Heart Disease Prediction using Machine Learning Algorithms

Sehar Randhawa
Indian School Al Ghubra
Muscat, Oman

***Abstract*—Heart disease is one of the biggest causes of morbidity and mortality among the population of the world. Prediction of cardiovascular disease is regarded as one of the most important subjects in the section of data analysis. The load of cardiovascular disease is rapidly increasing all over the world over the past few years. Many researches have been conducted in an attempt to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn, reduces the complications. Machine learning proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry. This project aims to predict future Heart Disease by analysing data of patients which classifies whether they have heart disease or not using the machine-learning algorithm.**

***Index Terms*—Heart disease, prediction, Machine Learning, Logistic Regression, K-Nearest Neighbours**

I. Introduction

The number of deaths due to cardiovascular diseases increased by 41 % between 1990 and 2013, climbing from 12.3 million deaths to 17.3 million deaths globally. In addition to that, half of the deaths in the United States and other developed countries are due to the same issue[1]. Therefore, early detection of heart diseases is required to reduce the health complications. Machine learning has been widely used in the modern healthcare sector for diagnosing and predicting the presence of diseases using data models. Logistic regression is one such relatively used machine learning algorithm for studies involving risk assessment of complex diseases. Thus, the study intends to identify the most significant predictors of cardiovascular diseases and predict the overall risk by using logistic regression and KNN algorithms.

II. Background of the Study

The dataset which was used for the logistic regression analysis is available on the Kaggle from an ongoing cardiovascular study of Framingham, Massachusetts. The classification goal of this study is to predict whether the patient has 10-year risk of future heart diseases. The Framingham dataset consists of 4238 records of patients data and 15 attributes.The data analysis is carried out in Python programming by using Google Colab which is a flexible and powerful data science application software.

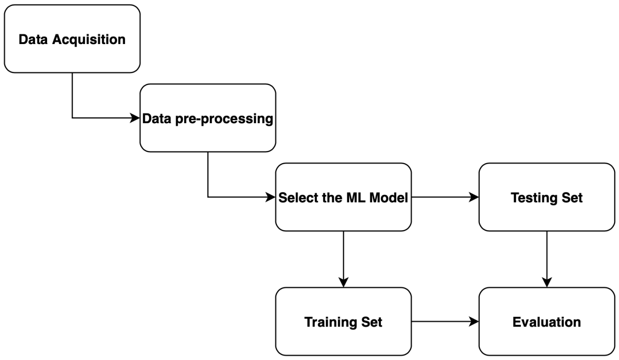
III. Machine Learning

Machine learning is widely used in almost many fields in the world including the healthcare sector. Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed [2].

Further, machine learning at its most basic is the practice of using algorithms to parse data, learn from it, and then make a determination or prediction about something in the world [3]. There are two major categories of problems often solved by machine learning i.e. regression and classification. Mainly, the regression algorithms are used for numeric data and classification problems include binary and multi-category problems [4].

Machine learning algorithms are further divided into two categories such as supervised learning and unsupervised learning [5]. Basically, supervised learning is performed by using prior knowledge in output values whereas unsupervised learning does not have predefined labels hence the goal of this is to infer the natural structures within the dataset [6]. Therefore, selection of machine learning algorithms need to be carefully evaluated.

IV. Methodology



V. Python Libraries

A. NumPy

A library for the Python programming language used to add support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

B. Pandas

A library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

C. Scikit-learn (also known as sklearn)

A machine learning library for the Python programming language that features various classification, regression and clustering algorithms including support-vector machines, random forests, gradient boosting and k-means. It is designed to interoperate with the Python numerical library NumPy.

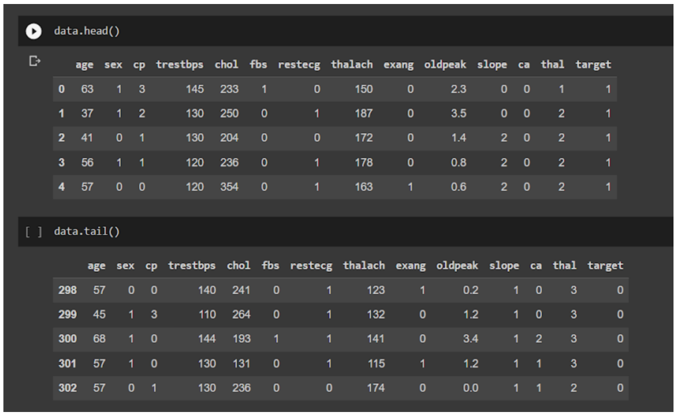
D. Matplotlib

Matplotlib is a python library used to create 2D graphs and plots by using python scripts. It has a module named pyplot which makes things easy for plotting by providing features to control line styles, font properties, formatting axes etc.

