Prediction System for Hearth Disease by Using Data Science Tools

Ömer Özgün Üngüder - 27523
Oguzhan Kaygusuz - 28159
Communication and Information Engineering
Hochschule Rhein-Waal
Oemer-Oezguen.Uengueder@hsrw.org
Oguzhan.Kaygusuz@hsrw.org

January 23, 2021

Abstract

Coronary Heart Diseases are the one of the most important problems of the modern world. Many people suffer from these diseases and most of cases are resulted in death. With the recent development in data science, many helpful systems could be developed to predict possible outcomes of various range of events. There are many different areas such as stock exchange market, self-driving cars, object detection and some disputable applications like DeepFake, where data science and machine learning tools are effectively being used.

This paper will focus on Coronary Heart Diseases, the factors effecting the disease and creating a model to predict potential patience by using data set obtained from clinical studies.

Contents

1 Theoretical Background

1.1 Coronary Heart Diseases

Coronary Heart Diseases (CHD) is considered as the leading factor causing the cardiovascular diseases in all around the world.Chatterjee et al.

(2014) It is sometimes called Ischaemic Heart Disease (IHC) or Coronary Artery Disease (CAD). In a healthy body the main responsibility of the artery system is to carry the blood from the organs to the cells. But under some circumstances some organic materials are piled up in the coronary arteries. With time, this piling up causes the coronary arteries to being narrowed and as a result blood flow is reduced. See Figure 1.

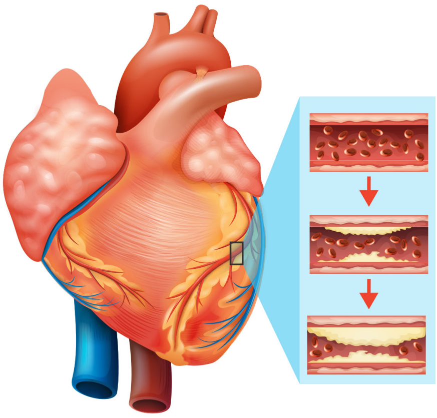


Figure 1: The blockage of an artery https://www.health.harvard.edu/heart-health/a-closer-look-at-heart-disease-risk

This blockage have many reverse effect on human health which can cause death.

According to the latest data in the global scale, in 2017 approximately 17,8 million people died from Cardiovascular Diseases(CVD). For the same year, it is estimated that almost 126,5 million people were living with CHD.Virani et al. (2020)

1.2 Risk Factors of CVD

There are many risk factors such as Sex, Age, Cholesterol, Diabetes an etc, which are having impact on CVD. Some examples of above-mentioned factors were listed and briefly explained in the following sub sections.

1.2.1 Sex

According to the World Health Organization (WHO) and analysis of Sophie H. Bots and her friend CHD mortality rate decreased between 1980 and 2010 for both genders.(see Figure 2, 3, 4, 5) However, mortality rate among men is still higher than women. Bots et al. (2017)

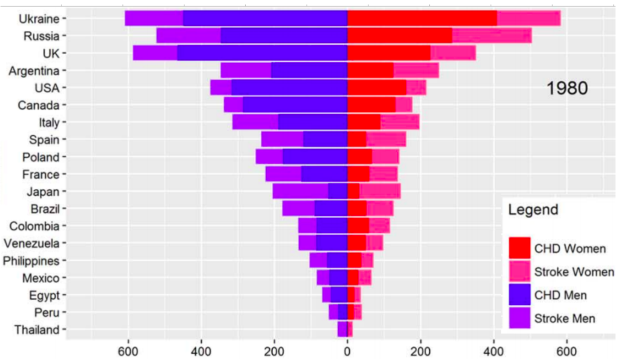


Figure 2: Mortality rates (per 100.000) among men and women in 1980 Bots et al. (2017)

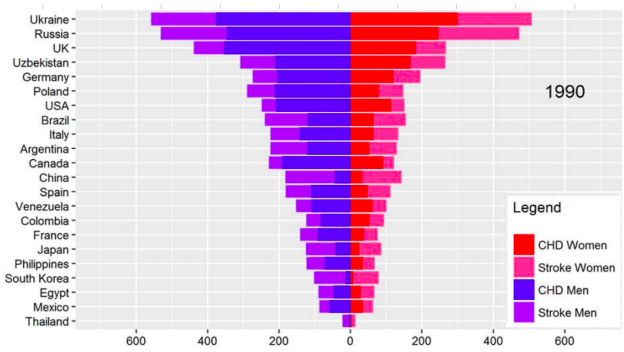


Figure 3: Mortality rates (per 100.000) among men and women in 1990 Bots et al. (2017)

