of scientific theories and methods is being increasingly needed to help people obtain valuable information (knowledge) from the growing volumes of digital data. This is followed by a Cross-industry Standard Data Mining Procedure (CRISP-DM), and the most common approach followed is Information Discovery in Databases (KDD). The KDD methodology is used for the analysis.

KDD is concerned with designing the various methods and techniques for making the data meaningful. The fundamental problem that the KDD approach solves is that of translating low-level data into other forms that may be simpler, more abstract or more useful. The typical way in which data is translated into information is based on analysis and interpretation by hand. Analysis of the data has become important in every region. KDD was built to overcome the old approach.[1]

KDD is ideal for use and has 5 essential phases. The first stage is a collection of data. The data which will be used for the research project is selected at this point. However, the aim and scope are often specified before the data is selected. The 2nd stage is the data pre-processing. Until the data that will be used is chosen. In KDD pre-processing is a key step. The 3rd stage is called transformation It involves data washing, deletion of unnecessary data, imputation of missing data. It ensures enough data for the algorithm to make the algorithm work better. When the unnecessary data has been filtered out and the data has now been merged it must be converted. Data transformation transforms data as per the algorithm or model, from one type to another. This also involves converting data from a complex format into a simple one. 4th stage is called as DM. The proposed model or algorithm for ML is implemented at this point. The dataset is broken down into a check and train. The model is trained using the training dataset. When the testing has been completed, the software runs on a new data called the test that has never been used on the software. The 5th stage is called as interpretation. It is a critical point, as well. It is from this stage that knowledge is gained, and insights are learned [15].

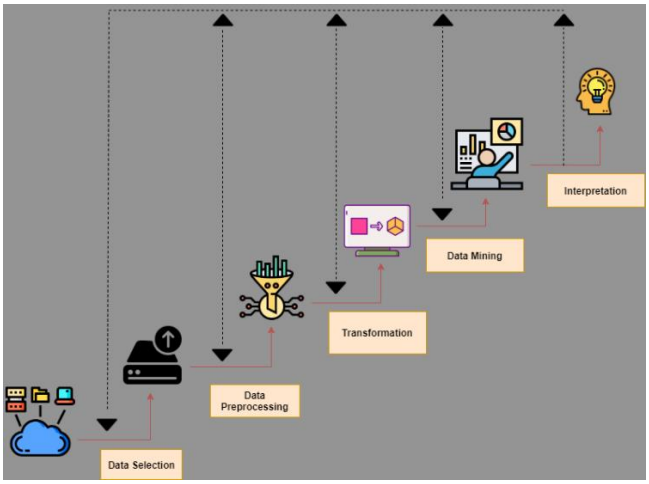

Fig. 1. KDD workflow

The data used to perform this research project comes from Kaggle. The dataset is made up of 70,000 rows, and 14 columns. The columns are Id, age day, age year, gender, height, weight, high blood pressure, low blood pressure, cholesterol, glucose, smoking, active and cardio alcohol use. Gender has attributes 1 and 2 where 1 is men's and 2 females. The cholesterol levels are 3 where 1 is normal, 2 above average and 3 above average. Smoking and alcohol are both binary values of 0 for no smoking and 1 for smoking. The active value is also binary where 0 is an inactive lifestyle and 1 is an active lifestyle. The final variable is the cardio result and is again a binary value where 0 is no CVD, and 1 is yes. The data is non-biased data that overcome different types of prejudice. Test and train selection will be done randomly by the split rule. The information provided as a result of CVD is almost equal in the number of cases.[2]

Figure 2 shows the correlation matrix which is the correlation coefficient between the variables. Variables can be correlated either positive or negative

1 https://medium.com/datadriveninvestor/data-science-project-management-methodologies-f6913c6b29eb

