DETERMINE THE RISK FACTOR OF HEART DISEASE BASED ON MEDICAL DATASET.

Dr.M.Rajeswari
Assistant Professor
Department of B.Com Business Analytics
PSGR Krishnammal College for Women, Coimbatore, India
rajeshwarim@psgrkcw.ac.in

M.Kaviya[2],
UG Scholar,
Department of B.Com(Business Analytics),
PSGR Krishnammal College for Women, Coimbatore, India.
kaviyamani27.km@gmail.com

ABSTRACT:

This paper mainly aims about risk factor of heart disease using medical dataset. The symptoms of heart disease are mainly Chest pain, chest tightness, shortness of breath and chest discomfort (angina).Machine learning (ML) has proved to be vital in predicting heart disease. Using their medical dataset is also a way to predict heart disease.

Keywords- Heart disease, Machine learning (ML), Medical dataset.

I.INTRODUCTION

The term "heart disease" refers to several types of heart conditions. In the United States, the most common type of heart disease is coronary artery disease (CAD), which can lead to heart attack. There are many types of heart disease, and each one has its own symptoms and treatment. For some, lifestyle changes and medicine can make a huge difference in improving your health. For others, you may need surgery to make your ticker work well again. Logistic Regression is used to predict the chances of affecting heart disease by comparing the medical data set of an individual with the given data set. CAD is the most common heart problem. With CAD, you may get blockages in your coronary arteries – the vessels that supply blood to your heart. That can lead to a decrease in the flow of blood to your heart muscle, keeping it from getting the oxygen it needs. The disease usually starts as a result of atherosclerosis, a condition sometimes called hardening of the arteries. Coronary heart disease can give you pain in your chest, called angina, or lead to a heart attack.

II.RELATED WORKS:

Many old physicians thought that high BP was necessary to force blood through the stiffened arteries of older persons and that it was a normal element of aging. The medical community believed that a permissible systolic BP was 100 plus the participant's age in millimeters of mercury [1]

For those aged >70 years, some considered the acceptable upper limits of normal BP to be 210 mmHg systolic and 120 mmHg diastolic.[2]

The cardiovascular hazard of hypertension was believed to derive chiefly from the diastolic pressure component. Consequently, elevated systolic pressure was considered harmless, especially in the elderly.[3]

Blood pressure is now recognized universally that hypertension increases atherosclerotic CVD incidence; the risk burden is 2–3-fold. CAD is the most common sequelae for hypertensive patients of all ages.[4]

Hypertension predisposes to all clinical manifestations of CHD including myocardial infarction, angina pectoris, and sudden death. Even high normal BP values are associated with an increased risk of CVD.[5]

The risk ratio for intracerebral hemorrhage was greater than for atherothrombotic brain infarction.[6]

It was found that hypertension was as strong a risk for atherothrombotic brain infarction as intracerebral hemorrhage.[7]

Framingham showed that the preponderance of hypertension-related strokes were atherothrombotic brain infarctions whether the hypertension was severe or mild. The proportion of strokes due to hemorrhage in mild hypertension was identical to that for severe hypertension.[8]

The Seventh JNC on hypertension established that those with BP of 120–139/80–89 mmHg are pre hypertensives, that is, these individuals may become hypertensives in the future. Starting as low as 115/75 mmHg, the risk of heart attack and stroke doubles for every 20-point jump in systolic BP or every 10-point rise in diastolic BP for adults aged 40–70.[9]

The presence of other risk factors for CVD such as high cholesterol, obesity, and diabetes is seen more in people with prehypertension than in those with normal blood pressure. The CVD risk in pre hypertensives increases with the number of associated risk factors present. Therefore, prehypertension confers a greater risk for CVD.[10]

The other major risk for CVD was cholesterol. In 1953, an association between cholesterol levels and CHD mortality was reported in various populations.[11]

It was shown that changes in cholesterol levels were associated with changes in CVD incidence rate.[12]