1.2.2 Age

One of the important studies to reveal and prove the effect of aging in CHD belongs to Donald M Lloyd-Jones and his friend. In 1999, they have published an article in The Lancet by assessing 7733 participants of the Framingham Heart Study. According to their study, CHD risk before 40 years was very low (in men 1.2% and in women 0.2&). However, after 40 years lifetime risk dramatically increases in both women and men (in men 48.6% and in women 31.7%), while risk is decreasing by getting aging.Lloyd-Jones et al. (1999) (See Figure 6)

1.2.3 Diabetes

One of the important studies to detect and analyze the risk factors leading to CHD was published in 2006 by Rachel Huxley and her friends. In the study, studies published between 1966 and March 2005 were used to perform a meta analysis. In total 447,064 patients were identified in these studies.

According to the result of the study, diabetic

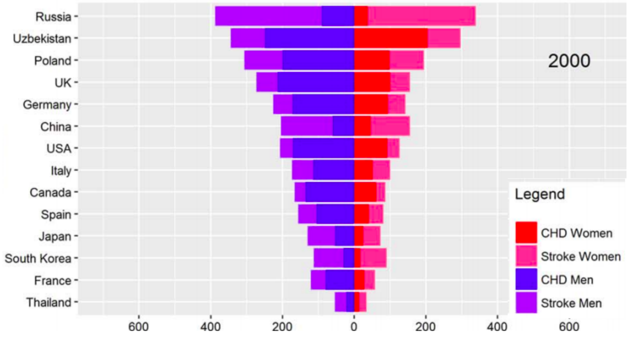


Figure 4: Mortality rates (per 100.000) among men and women in 2000 Bots et al. (2017)

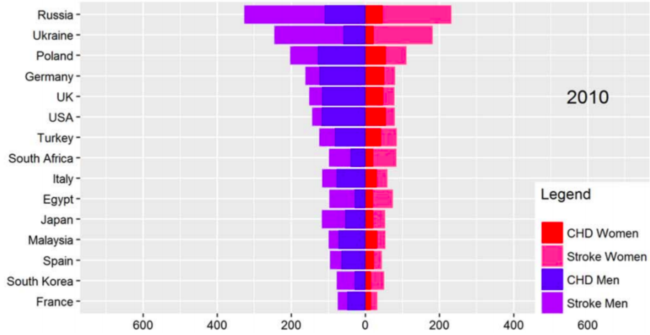


Figure 5: Mortality rates (per 100.000) among men and women in 2010 Bots et al. (2017)

patients show more fatal CHD rate than those without diabetes. The rate among the diabetics is 5.4%, whereas rate for those without diabetes is 1.6%.Huxley et al. (2005)

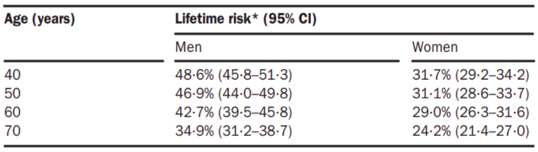
1.2.4 Smoking

Smoking is considered as a preventable risk factor for the CHD. Number of cigarettes smoked was associated with the increasing risk of developing CHD Willett et al. (1987)

1.2.5 Hypertension

Hypertension could be considered as increased blood pressure. As a result vessels are exposed to high blood pressure which harm the surface of arteries.

In the study of Stamler and his friends, 356,222 men were, who had no history for the disease, screened for trail and they detected a strong relation between blood pressure and CHD. Stamler et al. (1989)



Age (years)	Lifetime risk* (95% CI)	
	Men	Women
40	48·6% (45·8–51·3)	31·7% (29·2–34·2)
50	46·9% (44·0–49·8)	31·1% (28·6–33·7)
60	42·7% (39·5–45·8)	29·0% (26·3–31·6)
70	34·9% (31·2–38·7)	24·2% (21·4–27·0)

Figure 6: Lifetime risk of first coronary heart disease event at different ages reached free of coronary heart disease Lloyd-Jones et al. (1999)

