

PROFESSIONAL TRAINING REPORT

At

Sathyabama Institute of Science and Technology (Deemed to be University)

Submitted in partial fulfillment of the
requirements for the award of
Bachelor of Engineering Degree in Computer Science and Engineering

by

D. Satvik kumar (38110114)

B. Shameer (38110083)



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with Grade “A” by NAAC

JEPPIAAR NAGAR, RAJIV GANDHI SALAI

CHENNAI – 600119, TAMILNADU

MARCH - 2022



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)



Accredited with Grade “A” by NAAC

(Established under Section 3 of UGC Act, 1956)

JEPPIAAR NAGAR, RAJIV GANDHI SALAI

CHENNAI– 600119

www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **D.Satvik Kumar (Reg. No: 38110114) & B.Shameer (Reg. No: 38110083)** who carried out the project entitled “**Heart Disease Identification Method using Machine Learning Classification**” under my supervision from ----- 2021 to ----- 2022.

Internal Guide

Dr. T. Judgi M.E.,Ph.D.,

Head of the Department

Dr. L. Lakshmanan M.E., Ph.D.,

Dr. S. Vigneshwari M.E., Ph.D.,

Submitted for Viva voce Examination held on

InternalExaminer

ExternalExaminer

DECLARATION

I, D.Satvik Kumar & B.Shameer hereby declare that the project report entitled **Heart Disease Identification Method using Machine Learning Classification** done by me under the guidance of **Dr. Judgi.T M.E., Ph.D.**, is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering Degree in Computer Science and Engineering.

DATE: --/--/----

PLACE: Chennai

D. Satvik Kumar

&

B. Shameer

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. Sasikala M.E., Ph.D, Dean, School of Computing, Dr. S. Vigneshwari, M.E., Ph.D. and Dr. L. Lakshmanan, M.E., Ph.D., Heads of the Department of Computer Science and Engineering** for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. T.Judgi M.E, Ph.D.**, for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

Heart Disease Identification Method Using Machine Learning Classification

ABSTRACT:

Heart Disease prediction is one of the most complicated tasks in medical field. Data science plays a crucial role in processing huge amount of data in the field of healthcare. As heart disease prediction is a complex task, there is a need to automate the prediction process to avoid risks associated with it and alert the patient well in advance. To predict the Heart-Disease and patient's risk level, K-nearest neighbors (KNN), Logistic Regression (LR) and Random Forest (RF) using Machine learning Classification. Furthermore, the leave one subject out cross-validation method has been used for learning the best practices of model assessment and for hyper parameter tuning. The performance measuring metrics are used for assessment of the performances of the classifiers. The performances of the classifiers have been checked on the selected features as selected by features selection algorithms.

Table of contents

Chapter No :	Title	Page No
	Abstract	5
1.	Introduction	7
2.	Literature Review	7-20
	2.1 Heart disease identification method using Using machine learning classification in E-Healthcare	
	2.2 early and accurate detection and diagnosis Using intelligent computational model	
	2.3 Detection of cardiovascular disease using Machine learning classification models	
	2.4 Heart disease identification method using Machine learning classification in E-Healthcare	
3.	Existing System	20-22
	I. Proposed System	
	II. Hardware requirements	
	III. Software requirements	
4.	Machine Learning	22-34
	I. 7 Steps of Machine Learning	
	II. Understanding the ML workflow	
5.	System Design	34-50
	I. Module Names	
	II. Algorithm Used	
6.	Conclusion	50
	Reference Papers	51

Chapter 1

INTRODUCTION

Heart disease (HD) is the critical health issue and numerous people have been suffered by this disease around the world .The HD occurs with common symptoms of breath shortness, physical body weakness and, feet are swollen. Researchers try to come across an efficient technique for the detection of heart disease, as the current diagnosis techniques of heart disease are not much effective in early time identification due to several reasons, such as accuracy and execution time. The diagnosis and treatment of heart disease is extremely difficult when modern technology and medical experts are not available. The effective diagnosis and proper treatment can save the lives of many people. According to the European Society of Cardiology, 26 million approximately people of HD were diagnosed and diagnosed 3.6 million annually. Most of the people in the United States are suffering from heart disease Diagnosis of HD is traditionally done by the analysis of the medical history of the patient, physical examination report and analysis of concerned symptoms by a physician. But the results obtained from this diagnosis method are not accurate in identifying the patient of HD. Moreover, it is expensive and computationally difficult to analyze. Thus, to develop a non-invasive diagnosis system based on classifiers of machine learning to resolve these issues. Expert decision system based on machine learning classifiers and the application of artificial fuzzy logic is effectively diagnosis the HD as a result, the ratio of death decrease.

OBJECTIVES OF THE PROJECT:

- To measured with the help of graphical notations such as pie chart, bar chart and line chart. The data can be given in a dynamical data.
- The main objective of this research is to develop an Intelligent System using machine learning technique, namely, Naive Bayes, KNN, Random forest Decision tree.
- It is implemented as web based application in this user answers the predefined questions

Chapter 2

LITERATURE REVIEW

2.1 Heart Disease Identification Method using Machine Learning Classification in E-Healthcare

Abstract:

Coronary artery heart Disease (CAD) is caused by atherosclerosis in coronary arteries and results in cardiac arrest and heart attack. For diagnosis of CAD, angiography is used which is a costly time consuming and highly technical invasive method. Researchers are therefore, prompted for alternative methods such as machine learning algorithms that could use non-invasive clinical data for the heart Disease diagnosis and assessing its severity. In this study, we present a novel hybrid method for CAD diagnosis, including risk factor identification using correlation based feature subset (CFS) selection with particle swarm optimization (PSO) search method and K-Means clustering algorithms. Supervised learning algorithms such as multi-layer perceptron (MLP), multinomial logistic regression (MLR), fuzzy unordered rule induction algorithm (FURIA) and C4.5 are then used to model CAD cases. We tested this approach on clinical data consisting of 26 features and 335 instances collected at the Department of Cardiology, Indira Gandhi Medical College, and Shimla, India. MLR achieves highest prediction accuracy of 88.4 %. We tested this approach on benchmarked Cleveland heart Disease data as well. In this case also, MLR, outperforms other techniques. Proposed hybridized model improves the accuracy of classification algorithms from 8.3 % to 11.4 % for the Cleveland data. The proposed method is, therefore, a promising tool for identification of CAD improved prediction accuracy.

Introduction:

Data Mining is an analytic process designed to explore data (usually large amounts of data - typically business or market related) in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. Machine learning (ML) that improve naturally through experience. It is viewed as a subset of man-made brainpower. AI calculations construct a model dependent on example information, known as "preparing information", to settle on expectations or choices without being unequivocally customized to do as such. Illnesses under the coronary illness umbrella incorporate vein sicknesses, for example, coronary conduit infection; heart musicality issues (arrhythmia's); and heart abandons you're brought into the world with (inherent heart deserts), among others. The expression "coronary illness" is regularly utilized reciprocally with the expression "cardiovascular infection." Prescient demonstrating utilizes measurements to foresee outcomes.[1] Most frequently the occasion one needs to anticipate is later on, yet prescient displaying can be applied to an obscure occasion, paying little mind to happened.

Existing System

In Existing medical systems, including hospital management systems and decision making systems, focus on collecting and mining the entire medical data. The entire patient records are loaded and all factors are considered. The C4.5 model that we previously developed. On the basis of this new measure, a data mining algorithm was developed to mine the causal relationship between drugs and their associated with risk for CAD heart Disease.

Drawbacks

Existing systems have failed to utilize and understand the importance of misdiagnosis. A very important attribute which interconnects and addresses all these issues.

It varies from patient's medical history, climatic conditions, neighborhood, and various other factors.

Proposed system

In the proposed work user will search for the heart Disease diagnosis (heart Disease and treatment related information) by giving symptoms as a query in the search engine. These symptoms are preprocessed to make the further process easier to find the symptoms keyword which helps to identify the heart Disease quickly. CFS+PSO are a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification. This feature has been identified as the most suitable for the present system.

Advantage

It is easy to extract signatures from individual data instances, as their structures. Just collect the symptoms that enough to scaling samples.

Can easily predict the heart Disease level and severity easily using range level of queries.

The probability of vocabulary gap between diverse health seekers makes the data more consistent compared to other formats of health data.

Materials and Method

Data Visualization and Pre-Processing

The Wisconsin Prognostic Cleave Land Train Dataset is downloaded from the UCI Machine Learning Repository website and saved as a text file. This file is then imported into Excel spreadsheet and the values are saved with the corresponding attributes as column headers. The missing values are replaced with appropriate values.

Dimensionality Reduction (CFS+PSO)

Feature selection, by identifying the most salient features for learning, focuses a learning algorithm on those aspects of the data most useful for analysis and future prediction. Feature selection is a process of selecting a subset of relevant features from a large number of original features to achieve similar or better classification performance and improve the computation efficiency

Model for Cad Identification

Multinomial logistic regression model (MLR): Multinomial logistic regression is known by a variety of other names, including polytomous LR, multiclass LR, softmax regression, multinomial logit, maximum entropy (MaxEnt) classifier, and conditional maximum entropy model.

Risk Prediction

It provides reproducible and objective diagnosis, and hence can be a valuable adjunct tool in clinical practices. Then the data are

clustered using MLP, MLG, FURIA and C4.5 using all the features of CAD data. Risk forecast instruments are created to recognize patients in danger and to encourage doctor dynamic.

Classification Algorithms

Then the data are clustered using MLP, MLG, FURIA and C4.5 using all the features of CAD. In bunching, the thought isn't to foresee the objective class as in order, it's additionally attempting to assemble the comparative sort of things by thinking about the most fulfilled condition, all the things in a similar gathering should be comparable and ought not to be distinctive gathering things to not be comparative. Experiment results demonstrate the superiority of the proposed method with regard to prediction accuracy of CAD with the features selected by CFS & PSO, we need only a few clinical data to apply this model.

Conclusion

Clinical finding is a significant region of exploration which assists with recognizing the event of a coronary illness. The framework, utilizing different methods referenced, will thus uncover the root coronary illness alongside the arrangement of most plausible heart Diseases which have comparative side effects. The information base utilized is a portrayal data set so to decrease the dataset tokenization, separating and stemming is finished. The venture presents a novel mixture model to recognize and affirm CAD cases requiring little to no effort by utilizing clinical information that can be effectively gathered at clinics. Intricacy of framework is diminished by decreasing the dimensionality of the informational collection with PSO. It gives reproducible and target finding, and subsequently can be a significant extra device in clinical practices. Results are equivalently, encouraging and along these lines the proposed half and half technique will be useful in coronary illness diagnostics. Trial results exhibit the predominance of the proposed half breed technique concerning forecast precision of CAD.

2.2 Early and accurate detection and diagnosis of heart disease using intelligent computational model

Heart disease is a fatal human disease, rapidly increases globally in both developed and undeveloped countries and consequently, causes death. Normally, in this disease, the heart fails to supply a sufficient amount of blood to other parts of the body in order to accomplish their normal functionalities. Early and on-time diagnosing of this problem is very essential for preventing patients from more damage and saving their lives. Among the conventional invasive-based techniques, angiography is considered to be the most well-known technique for diagnosing heart problems but it has some limitations. On the other hand, the non-invasive based methods, like intelligent learning based computational techniques are found more upright and effectual for the heart disease diagnosis. Here, an intelligent computational predictive system is introduced for the identification and diagnosis of cardiac disease. In this study, various machine learning classification algorithms are investigated. In order to remove irrelevant and noisy data from extracted feature space, four distinct feature selection

algorithms are applied and the results of each feature selection algorithm along with classifiers are analyzed. Several performance metrics namely: accuracy, sensitivity, specificity, AUC, F1-score, MCC, and ROC curve are used to observe the effectiveness and strength of the developed model. The classification rates of the developed system are examined on both full and optimal feature spaces, consequently, the performance of the developed model is boosted in case of high varied optimal feature space. In addition, P-value and Chi-square are also computed for the ET classifier along with

each feature selection technique. It is anticipated that the proposed system will be useful and helpful for the physician to diagnose heart disease accurately and effectively.