E. Seaborn

Seaborn is a Python data visualisation library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

VI. Dataset



```
data.head()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

```
[ ] data.tail()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

Fig. 1. The Heart Disease Dataset by University of California, Irvine

VII. Machine Learning Algorithms used for Prediction

A. Logistic Regression

Logistic regression is a one of the machine learning classification algorithms for analysing a dataset in which there are one or more independent variables (IVs) that determine an outcome and also categorical dependent variable (DV). Linear regression uses output in continuous numeric whereas logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes. The logistic regression forms three types as below.

- Binary logistics regression (two possible outcomes in a DV)
- Multinomial logistics regression (three or more categories in DV without ordering)
- Ordinal logistics regression (three or more categories in DV with ordering)

B. K-Nearest Neighbours

KNN tries to find similarities between predictors and values that are within the dataset. KNN uses a non-parametric method as there is not a particular finding of parameters to a particular functional form. It does not make any type of assumptions about the features and output of the dataset. KNN is also called a lazy classifier as it memorises the training data and does not exactly learn and fix the weights. Hence most of the computing work occurs during the classification rather than training time. KNN usually works by just trying to see which class is the new feature near to and it just puts it to the class closest to that point.

1) Working of KNN Algorithm: Initially, we select a value for K in our KNN algorithm. Now we go for a distance measure. Let's consider the Euclidean distance here. Find the Euclidean distance of k neighbours. Now we check all the neighbours to the new point we have given and see which is nearest to our point. We only check for k-nearest here. Now we see which class there is the highest number obtained. The max number is chosen and we assign our new point to that class. In this way, we use the KNN algorithm.

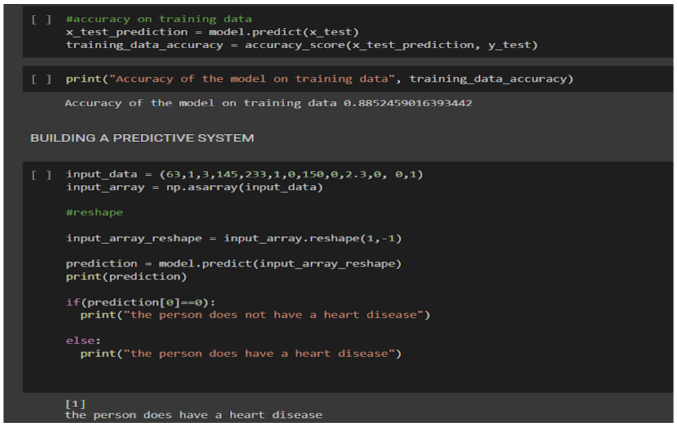
2) The ideal value of K in KNN: If we go for too small a value of k, there is a good chance we may have overfitting of data, that is the algorithm may perform reasonably well on training but not well on testing data. And, we also may encounter noise if we just use the small value of k, if we have large data.

3) Advantages and Disadvantages of KNN Algorithm:
Advantages:

- We can implement the algorithm with ease.
- It is very effective against noisy data by averaging k-nearest neighbours.
- Works well in case of large data.
- The decision boundaries that are formed can be of arbitrary shapes.

Disdvantages:

- Curse of dimensionality: Domination of distances by irrelevant attributes.
- Finding the correct value of k may be time expensive sometimes.
- Very high computation cost due to its distance measure.



```
[ ]  #accuracy on training data
     x_test_prediction = model.predict(x_test)
     training_data_accuracy = accuracy_score(x_test_prediction, y_test)
```

```
[ ]  print("Accuracy of the model on training data", training_data_accuracy)

     Accuracy of the model on training data 0.8852459016393442
```