1.2.6 Cholesterol

Cholesterol is an organic molecule produced by living animals and it is essential for the many cell activities. At the first two decades of people, cholesterol level remains same both for men and women. However, at the third and fourth decade cholesterol amount in the blood increases in men more than in women.van Lennep (2002)

Two different type of cholesterol Low-Density Lipoprotein Cholesterol (LDL) and High-Density Lipoprotein Cholesterol (HDL) are associated with the. The former has a negative effect on mortality rate from CHD, whereas the latter has reverse effect on the rate.Jacobs et al. (1990)

2 Dataset

Dataset, which is used in this study, has been gathered from the patients, who undergone angiography at the Cleveland Clinic in Cleveland, Ohio, the Hungarian Institute of Cardiology in Budapest, Hungary, the Veterans Administration Medical Center in Long Beach, California and the University Hospitals in Zurich and Basel, Switzerland.Detrano et al. (1989)

Dataset includes 303 instances and 75 attributes but in this study 13 of the attributes are used to create a model.Those are enumerated below.

1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholesterol in mg/dl
6. fasting blood sugar more than 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)

8. maximum heart rate achieved
9. exercise induced angina
10. oldpeak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) colored by fluoroscopy
13. Thalassemia: 3 = normal; 6 = fixed defect; 7 = reversible defect

3 Materials and Methods

Seven Machine Learning Classification Models have been used to predict the results. The best four of them have been selected for this study.These models are called Logistic Regression(LR),Random Forrest Classifier(RFC) and,Support Vector Machine(SVM),Naive Bayes(NB).

3.1 Machine Learning Models

3.1.1 Logistic Regression

Logistic Regression (LR) is one of the most important statistical techniques applied by statisticians and researchers for the analysis and classification of data sets.When we compare logistic regression and linear regression which may output continuous number values, logistic regression transforms its output using the logistic sigmoid function to give a probability value that can be mapped to two or more discrete classes. There are some advantages of LR. For example, it can indeed provide probabilities and extend to multiclass classification problems. Another advantage is that most of the methods used in LR model analysis follow the same principles used in linear regression.Maalouf (2011) The formula of LR is simply demonstrated in the following line.

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}} \tag{1}$$

As it is shown in the figure(figure 7). The sigmoid function maps any real value into another value between 0 and 1.It is used to map predictions to probabilities In machine learning.

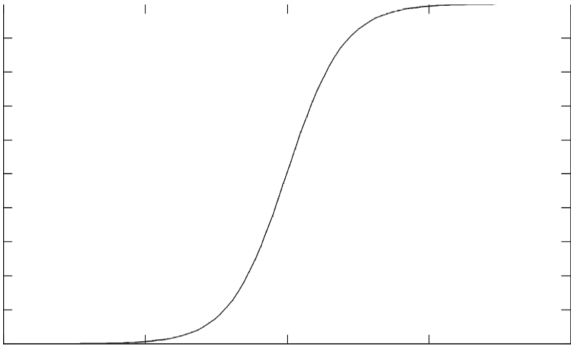


Figure 7: Logistic Response FunctionMaalouf (2011)

3.1.2 Random Forrest Classifier

Random forest is one of supervised learning algorithms. It is called forest beacuse it builds an ensemble of decision trees, usually trained with the “bagging” method. The basic concept of the bagging method is that a combination of learning models to get a better result.This combo classifier consists of several decision trees and merges them to get the best result. It principally applies bootstrap aggregating or bagging to able to learn. For example, X=x1,x2,x3,. . . ,xn with responses Y=x1,x2,x3,. . . ,xn which repeats the bagging from b=1 to B.

$$j = \frac{1}{B} \sum_{b=1}^{B} fb\left(x'\right) \tag{2}$$

The uncertainty of prediction on these trees is made with the help of standard deviation,Mohan et al. (2019)

$$\sigma = \sqrt{\frac{\sum_{\mathrm{b}=1}^{\mathrm{B}} \left(fb\left(x'\right) - \hat{f}\right)^2}{B-1}} \tag{3}$$