Heart disease is considered one of the most perilous and life snatching chronic diseases all over the world. In heart disease, normally the heart fails to supply sufficient blood to other parts of the body to accomplish their normal functionality. Heart failure occurs due to blockage and narrowing of coronary arteries. Coronary arteries are responsible for the supply of blood to the heart itself. A recent survey reveals that the United States is the most effected country by heart disease where the ratio of heart disease patients is very high. The most common symptoms of heart disease include physical body weakness, shortness of breath, feet swollen, and weariness

with associated signs, etc. The risk of heart disease may be increased by the lifestyle of a person like smoking, unhealthy diet, high cholesterol level, high blood pressure, deficiency of exercise and fitness, etc. Heart disease has several types in which coronary artery disease (CAD) is the common one that can lead to chest pain, stroke, and heart attack. The other types of heart disease include heart rhythm problems, congestive heart failure, congenital heart disease (birth time heart disease), and cardiovascular disease (CVD). Initially, traditional investigation techniques were used for the identification of heart disease, however, they were found complex. Owing to the non-availability of medical diagnosing tools and medical experts specifically in undeveloped countries, diagnosis and cure of heart disease are very complex. However, the precise and appropriate diagnosis of heart disease is very imperative to prevent the patient from more damage. Heart disease is a fatal disease that rapidly increases in both economically developed and undeveloped countries. According to a report generated by the World Health Organization (WHO), an average of 17.90 million humans died from CVD in 2016. This amount represents approximately 30% of all global deaths. According to a report, 0.2 million people die from heart disease

OPEN

1 Department of Computer Science, Abdul Wali Khan University Mardan, Mardan 23200, KP, Pakistan. 2 Department

of Electronic and Information Engineering, Jeonbuk National University, Jeonju 54896, South Korea. *email:

m.hayat@awkum.edu.pk; kitchong@jbnu.ac.kr

Scientific Reports | (2020) 10:19747 |

<https://doi.org/10.1038/s41598-020-76635-9>

www.nature.com/scientificreports/

annually in Pakistan. Every year, the number of victimizing people is rapidly increasing. European Society of Cardiology (ESC) has published a report in which 26.5 million adults were identified having heart disease and

3.8 million were identified each year. About 50–55% of heart disease patients die within the initial 1–3 years, and the cost of heart disease treatment is about 4% of the

overall healthcare annual budget. Conventional invasive-based methods used for the diagnosis of heart disease which were based on the medical history of a patient, physical test results, and investigation of related symptoms by the doctors. Among the conventional methods, angiography is considered one of the most precise technique for the identification of heart problems. Conversely, angiography has some drawbacks like high cost, various side effects, and strong technological knowledge. Conventional methods often lead to imprecise diagnosis and take more time due to human mistakes. In addition, it is a very expensive and computational intensive approach for the diagnosis of disease and takes time in assessment. To overcome the issues in conventional invasive-based methods for the identification of heart disease, researchers attempted to develop different non-invasive smart healthcare systems based on predictive machine learning techniques namely: Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naïve Bayes (NB), and Decision Tree (DT), etc.. As a result, the death ratio of heart disease patients has been decreased. In literature, the Cleveland heart disease dataset is extensively utilized by the researchers. In this regard, Robert et al. have used a logistic regression classification algorithm for heart disease detection and obtained an accuracy of 77.1%. Similarly, Wankhade et al. Have used a multi-layer perceptron (MLP) classifier for heart disease diagnosis and attained accuracy of 80%. Likewise, Allahverdi et al. Have developed a heart disease classification system in which they integrated neural networks with an artificial neural network and attained an accuracy of 82.4%. In a sequel, Awang et al. have used NB and DT for the diagnosis and prediction of heart disease and achieved reasonable results in terms of accuracy. They achieved an accuracy of 82.7% with NB and 80.4% with DT. Oyedodun and Olaniye have proposed a three-phase system for the prediction of heart disease using ANN. Das and Turkoglu have proposed an ANN ensemble-based predictive model for the prediction of heart disease. Similarly, Paul and Robin have used the adaptive fuzzy ensemble method for the prediction of heart disease. Likewise, Tomov et al. Have introduced a deep neural network for heart disease prediction and his proposed model performed well and produced good outcomes. Further, Manogaran and Varatharajan have introduced the concept of a hybrid recommendation system for diagnosing heart disease and their model has given considerable results. Alizadehsani et al. Have developed a non-invasive based model for the prediction of coronary artery disease and showed some good results regarding the accuracy and other performance assessment metrics. Amin et al. Have proposed a framework of a hybrid system for the identification of cardiac disease, using machine learning, and attained an accuracy of 86.0%. Similarly, Mohan et al. Have proposed another intelligent system that integrates RF with a linear model for the prediction of heart disease and achieved the classification accuracy of 88.7%. Likewise, Liaqat et al. Have developed an expert system that uses stacked SVM for the prediction of heart disease and obtained 91.11% classification accuracy on selected features. The contribution of the current work is to introduce an intelligent medical decision system for the diagnosis of heart disease based on contemporary machine learning algorithms. In this study, 10 different nature of machine learning classification algorithms such as Logistic Regression (LR), Decision Tree (DT), Naïve Bayes (NB), Random Forest (RF), Artificial Neural Network (ANN), etc. are implemented in order to select the best model for timely and accurate detection of heart disease at an early stage. Four feature selection algorithms, Fast Correlation-Based Filter Solution (FCBF), minimal redundancy maximal relevance (mRMR), Least Absolute Shrinkage and Selection Operator (LASSO), and Relief have been used for selecting

the vital and more correlated features that have truly reflect the motif of the desired target. Our developed system has been trained and tested on the Cleveland (S1) and Hungarian (S2) heart disease datasets which are available online on the UCI machine learning repository. All the processing and computations were performed using Anaconda IDE. Python has been used as a tool for implementing all the classifiers. The main packages and libraries used include pandas, NumPy, matplotlib, sci-kit learn (sklearn), and seaborn. The main contribution of our proposed work is given below:

- The performance of all classifiers has been tested on full feature spaces in terms of all performance evaluation matrices specifically accuracy.
- The performances of the classifiers are tested on selected feature spaces, selected through various feature selection algorithms mentioned above.
- The research study recommends that which feature selection algorithm is feasible with which classification algorithm for developing a high-level intelligence system for the diagnosing of heart disease patients.

The rest of the paper is organized as: “Results and discussion” section represents the results and discussion, “Material and methods” section describes the material and methods used in this paper. Finally, we conclude our proposed research work in “Conclusion” section.

Material and methods

The subsections represent the materials and the methods that are used in this paper.

Dataset.

The first and rudimentary step of developing an intelligent computational model is to construct or develop a problem-related dataset that truly and effectively reflects the pattern of the target class. Well organized and problem-related dataset has a high influence on the performance of the computational model. Looking at the significance of the dataset, two datasets i.e. the Cleveland heart disease dataset S1 and Hungarian heart disease dataset (S2) are used, which are available online at the University of California Irvine (UCI) machine learning repository and UCI Kaggle repository, and various researchers have used it for conducting their research studies. The S1 consists of 304 instances, where each instance has distinct 13 attributes along with the target labels and are selected for training. The dataset is composed of two classes, presence or absence of heart disease. The S2 is composed of 1025 instances in which 525 instances belong to positive class while the rest of 500 instances have negative class. The description of attributes of both the datasets is the same, and both have similar attributes. The complete description and information of the datasets with 13 attributes are given.

Proposed system methodology

The main theme of the developed system is to identify heart problems in human beings. In this study, four distant feature selection techniques namely: FCBF, mRMR, Relief, and LASSO are applied on the provided dataset in order to remove noisy, redundant features and select variant features, consequently may cause of enhancing the performance of the proposed model. Various machine learning classification algorithms are used in this study which include, KNN, DT, ETC, RF,

LR, NB, ANN, SVM, AB, and GB. Difference evaluation metrics are computed to assess the performance of classification algorithms. The methodology of the proposed system is carried out in five stages which include dataset preprocessing, selection of features, cross-validation technique, classification algorithms, and performance evaluation of classifiers. The framework of the proposed system is illustrated.

Results and discussion

This section of the paper discusses the experimental results of various contemporary classification algorithms. At first, the performance of all used classification models i.e. K-Nearest Neighbors (KNN), Decision Tree (DT), Extra-Tree Classifier (ETC), Random Forest (RF), Logistic Regression (LR), Naïve Bayes (NB), Artificial Neural Network (ANN), Support Vector Machine (SVM), Adaboost (AB), and Gradient Boosting (GB) along with full feature space is evaluated. After that, four feature selection algorithms (FSA): Fast Correlation-Based Filter (FCBF), Minimal Redundancy Maximal Relevance (mRMR), Least Absolute Shrinkage and Selection Operator (LASSO), and Relief are applied to select the prominent and high variant features from feature space. Furthermore, the selected feature spaces are provided to classification algorithms as input to analyze the significance.

2.3 Detection of Cardiovascular Disease using Machine Learning Classification Models.

Abstract:

Cardiovascular disease (CVD) is an all encompassing term for situations affecting the heart or blood vessels. This is commonly associated with an accumulation of fatty deposits within the arteries (atherosclerosis) and an increased risk of developing blood clots. Cardiovascular disease is considered one of the largest causes of morbidity and mortality in the world's population. Predicting and diagnosing the disease is a critical challenge in clinical data analysis and health care providers to prevent people from contracting such a disease and conserve lives. Healthcare industries collect massive amounts of data that contain some information related to heart disease diagnosis, which is serviceable in making effective decisions. Furthermore, AI algorithms and deep neural networks can be used to analyze and diagnose heart disease. The project intends to automatically detect cardiovascular disease using two datasets through a deep learning network and a variety of machine learning classification models. The performance evaluated based on the accuracy, precision, recall, and f-score for each of the models. Hence, the Random Forest model achieved the highest performance at 94% accuracy in the heart diseases dataset, while Gradient Boosting model achieved the highest performance at 73% accuracy, 73% Recall, 73% F1-score, and 74% Precision in Cardiovascular Disease Dataset.

INTRODUCTION :

Cardiovascular disease (CVD) is a family of disorders that includes coronary heart, cerebrovascular disease, peripheral arterial disease, and congenital heart disease. Each four out of 5 CVD cases are died due to heart attacks and strokes. According to the World Health Organization (WHO), over 17.9 million patients have died from heart diseases around the world, which makes Cardiovascular diseases (CVDs) the number one cause of death globally. Cardiovascular diseases are a group of diseases that causes a failure in a human's heart and blood vessels. It has many symptoms such

as Chest pain, fast heart rate, difficulty breathing, and dizziness. As a fact, CVDs are complex as the human body, and they are caused by different risk factors such as obesity, high blood pressure, and high Cholesterol. Diagnosing an infection with heart disease requires specialized cardiologists with complicated procedures and tests to figure the accurate and efficient treatment. Cardiovascular disease can be prevented by early diagnosis, followed by healthy eating, exercising, and avoiding alcohol consumption. In undeveloped countries, patients suffering from cardiovascular disease are diagnosed with a severe delay at times, or they are transported over long distances unnecessarily, with the increase in the cost of travel and treatment, which is a burden on them. To successfully diagnose the patients into has a heart disorder or not, AI science has been involved with other sciences to solve real-life problems automatically. Nowadays, AI algorithms are used in disease diagnosis and detection AI with medicine helped the doctors to make decisions in some complicated cases in addition to predicting the high-risk patients of getting infected by a disease such as heart disease or Cardiovascular disease Various studies have demonstrated an automatic detection to diagnose the severity of heart diseases using different machine learning techniques such as integrating multiple classification algorithms and Boosting algorithms to build strong automated prediction systems. In the research, seven models used which are Support Vector Machine, K-Nearest neighbor, Logistic Regression, Decision Tree, Naïve Bayes, Random Forest, Artificial Neural Network models and some of ensemble techniques with the cardiovascular disease (CVD) dataset. Diagnosis by means of artificial intelligence is useful in the lack of medical personnel in addition that it is cheap and fast and can help doctors in the diagnosis process more efficiently to save patient's lives in early stages. In section II, the related papers to the topic have been reviewed. Section III-A the dataset pre-processing. Then, the research methodology and approaches have been discussed IV where the evaluation metrics mentioned in V. The next section is presenting the results of the experiments VI. The conclusion of the research in section VII.

