LITERATURE SURVEY:

1. Multi class Prediction of Heart Disease Patients Using Big Data Analytics - Sarita Mishra, Manjusha Pandey, Siddharth Swarup Rautaray & Sabyasachi Chakraborty - 2022

The rapidly growing rate of illness and death is the result of many diseases. Another major factor is cardiovascular disease (CVD) due to heart failure. According to statistics from around the world, the highest rate of natural death is caused by heart problems. The number of deaths resulting from this can be controlled by the early detection of heart disease chances in a person. Big data and several machine learning technologies have made it possible to discover the chances of a cardiac issue in a person in much advance. Many data scientists have successfully exploited the big data available for heart disease patients and have developed prediction models using different algorithms that are non-invasive, accurate, and appear to be very effective in analyzing patients' characteristics and detecting the presence or absence of heart disease in them. However, to provide appropriate preventive measures and appropriate treatment to patients, it is not enough to detect the presence of CVD, but the degree of impact the disease has left on a person needs to be measured. In this paper, we have compared the performance of five different machine-based algorithms (Logistic Regression, Support Vector Machine, Random Forest, KNN, and Naïve Bayes) which are used to classify the cardiovascular disease into five different classes. 0–4) with the increasing value from 0. These algorithms are used in their most common ways and in the One-vs-All method with the best performance in the latest scenario. The results of this study showed that the KNN algorithm provided 99.56% best predictive accuracy with a combination of One-vs-all and Principal Component Analysis strategies that surpassed all other algorithms.

The main idea behind this is the model gives different accuracy levels for different K values and the best K value can be identified using an analyzing error rate or accuracy rate of the model. However, the size and the features of the dataset makes a big impact on the model to come up with a good accuracy rate. Nevertheless, the KNN algorithm does not work well with large datasets and does not work well with high dimensional data because it is hard to calculate the distance between each data point which is a drawback of the algorithm.

2. A Systematic Framework for Heart Disease Prediction Using Big Data Analytics-T Poongodi, R Indrakumari, S Janarthanan, P Suresh-Internet of Things, Artificial Intelligence and Blockchain Technology- 2021

Big data in deep insight derives from heterogeneous, longitudinal, complex, voluminous, and noisy data. The significant challenges in big data lie in searching, capturing, storing, analyzing, and sharing the data. Big data analytics is emerging as a promising technology in harnessing a massive amount of data which surpasses the processing capability of conventional systems. Big data is commonly characterized by volume, velocity, value, veracity, variety, and variability. With the big data, progression in healthcare communities leads to precise analysis of physiological or digitized clinical data benefits prior to detection of diseases, patient care, and healthcare community services. Heart disease is considered as the most life-threatening and deadliest disease that humans face across the world. The vital task in heart disease management is in processing extremely larger datasets and inferring knowledge to predict, prevent, and treat such chronic diseases. The idea is to identify the potential of big data analytics in predicting heart diseases and providing appropriate medicines and treatment for the heart patients. In the healthcare industry, the knowledge that is inferred can be utilized for predicting heart diseases in the early stages. Big data is generated from the user-generated content, mobile transactions, Internet clicks, social media, and genomics data especially created through corporate transactions or sensor networks. Moreover, the advances in using genomic data assist in sharing clinical data, drug discovery, EHR processing, patient registries, telemedicine, etc. K-means clustering algorithm is followed, and the data is visualized using Tableau Software. The diagnosis clinical parameters for heart disease prediction are age, gender, weight, chest pain, resting BP, resting ECG, cholesterol, etc. The chapter highlights the significant role of big data analytics predominantly in the healthcare industry for predicting heart diseases. Some of the challenges for implementing big data analytics in healthcare are discussed. The future directions in associating big data in healthcare for predicting heart diseases and personalizing medicine are also being investigated.

Due to lack of effective data governance procedures, capturing data is one of the biggest obstacles for healthcare organizations. To use data more efficient, it must be clean, precise, correctly formatted so that it can be used across various healthcare systems.

Most patient records are kept for fast and easy access in a centralized database these days, but the real problem lies when this information that needs to be shared with outside healthcare professionals. For most healthcare providers, data security is one of the top issues with constant hacking and security violations that need to be handled on a continuous basis.When dealing with highly sensitive data and even patient data, which is important, the healthcare industry must be very cautious. Not only can leakage of details prove costly to healthcare companies, but it is also unethical to disclose it without prior authorization.