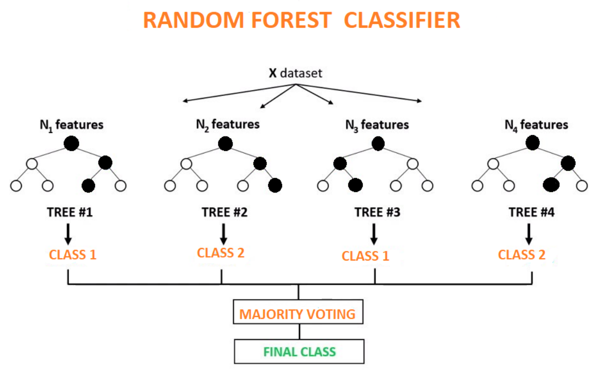


Figure 8: Demonstration of RFC

3.1.3 Support Vector Machine

SVM is one of the most widely used ML classification models nowadays.The main idea of support vector machine classification is that these measurements can be considered as a two-dimensional space. Then, Each case is represented by a data point in this space. By way of illustration, suppose that there are two classes in a two-dimensional case, a line can be now drawn to split these classes and to minimize the misclassification. More precisely, there are an infinite number of lines that can ensure an equally well separation between the data (Figure 8). A reasonable assumption behind support vector machine classification is that the best line is the one maximizing the margin between the two classes. In other terms, the line with a maximum distance to the closest data points from each of the two classes will be chosen by the model.

The mathematical representation of SVM was written in the following line.

$$\min_{w,b,\zeta} \frac{1}{2}||w||^2 + \sum_{i=1}^{m} \zeta_i s.t. y_i(w \cdot x_i + b) \geq 1 - \zeta_i,$$

$$\forall_i \in \{1, 2, \ldots, m\} \qquad (4)$$

After the separation line has been created, new cases can be automatically appointed to one of these classes depending on their position relative to the line.This method can be applied to any higher-dimensional space with the line becoming a plane in the three-dimensional case and a hyperplane for more than three dimensions. The closest data points of both classes to the hyperplane are called support vectors that determine the margin and proportionately the position and orientation of the hyperplane in n-dimensional space. (Figure 9).Dukart (2015)

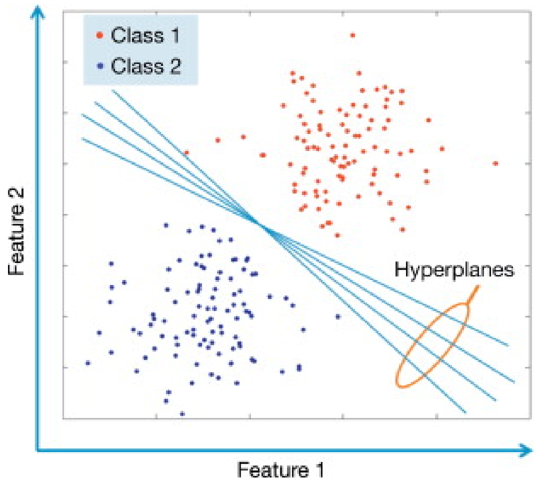


Figure 9: Results of support vector machine classification demonstrating that numerous hyperplanes can provide an equally good separation between the two classes. Dukart (2015)

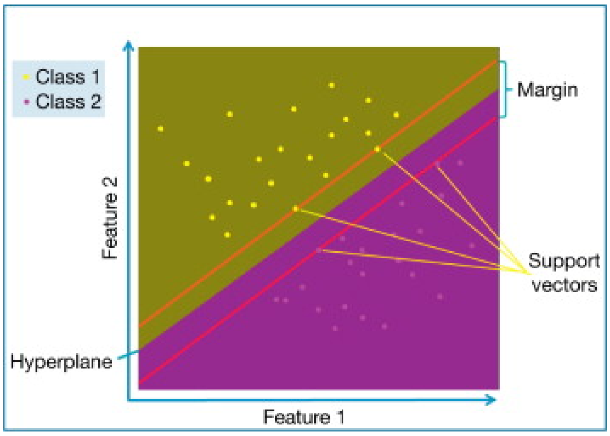


Figure 10: Results of support vector machine classification demonstrating the concept of support vectors and margin maximization. Dukart (2015)

3.1.4 Naive Bayes

Naive Bayes which is also called simple Bayes or independence Bayes is based on Bayes' theorem and an attribute independence assumption. Its competitive performance in classification is surprising, because the conditional independence assumption on which it is based, is rarely true in real world applications.Despite this unrealistic assumption, the resulting classifier known as naive Bayes is remarkably successful in practice,often competing with the other sophisticated techniques .Naive Bayes has proven effective in many practical applications, including text classification, medical diagnosis,and systems performance management.Zhang & Gao (2011)

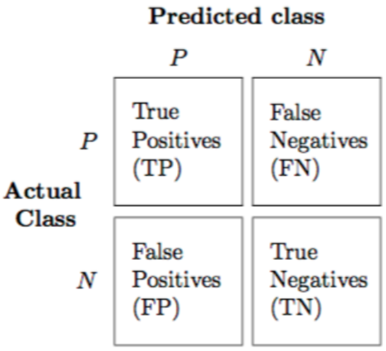
The mathematical representation of Gaussian Naive Bayes has been written in the following line.

