JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY

ANANTAPUR

DIABETES DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS

A project report Submitted to the

Jawaharlal Nehru Technological University Anantapur.

In partial fulfilment for the Award of the degree of

BACHELOR OF TECHNOLOGY
IN
ELECTRONICS AND COMMUNICATION ENGINEERING
By

D.VIJAY KUMAR
(H.T. No: 19JN1A0432)

D. KAMAL KALYAN
(H.T. No: 19JN1A0436)

D.VINAY KUMAR
(H.T. No: 19JN4A0428)

CH. SATISH CHANDRA
(H.T. No: 19JN1A0424)

E. THARUN
(H.T. No: 19JN1A0439)

Under the Esteemed Guidance of

Mr. P. RAVI KUMAR, M. Tech,
Assistant Professor

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

SREE VENKATESWARA COLLEGE OF ENGINEERING

NAAC 'A' Grade Accredited Institution, An ISO 9001::2015 Certified Institution (Approved by AICTE, New Delhi and Affiliated to JNTU, Anantapur) NORTHRAJUPALEM(VI), KODAVALURU(M), S.P.S.R NELLORE (DT) – 524 316

2019-2023

SREE VENKATESWARA COLLEGE OF ENGINEERING

NAAC 'A' Grade Accredited Institution, An ISO 9001::2015 Certified Institution
(Approved by AICTE, New Delhi and Affiliated to JNTU, Anantapur)
NORTHRAJUPALEM(VI), KODAVALURU(M), S.P.S.R NELLORE (DT) – 524 316

Department of
ELECTRONICS AND COMMUNICATION ENGINEERING

CERTIFICATE

Certified that the project entitled "**DIABETES DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS**" submitted by **D.KAMAL KALYAN(H.T. No. 19JN1A0436), D.VIJAY KUMAR (H.T. No. 19JN1A0432), CH. VINAY KUMAR (H.T. No. 19JN1A0428), CH. SATISH CHANDRA (H.T. No. 19JN1A0424), E. THARUN (H.T. No. 19JN1A0439)** in partial fulfilment for the award of degree of **BACHELOR OF TECHNOLOGY** in **ELECTRONICS AND COMMUNICATION ENGINEERING** by **Jawaharlal Nehru Technological University, Anantapur** during the academic year **2022-2023.** It is certified that all corrections/suggestions indicated for internal assessment have been incorporated in the report. The project report has been approved as it satisfies the academic requirements in respect of project work prescribed for the said degree.

Signature of the guide
Mr. P. Ravi Kumar

Signature of Head of the Dept.
Dr. D. Rajani

Principal
Dr. P. Kumar Babu

Internal Examiner

External Examiner

ACKNOWLEDGEMENT

We are highly thankful to **Mr. P. RAVI KUMAR, Assistant Professor** in the **Department of Electronics and Communication Engineering** for his inspiring guidance and for providing the background knowledge to deal with the problem at every phase of our project in a systematic manner.

We have immense pleasure in expressing our sincere thanks and deep sense of gratitude to **Dr. D. Rajani, Head of the Department** of Electronics and Communication Engineering for extending necessary facilities for the completion of the project.

With immense respect, we express our sincere gratitude to **Dr. P. Babu Naidu, Chairman,** and **Dr. P. Kumar Babu, Principal** for permitting us to take up our project work and to complete the project successfully.

We would like to express our sincere thanks to all faculty members of the Department for their continuous cooperation, which has given us the cogency to build-up adamant aspiration over the completion of our project.

Finally, we thank one and all who directly and indirectly helped us to complete the project successfully.

D. Kamal Kalyan (H.T. No. 19JN1A0436)
D. Vijay Kumar (H.T. No. 19JN1A0432)
CH. Vinay Kumar (H.T. No. 19JN1A0428)
CH. Satish Chandra (H.T. No. 19JN1A0424)
E. Tharun (H.T. No. 19JN1A0439)

Dept. of Electronics and Communication Engineering
Sree Venkateshwara College of Engineering,
NH5 Bypass Road, North Rajupalem, Nellore.

DECLARATION

We **D. KAMAL KALYAN (H.T. No. 19JN1A0436), D. VIJAY KUMAR (H.T. No. 19JN1A0432), CH. VINAY KUMAR (H.T. No. 19JN1A0428), CH. SATISH CHANDRA (H.T. No. 19JN1A0424) and E. THARUN (H.T. No. 19JN1A0439)** do hear by declare that the project entitled "**DIABETES DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS**" was carried out by our batch under the guidance of **Mr. P. RAVI KUMAR.** This Project work submitted to Jawaharlal Nehru Technological University, Anantapur in fulfillment of requirement for the award of **BACHELOR OF TECHNOLOGY in ELECTRONICS AND COMMUNICATIONENGINEERING** during the academic year 2021-2022.