3. Heart disease prediction using machine learning algorithms- Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain and Preeti Nagrath-2021

Day by day the cases of heart diseases are increasing at a rapid rate and it’s very Important and concerning to predict any such diseases beforehand. This diagnosis is a difficult task i.e. it should be performed precisely and efficiently. The research paper mainly focuses on which patient is more likely to have a heart disease based on various medical attributes. We prepared a heart disease prediction system to predict whether the patient is likely to be diagnosed with a heart disease or not using the medical history of the patient. We used different algorithms of machine learning such as logistic regression and KNN to predict and classify the patient with heart disease. A quite Helpful approach was used to regulate how the model can be used to improve the accuracy of prediction of Heart Attack in any individual. The strength of the proposed model was quiet satisfying and was able to predict evidence of having a heart disease in a particular individual by using KNN and Logistic Regression which showed a good accuracy in comparison to the previously used classifier such as naive bayes etc. So a quiet significant amount of pressure has been lift off by using the given model in finding the probability of the classifier to correctly and accurately identify the heart disease. The Given heart disease prediction system enhances medical care and reduces the cost. This project gives us significant knowledge that can help us predict the patients with heart disease It is implemented on the .pynb format.

Logistic Regression is a statistical analysis model that attempts to predict precise probabilistic outcomes based on independent features. On high dimensional datasets, this may lead to the model being over-fit on the training set, which means overstating the accuracy of predictions on the training set and thus the model may not be able to predict accurate results on the test set. This usually happens in the case when the model is trained on little training data with lots of features. So on high dimensional datasets, Regularization techniques should be considered to avoid over-fitting (but this makes the model complex). Very high regularization factors may even lead to the model being under-fit on the training data. Non linear problems can't be solved with logistic regression since it has a linear decision surface. Linearly separable data is rarely found in real world scenarios. So the transformation of non linear features is required which can be done by increasing the number of features such that the data becomes linearly separable in higher dimensions.

4. Accurate prediction of heart disease based on bio system using regressive learning based neural network classifier -A. Sheryl Oliver, Kavithaa Ganesan, S. A. Yuvaraj, T. Jayasankar, Mohamed Yacin Sikkandar, N. B. Prakash - 2021

Heart disease diagnosis is a very hard task in the medical field, so the mortality rate is increased every day. Also, the diagnosing process is implemented in recent times to predict heart disease. The method of diagnosing a disease in the medical field can be regarded not only as a new unknown situation to obtain clinical data and data collected from clinical experience, but also as a decision-making process as well as a doctor's diagnosis. The detection of heart abnormalities mainly depends on the examination of the ECG signal at the appropriate sampling period. The data is trained and tested must include more data to get the data as features. These properties are an accurate measure of the diagnosis of heart disease. The conventional system is having some problems like processing time is high, and it gives low accuracy, so the proposed Regressive Learning-Based Neural Network Classifier (RLNNC) system is implemented. The proposed system RLNNC presents a fully automated algorithm for the classification of heart disease, based on the Regressive Learning-Based Neural Network Classifier (RLNNC) and automated initial seed detection. With the advancement of machine learning and information technology, the development of an automated system. This can be predicted the same on this basis for patients with heart disease, and the drug occurs for the benefit of detecting and analyzing the heart disease. Analysis has shown that the proposed Regressive Learning-Based Neural Network Classifier (RLNNC) based techniques promote greater efficiency and higher accuracy than traditional methods.

Artificial neural networks require processors with parallel processing power, in accordance with their structure. For this reason, the realization of the equipment is dependent. Unexplained behavior of the network: This is the most important problem of ANN. When ANN produces a probing solution, it does not give a clue as to why and how. This reduces trust in the network. There is no specific rule for determining the structure of artificial neural networks. Appropriate network structure is achieved through experience and trial and error. ANNs can work with numerical information. Problems have to be translated into numerical values before being introduced to ANN. The display mechanism to be determined here will directly influence the performance of the network . This depends on the user's ability. The network is reduced to a certain value of the error on the sample means that the training has been completed. This value does not give us optimum results.