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \tag{5}$$

3.2 Other Statistical Tools

Two comparison models have been applied to be able to see the results. These are called Confusion Matrix and Cross-Validation Score and have been explained briefly under this title.

3.2.1 Confusion matrix



		Predicted class	
		P	*N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Figure 11: Confusion matrix for binary classificationKulkarni et al. (2020)

Confusion matrix is one of the most popular measures which is used while solving classification problems. It performs on binary classification as well as for multi-class classification problems.An example of a confusion matrix for binary classification is shown in the table (figure 10). Confusion matrices represent counts from predicted and actual values. The output "TN" stands for True Negative which shows the number of negative examples classified accurately. Similarly, "TP" stands for True Positive which indicates the number of positive examples classified accurately. The term "FP" shows False Positive value, i.e., the number of actual negative examples classified as positive; and "FN" means a False Negative value which is the number of actual positive examples classified as negative. One of the most commonly used metrics while performing classification is accuracy. The accuracy of a model (through a confusion matrix) is calculated using the given formula below.Kulkarni et al. (2020)

$$Accuracy = \frac{TP + TF}{TP + TF + FP + FN} \tag{6}$$

3.2.2 K-fold Cross-Validation

Cross-validation is a procedure used to evaluate machine learning models on a data sample.The function has a single parameter "k" that refers to the number of groups that a given data sample is to be split into. In other words, when a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation(This value has been chosen for this study as well).It is primarily used in applied machine learning to estimate the accuracy of a machine learning model on data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.It generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split. Cross-validation is one of the most widely used data resampling methods to estimate the true prediction error of models and to adjust model parameters.Yadav & Shukla (2016)

3.2.3 Standart Division

The standard deviation is a summary measure of the differences of each observation from the mean.In other words, it is a measure of how close the numbers are to the mean. For instance, blood sugar levels of the study sample should be measured from the same population in order to understand blood sugar levels of the population. The findings of this sample are best characterized by two parameters; mean and SD. It is the center of distribution of observations (central tendency). Other parameter, SD tells us distribution of individual observations about the mean. Namely, it characterizes typical distance of an observation from distribution center or middle value. If observations are more distributed, then there will be more variability. Thus, a low SD indicates less variability while high SD means more spread out of data. Mathematically, the SD is represented in the formula below.Barde & Barde (2012)

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{7}$$

4 Conclusion

The data has been inspected regarding 13 categories.303 participants had participated in this study. The number of men participants is 207. In contrast, the number of women participants is 96. The average age had been founded 54.36 with 9.08 SD. The age of the youngest participant is 29, whereas the age of the oldest is 77.The percentage of Patients who Have Heart Disease is 54.46%.As it is shown in the figure(figure 11), The participants with heart diseases have a bigger tendency to be young than the participants that have no heart diseases. At the same time, Age disperse of the participants with heart diseases is more than the participants that have no heart diseases.

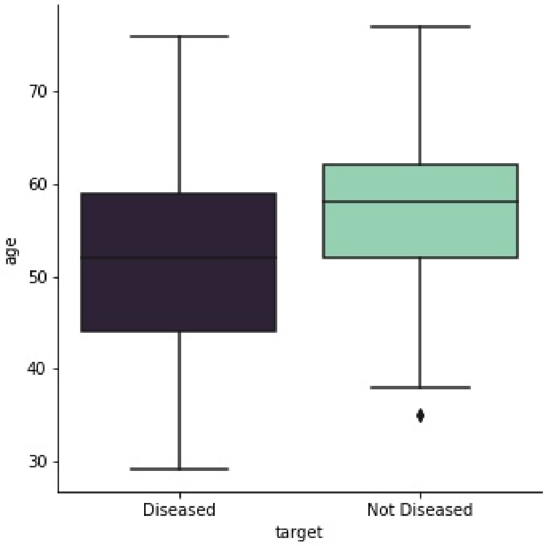


Figure 12: The diagram regarding age, Patients who have Heart Disease, and Patients who do not have Heart Disease.