2 https://www.kaggle.com/raminhashimzade/cardio-disease

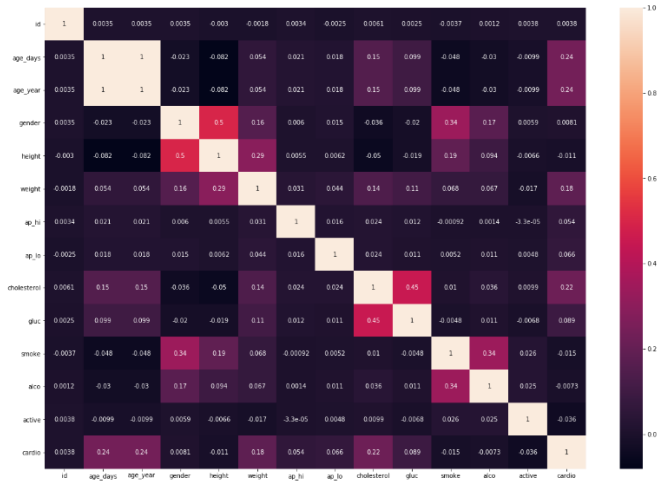

Fig. 2. Correlation Matrix

From figure 2, it is seen that the highest correlation is among age by days and age by years. But since age by years is extracted from age by days the correlation is not considered. The next high correlation is among height and gender with a value of 0.5.

Figure 3 shows a donut chart with a distinct count of height by gender. The division among the gender shows how much percent of that gender has CVD and how much doesn't have CVD along with the average height of the respective class. Females who don't have CVD have an average height of 161.61 cm while males who don't have CVD have an average height of 169.82 cm. Male who have CVD have an average height of 170.07 cm while females who have CVD have an average height of 161.10 cm.

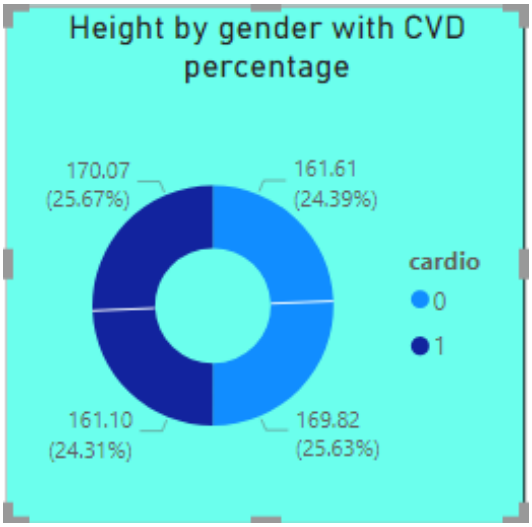

Fig. 3. Donut Chart of Height by Gender

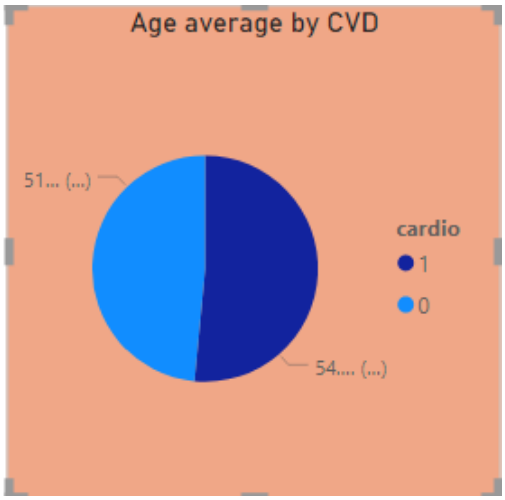

Fig. 4. Pie Chart of Age by CVD

Figure 4 shows that an almost equal number of people have CVD and don't have CVD. People who don't have CVD have an average age of 51.73 years while people who have CVD have an average age of 54.94%.

The DM algorithm selected for the project is RF. RF is made up of many individual decision trees that function as an ensemble. Every single tree in the random forest spits out a class prediction and the class with the most votes is the prediction of the model. The basic idea behind RF is a simple but powerful one — the wisdom of crowds. Many fairly uncorrelated models (trees) working as a committee would surpass any of the component models.[3]