BUILDING A PREDICTIVE SYSTEM

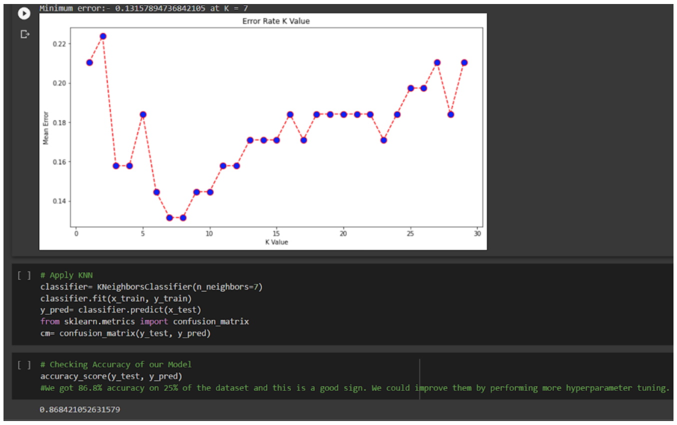
```
[ ]  input_data = (63,1,3,145,233,1,0,150,0,2.3,0, 0,1)
     input_array = np.asarray(input_data)

     #reshape

     input_array_reshape = input_array.reshape(1,-1)

     prediction = model.predict(input_array_reshape)
     print(prediction)

     if(prediction[0]==0):
       print("the person does not have a heart disease")

     else:
       print("the person does have a heart disease")


     [1]
     the person does have a heart disease
```

Fig. 2. Result of Logistic Regression

VIII. Comparing Both Algorithms

A. Accuracy Result of Logistic Regression

B. Accuracy Result of K-Nearest Neighbours



```
Minimum error:- 0.13157894736842105 at K = 7
```

```
[ ]  # Apply KNN
     classifier= KNeighborsClassifier(n_neighbors=7)
     classifier.fit(x_train, y_train)
     y_pred= classifier.predict(x_test)
     from sklearn.metrics import confusion_matrix
     cm= confusion_matrix(y_test, y_pred)
```

```
[ ]  # Checking Accuracy of our Model
     accuracy_score(y_test, y_pred)
     #We got 86.8% accuracy on 25% of the dataset and this is a good sign. We could improve them by performing more hyperparameter tuning.

     0.868421052631579
```

Fig. 3. Result of K-Nearest Neighbours

IX. Result

KNN is actually a lazy classifier. It memorises the training data and does not exactly learn and fix the weights. KNN is a non-parametric model, whereas Logistic Regression is a parametric model. KNN is comparatively slower than Logistic Regression. KNN supports non-linear solutions where Logistic Regression supports only linear solutions. Logistic Regression can derive confidence level (about its prediction), whereas KNN can only output the labels. In our results too Regression gave an accuracy of a whooping 88.5% whereas KNN's accuracy was limited to 86.7%, which stand alone is a feat in itself, but still lesser than our Regression Model.

X. Conclusion

Manually determining the odds of cardiovascular disease based on risk factors can be hard. Using Machine learning techniques we can predict the outcome with the help of existing data. But still, we can't always trust the machine. As you can see from this prediction, we got some good accuracy of 86-89% irrespective of the machine learning algorithms we used, however good that number might seem to be, but that still leaves us with a 11-14% uncertainty. And when lives are at stake, one can't take even an iota of risk. So, the only way to prevent heart disease is to stay healthy.

References

References

[1] Mozaffarian, D., Benjamin, E., Go, A., Arnett, D., Blaha, M.Cushman, M. et al. (2015). Heart Disease and Stroke Statistics—2015, Update. Circulation, 131(4). doi: 10.1161/cir.0000000000000152.

[2] Das, S., Dey, A., Pal, A., & Roy, N. (2015). Applications of Artificial Intelligence in Machine Learning: Review and Prospect. International Journal of Computer Applications, 115(9), 31-41. doi: 10.5120/20182-2402

[3] Abduljabbar, R., Dia, H., Liyanage, S., & Bagloee, S. (2019). Applications of Artificial Intelligence in Transport: An Overview. Sustainability, 11(1), 189. doi: 10.3390/su11010189

[4] Strecht, Pedro & Cruz, Luís & Soares, Carlos & Moreira, João & Abreu, Rui. (2015). A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance https://www.researchgate.net/publication/278030689_A_Comparative_Study_of_Classification_and_Regression_Algorithms_for_Modelling_Students'_Academic_Performance

[5] Ng, A. Y. and Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In NIPS 14, pp. 841–848.