Project associates,

D. Kamal kalyan	(H.T. No. 19JN1A0436)
D. Vijay Kumar	(H.T. No. 19JN1A0432)
CH. Vinay Kumar	(H.T. No. 19JN1A0428)
CH. Satish Chandra	(H.T. No. 19JN1A0424)
E. Tharun	(H.T. No. 19JN1A0439)

CONTENTS

LIST OF FIGURES

ABBREVATIONS

KNN	-	K-NEAREST NEIGHBOUR ALGORITHM
SVM	-	SUPPORT VECTOR MACHINE
APCA	-	ADAPTIVE PRINCIPAL COMPONENT ANALYSIS
HOMED	-	HYBRID OUTLINE MODEL FOR EARLY DETECTION
PCA	-	PRINCIPAL COMPONENT ANALYSIS
PIDD	-	PIMA INDIAN DIABETES DATASET
ML	-	MACHINE LEARNING
CDSS	-	CLINICAL DECISION SUPPORT SYSTEM
IDF	-	INTERNATIONAL DIABETES FEDERATION
PD	-	PRE-DIABETES
CNN	-	COVOLUTION NEURAL NETWORK
ANN	-	ARTIFICIAL NEURAL NETWORK
DNN	-	DEEP NEURAL NETWORK
IBL	-	INSTANCE BASED LEARNING
GDS	-	GESTATIONAL DIABETES MELLITUS

ABSTRACT

Diabetes is among critical diseases and lots of people are suffering from this disease. Age, obesity, lack of exercise, hereditary diabetes, living style, bad diet, high blood pressure, etc. can cause Diabetes. People having diabetes have high risk of diseases like heart disease, kidney disease, stroke, eye problem, nerve damage, etc. Current practice in hospital is to collect required information for diabetes diagnosis through various tests and appropriate treatment is provided based on diagnosis.

Big Data Analytics plays a significant role in healthcare industries. Healthcare industries have large volume databases. Using big data analytics one can study huge datasets and find hidden information, hidden patterns to discover knowledge from the data and predict outcomes accordingly. In existing method, the classification and prediction accuracy are not so high.

This project deals with the prediction of Diabetes Disease by performing an analysis of five supervised machine learning algorithms, i.e., K-Nearest Neighbors, Naïve Baye, Decision Tree Classifier, Random Forest and Support Vector Machine. Further, by incorporating all the present risk factors of the dataset, we have observed a stable accuracy after classifying and performing cross-validation. We managed to achieve a stable and highest accuracy of 76% with KNN classifier and remaining all other classifiers also give a stable accuracy of above 70%.

We analyzed why specific Machine Learning classifiers do not yield stable and good accuracy by visualizing the training and testing accuracy and examining model overfitting and model underfitting. The main goal of this project is to find the most optimal results in terms of accuracy and computational time for Diabetes disease prediction.

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

In several countries including United States (Raghupathi & Raghupathi 2020), China (Zhou *et al.* 2017), UK (Grosios *et al.* 2010) healthcare is a thriving domain that has industrialized as a leading sector related to economic growth and employment as well as functional overheads. From the study of the global spending on healthcare, it is anticipated that the general healthcare expenditure will reach double in the following 5 years (Barlow *et al.* 2021). In India, public health spending is expected to increase to ₹486000 crores by 2022 from ₹267000 crores in the year 2018 (Ministry of Health & Welfare 2017). The overall financial outlay to health and wellbeing is expected to rise 45% from the financial year 2018 to 2022. Non-value added and unproductive activities including diagnostic and medication errors, improper usage of antibiotics, readmissions, and deception create a substantial amount of spending for care. Approximately, 5.2 million Indians demise annually due to clinical faults and the consequences of malicious activities (Patel *et al.* 2018).

A suitable Clinical Decision Support System (CDSS) can handle these problems and accelerate the transformation towards a value-based clinical system to deliver passable care and high services (Sutton *et al.* 2020). Digital therapeutics has been recognized to be a renowned intervention for treatment in disease control, and both medical professionals and patients are benefitting from CDSS. At present, the healthcare industry is adopting information and communication technologies for its organizational venture to deliver high-quality care services in both commercial and technical aspects (Sheikh *et al.* 2021).

CDSS is a system for increasing the quality of care by calculating the risk of disease progression and medical procedures. It supports medical professionals and patients in making appropriate clinical decisions. It has made obstacles associated with disease management and the threat of high morbidity and mortality complications. Maintaining plasma sugar levels, logically controlling nutrition, and

effectively regulating the treatment plan are all critical for COVID-19 patients (Lou *et al.* 2020).