DATASETS

The research intends to use two datasets, and for heart disease detection and diagnosis by using 9 classification algorithms.

Cardiovascular Disease Dataset :

In this research, the cardiovascular disease dataset used to diagnose and detect heart disease and compare the results with previous research. It contains a large number of patient's information and their medical records. The dataset collected by Kaggle from three different resources which are Objective where the data collected as facts about cardiovascular diseases, Examination is the results of some medical tests, and Subjective represents the information provided by the patient. The dataset utilized for training, testing, and validation sets. The proposed data is publicly available on the website. The shape of the cardiovascular diseases' dataset.

Heart Disease Dataset :

The second dataset in the research is the Heart Disease Dataset , It is a combination of the most popular datasets for heart disease prediction which are: Cleveland, Hungarian, Switzerland, Long Beach VA, and Stat log (Heart) Data Set. It contains 1190 records, 10 features, and one target. The dataset gathered by Anu Siddhartha in the year 2020. Heart Disease Dataset is a public dataset available on the IEEE Data port website.

RESEARCH METHODOLOGY:

This paper intends to employ different classification techniques which are Support Vector Machine (SVM), Decision Tree Classifier (DT), Logistic Regression, K-Nearest Neighbor (KNN), and Naïve Bayes (NB). In addition, implementing a deep network and measuring the impact of various optimization learning methods to detect cardiovascular diseases in order to find the best classifier. The steps to carry out this research.

Data Pre processing

In terms of improving the models' performance, feature selection is one of the main steps where only the most correlated features kept and used to train the model and unnecessary data removed to avoid time-consuming. Based on the data analysis, the dataset split into 90:10 for the training set, and testing set, respectively using cross-validation. In the Data processing phase, the Body Mass Index has been calculated to check the patient's health condition, where the normal value of the BMI is between 18 to 25 and the abnormal value is over 25 or under 18. The BMI values are calculated by using the formula below. After that, the weight and height features removed. According to the American Heart Association (AHA), The blood pressure measured by the Systolic and Diastolic rate. It is categorized into five classes of severity. The blood pressure calculated for each record and specified the blood pressure level.

Classification Algorithms :

Classification algorithms are the most widely used techniques in the machine learning field. In general, the models in machine learning are divided into three general approaches based on their learning style, Supervised learning, Unsupervised Learning, and Reinforcement Learning. Classification is a way of learning when the machine learns how to assign the labels to each class of the data. There are several models known as classification algorithms such as logistic regression, Naive Bayes, random forest, decision tree, gradient-boosted tree, and linear perceptron. This research reviews the implementation of seven classification algorithms on the dataset and some ensemble classifiers.

RESULTS AND DISCUSSION :

Based on experiments of implementing deep learning networks and a variety of classification machine learning algorithms to automatically detect cardiovascular disease using the cardiovascular disease dataset this section explains the results obtained with cross validation (10 KFold) 90 for training and 10 for testing. The following Table V shows the results obtained from the classifier after data processing and features selection. The Gradient Boosting model showed the highest performance in predicting the disease followed by the ANN at 70% accuracy. The lowest model was DT at 62% accuracy.

2.4 Heart Disease Identification Method Using Machine Learning classification in E-Healthcare

ABSTRACT:

Heart disease is one of the complex diseases and globally many people suffered from this disease. On time and efficient identification of heart disease plays a key role in

healthcare, particularly in the field of cardiology. In this article, we proposed an efficient and accurate system to diagnosis heart disease and the system is based on machine learning techniques. The system is developed based on classification algorithms includes Support vector machine, Logistic regression, Artificial neural network, K-nearest neighbor, Naïve bays, and Decision tree while standard features selection algorithms have been used such as Relief, Minimal redundancy maximal relevance, Least absolute shrinkage selection operator and Local learning for removing irrelevant and redundant features. We also proposed novel fast conditional mutual information feature selection algorithm to solve feature selection problem. The features selection algorithms are used for features selection to increase the classification accuracy and reduce the execution time of classification system. Furthermore, the leave one subject out cross-validation method has been used for learning the best practices of model assessment and for hyper parameter tuning. The performance measuring metrics are used for assessment of the performances of the classifiers. The performances of the classifiers have been checked on the selected features as selected by features selection algorithms. The experimental results show that the proposed feature selection algorithm (FCMIM) is feasible with classifier support vector machine for designing a high-level intelligent system to identify heart disease. The suggested diagnosis system (FCMIM-SVM) achieved good accuracy as compared to previously proposed methods. Additionally, the proposed system can easily be implemented in healthcare for the identification of heart disease.

INTRODUCTION:

Heart disease (HD) is the critical health issue and numerous people have been suffered by this disease around the world. The HD occurs with common symptoms of breath shortness, physical body weakness and, feet are swollen. Researchers try to come across an efficient technique for the detection of heart disease, as the current diagnosis techniques of heart disease are not much effective in early time identification due to several reasons, such as accuracy and execution time. The diagnosis and treatment of heart disease is extremely difficult when modern technology and medical experts are not available. The effective diagnosis and proper treatment can save the lives of many people. According to the European Society of Cardiology, 26 million approximately people of HD were diagnosed and diagnosed 3.6 million annually. Most of the people in the United States are suffering from heart disease. Diagnosis of HD is traditionally done by the analysis of the medical history of the patient, physical examination report and analysis of concerned symptoms by a physician. But the results obtained from this diagnosis method are not accurate in identifying the patient of HD. Moreover, it is expensive and computationally difficult to analyze. Thus, to develop a non-invasive diagnosis system based on classifiers of machine learning to resolve these issues. Expert decision system based on machine learning classifiers and the application of artificial fuzzy logic is effectively diagnosis the HD as a result, the ratio of death decreases and. The Cleveland heart disease data set was used by various researchers for the identification problem of HD. The machine learning predictive models need proper data for training and testing. The performance of machine learning model can be increased if balanced dataset is use for training and testing of the model. Furthermore, the model predictive capabilities can improved by using proper and related features from the data. Therefore, data balancing and feature selection is significantly important for model performance improvement. In literature

various diagnosis techniques have been proposed by various researchers, however these techniques are not effectively diagnosis HD. In order to improve the predictive capability of machine learning model data preprocessing is important for data standardization. Various Preprocessing techniques such removal of missing feature value instances from the dataset, Standard Scalar (SS), Min-Max Scalar etc. The feature extraction and selection techniques are also improve model performance. Various feature selection techniques are mostly used for important feature selection such as, LASSO, Relief, LLBFS, PCA, GA, and optimization methods, such as ACO, FFO, BFO etc. Similarly Yun et al. presented different techniques for different type of feature selection, such as feature selection for high-dimensional small sample size data, large-scale data, and secure feature selection. They also discussed some important topics for feature selection have emerged, such as stable feature selection, multi-view feature selection, distributed feature selection, multi-label feature selection, online feature selection, and adversarial feature selection. Jundong et al. Discussed the challenges of feature selection for big data. It is necessary to decrease the dimensionality of data for various learning tasks due to the curse of dimensionality. Feature selection has great influence in numerous applications such as building simpler, increasing learning performance, creating clean and understandable data. The feature selection from big data is challenging job and create big problems because big data has many dimensions. Further, challenges of feature selection for structured, heterogeneous and streaming data as well as its scalability and stability issues. For big data analytics challenges of feature selection is very important to resolved. In designed unsupervised hashing scheme, called topic hyper graph hashing , to report the limitations. Topic hyper graph hashing effectively mitigates the semantic shortage of hashing codes by exploiting auxiliary texts around images. The proposed Topic hyper graph hashing can achieve superior performance equaled with numerous state-of-the art approaches, and it is more appropriate for mobile image retrieval. The feature selection algorithms are classified into three type such as filter based, wrapper based and embedded based. All these feature selection mechanisms have some advantages and limitations in certain cases. The filter based method measures the relevance of a feature by correlation with the dependent variable while the wrapper feature selection algorithm measure the usefulness of a subset of features by actually training the classifier on it. The filter method is less computationally complex than wrapper method. The feature set selected by the filter is general and can be applied to any model and it is independent of a specific model. In feature selection global relevance is of greater importance. On another hand suitable machine learning model is necessary for good results. Obviously, a good machine learning model is a model that not only performs well on data seen during training (else a machine learning model could simply learn the training data), but also on unseen data. To evaluate all classifiers on data and that they get, on average, 50% of the cases right. Furthermore, appropriate cross validation techniques and performance evaluation metrics are critical necessary for a model when model is train and test on dataset. We proposed a machine learning based diagnosis method for the identification of HD in this research work. Machine learning predictive models include ANN, LR, K-NN, SVM, DT, and NB are used for the identification of HD. The standard state of the art features selection algorithms, such as Relief, Minimal-Redundancy-Maximal-Relevance (mRMR), Least-absolute-shrinkage-selection-operator (LASSO) and Local-learning-based-features-selection (LLBFS) have been used to select the features. We also proposed fast conditional

mutual information (FCMIM) features selection algorithm for features selection. Leave-one-subject-out cross-validation (LOSO) technique has been applied to select the best hyper parameters for best model selection. Apart from this, different performance assessment metrics have been used for classifiers performances evaluation. The proposed method has been tested on Cleveland HD dataset. Furthermore, the performance of the proposed technique have been compared with state of the art existing methods in the literature, such as NB, Three phase ANN diagnosis system, Neural network ensembles(NNE), ANN-Fuzzy-AHP diagnosis system(AFP), Adaptive-weighted-Fuzzy-system ensemble (AWFSE). The research study has the following contributions.

- Firstly, the authors try to address the problem of features selection by employing pre-processing techniques and standard state of the art four features selection algorithms such as Relief, mRMR, LASSO, and LLBFS for appropriate subset of features and then applied these features for effective training and testing of the classifiers that identify which feature selection algorithm and classifier gives good results in term of accuracy and computation time.
- Secondly, the authors proposed fast conditional mutual information (FCMIM) FS algorithm for feature selection and then these features are input to classifiers for improving prediction accuracy and reducing computation time. The classifiers performances have been compared on features selected by the standard state of the art FS algorithms with the selected features of the proposed FS algorithm.

Thirdly, identify weak features from the dataset which affect the performance of the classifiers. Finally, suggests that heart disease identification system (FCMIM-SVM) effectively identify the HD. The paper remaining sections are structured as follows. The literature related to the problem has been discussed in section 2. In section 3 the dataset and the theoretical and mathematical knowledge of feature selection and classification algorithms are discussed in details. Additionally, discuss the technique of cross-validation and performance measuring metrics. In section 4 results of all experiments are analyzed and discussed in details. The last section 5 the conclusion and future direction of the research work have been explored in details.