III. METHADOLOGY

A.DATA PREPROCESSING

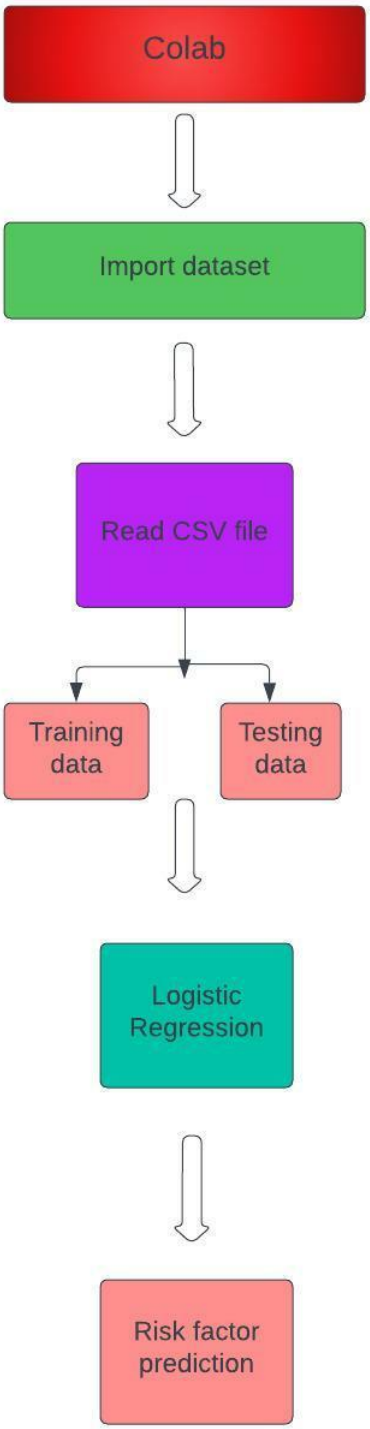
Data preprocessing is a manipulation or dropping of data before it is used in order to ensure or enhance performance. It is a process of transforming raw data into an understandable format. We use logical regression algorithm to predict heart disease dataset.

B.LOGICAL REGRESSION

This supervised learning method gives the probability of a target variable. Logistic regression uses fairly common machine learning algorithm that's accustomed predict categorical outcomes. Logistic regression uses classification algorithm, used when the worth of the target variable is categorical in nature. Logistic regression is most ordinarily used when the information in question has binary output, so when it belongs to at least one class or another, or is either a 0 or 1. The following steps for logistic regression:

- **STEP 1:** Generate a dataset and download necessary packages.
- **STEP 2:** Splinter the dataset into test and training dataset. Training set - used to train the model. Testing set – describes the evaluation of the models.
- **STEP 3**: Visualization gives a better scope of interactivity of the algorithm to convey a better understanding of the data set.
- **STEP 4:** Define a prediction value using logistic regression.

C.FLOWCHART


Colab
Import dataset
Read CSV file
Training data
Testing data
Logistic Regression
Risk factor prediction

IV.RESULT

```
plt.show()

/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43:
FutureWarning: Pass the following variable as a keyword arg: x. From
version 0.12, the only valid positional argument will be `data`, and
passing other arguments without an explicit keyword will result in an
error or misinterpretation.
  FutureWarning

AxesSubplot(0.125,0.125;0.775x0.755)
```

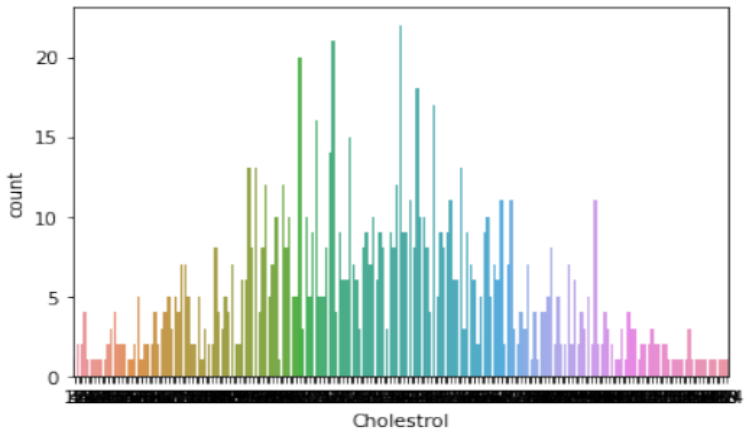


Fig1

Fig1 represents the cholestral level.

```
     predicted_value  KnowO/P
313                0      272
727                2      245
204                4      186
442                0      265
400                2      255
..               ...      ...
498                0      258
178                4      236
970                2      199
568                4      189
703                2      294

[300 rows x 2 columns]

#STATUS
#accuracy for HeartDisease
metrics.accuracy_score(test_Y,predicted_value)

0.53

#prediction heart disease
pd.DataFrame({'predicted_value':predicted_value,'KnowO/P':test_Y})

     predicted_value  KnowO/P
313                1        1
727                1        1
204                0        0
442                1        1
400                1        1
..               ...      ...
498                1        1
178                0        0
970                0        0
```

```
970                0        2
568                0        1
703                0        2

[300 rows x 2 columns]

#STATUS
#accuracy for cholestrol
metrics.accuracy_score(test_Z,predicted_value)#STATUS
#accuracy for cholestrol

0.13
```

Fig 2

Fig 2 represents the accuracy level of the score test

One of the common Heart disease symptom is chest pain. Almost 78% of the case, chest pain was the common symptom but some people will not experience chest pain and some have other symptoms. Using Logistic Regression risk factors are compared and the chances of affecting heart disease is predicted. In the above picture, the known output in the dataset may be above or below 95% but the predicted values will be above or below 95% by comparing the attributes.