The age distribution of patients regarding their gender has been shown in the figure(figure 12). In other terms, The disperse of male and female participants regarding their age.

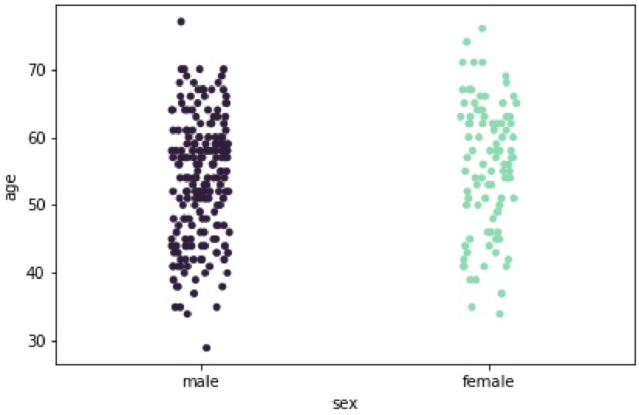


Figure 13: The diagram regarding age and sex

It is shown in the figure that the distribution the Patients who have Heart Disease, and Patients who do not have Heart Disease with respect to their genders.

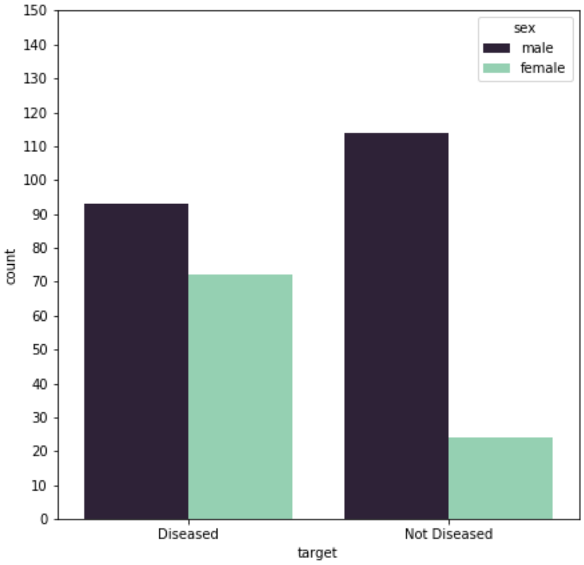


Figure 14: The diagram shows the nu,ber of patients regarding gender ,and the Patients who have Heart Disease, and Patients who do not have Heart Disease.

4.1 Logistic Regression

When the results have been compared to the other results in the other models. Logistic Regression is one of the best models. It has a satisfying value on the confusion matrix that is shown in the figure (figure11). Furthermore, it has the best cross-validation values..

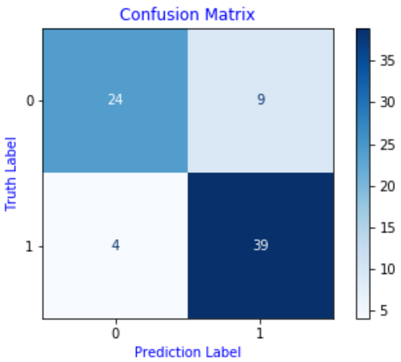


Figure 15: The accuracy of the chosen data sample on logistic Regression has been found as %82.89.

4.1.1 K-fold Cross-Validation

When has been the mean accuracy of the cross-validation value calculated, 82.39 percent has been found. Additionally, The Standard Deviation of this value has been found as 4.34 percent.

4.2 Random Forrest Classifier

When the results have been compared to the other results in the other models. Random Forrest Classifier is one of the best models. It has the reliable cross-validation values that is shown in the table below.

4.2.1 K-fold Cross-Validation of Random Forrest Classifier

For each n-estimator, a parameter of the RFC model, 10 fold cross-validation had been applied separately to get the best result. The mean of 10 tests had been calculated 82,71 percent. Additionally, The Standard Deviation of this value has been found as 7,79 percent.

4.3 Support Vector Machine

Support Vector Machine has satisfying value on the confusion matrix that is shown in figure(figure13).Moreover its values on k-fold cross validation is quite suitable.