The only way for the diabetic patient to live with this disease is to preserve the blood glucose level as normal as possible (i.e., fasting blood glucose ≥ 126mg/dl (7.0mmol/l) or 2–h blood glucose during the 75-g oral glucose tolerance test ≥ 200mg/dl (11.1mmol/l)) without extreme higher or lower levels (Buysschaert *et al.* 2016), and this is achieved when the patient undergoes a proper medication which may include consuming oral drugs or some form of insulin, exercise, and nutrition. Furthermore, managing DM is also a perplexing, expensive, and difficult task for medical experts. There are huge vital data to store about the patients and diseases that support the doctors in making optimum clinical verdicts to improve the life expectancy of the patients. This makes the conventional machine learning-based diagnosis approaches stop or degrades the training procedures as the algorithm becomes susceptible to over-fitting due to the huge irrelevant and redundant attributes in a high-dimensional database.

Several DDS use ML techniques to identify DM and disease management systems. However, the inherent characteristics such as trapping into local optimum solution, lack of privacy, missing values in the input dataset, and deficiency of incremental classification are major issues related to conventional machine learning-based diabetes classification algorithms. Conventional ML-based DDS has been originally developed as non-incremental (offline) methods that are trained with predefined datasets before they can be employed to solve the diabetes classification problems. Furthermore, traditional supervised machine learning approaches often cannot be performed incrementally and therefore they need to collect the entire database again to classify new cases effectively. Besides, since medical databases need continuous data appraisal, it is essential to incrementally appraise the learned models to reduce data classification processing time and improve storage complexity efficiency.

Sensitive user health data are supposed to be managed and shared using

healthcare informatics ion in the form of EHR to support care delivery and disease management system (Zheng *et al.* 2017). Patients are unwilling to reveal their information other than the treatment tenacities and they want to be effectively well-versed about data communication. Hence, it is very imperative to secure sensitive information of patients for enabling them to continue sharing data related to their health condition. By considering these issues, this work targets to develop a secured online diabetes classification model where the accuracy of the classifier is increased significantly while preserving the privacy of the sensitive user data.

For collecting and processing data regarding the DM disease, it is essential in developing and deploying approaches that allow for the accurate selection of the most pertinent intangible size and their configurations, the assimilation of current healthcare information from various sources, and the construction of structures that denote the medical expert's inference system.

It is essential to consider that these techniques should deliver comprehensible data representation models that help doctors make appropriate verdicts according to information, signs, and comorbidities from the diabetics. The best solution that considers these demands is ontology.

Bearing the above issues in mind, this study proposes three diabetes classification frameworks based on (i) a hybrid optimizer-based SVM using CSA and BGWO; (ii) a hybrid online model, called HOMED using an ISVM and an APCA method; and (iii) an Ont-SVM using Kronecker product and CSA-based parameter optimization techniques. In the first work, a hybrid optimizer using SVM is proposed to develop an effective DDS. This system assimilates a CSA and BGWO for exploiting the entire potential of SVM in the DDS. The performance of the established classifier (i.e., CS-BGWO-SVM) is cautiously analysed on the real-life datasets including the PIDD and the DTD. To assess the performance of the proposed classifier, its enactment is compared with several state-of-the-art classifiers using SVM in terms of performance measures including classification accuracy, IoU, specificity, sensitivity, and the AUC. The empirical results reveal that CS-BGWO-SVM can be considered as a more

effective classifier with greater detection performance. Also, the Wilcoxon statistical test is carried out to find out whether the classifier provides a considerable improvement with respect to evaluation metrics or not.

In the second work, a hybrid online model for early detection of diabetes disease is proposed to integrate an APCA method for missing value imputation, data clustering, and attribute extraction with an ISVM for classification. The performance of HOMED is evaluated on PIDD based on different evaluation metrics. The experimental results prove that HOMED significantly improves the predictive accuracy and reduces processing overhead related to other offline approaches.

In the third approach, this work targets to develop a secured diabetes classification model where the privacy of the sensitive user data is protected. The recommended model protects user data using Kronecker product-based CSA optimization approach. Then, the domain ontology is developed to detect DM by determining clinical resemblances between the trained dataset with ontological rules. Hence, this work implements the Ont-SVM classification model by integrating the concept of privacy protection, ontology, and the SVM-based classification approach. The empirical analysis on the PIDD demonstrates that the presented Ont-SVM outdoes other related state-of-the-art methods found in the literature with respect to performance measures. The effectiveness of all three approaches is appraised on real-world datasets collected from open-source repositories http://www.ics.uci.edu/~mlearn/ML.Repositpry.html and Data. World datasets repository. https://data.world/). These proposed algorithms can assist clinicians to facilitate secured and precise incremental classification of DM with better accuracy and other performance measures.

