International Journal of Preventive Medicine and Health (IJPMH)
ISSN: 2582-7588(Online), Volume-1 Issue-2, May 2021

Prediction of Heart Stroke using A Novel Framework – PySpark

CH Sai Harish, G Krishna Vamsi, G J P Akhil, J N V Hari Sravan, V Mounika Chowdary

Abstract: ***Heart diseases are one of the most challenging problems faced by the Health Care sectors all over the world. These diseases are very basic now a days. With the expanding count of deaths because of heart illnesses, the necessity to build up a system to foresee heart ailments precisely. The work in this paper focuses on finding the best Machine Learning algorithm for identification of heart diseases. Our study compares the precision of three well known classification algorithms, Decision Tree and Naïve Bayes, Random Forest for the prediction of heart disease by making the use of dataset provided by Kaggle. We utilized various characteristics which relate with this heart diseases well, to find the better algorithm for prediction. The result of this study indicates that the Random Forest algorithm is the most efficient algorithm for prediction of heart disease with accuracy score of 97.17%.***

Keywords: Machine Learning, Decision Tree, Naive Bayes, Random Forest, Machine Learning, Heart Disease Prediction, Kaggle.

I. INTRODUCTION

Heart disease or cardiovascular disease is a condition which involves the narrowing or the blockage of blood vessels in the heart which cause problems or failures in the human cardiovascular system. It causes many abnormal medical conditions like Hypertension, Cardiac Arrest, Arrhythmia, Stroke and Heart failure to name a few. It is one of the most challenging problem for the health care sector because of the complexity involved in the detection, diagnosis, and treatment of this condition.

CVD's cause 17.9 million deaths worldwide every year which is approximately equal to 31% of all deaths in the world according to WHO [1]. The occurrence heart disease in a person depends on a wide variety of factors like age, gender, type of the work, body mass index etc.

Manuscript received on March 31, 2021.
Revised Manuscript received on April 06, 2021.
Manuscript published on May 10, 2021.
* Correspondence Author
Mr. Chitluri Sai Harish*, B. Tech, Research Associate, Department of Computer Science and Engineering from KL University.
Mr. G gnana krishna vamsi, B. Tech, Research Associate, Department of Computer Science and Engineering from KL University.
Mr. G jaya phani akhil, B. Tech, Research Associate, Department of Computer Science and Engineering from KL University.
Mr. J n v hari sravan, B. Tech, Research Associate, Department of Computer Science and Engineering from KL University.
Ms. V mounika chowdary, Assistant Faculty and Project guide for the undergraduate Students, Department of Computer Science and Engineering from KL University.

Because of this, it may be a bit difficult to find the correct cause of the disease and also to predict whether the person is prone to the disease or not, based on the conditions mentioned above.Machine learning can be of great help in the prediction of heart disease when certain data about a person is given. The work proposed in this paper focus mainly on various data mining practices that are employed in heart disease prediction. We used Kaggle and PySpark for obtaining and analyzing the data respectively, which is helpful in increasing the accuracy of the machine learning algorithms.

SPARK FRAMEWORK:

Spark is an open-source cluster computing framework used to increase the speed of computing and data processing for huge datasets, which is commonly known as Big Data. It is a computational engine, fundamentally, that works with very large amount of data by processing them using bath and parallel system processing mechanisms. It is an extended version of Hadoop Map Reduce and is useful in a larger number of types of computations, which includes interactive queries and stream processing. SQL queries, Streaming data, Machine learning (ML), and Graph algorithms are some of it is application fields. APIs in Java, Scala, Python and R are also provided by it which helps users proficient in different coding languages. In-memory cluster computing, which is the primary feature of spark, increases the processing speed of an application.

PYSPARK:

It is the Python API to support Apache Spark. It is widely used by data scientists to work on Resilient Distributed Datasets (RDD). As Spark is written in SCALA and some programmers find it difficult to use it by coding in SCALA. PySpark helps programmers to use python instead of SCALA, and as python contains many different libraries, it is easier to work on huge datasets. PySpark links Python API to the spark core and initializes the Spark context.