5. Artificial Intelligence-Based Ensemble Model for Rapid Prediction of Heart Disease -Navya Harika, Sita Rama Swamy & Nilima- 2021

Heart disease is the leading cause of mortality among men and women. Accurate and rapid diagnosis of heart disease will assist in saving many lives. To develop a novel ensemble framework based on heterogeneous classifiers namely support vector machine (SVM), Naïve Bayes (NB), and artificial neural networks (ANN) for rapid prediction of heart disease. The present study also verifies the most accurate algorithm among all three. Data are collected from the UCI machine learning repository. After pre-processing, the data were divided into training and test data in a ratio of 80:20. Using the training data, the three contributing algorithms were trained by providing heart disease status. The algorithms were tested with the unseen data instances and hence evaluated for accuracy. The ensemble technique uses the results from individual classifiers and yields a result based on majority voting method. The ensemble model was observed to predict heart disease with an accuracy of 87.05% followed by ANN (84.74%), NB (81.35%) and SVM (79.66%). Among the individual classifiers, ANN had the least miss-classification rate and performed best in terms of all other model diagnostics. The use of the proposed ensemble classifier is recommended to predict the heart condition to have better accuracy and least miss-classification.

Naive Bayes assumes that all predictors (or features) are independent, rarely happening in real life. This limits the applicability of this algorithm in real-world use cases. This algorithm faces the ‘zero-frequency problem’ where it assigns zero probability to a categorical variable whose category in the test data set wasn’t available in the training dataset. It would be best if we used a smoothing technique to overcome this issue. Its estimations can be wrong in some cases, so you shouldn’t take its probability outputs very seriously.

6. Heart disease prediction using machine learning techniques: A systematic review- Kiranjit Kaur , Munish Saini-2020

The key task within the healthcare field is usually the diagnosis of the disease. In case, a disease is actually diagnosed at earlier stage, then many lives might be rescued. Machine learning classification techniques can considerably help the healthcare field just by offering a precise and easy diagnosis of various diseases. Consequently, saving time both formed ical professionals and patients. As heart disease is usually the most recognized killer in the present day, it might be one of the most challenging diseases to diagnose. In this paper, we provide a survey of the various machine learning classification techniques that have been proposed to assist the healthcare professionals in diagnosing the cardiovascular disease. We started by giving the overview of various machine learning techniques along with describing brief definitions of the most commonly used classification techniques to diagnose heart disease. Then, we review representable research works on employing machine learning classification techniques in this field. Furthermore, a detailed comparison table of the surveyed papers is actually presented.

The main drawback is that some datasets will have few to no values in the other classes, with the majority of the data values falling into only one or two classifications. The main drawback of the quantile classification process is that features assigned to the same class might have drastically different values, especially if the data are not distributed uniformly over the class's range. Natural breaks approach might have the drawback of producing classes with wildly disparate number ranges. Another drawback is that because the class ranges are so unique to each dataset, it might be challenging to compare two or more maps made using the natural breaks classification approach.

7. Prediction of heart disease and classifiers' sensitivity analysis-Khaled Mohamad Almustafa-02 July 2020

Heart disease (HD) is one of the most common diseases nowadays, and an early diagnosis of such a disease is a crucial task for many health care providers to prevent their patients for such a disease and to save lives. In this paper, a comparative analysis of different classifiers was performed for the classification of the Heart Disease dataset in order to correctly classify and or predict HD cases with minimal attributes. The set contains 76 attributes including the class attribute, for 1025 patients collected from Cleveland, Hungary, Switzerland, and Long Beach, but in this paper, only a subset of 14 attributes are used, and each attribute has a given set value. The algorithms used K- Nearest Neighbor (K-NN), Naive Bayes, Decision tree J48, JRip, SVM, Ada-boost, Stochastic Gradient Decent (SGD) and Decision Table (DT) classifiers to show the performance of the selected classifications algorithms to best classify, and or predict, the HD cases.