Num.	Accuracy	SD
1	%84,11	%8,04
2	%81,94	%9,82
3	%83,30	%7,72
4	%80,65	%9,30
5	%81,39	%8,55
6	%84,60	%3,46
7	%83,73	%6,78
8	%82,71	%10,00
9	%79,72	%7,38
10	%85	%6,87
Mean	%82,71	%7,79

Table 1: Accuracy and SD values of RFC

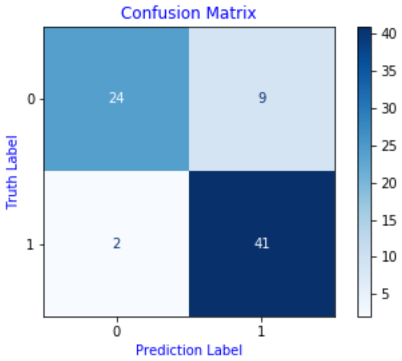


Figure 16: The accuracy of the chosen data sample on logistic Regression has been found as %85.53

4.3.1 K-fold Cross-Validation of Support Vector Machine

When has been the mean accuracy of the cross-validation value calculated, 81.05 percent has been found. Besides that, The Standard Deviation of this value has been found as 6.25 percent.

4.4 Naive Bayes

The result of Naive Bayes is also quite impressive. It has a satisfying value on the confusion matrix that is shown in figure(figure 14). In addition, it has a good cross-validation value.

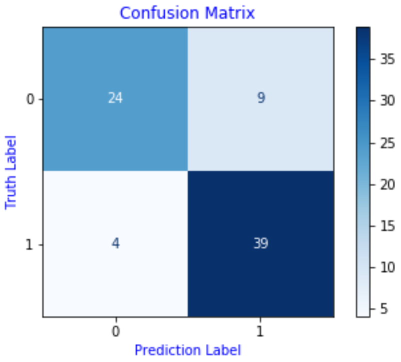


Figure 17: The accuracy of the chosen data sample on logistic Regression has been found as %82.89

4.4.1 K-fold Cross-Validation of Naive Bayes

When has been the mean accuracy of the cross-validation value calculated, 81.90 percent has been found. Additionally, The Standard Deviation of this value has been found as 7.03 percent.

4.4.2 Summary of Conclusion

When the 4 machine learning models were compared to each other for this study. It could be said that LR and RFC are the best models regarding their k-fold Cross Validation accuracy and SD values. A parameter(n-estimator) of RFC was changed by a function per every case to get the best values. Therefore, We had different results on RFC per every case. Although it looks like RFC had created the best result in the competition, the result of RFC is the mean of 10 tests. Because RFC didn't have produced better results for each test than LR, it is impossible to say that the result of RFC is better than the result of LR. When we considered dynamic results of RFC, it can be concluded that LR had produced the best results that had been found %82.39 for accuracy with %4.34 SD depends on k fold validation and those values are stable.Although The other two models didn't have produced better results than RFC and LR, they can not be considered bad. While The accuracy result for NB is %81.90 with % 7.03 SD, it had been found %81.05 with % 6.25 SD for SVM.

5 Discussion

Our study was investigated on 303 participants, with more patients maybe had better results. The classic machine learning models had been used in this study. But, The hybrid models can be used to make the results better.Mohan et al. (2019). Some columns can be dropped as well as some columns can be generated depending on the main categories to be able to create much better models. Our study indicated that As it is shown in the figure(figure 11), The participants with heart diseases have a bigger tendency to be young than the participants that have no heart diseases. There are studies in the literature that claim the opposite of it.Lloyd-Jones et al. (1999) As much as we had avoided doing overfitting, we may have done it in LR. It is hard to make overfitting for RFC because of its structure.

6 Appendix

References

Barde, P. & Barde, M. (2012), ‘What to use to express the variability of data: Standard deviation or standard error of mean?’, *Perspectives in Clinical Research* **3**(3), 113.

Bots, S. H., Peters, S. A. E. & Woodward, M. (2017), ‘Sex differences in coronary heart disease and stroke mortality: a global assessment of the effect of ageing between 1980 and 2010’, *BMJ Global Health* **2**(2), e000298.

Chatterjee, K., Anderson, M., Heistad, D. & Kerber, R. E. (2014), *Manual of Coronary Heart Diseases*, Jaypee Brothers Medical Publishers.