1.1.1 DIABETES MELLITUS

DM has been investigated for several centuries. The 5th-century doctor Aretaeus first introduced the word ―diabetes‖ to define the disease as a ―melting down of limbs and flesh into urine‖. Indian doctors during the 5th

century, BC designated the sweet, honey-like flavor of urine in polyuric patients that enticed ants and other insects, but the term ―Mellitus‖ (Latin for ―honey‖) was appended in the 17th.century. DM is a long-term and non-communicable disease with relative and/or complete insulin shortage. The source of diabetes can differ significantly but always take in the defect of the pancreas to discharge sufficient insulin or the body ‘s cells do not respond well to insulin or in both at some point in the disease period.

DM development is sturdily connected with the major disorders in the metabolism of glucose, carbohydrate, protein, and fat. Besides, it is related to declined life expectancy, substantial morbidity owing to diabetes-oriented microvascular problems (e.g., neuropathy, nephropathy, and retinopathy), macrovascular problems (e.g., peripheral vascular disease, stroke, and heart disease), infections, and other consequences (e.g., COVID-19, nonalcoholic fatty liver disease, dental disease, dyslipidemia, and psychosocial problems), and reduced quality of life.DM can be classified into 4categories as given below:

1. **Type 1 Diabetes mellitus (T1DM):** Juvenile diabetes or T1DM distresses adults and children (i.e., 18 years old or below). It is primarily due to the scarcity of insulin secreted by the β-cells in the pancreas that does not produce sufficient insulin. According to National Diabetes Statics Report (Patel & Macerollo 2010), T1DM finds around 5% to 10% of all identified cases worldwide. Patients with T1DM require external insulin treatment regularly to regulate plasma sugar levels. The development of T1DM is related to genetics and ecological aspects.

2. **Type 2 Diabetes mellitus (T2DM):** It is non-insulin-dependent diabetes and accounts for 90% to 95% of DM cases found in adult individuals (Patel & Macerollo 2010). T2DM is driven by insulin opposition, fat, and liver cells which are scarce to consume insulin effectively. The hazard of T2DM problems is associated with age,

obesity, sloth, unhealthy diet, history of gestational diabetes, heredity, and devastation of glucose metabolism.

3. **Prediabetes (PD):** It is a milder form of DM wherein patients have high plasma sugar. It is also called reduced glucose tolerance. Additionally, individuals with prediabetes have a high risk of having T2DM. It can be diagnosed by a simple laboratory test (Patel & Macerollo 2010).

4. **Gestational diabetes mellitus (GDM):** GDM is impaired glucose tolerance diagnosed with onset or first recognition during gestation that naturally resolves after delivery. According to the National Diabetes Statics Report, 50% of women with GDM continue to have more plasma glucose level and are consecutively diagnosed as diabetes, mostly T2DM (Patel & Macerollo 2010). Besides, the children of women who had GDM during pregnancy may be vulnerable of suffering from diabetes and being overweight later.

All the above-mentioned types of diabetes lead to distressing hitches if not managed effectively. There are several difficulties in the effective disease management of DM since economic and personal overheads are associated with DM treatment. Its enduring significances interpret into human suffering and economic overheads. On the other hand, inclusive diabetes care can reduce the progress of problems, healthcare outlays, and maximize the quality of life. Hence, there is no qualm that DM demands inordinate attention. At the same time, ML-based DDS approaches can help individuals make a primary decision about DM based on their regular screening test results, and it can act as a reference for medical professionals.

1.1.2 SIGNIFICANCE OF DM DIAGNOSIS SYSTEM

As mentioned earlier, DM has become a vital concern in the medical domain to determine appropriate solutions to combat the disease. It is important

to intervene not only to treat but also to prevent as well as make an early disease diagnosis. The algorithms applied in data mining, ML or any other domain of artificial intelligence achieve prognostic modelling in DM detection, that is, the utilization of data and knowledge to calculate impending outcomes using previous data. These techniques are employed for the timely prediction and detection of DM to minimize the rate of morbidity. The most general signs of DM include hyperglycemia, anomalous metabolism, and related risk for problems affecting the nervous system, kidneys, and eyes, which are main parts of the human body. Such signs are employed to collect clinical data, and then the classification is carried out according to those symptoms.

Essentially, the process of diabetes detection is a data classification problem. Data classification is generally described as the method of labelling hidden data patterns (Li *et al.* 2017). From the ML viewpoint, classification is the problem of classifying a collection of data samples into appropriate categories based on the learning result of a subset of the dataset whose correct class is recognized. Many clinicians and scholars have proposed different DDS using innovative technologies to detect diabetes mellitus. As technology continues to progress, the development of such system is becoming more effective. Generally, predicting diabetes disease is a captious task. These approaches exploit the technological advances that allow machines to learn.