Decision Tree Algorithm:

This is a type of predictive modeling algorithm employed mainly for statistics, data mining and for classification and regression problems in machine learning. It has a flowchart type structure containing internal nodes, leaf nodes and branches. The internal nodes, leaf nodes represent and branches represent test on features, class label, conjunctions of features which lead to the class labels respectively. The classification rules are represented by the path from the root to the leaf.

Naive Bayes Algorithm:

This is the simplest machine learning algorithm based on Bayes Theorem and is used as a probabilistic classifier in machine learning. It assumes that the occurrence of a certain feature is independent of the occurrence of other features. It learns the probability of objects with specific features belonging to a specific group.

Random Forest Algorithm:

It is a type of Classification Algorithm comprising of numerous choices of trees. It utilizes the bagging and feature randomness when fabricating every individual tree to attempt to make an uncorrelated collection of trees whose forecast by advisory group is more precise than that of any single tree.

II. RELATED WORK

Heart diseases are caused by a wide range of factors. Some people might not suffer from heart disease even though they have bad lifestyle and some people might suffer even though they have a good lifestyle. Identifying the relationship between those factors and predicting weather a person will get disease or not is a difficult task. Abbreviations and Acronyms. Therefore, Machine Learning methods were employed to help in this process.

- In [2], the researchers analyzed and compared three popular big data processing platforms which are Apache Hadoop, Apache Spark and Apache Flink for healthcare data analysis applications. They found that the Apache Flink permits clients to store data in memory and use that data on various occasions multiple times. It also provides a complex Fault Tolerance mechanism. Apache Hadoop environment is simple, has error detection and scalability management based on clusters. But it has a slow response time for complex analysis and flow processing applications. Apache Spark on the other hand can process big data sets in the memory with a great processing speed.
- In [3] the authors performed an analysis on Apache Spark. They explained its features, key components and abstractions. They showed the contents of Apache Spark which can be used for the design and implementation of big data algorithms and pipelines for machine learning and stream processing.
- In [4], the authors analyzed and surveyed different machine learning models such as SVM, KNN, DT etc.
- In [5] the authors used two algorithms viz Naive Bayes and Decision Tree algorithms to predict heart disease in humans. Their results showed that Naive Bayes algorithm had better accuracy on small datasets whereas Decision Tree algorithm performed better on large datasets.

III. METHODOLOGY

Our Proposed Methodology comprises of 5 Stage Process.

At First, we will stack the information from the Dataset Collected in Kaggle, later we will Preprocess the Data, Required Data Transformation Techniques and utilization of Various ML Algorithms and Resulting of different standard Metrics.

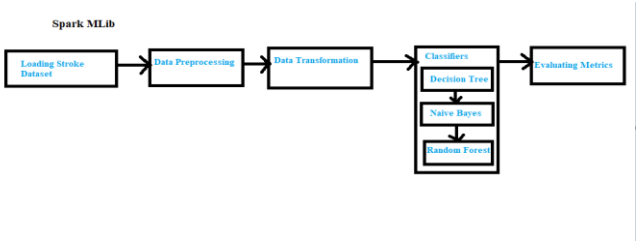

Fig 3.0 Process Flow Diagram

Our System has four stages in Heart Disease prediction. Each stage is explained below as follows:

1) Data collection and Analysis
2) Data pre-processing
3) Training and Testing the ML Models
4) Evaluating the ML Models

Stage 1: Data collection and Analysis:

- First, the data was acquired from the Kaggle repository dataset. It had data on the following indices:

1. Age
2. Gender
3. Hypertension
4. Heart Disease
5. Work Type
6. Marital Status
7. Residence type
8. Average Glucose Level
9. BMI
10. Smoking Status
11. Stroke

Table 1. Features name and description of stroke dataset

#num	Features	Description
1	Age	Age
2	Gender	Male and Female
3	Hypertension	Hypertension
4	Heart Disease	1 Has heart disease 0 Does not have heart disease
5	Ever_married	1 means Married 0 means Not married
6	Work_type	Children Private Never worked Govt job Self employed
6	Residence_type	Rural Urban
7	Avg_glucose_level	Average glucose level
8	bmi	Body mass index
10	smoking_status	Never smoked Formerly smoked

Fig 3.1 Dataset and attributes

- Next, number of persons with and without a heart disease was obtained.
- Finally, the and number of persons with and without a heart disease for different categories like age, gender and work type was obtained.