CONCLUSION:

In this study, an efficient machine learning based diagnosis system has been developed for the diagnosis of heart disease. Machine learning classifiers include LR, K-NN, ANN, SVM, NB, and DT are used in the designing of the system. Four standard feature selection algorithms including Relief, MRMR, LASSO, LLBFS, and proposed a novel feature selection algorithm FCMIM used to solve feature selection problem. LOSO cross-validation method is used in the system for the best hyper parameters selection. The system is tested on Cleveland heart disease dataset. Further more, performance evaluation metrics are used to check the performance of the identification system. According to Table 15 the specificity of ANN classifier is best on Relief FS algorithm as compared to the specificity of MRMR, LASSO, LLBFS, and FCMIM feature selection algorithms. Therefore for ANN with relief is the best predictive system for detection of healthy people. The sensitivity of classifier NB on selected features set by LASSO FS algorithm also gives the best result as compared to the sensitivity values of Relief FS algorithm with classifier SVM (linear). The classifier Logistic Regression MCC is 91% on selected features selected by FCMIM FS algorithm. The processing time of Logistic Regression

with Relief, LASSO, FCMIM and LLBFS FS algorithm best as compared to MRMR FS algorithms, and others classifiers. Thus the experimental results show that the proposed features selection algorithm select features that are more effective and obtains high classification accuracy than the standard feature selection algorithms. According to feature selection algorithms, the most important and suitable features are Thallium Scan type chest pain and Exercise-induced Angina. All FS algorithms results show that the feature Fasting blood sugar (FBS) is not a suitable heart disease diagnosis. The accuracy of SVM with the proposed feature selection algorithm (FCMIM) is 92.37% which is very good as compared previously proposed methods as shown in Table 17. Further, the performance of machine learning based method FCMIM SVM is high then Deep neural network for detection of HD. The novelty of the study is developing a diagnosis system for identification of heart disease. In this study, four standard feature selection algorithms along with one proposed feature selection algorithm is used for features selection. LOSO CV method and performance measuring metrics are used. The Cleveland heart disease dataset is used for testing purpose. As we think that developing a decision support system through machine learning algorithms it will be more suitable for the diagnosis of heart disease. Furthermore, we know that irrelevant features also degrade the performance of the diagnosis system and increased computation time. Thus another innovative touch of our study to used features selection algorithms to selects the appropriate features that improve the classification accuracy as well as reduce the processing time of the diagnosis system. In the future, we will use other features selection algorithms, optimization methods to further increase the performance of a predictive system for HD diagnosis. The controlling and treatment of disease is significance after diagnosis, therefore, i will work on treatment and recovery of diseases in future also for critical disease such as heart, breast, Parkinson, diabetes.

Chapter 3

EXISTING SYSTEM:

In Existing medical systems, including hospital management systems and decision making systems, focus on collecting and mining the entire medical data. The entire patient records are loaded and all factors are considered. The C4.5 model that we previously developed. On the basis of this new measure, a data mining algorithm was developed to mine the causal relationship between drugs and their associated with risk for CAD heart Disease. heart musicality issues (arrhythmia's); and heart abandons you're brought into the world with (inherent heart deserts), among others. The expression "coronary illness" is regularly utilized reciprocally with the expression "cardiovascular infection." Prescient demonstrating utilizes measurements to foresee outcomes. Most frequently the occasion one needs to anticipate is later on, yet prescient displaying can be applied to an obscure occasion, paying little mind to happened.

PROBLEM STATEMENT:

Existing systems have failed to utilize and understand the importance of misdiagnosis. A very important attribute which interconnects and addresses all these issues.

It varies from patient's medical history, climatic conditions, neighborhood, and various other factors

Propose system:

In the proposed work user will search for the heart Disease diagnosis (heart Disease and treatment related information) by giving symptoms as a query in the search engine. These symptoms are preprocessed to make the further process easier to find the symptoms keyword which helps to identify the heart Disease quickly.

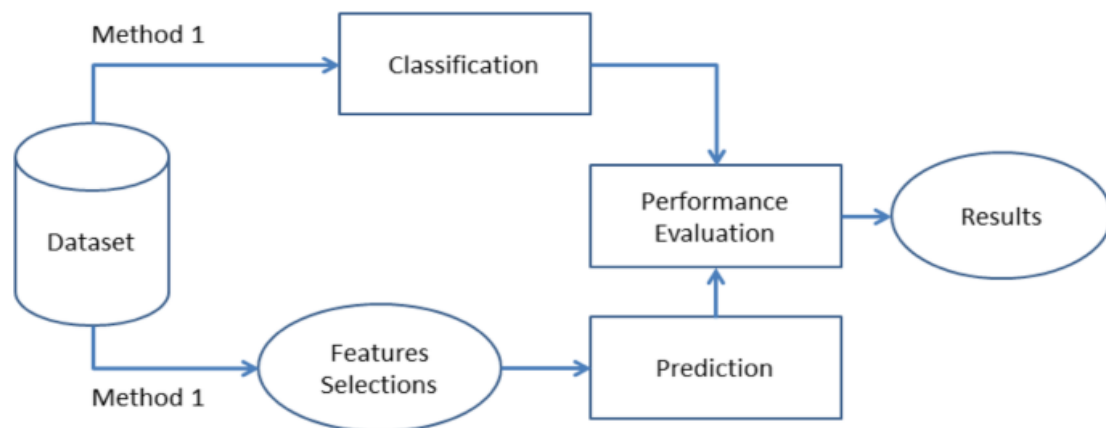
CFS+PSO are a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification.

This feature has been identified as the most suitable for the present system.

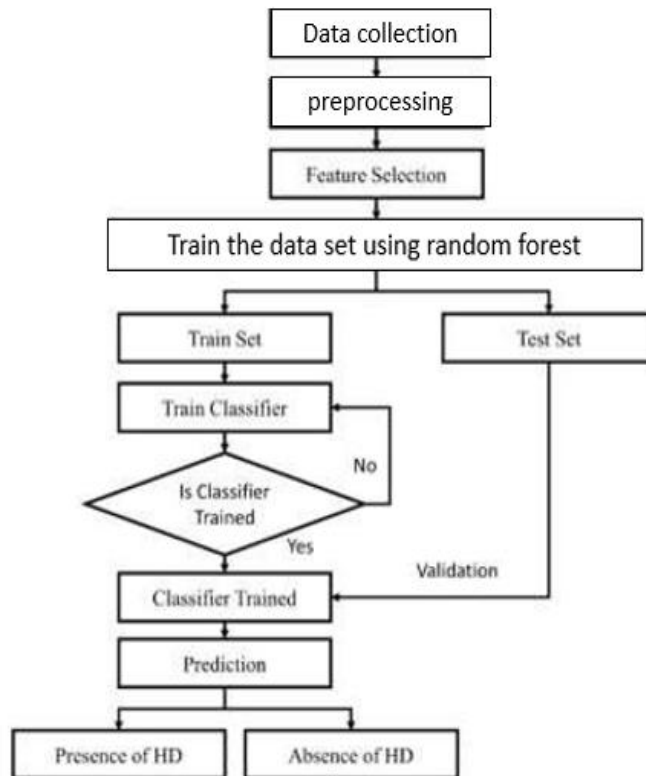
Advantage:

- It is easy to extract signatures from individual data instances, as their structures. Just collect the symptoms that enough to scaling samples.
- Can easily predict the heart Disease level and severity easily using range level of queries.
- The probability of vocabulary gap between diverse health seekers makes the data more consistent compared to other formats of health data.

Block diagram



Flow diagram



Hardware requirements:

- System : Pentium i3 Processor.
- Hard Disk : 500 GB.
- Monitor : 15'' LED
- Input Devices : Keyboard, Mouse
- Ram : 2 GB

Software requirements:

- Operating system : Windows 10.
- Coding Language : Python

Chapter 4

Machine learning

What are the 7 steps of machine learning?

7 Steps of Machine Learning

- Step 1: Gathering Data. ...

- Step 2: Preparing that Data. ...
- Step 3: Choosing a Model. ...
- Step 4: Training. ...
- Step 5: Evaluation. ...
- Step 6: Hyper parameter Tuning. ...
- Step 7: Prediction.

Introduction:

In this blog, we will discuss the workflow of a Machine learning project this includes all the steps required to build the proper machine learning project from scratch.

We will also go over data pre-processing, data cleaning, feature exploration and feature engineering and show the impact that it has on Machine Learning Model Performance. We will also cover a couple of the pre-modelling steps that can help to improve the model performance.

Python Libraries that would be need to achieve the task:

1. Numpy
2. Pandas
3. Sci-kit Learn
4. Matplotlib

Understanding the machine learning workflow

We can define the machine learning workflow in 3 stages.

1. Gathering data
2. Data pre-processing
3. Researching the model that will be best for the type of data
4. Training and testing the model

5. Evaluation

Okay but first let's start from the basics

What is the machine learning Model?

The machine learning model is nothing but a piece of code; an engineer or data scientist makes it smart through training with data. So, if you give garbage to the model, you will get garbage in return, i.e. the trained model will provide false or wrong prediction

1. Gathering Data

The process of gathering data depends on the type of project we desire to make, if we want to make an ML project that uses real-time data, then we can build an IoT system that using different sensors data. The data set can be collected from various sources such as a file, database, sensor and many other such sources but the collected data cannot be used directly for performing the analysis process as there might be a lot of missing data, extremely large values, unorganized text data or noisy data. Therefore, to solve this problem Data Preparation is done.

We can also use some free data sets which are present on the internet. **Kaggle** and **UCI Machine learning Repository** are the repositories that are used the most for making Machine learning models. Kaggle is one of the most visited websites that is used for practicing machine learning algorithms, they also host competitions in which people can participate and get to test their knowledge of machine learning.

2. Data pre-processing

Data pre-processing is one of the most important steps in machine learning. It is the most important step that helps in building machine learning models more accurately. In

machine learning, there is an 80/20 rule. Every data scientist should spend 80% time for data pre-processing and 20% time to actually perform the analysis.

What is data pre-processing?

Data pre-processing is a process of cleaning the raw data i.e. the data is collected in the real world and is converted to a clean data set. In other words, whenever the data is gathered from different sources it is collected in a raw format and this data isn't feasible for the analysis.

Therefore, certain steps are executed to convert the data into a small clean data set, this part of the process is called as data pre-processing.

Why do we need it?

As we know that data pre-processing is a process of cleaning the raw data into clean data, so that can be used to train the model. So, we definitely need data pre-processing to achieve good results from the applied model in machine learning and deep learning projects. Most of the real-world data is messy, some of these types of data are:

1. **Missing data:** Missing data can be found when it is not continuously created or due to technical issues in the application (IOT system).
2. **Noisy data:** This type of data is also called outliers, this can occur due to human errors (human manually gathering the data) or some technical problem of the device at the time of collection of data.
3. **Inconsistent data:** This type of data might be collected due to human errors (mistakes with the name or values) or duplication of data.

Three Types of Data

1. Numeric e.g. income, age

2. Categorical e.g. gender, nationality

3. Ordinal e.g. low/medium/high

How can data pre-processing be performed?

These are some of the basic pre — processing techniques that can be used to convert raw data.

1. **Conversion of data:** As we know that Machine Learning models can only handle numeric features, hence categorical and ordinal data must be somehow converted into numeric features.

2. **Ignoring the missing values:** Whenever we encounter missing data in the data set then we can remove the row or column of data depending on our need. This method is known to be efficient but it shouldn't be performed if there are a lot of missing values in the dataset.