1.1.3 ML TECHNIQUES: THE FUTURE FOR DIABETES CARE

There has been a colossal impact in the healthcare sector with the proliferation of cutting-edge technologies including artificial intelligence, data mining, machine learning, Internet-of-medical things, etc. These techniques have demonstrated their effectiveness in disease prediction from massive amounts of healthcare data (Singla *et al.* 2019). They are used to perform prognostic modeling, that is, the exploitation of data and statistics to calculate imminent results from past experiences. Implementation of ML techniques can greatly improve the reach of diabetes care thus making it more effective. ML

algorithms find extensive use in 4 major fields in diabetes care, such as automatic retinal screening, medical decision support, prognostic population risk stratification, and patient self-management tools (van Gemert-Pijnen *et al.* 2011; Dankwa-Mullan *et al.* 2019). It led to a transformation in the diabetic disease management system using data-driven healthcare care. It has transformed the way diabetes is prevented, identified, and managed, which can help in reducing the worldwide prevalence of 8.8% (Ellahham 2020).

Various ML techniques (e.g., naïve Bayes (NB), Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), SVM, etc.) and some deep learning techniques (e.g., Multilayer Perceptron (MLP), Artificial Neural Networks (ANN), Recurrent Neural Networks (RNN), Deep Neural Network (DNN), Convolutional Neural Networks (CNN), deep belief network, etc.) have been constructively used in practice to detect diabetes (Sharma & Shah 2021). Numerous researchers and clinicians have intensified efforts to increase classification performance to achieve superior results when detecting or diagnosing DM. Due to its strong potential to distinguish the _inherent features of different classes of data samples SVM is frequently employed ML technique for the classification of DM (Sharma & Shah 2021).

The SVM-based classification includes two phases including learning and testing. In the learning phase, this classification algorithm detects the predictive features by training data samples. During the testing process, the system classifies the newly arrived data sample according to those features and labels them to the equivalent category. SVM is a non-parametric classifier able to solve regression and classification problems through linear and non-linear functions. It is a discriminative classification algorithm to determine abnormalities of biological signs due to its superiority and strong potential to handle the non-linear and high-dimensional database in the domain of the medical sector (Vapnik 1995).The basic concept behind this classification algorithm is to distinguish the new testing dataset into their suitable classes according to the recognized learning dataset.

1.2 RESEARCH MOTIVATION

DM is spiraling out of the control. It is a significant global challenge to the well-being and health of people, families, and societies. Based on the IDF report released in 2019, one in ten adults are living with DM worldwide and almost half are unidentified. Besides, the high incidence of DM makes it vital comorbidity in patients with COVID-19. In the year 2019, the top three countries with the maximum number of people with DM are China (116.4 million), India (77.0 million), and the USA (31.0 million). This trend is projected to endure in 2030 and 2045, with China (140.5 and 147.2 million) and India (101.0 and 134.2 million) continuing to have the maximum outlay of DM (Rajendra & Mohan 2021).

India has faced a recent sharp upsurge in prediabetes/diabetes where the occurrence of DM is projected at 8% - 10% with a somewhat higher prevalence in urban areas as compared to rural areas (Singla *et al.* 2019). On the other hand, the incidence of DM in the city of Delhi has been estimated at around 27% and it has been witnessed that 46% or more individuals have prediabetes (Deepa *et al.* 2015). A similar incidence has been observed in three other metropolitan cities. Based on the research report released by Singla *et al.* (2018) the maximum prevalence of DM is in the age group of 30 – 34 years. Such a timely and accurate detection of DM would be a vast outlay on the medical industry. The deficiency of healthcare services and specially qualified physicians are a barrier for the global healthcare industry. The utilization of ML in DDS can help in bridging this large gap. Uniformity of care (or minimum standard care) is another issue observed in India (Singla *et al.* 2019). Since numerous diabetic patients are being treated by primary health care doctors and owing to the absence of any audit of these practices, average number of people with DM in India stay at about 9% (Mohan *et al.* 2014). This envisages the possibly ever-increasing problem of impediments resulting from the poor diabetes control and also paves a way to make things better with the help of ML technique At present, most healthcare institutions are adopting ML techniques to provide high-quality medical services to their patients. The speed, effectiveness, dependability, and accuracy of the DDS can be ameliorated using

ML classification algorithms (Sharma & Shah 2021). At the same time, a huge volume of data is collected using this system continuously.