Stage 2: Data pre-processing:

- It was found that some categorical data was missing in the dataset.
- Some columns did not have the value of BMI while some did not have the value of the smoking status.
- The missing BMI values were filled with the mean BMI value of the data and the smoking status was filled with the string "No Info".
- The columns were indexed and encoded using String Indexer and One Hot Encoder respectively.
- Next, all the columns were combined into a single vector using Vector Assembler for the training of ML models.

Stage 3: Training and Testing the ML Models:

- There a complex bunch of stages, that are needed to be performed to process data. To wrap all of that Spark ML represents such a workflow as a Pipeline, which consists of a sequence of Pipeline Stages to be run in a specific order.
- The dataset is split into Training and Testing Dataset Using Random Splitter Function.
- Two models viz Decision Tree and Naive Bayes models are trained with the training data.
- The vector column previously generated is given as input for the model for training.

Stage 4: Evaluating the ML Models:

- The Models are evaluated using Multiclass Classification Evaluator.
- A confusion matrix has been used to assess the performance of the algorithms used.
- This matrix measures the performance of a model on a set of test data. Its accuracy, precision, recall, and f-measure are also measured
- It outputs two types of correct predictions and two types of incorrect predictions for the classifier which are True Positive, False Positive, True Negative and False Negative.

	Predicted Class 0	Predicted Class 1
Actual Class 0	TP	FN
Actual Class 1	FP	TN

Fig 3.2 Confusion Matrix

IV. CONCLUSION AND FUTURE SCOPE

We have been identified that the Random Forest algorithm is more efficient in the prediction of heart diseases. Its accuracy was found to be 97.17%. In the future, this work can be upgraded by building up a web application based on the Random Forest algorithm and using a bigger dataset when contrasted with the one utilized in this examination. This will help in giving better outcomes and help healthcare experts in the prediction of coronary illnesses adequately and productively.

V. RESULTS

The Following are the Results that we got from the evaluation of the two algorithms on the test data.

Algorithm	True Positive	False Positive	True Negative	False Negative
Decision Tree	1350	61	800	22
Naïve Bayes	1265	192	628	148
Random Forest	1380	45	790	18

Tab 4.1 Values Obtained for confusion matrix for different Algorithms.

Algorithm	Precision	Recall	F-Measure	Accuracy
Decision Tree	0.945	0.981	0.950	94.24%
Naïve Bayes	0.868	0.895	0.881	86.04%
Random Forest	0.968	0.987	0.967	97.17%

Tab 4.2 Standard Metrics of Various Algorithms

ACKNOWLEDGMENT

We express sincere thanks to our project supervisor **Ms. V Mounika Chowdary** for her novel association of ideas, encouragement, appreciation and intellectual zeal which motivated us to venture this project successfully. We express sincere gratitude to our project coordinator **Dr. C Karthikeya**n for his constant motivation provided in successful completion of our project. We express the sincere gratitude to our beloved Principal **Dr. K Subbarao** for motivating us towards our academic growth. It is magnificent pleasure for us to express gratitude to our honorable President Sri. Koneru Satyanarayana for giving the opportunity and platform with facilities in accomplishing the project. Finally, it is pleased to acknowledge the indebtedness to all those who devoted themselves directly or indirectly to make this project report successful.