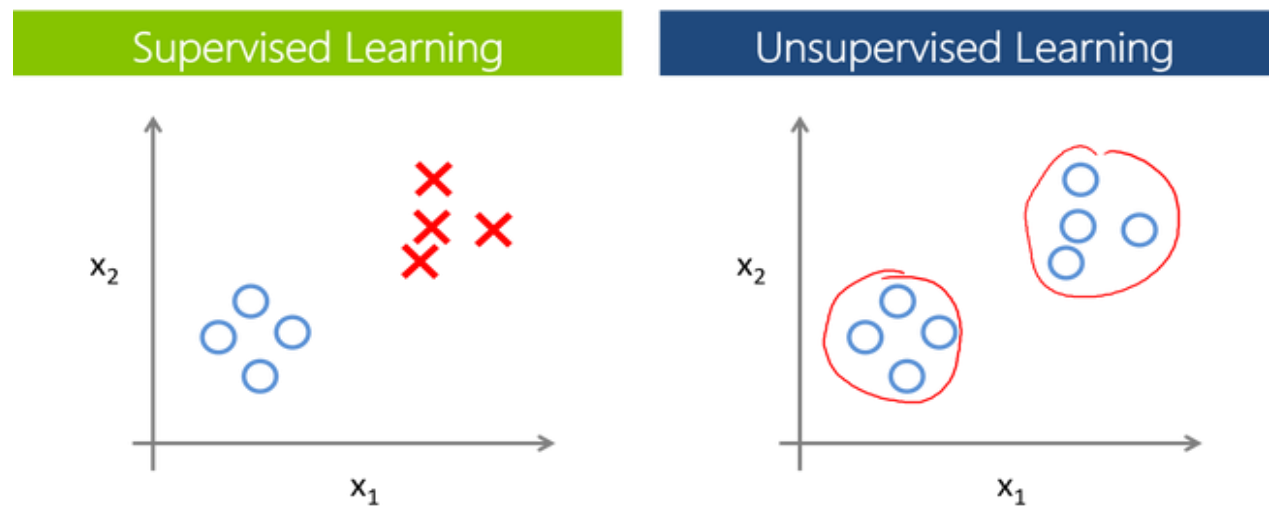
3. **Filling the missing values:** Whenever we encounter missing data in the data set then we can fill the missing data manually, most commonly the mean, median or highest frequency value is used.

4. **Machine learning:** If we have some missing data then we can predict what data shall be present at the empty position by using the existing data.

5. **Outliers detection:** There are some error data that might be present in our data set that deviates drastically from other observations in a data set. [Example: human weight = 800 Kg; due to mistyping of extra 0]

Researching the model that will be best for the type of data

Our main goal is to train the best performing model possible, using the pre-processed data.



Supervised Learning:

In Supervised learning, an AI system is presented with data which is labelled, which means that each data is tagged with the correct label.

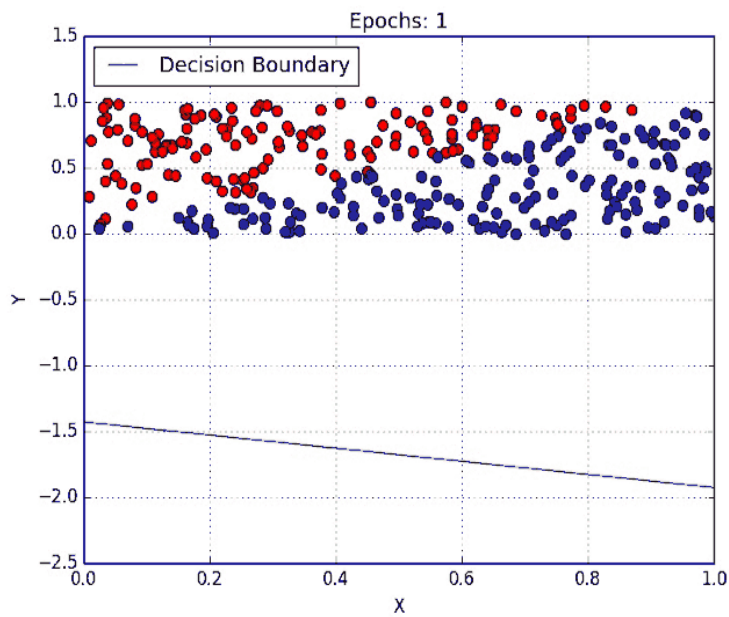
The supervised learning is categorized into 2 other categories which are

“**Classification**” and “**Regression**”.

Classification:

Classification problem is when the target variable is **categorical** (i.e. the output could be classified into classes — it belongs to either Class A or B or something else).

A classification problem is when the output variable is a category, such as “red” or “blue”, “disease” or “no disease” or “spam” or “not spam”.



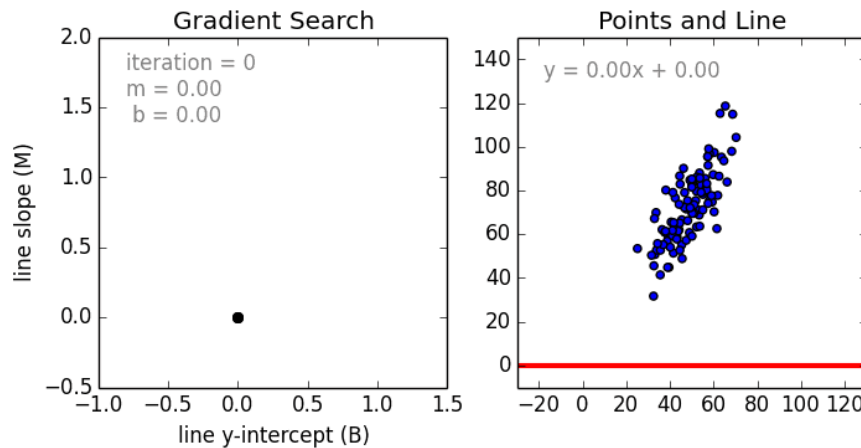
As shown in the above representation, we have 2 classes which are plotted on the graph i.e. red and blue which can be represented as ‘setosa flower’ and ‘versicolor flower’, we can image the X-axis as their ‘Sepal Width’ and the Y-axis as the ‘Sepal Length’, so we try to create the [best fit line](#) that separates both classes of flowers.

These some most used classification algorithms.

- **K-Nearest Neighbor**
- **Random Forest**
- **Logistic Regression**

Regression:

While a **Regression** problem is when the target variable is **continuous** (i.e. the output is numeric).



As shown in the above representation, we can imagine that the graph's X-axis is the 'Test scores' and the Y-axis represents 'IQ'. So we try to create the best fit line in the given graph so that we can use that line to predict any approximate IQ that isn't present in the given data.

These are some of the most used regression algorithms.

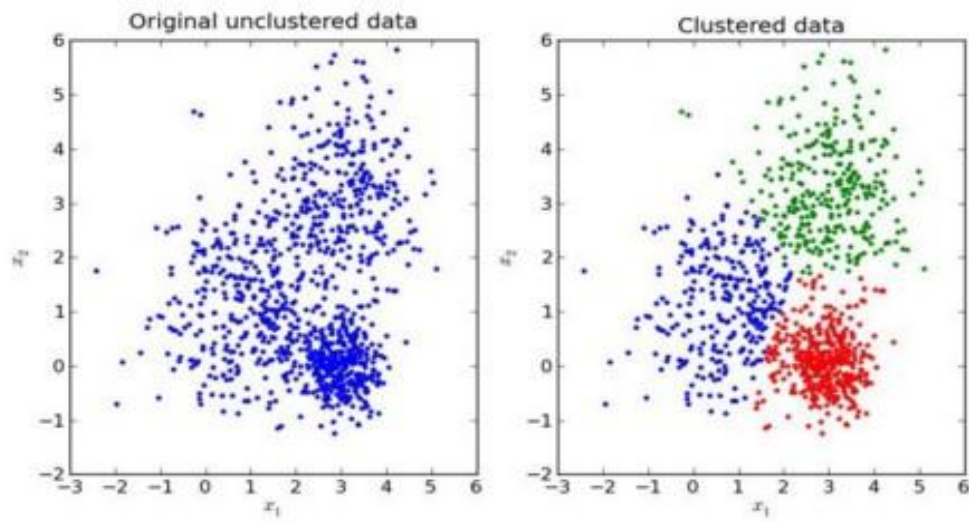
- **Linear Regression**
- **Random Forest**
- **Gaussian Progresses Regression**
- **Ensemble Methods**

Unsupervised Learning:

The unsupervised learning is categorized into 2 other categories which are "**Clustering**" and "**Association**".

Clustering:

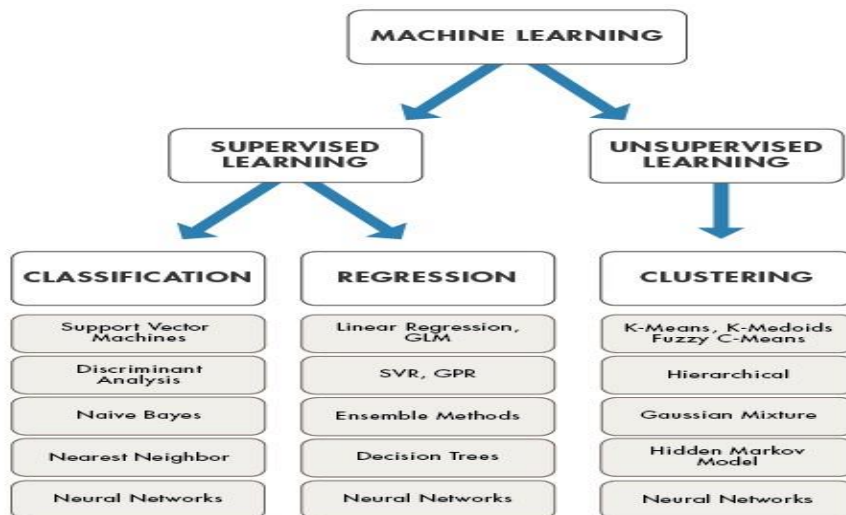
A set of inputs is to be divided into groups. Unlike in classification, the groups are not known beforehand, making this typically an unsupervised task.



Methods used for clustering are:

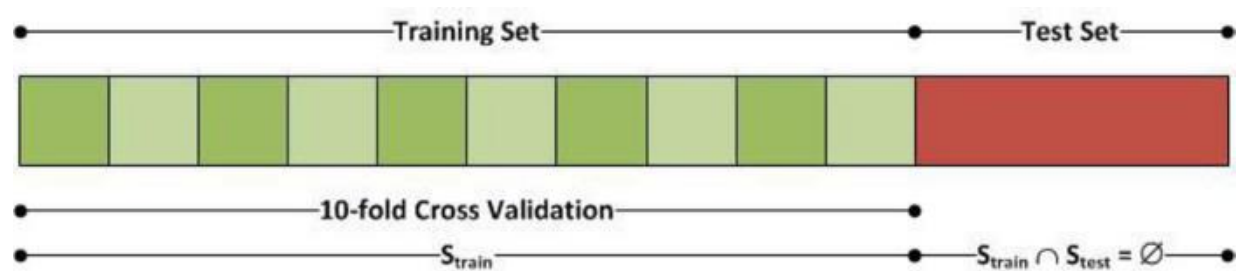
- **Gaussian mixtures**
- **K-Means Clustering**
- **Boosting**
- **Hierarchical Clustering**
- **K-Means Clustering**
- **Spectral Clustering**

Overview of models under categories:



4. Training and testing the model on data

For training a model we initially split the model into 3 three sections which are ‘**Training data**’, ‘**Validation data**’ and ‘**Testing data**’. You train the classifier using ‘**training data set**’, tune the parameters using ‘**validation set**’ and then test the performance of your classifier on unseen ‘**test data set**’. An important point to note is that during training the classifier only the training and/or validation set is available. The test data set must not be used during training the classifier. The test set will only be available during testing the classifier.