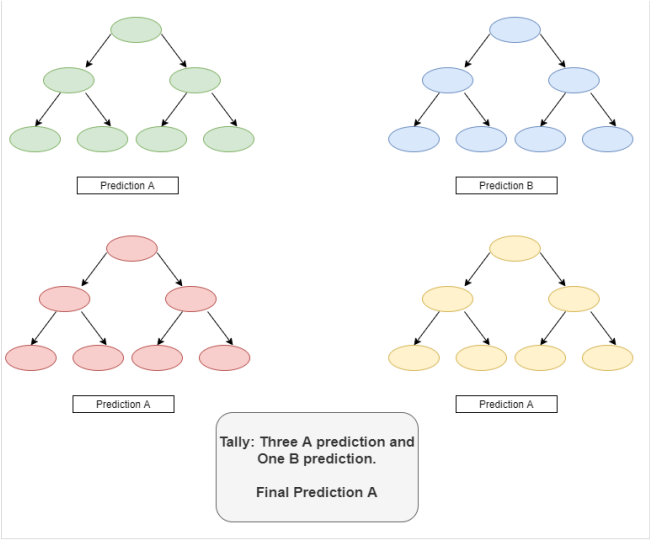

Fig. 5. RF prediction process

As seen in Figure 4, RF prediction is based on the majority vote gathered by the multiple DTs. The secret to this is the weak correlation between models. Just as investments with low correlations combine to create a portfolio greater than the sum of its components, uncorrelated models can generate predictions of the ensemble that are more accurate than any of the individual predictions. The explanation for this wonderful effect is that the trees shield one another from their mistakes.

Although some trees will be wrong, many other trees would be right so that the trees will move in the right direction as a group.

Few advantages of using a RF are as follows:

- Predictive performance can compete with the best supervised algorithms
- They provide a confident estimate of the importance of the feature
- They offer efficient test error estimates without incurring the cost of repeated model training linked to cross-validation.[4]

III. Results

The results are being calculated by the output of the confusion matrix. The confusion matrix is a table often used to describe the classification model (or "classifier") performance on a set of test data for which true values are known. It allows the visualisation of an algorithm's performance. A confusion matrix is a summary of the results of a prediction over a classification issue. The number of predictions that are correct and incorrect is summarized with count values and broken down by class. The confusion matrix shows how confused your model of classification becomes when it makes predictions. It gives us insight not only into the mistakes made by a classifier but, more importantly, the types of mistakes made.[5]

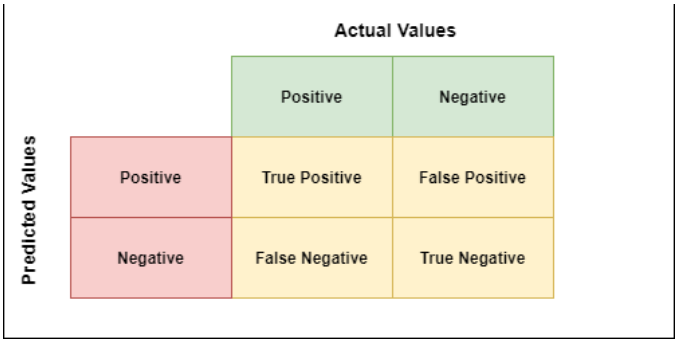

Fig. 6. Confusion Matrix

Figure 5 shows the details of the confusion matrix. The meaning of the values is as follows:

- True Positive (TP): Predictive positive and it is true
- True Negative (TN): Predicted negative and it is true
- False Positive (FP): Predicted positive and it is false
- False Negative (FN): Predicted negative and it is false

Some of the result evaluation standards are calculated using TP, TN, FP and FN and are shows below:

- Recall: Of all the positive classes, how much we were correctly predicted. It should be as high as possible. Recall = TP / (TP + FN)
- Precision: Of all the positive classes we rightly predicted how many are actually positive. It should be as high as possible. It shows the accuracy of the model Precision = TP/ (TP + FP)
- F measure: It is difficult to compare 2 models with low accuracy and high recall or vice versa. Comparing two models with low precision and high recall is difficult, or vice versa. So, we use f measure to make them comparable. F measure helps to simultaneously measure both Recall and Precision. F measure = 2*Recall*Precision / Recall + Precision.[6]

For n estimator = 300, the accuracy of the training model was 99.98%. However, the accuracy of the test dataset when for the model was 72%. The confusion matrix for the test dataset is shown in figure 6.

```
print('Accurancy: {:.0f}%'.format(model_rf.score(test, target_test)*100))

DecisionTree
[[6475 2270]
 [2531 5968]]
Accurancy: 72%
```

Fig. 7. Dataset Confusion Matrix

From figure 6, the TP value is 64.75%. It means 64.75% times CVD's were predicted positive and it was true. But as we are working on medical data, the most important value of all is TP and FN. TP is saying the person has CVD and it is true and this should be high. While FN is saying the person, he/she doesn't have CVD but, they have CVD. This should be as low as possible because if the person is said that they don't have CVD they won't do the treatment and would possibly result in death.