1.3 PROBLEM STATEMENT

Accurate disease detection and efficient disease management are two important issues in the medical industry and have a positive impact on an individual ‘s health condition. Healthcare organizations are not exposed much to the advanced information and communication technological contexts. They exploit the same clinical procedures for years. This frame of mind has created a frozen culture, and it is very difficult to utilize new procedures in medical sector. Large pandemic outbreaks like COVID-19 also impose new technical hitches for medical professionals, healthcare providers, investigators, and the public, as it generates a large volume of confidential data in the clinical decision support system.

Conventional approaches for data processing, storage, management, retrieval, and investigation are insufficient for the zettabytes of clinical data generated every year. Additionally, since clinical datasets become more assorted and intricate, cutting-edge data mining and computing methods are mandatory to make the data useful. Moreover, result reproducibility and data-communication strategies remain to be important deliberated issues. Most of the medical systems, for instance, intensive care units, diagnosis divisions, case administration units, medical trials, etc., all have got their own tactics for collecting data. Until today most of them utilize paper charts, paper folders, and fax to exchange information.

As mentioned previously, DM has been a severe global health hazard for many decades. Numerous studies have observed the unabated growing number of diabetics worldwide. It is a non-communicable syndrome characterized by higher plasma sugar levels and related to complaints of fat, protein, and carbohydrate disorders. Moreover, it worsens almost all organs of the body. The applications used for diabetes management mandate continuous medical treatment to reduce the long-haul risk and to prevent lethal complicate.

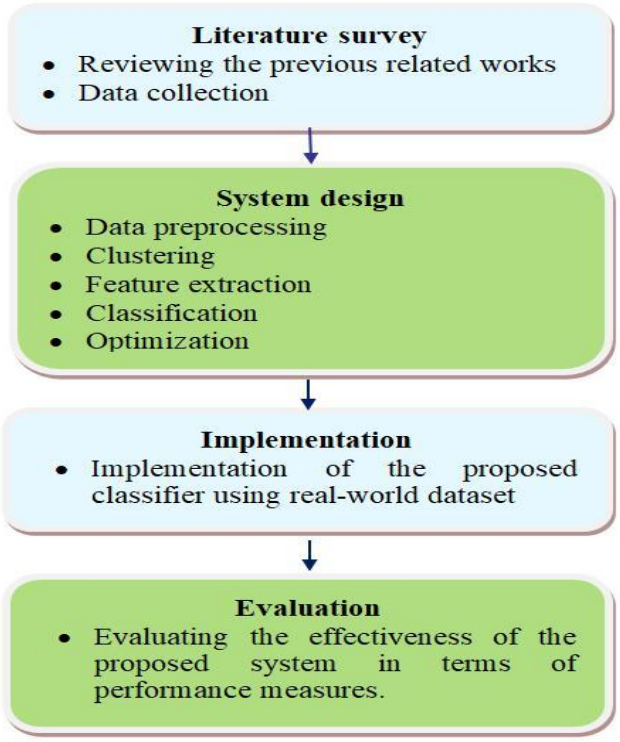

Figure 1.1: The research phases of the proposed system

1.4 RESEARCH CONTRIBUTION

The key contributions of this research are to design and implement an accurate diabetes detection and classification model with improved classification performance measures such as accuracy, IoU, precision, sensitivity, specificity, AUC, false-positive rate, false-negative rate, false discovery rate, PPV, and NPV.

This study targets to realize diabetes detection while preserving the confidentiality of the sensitive data. Hence, this work proposes and implements an efficient DDS using an SVM classification algorithm that is more suitable for handling this problem.

The selection of an SVM-based classifier is carried out through an extensive literature survey. The predictive performance is further improved by developing and implementing new bio-inspired optimizers. The effectiveness of these proposed approaches are cautiously evaluated on real-world datasets. The effectiveness of the proposed methods is confirmed by relating its performance

with other state-of-the-art ML-based classifiers with respect to evaluation metrics.

The key contributions of this work are summarized as follows:

1. An extensive literature survey with the germane theoretical contextual background of previous studies used to diagnose diabetes mellitus is carried out. The various state-of-the-art methods that are used for classifying DM are reviewed.

2. An optimized SVM classifier with hybrid optimization algorithms is selected for DDS to classify the input samples. This proposed system integrates CSA and BGWO optimization algorithms effectively to achieve better classification accuracy.

3. A hybrid online model is developed for early identification of diabetes disease using an adaptive PCA method for missing value imputation, data clustering, feature selection, and an incremental support vector machine algorithm for classification.

4. An ontology based SVM classifier for the diagnosis of diabetes is developed by integrating ontology, privacy protection, and SVM classification algorithm.

5. The domain ontology is developed for detecting DM by analyzing clinical resemblances between the trained dataset and ontological rules.