8. A new Internet of Things architecture for real-time prediction of various diseases using machine learning on big data environment - Abderrahmane Ed-daoudy – 2019

A number of technologies enabled by Internet of Thing (IoT) have been used for the prevention of various chronic diseases, continuous and real-time tracking system is a particularly important one. Wearable medical devices with sensor, health cloud and mobile applications have continuously generating a huge amount of data which is often called as streaming big data. Due to the higher speed of the data generation, it is difficult to collect, process and analyze such massive data in real-time in order to perform real-time actions in case of emergencies and extracting hidden value. using traditional methods which are limited and time-consuming. Therefore, there is a significant need to real-time big data stream processing to ensure an effective and scalable solution. In order to overcome this issue, this work proposes a new architecture for real-time health status prediction and analytics system using big data technologies. The system focus on applying distributed machine learning model on streaming health data events ingested to Spark streaming through Kafka topics. Firstly, we transform the standard decision tree (DT) (C4.5) algorithm into a parallel, distributed, scalable and fast DT using Spark instead of Hadoop MapReduce which becomes limited for real-time computing. Secondly, this model is applied to streaming data coming from distributed sources of various diseases to predict health status. Based on several input attributes, the system predicts health status, send an alert message to care providers and store the details in a distributed database to perform health data analytics and stream reporting. We measure the performance of Spark DT against traditional machine learning tools including Weka. Finally, performance evaluation parameters such as throughput and execution time are calculated to show the effectiveness of the proposed architecture. The experimental results show that the proposed system is able to effectively process and predict real-time and massive amount of medical data enabled by IoT from distributed and various diseases.

9. Prediction of Coronary Heart Disease using Machine Learning: An Experimental Analysis - Amanda H. Gonsalves – July 2019

The field of medical analysis is often referred to be a valuable source of rich information. Coronary Heart Disease (CHD) is one of the major causes of death all around the world therefore early detection of CHD can help reduce these rates. The challenge lies in the complexity of the data and correlations when it comes to prediction using conventional techniques. The aim of this research is to use the historical medical data to predict CHD using Machine Learning (ML) technology. The scope of this research is limited to using three supervised learning techniques namely Naïve Bayes (NB), Support Vector Machine (SVM) and Decision Tree (DT), to discover correlations in CHD data that might help improving the prediction rate. Using the South African Heart Disease dataset of 462 instances, intelligent models are derived by the considered ML techniques using 10-fold cross validation. Empirical results using different performance evaluation measures report that probabilistic models derived by NB are promising in detecting CHD.

10. Predictive System: Comparison of Classification Techniques for Effective Prediction of Heart Disease - Debjani Panda & Satya Ranjan Dash– 27 September 2019

Today’s world is challenging to most of its people with major concerns for keeping up a good health. Among these challenges, one of the most haunting ones is heart disease. Worldwide, the maximum number of deaths is related to heart diseases. Most of the affected people suffering from heart-related diseases are unaware of their health conditions, and cases are reported at a very later stage, which becomes challenging for doctors to advise them proper treatment and medication with lifestyle changes. This research work aims in comparing classification techniques in finding out which is the most efficient one to predict the disease in less time. Mining important factors and analyzing the relativity between them help in predicting if the patient is having heart disease. The classification techniques used are SVM, Decision Tree, Naïve Bayes, KNN, Random Forest, Ensemble Classification (Extra Trees) and Logistic Regression.

10. Learning Classifier Methods in Prediction of Heart Disease -D Parvathinathan - 2018

Cardiovascular Disease (CVD) refers to any condition involving narrow or blocked blood vessels, which can lead to a heart attack, chest pain or stroke. CVD is a chronic illness and the leading cause of death for both men and women. CVD can attack a person instantly resulting in high healthcare costs and, in some cases, can result in death. A serious and important challenge facing medical practitioners is the ability to accurately diagnose patients with CVD early on. In recent years, medical practitioners have sought the help of computer scientists in order to apply advanced data mining techniques, which can facilitate decision support and help accurately diagnose CVD soon. In this thesis, three data mining techniques are evaluated for their accuracy in predicting CVD. The techniques implemented and analyzed are Logistical Regression, Naive Bayes and Artificial Neural Networks using Multi-Layer Perceptron (MLP). Results show that Logistic regression predicted the presence of CVD with an accuracy of 88.6%, while Naive Bayes predicted CVD with an accuracy rate of 83% and Neural Networks with an accuracy of 80%.