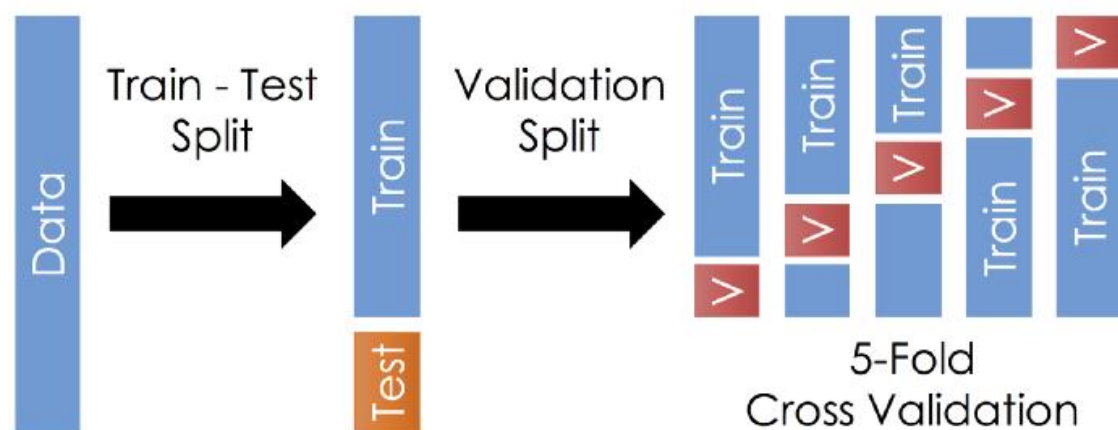


Training set:

The training set is the material through which the computer learns how to process information. Machine learning uses algorithms to perform the training part. A set of data used for learning, that is to fit the parameters of the classifier.

Validation set:

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. A set of unseen data is used from the training data to tune the parameters of a classifier.



Once the data is divided into the 3 given segments we can start the training process.

In a data set, a training set is implemented to build up a model, while a test (or validation) set is to validate the model built. Data points in the training set are excluded from the test (validation) set. Usually, a data set is divided into a training set, a validation set (some people use 'test set' instead) in each iteration, or divided into a training set, a validation set and a test set in each iteration.

The model uses any one of the models that we had chosen in step 3/ point 3. Once the model is trained we can use the same trained model to predict using the testing data i.e. the unseen data. Once this is done we can develop a confusion matrix, this tells us how well our model is trained. A confusion matrix has 4 parameters, which are '**True positives**', '**True Negatives**', '**False Positives**' and '**False Negative**'. We prefer that

we get more values in the True negatives and true positives to get a more accurate model. The size of the Confusion matrix completely depends upon the number of classes.

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

True positives : These are cases in which we predicted TRUE and our predicted output is correct.

True negatives : We predicted FALSE and our predicted output is correct.

False positives : We predicted TRUE, but the actual predicted output is FALSE.

False negatives : We predicted FALSE, but the actual predicted output is TRUE.

We can also find out the accuracy of the model using the confusion matrix.

$$\text{Accuracy} = (\text{True Positives} + \text{True Negatives}) / (\text{Total number of classes})$$

i.e. for the above example:

$$\text{Accuracy} = (100 + 50) / 165 = 0.9090 \text{ (90.9\% accuracy)}$$

5. Evaluation

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future.

To improve the model we might tune the hyper-parameters of the model and try to improve the accuracy and also looking at the confusion matrix to try to increase the number of true positives and true negatives.

Conclusion

In this blog, we have discussed the workflow a Machine learning project and gives us a basic idea of how a should the problem be tackled.

Chapter 5 **SYSTEM DESIGN**

Module name:

- Dataset collection
- Pre-Processing
- Feature Extraction
- Model training
- Testing model
- Performance Evaluation
- Prediction

Dataset collection:

Collecting data allows you to capture a record of past events so that we can use data analysis to find recurring patterns. From those patterns, you build predictive models using machine learning algorithms that look for trends and predict future changes.

Predictive models are only as good as the data from which they are built, so good data collection practices are crucial to developing high-performing models.

The data need to be error-free (garbage in, garbage out) and contain relevant information for the task at hand. For example, a loan default model would not benefit from tiger population sizes but could benefit from gas prices over time.

In this module, we collect the data from kaggle dataset archives. This dataset contains the information of divorce in previous years.

Pre-Processing:

The Wisconsin Prognostic Cleave Land Train Dataset is downloaded from the UCI Machine Learning Repository website and saved as a text file.

This file is then imported into Excel spreadsheet and the values are saved with the corresponding attributes as column headers.

The missing values are replaced with appropriate values.

Feature Extraction:

This is done to reduce the number of attributes in the dataset hence providing advantages like speeding up the training and accuracy improvements.

In machine learning, pattern recognition, and image processing, feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations. Feature extraction is related to dimensionality reduction

When the input data to an algorithm is too large to be processed and it is suspected to be redundant (e.g. the same measurement in both feet and meters, or the repetitiveness of images presented as pixels), then it can be transformed into a reduced set of features (also named a feature vector).

Determining a subset of the initial features is called feature selection. The selected features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced representation instead of the complete initial data.

Model training:

A training model is a dataset that is used to train an ML algorithm. It consists of the sample output data and the corresponding sets of input data that have an influence on the output.

The training model is used to run the input data through the algorithm to correlate the processed output against the sample output. The result from this correlation is used to modify the model.

This iterative process is called “model fitting”. The accuracy of the training dataset or the validation dataset is critical for the precision of the model.

Model training in machine language is the process of feeding an ML algorithm with data to help identify and learn good values for all attributes involved.

There are several types of machine learning models, of which the most common ones are supervised and unsupervised learning.

In this module we use supervised classification algorithms like linear regression to train the model on the cleaned dataset after dimensionality reduction.

Testing model:

In this module we test the trained machine learning model using the test dataset

Quality assurance is required to make sure that the software system works according to the requirements. Were all the features implemented as agreed? Does the program behave as expected? All the parameters that you test the program against should be stated in the technical specification document.

Moreover, software testing has the power to point out all the defects and flaws during development. You don't want your clients to encounter bugs after the software is released and come to you waving their fists. Different kinds of testing allow us to catch bugs that are visible only during runtime.

Performance Evaluation:

In this module, we evaluate the performance of trained machine learning model using performance evaluation criteria such as F1 score, accuracy and classification error.

In case the model performs poorly, we optimize the machine learning algorithms to improve the performance.

Performance Evaluation is defined as a formal and productive procedure to measure an employee's work and results based on their job responsibilities. It is used to gauge the amount of value added by an employee in terms of increased business revenue, in comparison to industry standards and overall employee return on investment (ROI).

All organizations that have learned the art of "winning from within" by focusing inward towards their employees, rely on a systematic performance evaluation process to measure and evaluate employee performance regularly.

Ideally, employees are graded annually on their work anniversaries based on which they are either promoted or are given suitable distribution of salary raises

Performance evaluation also plays a direct role in providing periodic feedback to employees, such that they are more self-aware in terms of their performance metrics.

Prediction:

"Prediction" refers to the output of an algorithm after it has been trained on a historical dataset and applied to new data when forecasting the likelihood of a particular outcome, such as whether or not a customer will churn in 30 days.

The algorithm will generate probable values for an unknown variable for each record in the new data, allowing the model builder to identify what that value will most likely be.

The word "prediction" can be misleading. In some cases, it really does mean that you are predicting a future outcome, such as when you're using machine learning to determine the next best action in a marketing campaign.

Other times, though, the "prediction" has to do with, for example, whether or not a transaction that already occurred was fraudulent.

In that case, the transaction already happened, but you're making an educated guess about whether or not it was legitimate, allowing you to take the appropriate action.

In this module we use trained and optimized machine learning model to predict whether the patient give the answer some questions .

Algorithm used:

- Logistic Regression
- K-Nearest Neighbour
- Random forest

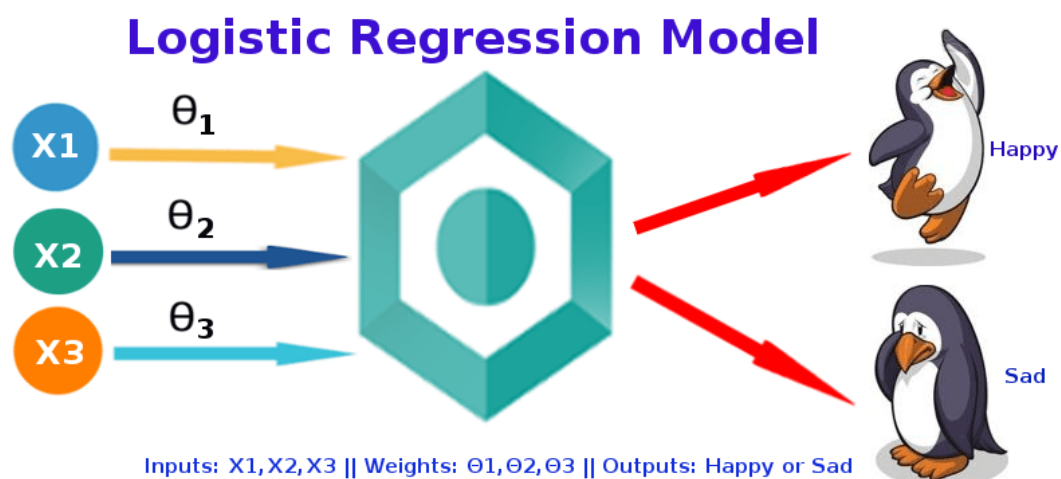
What is Logistic Regression?

I hope the preceding discussion has provided us with a better understanding of machine learning and its various types. Logistic Regression is a Machine Learning method that is used to solve classification issues. It is a predictive analytic technique that is based on the probability idea. The classification algorithm Logistic Regression is used to predict the likelihood of a categorical dependent variable. The dependant variable in logistic regression is a binary variable with data coded as 1 (yes, True, normal, success, etc.) or 0 (no, False, abnormal, failure, etc.).

The goal of Logistic Regression is to discover a link between characteristics and the likelihood of a specific outcome. For example, when predicting whether a student passes or fails an exam based on the number of hours spent studying, the response variable has two values: pass and fail.

A Logistic Regression model is similar to a Linear Regression model, except that the Logistic Regression utilizes a more sophisticated cost function, which is known as the “Sigmoid function” or “logistic function” instead of a linear function.

Many people may have a question, whether Logistic Regression is a classification or regression category. The logistic regression hypothesis suggests that the cost function be limited to a value between 0 and 1. As a result, linear functions fail to describe it since it might have a value larger than 1 or less than 0, which is impossible according to the logistic regression hypothesis.



To answer this question, Logistic Regression is considered a regression model also. This model creates a regression model to predict the likelihood that a given data entry belongs to the category labeled “1.” Logistic regression models the data using the sigmoid function, much as linear regression assumes that the data follows a linear distribution.

Why the name Logistic Regression?

It's called ‘Logistic Regression’ since the technique behind it is quite similar to Linear Regression. The name “Logistic” comes from the Logistic function, which is utilized in this categorization approach.

Logit function to Sigmoid Function – Logistic Regression:

Logistic Regression can be expressed as, where $p(x)/(1-p(x))$ is termed odds, and the left-hand side is called the logit or log-odds function. The odds are the ratio of the chances of success to the chances of failure. As a result, in Logistic Regression, a linear combination of inputs is translated to $\log(\text{odds})$, with an output of 1.

The following is the inverse of the aforementioned function

This is the Sigmoid function, which produces an S-shaped curve. It always returns a probability value between 0 and 1. The Sigmoid function is used to convert expected values to probabilities. The function converts any real number into a number between 0 and 1. We utilize sigmoid to translate predictions to probabilities in machine learning.

Types of Logistic Regression:

In general, it can be classified into,

1. Binary Logistic Regression – two or binary outcomes like yes or no
2. Multinomial Logistic Regression – three or more outcomes like first, second, and third class or no class degree
3. Ordinal Logistic Regression – three or more like multinomial logistic regression but here with the order like customer rating in the supermarket from 1 to 5

Requirements for Logistic Regression to work well

This model can work for all the datasets, but still, if you need good performance, then there will be some assumptions to consider,

The dependant variable in binary logistic regression must be binary.