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.-J., Sandhu, S., Guppy, K. H., Lee, S. & Froelicher, V. (1989), ‘International application of a new probability algorithm for the diagnosis of coronary artery disease’, *The American Journal of Cardiology* **64**(5), 304–310.

Dukart, J. (2015), Basic concepts of image classification algorithms applied to study neurodegenerative diseases, *in* ‘Brain Mapping’, Elsevier, pp. 641–646.

Huxley, R., Barzi, F. & Woodward, M. (2005), ‘Excess risk of fatal coronary heart disease associated with diabetes in men and women: meta-analysis of 37 prospective cohort studies’, *BMJ* **332**(7533), 73–78.

Jacobs, D. R., Mebane, I. L., Bangdiwala, S. I., Criqui, M. H. & Tyroler, H. A. (1990), ‘High density lipoprotein cholesterol as a predictor of cardiovascular disease mortality in men and women: the follow-up study of the Lipid Research Clinics Prevalence Study’, *American Journal of Epidemiology* **131**(1), 32–47.

Kulkarni, A., Chong, D. & Batarseh, F. A. (2020), Foundations of data imbalance and solutions for a data democracy, *in* ‘Data Democracy’, Elsevier, pp. 83–106.

Lloyd-Jones, D. M., Larson, M. G., Beiser, A. & Levy, D. (1999), ‘Lifetime risk of developing coronary heart disease’, *The Lancet* **353**(9147), 89–92.

Maalouf, M. (2011), ‘Logistic regression in data analysis: an overview’, *International Journal of Data Analysis Techniques and Strategies* **3**(3), 281.

Mohan, S., Thirumalai, C. & Srivastava, G. (2019), ‘Effective heart disease prediction using hybrid machine learning techniques’, *IEEE Access* **7**, 81542–81554.

Stamler, J., Neaton, J. D. & Wentworth, D. N. (1989), ‘Blood pressure (systolic and diastolic) and risk of fatal coronary heart disease.’, *Hypertension* **13**(5_Suppl), I2–I2.

van Lennep, J. R. (2002), ‘Risk factors for coronary heart disease: implications of gender’, *Cardiovascular Research* **53**(3), 538–549.

Virani, S. S., Alonso, A., Benjamin, E. J., Bittencourt, M. S., Callaway, C. W., Carson, A. P., Chamberlain, A. M., Chang, A. R., Cheng, S., Delling, F. N., Djousse, L., Elkind, M. S. V., Ferguson, J. F., Fornage, M., Khan, S. S., Kissela, B. M., Knutson, K. L., Kwan, T. W., Lackland, D. T., Lewis, T. T., Lichtman, J. H., Longenecker, C. T., Loop, M. S., Lutsey, P. L., Martin, S. S., Matsushita, K., Moran, A. E., Mussolino, M. E., Perak, A. M., Rosamond, W. D., Roth, G. A., Sampson, U. K. A., Satou, G. M., Schroeder, E. B., Shah, S. H., Shay, C. M., Spartano, N. L., Stokes, A., Tirschwell, D. L., VanWagner, L. B. & Tsao, C. W. (2020), ‘Heart Disease and Stroke Statistics 2020 Update: A Report From the American Heart Association’, *Circulation* **141**(9).

Willett, W. C., Green, A., Stampfer, M. J., Speizer, F. E., Colditz, G. A., Rosner, B., Monson, R. R., Stason, W. & Hennekens, C. H. (1987), ‘Relative and absolute excess risks of coronary heart disease among women who smoke cigarettes’, *New England Journal of Medicine* **317**(21), 1303–1309.

Yadav, S. & Shukla, S. (2016), Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification, *in* ‘2016 IEEE 6th International Conference on Advanced Computing (IACC)’, IEEE.

Zhang, W. & Gao, F. (2011), ‘An improvement to naive bayes for text classification’, *Procedia Engineering* **15**, 2160–2164.

List of Figures

Acronyms

CHD - Coronary Heart Disease
IHD - Ischaemic Heart Disease
CAD - Coronary Artery Disease
CVD - Cardiovascular Disease
LDL - Low-Density Lipoprotein Cholesterol
HDL - High-Density Lipoprotein Cholesterol
WHO - World Health Organization
ML - Machine Learning
LR - Logistic Regression
SVM - Support Vector Machine
RFC - Random Forrest Classifier
NB - Naive Bayes