CHAPTER 2

BACKGROUND AND LITERATURE REVIEW

This chapter sets the nitty-gritty for the methodological innovations provided in the following chapters. It contains demonstration and quantitative studies employed in traditional diabetes classification approaches. Indeed, innumerable DDS and prediction techniques have been employed to detect, investigate, and understand diabetes to generate quicker and more accurate insights for making therapeutically useful decisions in prognosis and treatment planning. There are numerous research works, which have been used for the detection of DM, found in the literature. In this chapter, the related research works which focus on detecting and predicting diabetes datasets are discussed. Besides, this survey considers different online DDS to provide a comprehensive overview and the theoretical context of previous incremental diabetes detection systems. Then, the state-of-the-art in ML based diabetes classification techniques is explored to justify their application in the present study. This chapter also throws light on the feebleness of the extant methods that make them inappropriate for a DDS.

2.1 DIABETES DIAGNOSIS SYSTEM

Medical professionals, particularly physicians and nurses are stunned due to an enormous and startling increase in the number of patients during this COVID-19 pandemic. In such cases, artificial intelligence approaches could be employed to identify an individual with serious infections. Especially, diseases that upturn the threat of death and hospitalization in coronavirus patients, including heart disease, high blood pressure, and diabetes mellitus should be identified at an early stage. The data-intensive characteristic of diabetes identification and therapy makes it apt for integrating ML techniques to improve results and select advanced methods. In the selected primary studies of DDS, the researchers and clinicians used different ML techniques for providing accurate

results with improved performance. However, there is still a need to carry out a survey to explore which method is optimal. This survey targets to find fields of recommendations and opportunities in the classification of DM using ML techniques. It also studies the best evaluation measures, the databases employed to design the models, and the optimization methods applied to increase the performance of the classification process.

Various diabetes identification approaches using artificial intelligence and ML techniques have been designed and implemented against diabetes datasets are found in literature, but there is still a need to carry out a survey to find out the optimum one. The key objective of this work is to develop an efficient diabetes detection model by exploiting ML-based approaches including SVM. Generally, predicting diabetes disease is a captious task. Modern DDSs exploit the advanced cutting-edge technologies that allow machines to learn. The challenge related to developing such systems is to focus on a different set of data patterns and necessarily create new implications from the early statistics, with or without human interventions. The most appropriate

ML algorithm for designing DDS is generally designated based on their classification performance and accuracy.

2.2 BASICS OF ML APPROACHES

Recently, ML algorithms gained great attention from researchers and clinicians due to the accessibility of huge databases integrated with the enhancement in algorithms and the increased processing capacity of the systems. Indeed, ML is the technical domain handling the ways in which machines learn from past results. Several researchers often used the terms ―machine learning‖ and ―artificial intelligence‖ interchangeably (however they are not the same), since the ability of learning is the key feature of an object known as intelligence. More precisely, the objective of artificial intelligence is to create an intelligent agent or assistant that employs various ML approach-based solutions. The

purpose of ML is to design computerized models that can familiarize and learn from their previous statistics. A more comprehensive and prescribed definition of ML is given by Mitchell (1997). ―A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.‖

Currently, ML approaches are effectively employed for regression, clustering, classification, or dimensionality reduction processes of huge particularly high-dimensional datasets. Furthermore, ML has demonstrated to have exceptional aptitudes in various domains including playing go online (Omidshafiei *et al.* 2020), autonomous vehicles (Bachute & Subhedar 2021), image processing (Koulali *et al.* 2021), etc. Thus, large parts of our everyday life, for instance, web searches, image and speech recognition, email/spam filtering, fraud detection, credit scores, and many more are driven by efficient ML techniques (Sharma *et al.* 2021). ML approaches are categorized into three broad classes as follows (Sharma *et al.* 2021):

(i) predictive or supervised learning, where the system deduces a function from categorized learning dataset;

(ii) descriptive or unsupervised learning, where the training phase attempts to deduce the format of unknown data sample;

(iii) reinforcement learning, where the system communicates with a dynamic environment to collect data.