Only the variables that are relevant should be included.

The independent variables must be unrelated to one another. That is, there should be minimal or no multicollinearity in the model.

The log chances are proportional to the independent variables.

Large sample sizes are required for logistic regression.

Decision Boundary – Logistic Regression

A threshold can be established to forecast which class a data belongs to. The derived estimated probability is categorized into classes based on this threshold.

If the predicted value is less than 0.5, categorize the particular student as a pass; otherwise, label it as a fail. There are two types of decision boundaries: linear and non-linear. To provide a complicated decision boundary, the polynomial order can be raised.

Why can't the cost function that was used for linearity be used for logistics?

The cost function for linear regression is mean squared error. If this is utilized for logistic regression, the function of parameters will be non-convex. Only if the function is convex will gradient descent lead to a global minimum.

Cost function – Linear Regression Vs Logistic Regression

Linear regression employs the Least Squared Error as the loss function, which results in a convex network, which we can then optimize by identifying the vertex as the global minimum. For logistic regression, however, it is no longer a possibility. Because the hypothesis has been modified, calculating Least Squared Error using the sigmoid function on raw model output will result in a non-convex graph with local minimums.

What is cost function? Cost functions are used in machine learning to estimate how poorly models perform. Simply put, a cost function is a measure of how inaccurate the model is in estimating the connection between X and y . This is usually stated as a difference or separation between the expected and actual values. A machine learning model's goal is to discover parameters, weights, or a structure that minimizes the cost function.

A convex function indicates there will be no intersection between any two points on the curve, but a non-convex function will have at least one intersection. In terms of cost functions, a convex type always guarantees a global minimum, whereas a non-convex type only guarantees local minima.

Advantages of Logistic Regression

1. Overfitting is less likely with logistic regression, although it can happen in high-dimensional datasets. In these circumstances, regularization (L1 and L2) techniques may be used to minimize over-fitting.
2. It works well when the dataset is linearly separable and has good accuracy for many basic data sets.
3. It is more straightforward to apply, understand, and train.
4. The inferences regarding the relevance of each characteristic are based on the anticipated parameters (trained weights). The association's orientation, positive or negative, is also specified. As a result, logistic regression may be used to determine the connection between the characteristics.
5. Unlike decision trees or support vector machines, this technique allows models to be readily changed to incorporate new data. Stochastic gradient descent can be used to update the data.
6. It is less prone to over-fitting in a low-dimensional dataset with enough training instances.
7. When the dataset includes linearly separable characteristics, Logistic Regression shows to be highly efficient.
8. It has a strong resemblance to neural networks. A neural network representation may be thought of as a collection of small logistic regression classifiers stacked together.
9. The training time of the logistic regression method is considerably smaller than that of most sophisticated algorithms, such as an Artificial Neural Network, due to its simple probabilistic interpretation.
10. Multinomial Logistic Regression is the name given to an approach that may easily be expanded to multi-class classification using a softmax classifier.

Disadvantages of Logistic Regression

1. Logistic Regression should not be used if the number of observations is fewer than the number of features; otherwise, it may result in over fitting.
2. Because it creates linear boundaries, we won't obtain better results when dealing with complex or non-linear data.
3. It's only good for predicting discrete functions. As a result, the Logistic Regression dependent variable is restricted to the discrete number set.
4. The average or no multi collinearity between independent variables is required for logistic regression.

5. Logistic regression needs a big dataset and enough training samples to identify all of the categories.
6. Because this method is sensitive to outliers, the presence of data values in the dataset that differs from the anticipated range may cause erroneous results.
7. Only significant and relevant features should be utilized to construct a model; otherwise, the model's probabilistic predictions may be inaccurate, and its predictive value may suffer.
8. Complex connections are difficult to represent with logistic regression. This technique is readily outperformed by more powerful and sophisticated algorithms such as Neural Networks.
9. Because logistic regression has a linear decision surface, it cannot address nonlinear issues. In real-world settings, linearly separable data is uncommon. As a result, non-linear features must be transformed, which may be done by increasing the number of features such that the data can be separated linearly in higher dimensions.
10. Based on independent variables, a statistical analysis model seeks to predict accurate probability outcomes. On high-dimensional datasets, this may cause the model to be over-fit on the training set, overstating the accuracy of predictions on the training set, and so preventing the model from accurately predicting outcomes on the test set. This is most common when the model is trained on a little amount of training data with many features. Regularization strategies should be explored on high-dimensional datasets to minimize over-fitting (but this makes the model complex). The model may be under-fit on the training data if the regularization parameters are too high.

Application of Logistic Regression

All use cases where data must be categorized into multiple groups are covered by Logistic Regression. Consider the following illustration:

11. Fraud detection in Credit card
12. Email spam or ham
13. Sentiment Analysis in Twitter analysis
14. Image segmentation, recognition, and classification – X-rays, Scans
15. Object detection through video
16. Handwriting recognition
17. Disease prediction – Diabetes, Cancer, Parkinson etc...

When do we use KNN algorithm?

KNN can be used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry. To evaluate any technique we generally look at 3 important aspects:

1. Ease to interpret output
2. Calculation time
3. Predictive Power

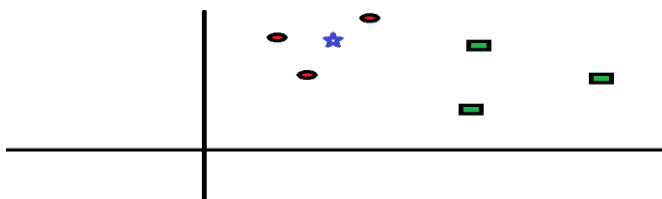
Let us take a few examples to place KNN in the scale :

	Logistic Regression	CART	Random Forest	KNN
1. Ease to interpret output	2	3	1	3
2. Calculation time	3	2	1	3
3. Predictive Power	2	2	3	2

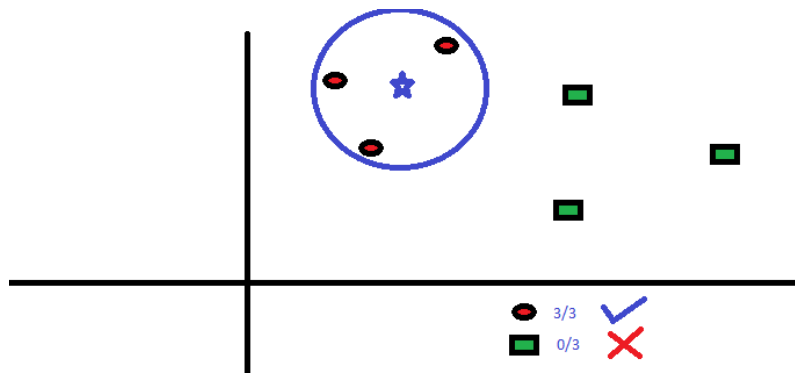
KNN algorithm fairs across all parameters of considerations. It is commonly used for its easy of interpretation and low calculation time.

How does the KNN algorithm work?

Let's take a simple case to understand this algorithm. Following is a spread of red circles (RC) and green squares (GS) :



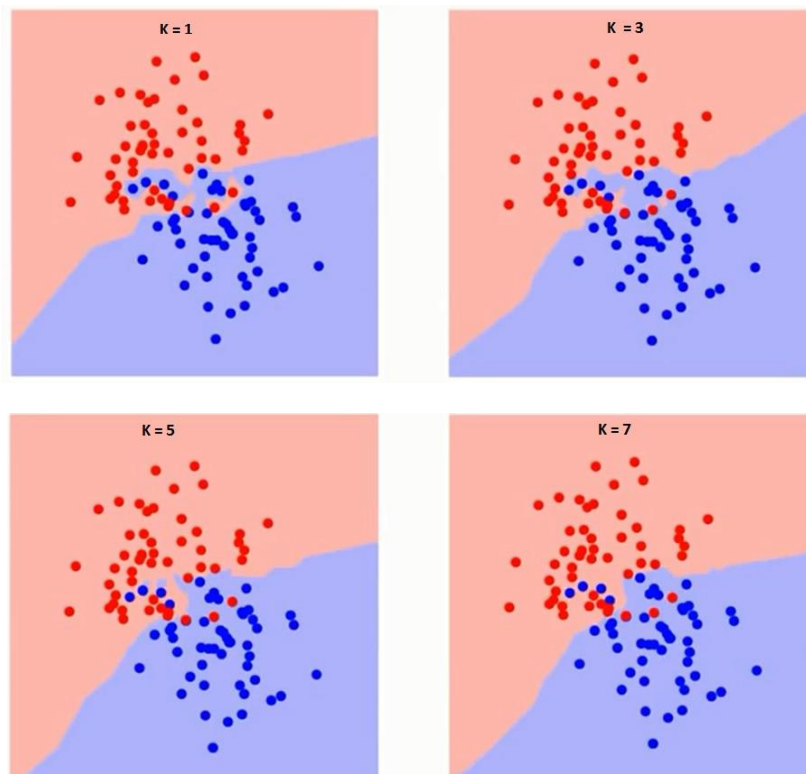
You intend to find out the class of the blue star (BS). BS can either be RC or GS and nothing else. The “K” in KNN algorithm is the nearest neighbor we wish to take the vote from. Let's say $K = 3$. Hence, we will now make a circle with BS as the center just as big as to enclose only three data points on the plane. Refer to the following diagram for more details:



The three closest points to BS is all RC. Hence, with a good confidence level, we can say that the BS should belong to the class RC. Here, the choice became very obvious as all three votes from the closest neighbor went to RC. The choice of the parameter K is very crucial in this algorithm. Next, we will understand what are the factors to be considered to conclude the best K.

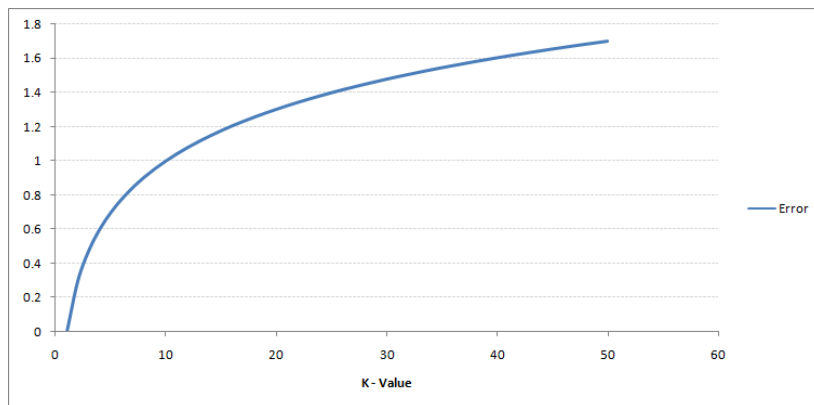
How do we choose the factor K?

First let us try to understand what exactly does K influence in the algorithm. If we see the last example, given that all the 6 training observation remain constant, with a given K value we can make boundaries of each class. These boundaries will segregate RC from GS. In the same way, let's

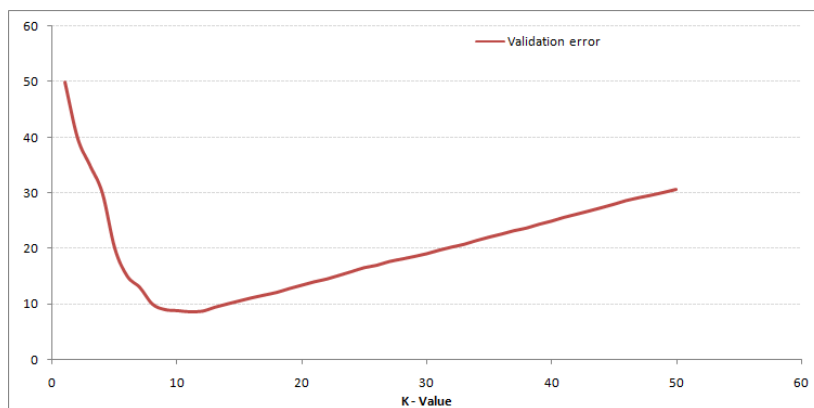


try to see the effect of value “K” on the class boundaries. The following are the different boundaries separating the two classes with different values of K.