IV. Other Relevant Features of the Analysis

After the model has applied to the train data the accuracy of the model was 99.98% but when the same model has applied to the train data the accuracy drastically reduced to 72%. This is due to a overfit model or overfitting. The model has learned specifics that help it to perform better in training data which does not apply to the larger data population and therefore leads to worse test performance. Overfitting can be overcome by performing the following:

- Cross-validation: It is an effective preventive measure against overfitting. The idea is clever: To generate multiple mini train-test splits, use the initial training data. Tune your model using these splits.
- Train with more data: It doesn't always work but training with more data can help algorithms better detect the signal.
- Remove features: Some algorithms have the selection of functions built-in. Users can manually improve their generalizability for those who don't, by removing irrelevant input features.
- Regularization: It refers to a wide range of techniques that artificially compel the model to be simpler.
- Ensembling: Ensembles are methods of machine learning to combine predictions from multiple individual models.[7]

4 https://www.oreilly.com/library/view/hands-on-machine-learning/9781789346411/e17de38e-421e-4577-afc3-efdd4e02a468.xhtml
5 https://www.geeksforgeeks.org/confusion-matrix-machine-learning/
6 https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62
7 https://elitedatascience.com/overfitting-in-machine-learning

V. References

[1] S. S and S. D, "Early Forecasting and Prevention of Cardio Vascular Disease Based on Human Life Style Factors," in *2018 Conference on Emerging Devices and Smart Systems (ICEDSS)*, Tiruchengode, India, 2018.

[2] C. Guo, J. Zhang, Y. Liu, Y. Xie, Z. Han and J. Yu, "Recursion Enhanced Random Forest With an Improved Linear Model (RERF-ILM) for Heart Disease Detection on the Internet of Medical Things Platform,," *IEEE Access,* vol. 8, pp. 59247-59256, 2020.

[3] A. Gupta, R. Kumar, H. S. Arora and B. Raman, "MIFH: A Machine Intelligence Framework for Heart Disease Diagnosis," *IEEE Access,* vol. 8, pp. 14659-14674, 2020.

[4] S. Alty, S. Millasseau, P. Chowienczyc and A. Jakobsson, "Cardiovascular disease prediction using support vector machines," in *46th Midwest Symposium on Circuits and Systems*, Cairo, 2003.

[5] K. G. Dinesh, K. Arumugaraj, K. D. Santhosh and V. Mareeswari, "Prediction of Cardiovascular Disease Using Machine Learning Algorithms," in *International Conference on Current Trends towards Converging Technologies (ICCTCT)*, Coimbatore, 2018 .

[6] Bhuvaneswari Amma N G, "An intelligent approach based on Principal Component Analysis and Adaptive Neuro Fuzzy Inference System for predicting the risk of cardiovascular diseases," *2013 Fifth International Conference on Advanced Computing (ICoAC)*, Chennai, 2013, pp. 241-245.

[7] Y. Talatov and D. S. Ripka, "Diagnostics of the Cardiovascular System Based on Neural Networks," in *2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, St. Petersburg and Moscow, Russia, 2020.

[8] G. Cugliari, S. Benevenuta, S. Guarrera, C. Sacerdote, S. Panico, V. Krogh, R. Tumino, P. Vineis, P. Fariselli and G. Matullo, "Improving the prediction of cardiovascular risk with machine-learning and DNA methylation data," in *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, Siena, Italy, 2019.

[9] N. S. Rajliwall, R. Davey and G. Chetty, "Cardiovascular Risk Prediction Based on XGBoost," in *2018 5th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)*, Nadi, Fiji, 2018.

[10] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques,," *IEEE Access,* vol. 7, pp. 81542-81554, 2019.

[11] J. Maiga, G. G. Hungilo and Pranowo, "Comparison of Machine Learning Models in Prediction of Cardiovascular Disease Using Health Record Data," in *2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, Jakarta, Indonesia, 2019 .

[12] G. Choudhary and S. N. Singh, "Prediction of Cardiovascular Disease using Data Mining Technique," in *2019 4th International Conference on Information Systems and Computer Networks (ISCON)*, Mathura, India, 2019.

[13] N. Louridi, M. Amar and B. E. Ouahidi, "Identification of Cardiovascular Diseases Using Machine Learning," in *2019 7th Mediterranean Congress of Telecommunications (CMT)*, Fès, Morocco, 2019 .

[14] G. Suseendran, N. Zaman, M. Thyagaraj and R. K. Bathla, "Heart Disease Prediction and Analysis using PCO, LBP and Neural Networks," in *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, Dubai, United Arab Emirates, 2019 .

[15] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Magazine,* vol. 17, no. 3, p. 37, 1996.