REFERENCES

1. "Coronary Diseases", https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1
2. Elham Nazari, Mohammad Hasan Shahriari, Hamed Tabesh "Big Data Analysis in Healthcare: Apache Hadoop, Apache spark and Apache Flink", Frontiers in Health Informatics, September 2019.
3. Salman Salloum, Ruslan Dautov , Xiaojun Chen, Patrick Xiaogang Peng, Joshua Zhexue Huang "Big Data Analysis on Apache Spark", International Journal of Data Science and Analytics, October 2016.
4. V V Ramalingam ,Ayantan Dandapath, M Karthik Raja," Heart Disease Prediction using Machine Learning Techniques: A Survey", October 2018.
5. N. Rajesh, T. Maneesha , Shaik Hafeez, Hari Krishna, "Prediction of Heart Disease Using Machine Learning Algorithms", IJET, May 2018.
6. Al-Talqani, H.M., Dyslipidemia and Cataract in Adult Iraqi Patients. EC Ophthalmology, 2017. 5: p. 162-171.
7. McKinley, R., et al., Fully automated stroke tissue estimation using random forest classifiers (FASTER). Journal of Cerebral Blood Flow & Metabolism, 2017.
8. Jos Timanta Tarigan, C.L.G., Elviawaty Muisa Zamzami, A REVIEW ON APPLYING MACHINE LEARNING IN GAME INDUSTRY International Journal of Advanced Science and Technology, 2019-09-27 28(2).
9. Saiteja Myla, S.T.M., K Karthikeya ,Preetham.B , SK Hasane Ahammad, The Rise of "Big Data" in the field of Cloud Analytics. International Journal of Advanced Science and Technology, 2019. 28(8).
10. Ara, A. and A. Ara, Beyond Hadoop: The Paradigm Shift of Data from Stationary to Streaming Data for Data Analytics.
11. Hadoop, Apache Hadoop [cited 2019; Available from: https://hadoop.apache.org/.
12. Spark, A. Apache Spark. [cited 2019; Available from: https://spark.apache.org/.
13. Ahmed, H., Heart disease identification from patients' social posts, machine learning solution on Spark. Future Generation Computer Systems, 2019.
14. Healthcare dataset stroke data. [cited 2019; Available from: https://www.kaggle.com/asaumya/healthcare-dataset-stroke-data.
15. Shanthi, D., G. Sahoo, and N. Saravanan, Designing an artificial neural network model for the prediction of thrombo-embolic stroke. International Journals of Biometric and Bioinformatics (IJBB), 2009. 3(1): p. 10-18.
16. Kansadub, T., et al. Stroke risk prediction model based on demographic data. in 2015 8th Biomedical Engineering International Conference (BMEiCON). 2015. IEEE.
17. Sung, S.-F., et al., Developing a stroke severity index based on administrative data was feasible using data mining techniques. Journal of clinical epidemiology, 2015. 68(11): p. 1292-1300.
18. Linder, R., et al., Two models for outcome prediction. Methods of information in medicine, 2006. 45(05): p. 536- 540
19. Khosla, A., et al. An integrated machine learning approach to stroke prediction. in Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining 2010. ACM.
20. Adam, S.Y., A. Yousif, and M.B. Bashir, Classification of ischemic stroke using machine learning algorithms. Int J Comput Appl ,2016 149(10): p. 26-31.
21. Cheng, C.-A., Y.-C. Lin, and H.-W. Chiu. Prediction of the prognosis of ischemic stroke patients after intravenous thrombolysis using artificial neural networks. in ICIMTH ,2014.
22. Swethalakshmi, H., et al. Online handwritten character recognition of Devanagari and Telugu Characters using support vector machines. 2006.

AUTHORS PROFILE

Mr. Chitluri Sai Harish B. Tech, is a research associate at the department of Computer Science and Engineering from KL University. His areas of Interest are towards Data Science, Machine Learning and Big Data Analytics.

Mr. G gnana krishna vamsi, B. Tech, assistant researcher at the department of Computer Science and Engineering from KL University. His areas of Interest are towards Data Science, Machine Learning and Big Data Analytics.

Mr. G jaya phani akhil B. Tech, assistant researcher at the department of Computer Science and Engineering from KL University. His areas of Interest are towards Data Science, Machine Learning and Big Data Analytics.

Mr. J n v hari sravan, B. Tech, assistant researcher at the department of Computer Science and Engineering from KL University. His areas of Interest are towards Data Science, Machine Learning and Big Data Analytics.

Ms. V mounika chowdary, is a Faculty researcher at the department of Computer Science and Engineering from KL University. Currently she is working as Assistant Faculty and Project guide for the undergraduate students in helping out with her ideas towards Data Science, Machine Learning and Big Data Analytics.