If you watch carefully, you can see that the boundary becomes smoother with increasing value of K. With K increasing to infinity it finally becomes all blue or all red depending on the total majority. The training error rate and the validation error rate are two parameters we need to access different K-value. Following is the curve for the training error rate with a varying value of K :



As you can see, the error rate at K=1 is always zero for the training sample. This is because the closest point to any training data point is itself. Hence the prediction is always accurate with K=1. If validation error curve would have been similar, our choice of K would have been 1. Following is the validation error curve with varying value of K:



This makes the story more clear. At K=1, we were over fitting the boundaries. Hence, error rate initially decreases and reaches a minima. After the minima point, it then increase with increasing K. To get the optimal value of K, you can segregate the training and validation from the initial dataset. Now plot the validation error curve to get the optimal value of K. This value of K should be used for all predictions.

Breaking it Down – Pseudo Code of KNN

We can implement a KNN model by following the below steps:

1. Load the data
2. Initialize the value of k
3. For getting the predicted class, iterate from 1 to total number of training data points
 1. Calculate the distance between test data and each row of training data.
Here we will use Euclidean distance as our distance metric since it's the most popular method. The other metrics that can be used are Chebyshev, cosine, etc.
 2. Sort the calculated distances in ascending order based on distance values
 3. Get top k rows from the sorted array
 4. Get the most frequent class of these rows
 5. Return the predicted class

End Notes

KNN algorithm is one of the simplest classification algorithm. Even with such simplicity, it can give highly competitive results. KNN algorithm can also be used for regression problems. The only difference from the discussed methodology will be using averages of nearest neighbors rather than voting from nearest neighbors. KNN can be coded in a single line on R. I am yet to explore how can we use KNN algorithm on SAS.

Random Forest:

Introduction

Random forest is a **Supervised Machine Learning Algorithm** that is **used widely in Classification and Regression problems**. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing **continuous variables** as in the case of regression and **categorical variables** as in the case of classification. It performs better results for classification problems.

Real Life Analogy

Let's dive into a real-life analogy to understand this concept further. A student named X wants to choose a course after his 10+2, and he is confused about the choice of course based on his skill set. So he decides to consult various people like his cousins, teachers, parents, degree students, and working people. He asks them varied questions like why he should choose, job opportunities with that course, course fee, etc. Finally, after consulting various people about the course he decides to take the course suggested by most of the people.



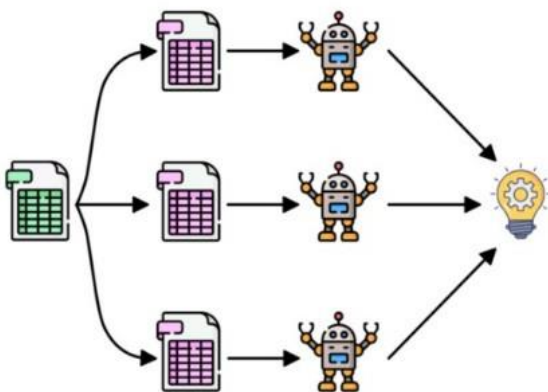
Working of Random Forest Algorithm

Before understanding the working of the random forest we must look into the ensemble technique. **Ensemble** simply means combining multiple models. Thus a collection of models is used to make predictions rather than an individual model.

Ensemble uses two types of methods:

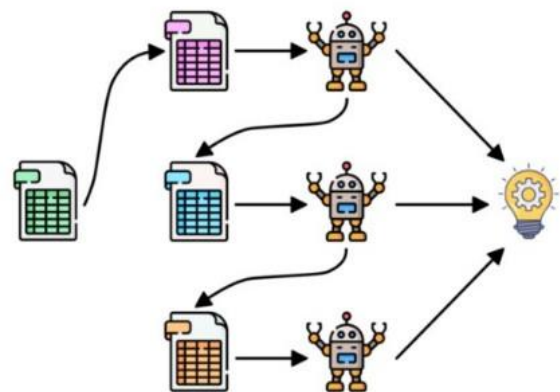
1. **Bagging**– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest.
2. **Boosting**– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST

Bagging



Parallel

Boosting

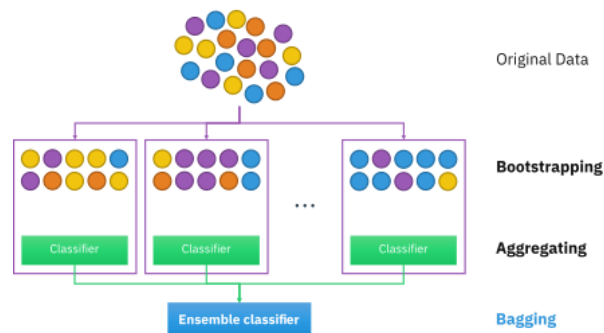


Sequential

As mentioned earlier, Random forest works on the Bagging principle. Now let's dive in and understand bagging in detail.

Bagging

Bagging, also known as **Bootstrap Aggregation** is the ensemble technique used by random forest. Bagging chooses a random sample from the data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as **row sampling**. This step of row sampling with replacement is called **bootstrap**. Now each model is trained independently which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting is known as **aggregation**.



Now let's look at an example by breaking it down with the help of the following figure. Here the bootstrap sample is taken from actual data (Bootstrap sample 01, Bootstrap sample 02, and Bootstrap sample 03) with a replacement which means there is a high possibility that each sample won't contain unique data.

Now the model (Model 01, Model 02, and Model 03) obtained from this bootstrap sample is trained independently. Each model generates results as shown. Now Happy emoji is having a majority when compared to sad emoji. Thus based on majority voting final output is obtained as Happy emoji.

Steps involved in random forest algorithm:

Step 1: In Random forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on **Majority Voting or Averaging** for Classification and regression respectively.

For example: consider the fruit basket as the data as shown in the figure below. Now n number of samples are taken from the fruit basket and an individual decision tree is constructed for each sample. Each decision tree will generate an output as shown in the figure. The final output is considered based on majority voting. In the below figure you can see that the majority decision tree gives output as an apple when compared to a banana, so the final output is taken as an apple.

Important Features of Random Forest

1. **Diversity-** Not all attributes/variables/features are considered while making an individual tree, each tree is different.
2. **Immune to the curse of dimensionality-** Since each tree does not consider all the features, the featurespace is reduced.
3. **Parallelization-** Each tree is created independently out of different data and attributes. This means that we can make full use of the CPU to build random forests.
4. **Train-Test split-** In a random forest we don't have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree.
5. **Stability-** Stability arises because the result is based on majority voting/averaging.

Difference Between Decision Tree & Random Forest

Random forest is a collection of decision trees; still, there are a lot of differences in their behavior.

Decision trees	Random Forest
1. Decision trees normally suffer from the problem of overfitting if it's allowed to grow without any control.	1. Random forests are created from subsets of data and the final output is based on average or majority ranking and hence the problem of overfitting is taken care of.
2. A single decision tree is faster in computation.	2. It is comparatively slower.
3. When a data set with features is taken as input by a decision tree it will formulate some set of rules to do prediction.	3. Random forest randomly selects observations, builds a decision tree and the average result is taken. It doesn't use any set of formulas.

Thus random forests are much more successful than decision trees only if the trees are diverse and acceptable.

Important Hyper parameters

Hyper parameters are used in random forests to either enhance the performance and predictive power of models or to make the model faster.

Following hyper parameters increases the predictive power:

1. **n_estimators**– number of trees the algorithm builds before averaging the predictions.
2. **max_features**– maximum number of features random forest considers splitting a node.
3. **mini_sample_leaf**– determines the minimum number of leaves required to split an internal node.

Following hyperparameters increases the speed:

1. **n_jobs**– it tells the engine how many processors it is allowed to use. If the value is 1, it can use only one processor but if the value is -1 there is no limit.
2. **random_state**– controls randomness of the sample. The model will always produce the same results if it has a definite value of random state and if it has been given the same hyper parameters and the same training data.
3. **oob_score** – OOB means out of the bag. It is a random forest cross-validation method. In this one-third of the sample is not used to train the data instead used to evaluate its performance. These samples are called out of bag samples.

Advantages and Disadvantages of Random Forest Algorithm

Advantages

1. It can be used in classification and regression problems.
2. It solves the problem of over fitting as output is based on majority voting or averaging.
3. It performs well even if the data contains null/missing values.
4. Each decision tree created is independent of the other thus it shows the property of parallelization.
5. It is highly stable as the average answers given by a large number of trees are taken.

6. It maintains diversity as all the attributes are not considered while making each decision tree though it is not true in all cases.
7. It is immune to the curse of dimensionality. Since each tree does not consider all the attributes, feature space is reduced.
8. We don't have to segregate data into train and test as there will always be 30% of the data which is not seen by the decision tree made out of bootstrap.

Disadvantages

1. Random forest is highly complex when compared to decision trees where decisions can be made by following the path of the tree.
 2. Training time is more compared to other models due to its complexity.
- Whenever it has to make a prediction each decision tree has to generate output for the given input data.

Summary

Now, we can conclude that Random Forest is one of the best techniques with high performance which is widely used in various industries for its efficiency. It can handle binary, continuous, and categorical data.

Random forest is a great choice if anyone wants to build the model fast and efficiently as one of the best things about the random forest is it can handle missing values.

Overall, random forest is a fast, simple, flexible, and robust model with some limitations.

Chapter 6

Conclusion:

Clinical finding is a significant region of exploration which assists with recognizing the event of a coronary illness. The framework, utilizing different methods referenced, will thus uncover the root coronary illness alongside the arrangement of most plausible heart Diseases which have comparative side effects. The information base utilized is a portrayal data set so to decrease the dataset tokenization, separating and stemming is finished. The venture presents a novel mixture model to recognize and affirm CAD cases requiring little to no effort by utilizing clinical information that can be effectively gathered at clinics. Intricacy of framework is diminished by decreasing the dimensionality of the informational collection with PSO. It gives reproducible and target finding, and subsequently can be a significant extra device in clinical practices. Results are equivalently, encouraging and along these lines the proposed half and half technique will be useful in coronary illness diagnostics. Trial results exhibit the predominance of the proposed half breed technique concerning forecast precision of CAD.

REFERENCE PAPERS:

- [1] S. I. Ansarullah and P. Kumar, “A systematic literature review on cardiovascular disorder identification using knowledge mining and machine learning method,” *Int. J. Recent Technol. Eng.*, vol. 7, no. 6S, pp. 1009–1015, 2019
- [2] A. U. Haq, J. P. Li, J. Khan, M. H. Memon, S. Nazir, S. Ahmad, G. A. Khan, and A. Ali, “Intelligent machine learning approach for effective recognition of diabetes in E-healthcare using clinical data,” *Sensors*, vol. 20, no. 9, p. 2649, May 2020
- [3] A. U. Haq, J. Li, M. H. Memon, M. H. Memon, J. Khan, and S. M. Marium, “Heart disease prediction system using model of machine learning and sequential backward selection algorithm for features selection,” in *Proc. IEEE 5th Int. Conf. Converg. Technol. (ICT)*, Mar. 2019, pp. 1–4
- [4] U. Haq, J. Li, M. H. Memon, J. Khan, and S. U. Din, “A novel integrated diagnosis method for breast cancer detection,” *J. Intell. Fuzzy Syst.*, vol. 38, no. 2, pp. 2383–2398, 2020.