V.CONCLUSION AND FURTHER WORK

In this paper, dataset incorporation, importing packages and visualization are performed in Colab notebook. Logical Regression is used to predict the risk factor of the heart disease. To have constant and feasible immunity every people should eat healthy and to stay fit.

REFERENCES:

1. D. Yach, C. Hawkes, C. L. Gould, and K. J. Hofman, "The global burden of chronic diseases: overcoming impediments to prevention and control," *The Journal of the American Medical Association*, vol. 291, no. 21, pp. 2616–2622, 2004.View at: Publisher Site | Google Scholar

2. J. Mackay and G. A. Mensah, *The Atlas of Heart Disease and Stroke*, WHO, 2004.

3. L. Ohno-Machado, P. Nadkarni, and K. Johnson, "Natural language processing: algorithms and tools to extract computable information from EHRs and from the biomedical literature," *Journal of the American Medical Informatics Association*, vol. 20, no. 5, p. 805, 2013.View at: Publisher Site | Google Scholar

4. J. Jonnagaddala, H.-J. Dai, P. Ray, and S.-T. Liaw, "Mining electronic health records to guide and support good clinical decision support systems," in *Improving Health Management through Clinical Decision Support Systems*, J. Moon and M. P. Galea, Eds., IGI-Global, 2015.View at: Google Scholar

5. Y.-C. Chang, H.-J. Dai, J. C.-Y. Wu, J.-M. Chen, R. T.-H. Tsai, and W.-L. Hsu, "TEMPTING system: a hybrid method of rule and machine learning for temporal relation extraction in patient discharge summaries," *Journal of Biomedical Informatics*, vol. 46, supplement, pp. S54–S62, 2013.View at: Publisher Site | Google Scholar

6. M. Fiszman, G. Rosemblat, C. B. Ahlers, and T. C. Rindflesch, "Identifying risk factors for metabolic syndrome in biomedical text," *AMIA Annual Symposium Proceedings*, vol. 2007, pp. 249–253, 2007.View at: Google Scholar

7. S. Goryachev, H. Kim, and Q. Zeng-Treitler, "Identification and extraction of family history information from clinical reports," *AMIA Annual Symposium Proceedings*, vol. 2008, pp. 247–251, 2008.View at: Google Scholar

8. J. Jonnagaddala, S. Liaw, P. Rayb, M. Kumarc, and H. Dai, "TMUNSW: identification of disorders and normalization to SNOMED-CT terminology in unstructured clinical notes," in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval '15)*, Association for Computational Linguistics, 2015.View at: Google Scholar

9. J. Jonnagaddala, S.-T. Liaw, P. Ray, M. Kumar, and H.-J. Dai, "HTNSystem: hypertension information extraction system for unstructured clinical notes," in *Technologies and Applications of Artificial Intelligence*, S.-M. Cheng and M.-Y. Day, Eds., vol. 8916 of *Lecture Notes in Computer Science*, pp. 219–227, Springer, 2014.View at: Publisher Site | Google Scholar

10. S. Kraus, C. Blake, and S. L. West, "Information extraction from medical notes," in *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics. Building Sustainable Health Systems*, IOS Press, 2007.View at: Google Scholar

11. G. K. Savova, J. J. Masanz, P. V. Ogren et al., "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, 2010.View at: Publisher Site | Google Scholar

12. H. Xu, S. P. Stenner, S. Doan, K. B. Johnson, L. R. Waitman, and J. C. Denny, "MedEx: a medication information extraction system for clinical narratives," *Journal of the American Medical Informatics Association*, vol. 17, no. 1, pp. 19–24, 2010.View at: Publisher Site | Google Scholar

13. Q. T. Zeng, S. Goryachev, S. Weiss, M. Sordo, S. N. Murphy, and R. Lazarus, "Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system," *BMC Medical Informatics and Decision Making*, vol. 6, article 30, 2006.View at: Publisher Site | Google Scholar

14. R. J. Byrd, S. R. Steinhubl, J. Sun, S. Ebadollahi, and W. F. Stewart, "Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records," *International Journal of Medical Informatics*, vol. 83, no. 12, pp. 983–992, 2014.View at: Publisher Site | Google Scholar

15. J. Jonnagaddala, S. T. Liaw, P. Ray, M. Kumar, N. W. Chang, and H. J. Dai, "Coronary artery disease risk assessment from unstructured electronic health records using text mining," *Journal of Biomedical Informatics*, In press.View at: Google Scholar

16. G. K. Savova, P. V. Ogren, P. H. Duffy, J. D. Buntrock, and C. G. Chute, "Mayo clinic NLP system for patient smoking status identification," *Journal of the American Medical Informatics Association*, vol. 15, no. 1, pp. 25–28, 2008.View at: Publisher Site | Google Scholar

17. N.-W. Chang, H.-J. Dai, J. Jonnagaddala, C.-W. Chen, and W.-L. Hsu, "A context-aware approach for progression tracking of medical concepts in electronic medical records," *Journal of Biomedical Informatics*, In press.View at: Google Scholar