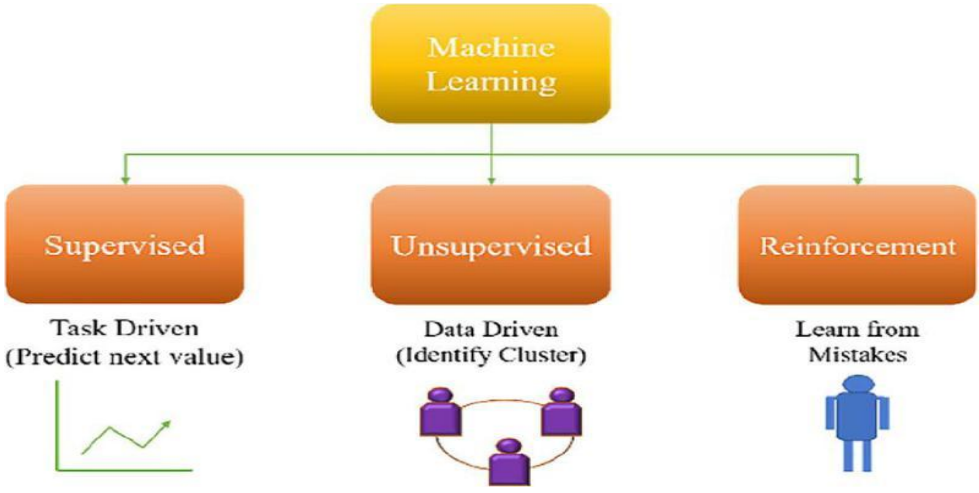

Figure 2.1 Types of machine learning approaches

1. **Supervised or predictive learning:** In supervised learning, the machine must ―learn‖ inductively a function known as an objective function, which is an expression of a model relating to the data (Sarker 2021). The target function is employed to calculate the dependent parameter (output) from a group of independent parameters (input).

The collection of feasible input parameters of the function is known as domains. Every instance in the domain is defined by a set of features or attributes SVM, ANN, Genetic Algorithms (GA), *k*-Nearest Neighbours (*k*-NN), Instance-Based Learning (IBL), DT, and rule learning are examples of supervised ML algorithms.

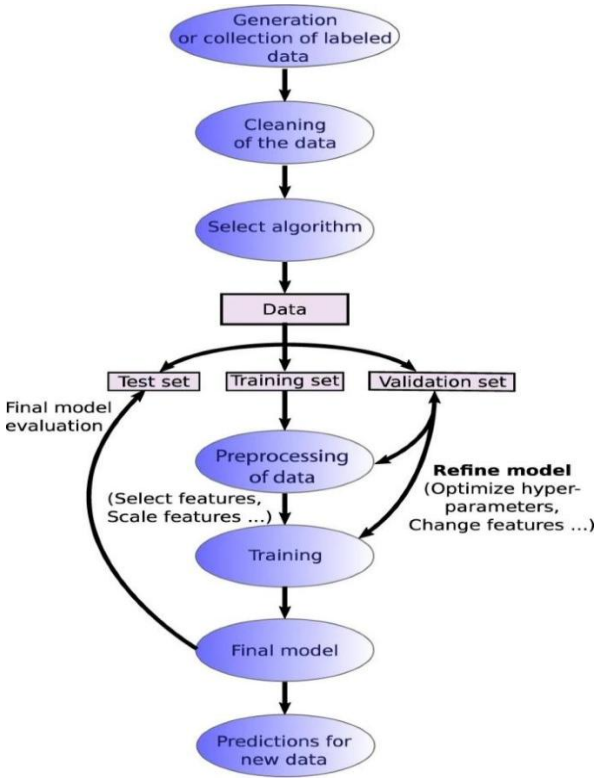

Figure 2.2 Structural outline of supervised ML approach

1. **Unsupervised or descriptive learning:** In this type of machine learning approach, the system attempts to learn the unknown pattern of data or correlation between parameters. In this case, learning data includes cases without any equivalent labels. More precisely, an unsupervised algorithm

learns some features of input data. After providing a new dataset, it exploits previously trained features to classify the data.

It is generally selected for data clustering and feature space reduction. Sparse Coding and Deep Belief Networks (DBNs) are the two renowned approaches to unsupervised ML models (Harrou *et al.* 2021).

4. **Reinforcement learning:** This approach focuses on the problem of learning a behavioral policy, a mapping from conditions or circumstances to activities (Botvinick *et al.* 2019). In this approach, the system tries to learn over direct contact with the atmosphere in order to increase the cumulative reward. The activities are based on the decision taken to facilitate the outcomes become more valuable at the output or anticipated satisfactory condition. But, the learner has no prior knowledge about the performance of the setting, and the only way to realize this is using trial and error. After providing the state values, it learns to decide which action to be considered for the current situation.

The current and upcoming condition is hinged on the decision taken by the learning process (i.e., action taken). This approach is completely based on two constraints: delayed output and trial and error searching. This approach is mostly used for independent systems, owing to its autonomous features regarding its environment.

2.3 APPLICATIONS OF ML APPROACHES

ML approaches are employed to efficiently categorize complex and high-dimensional databases. They enable investigators and clinicians to analyze unknown patterns of clinical datasets for envisaging the anticipated results and minimizing the overheads of classifying serious diseases (Ahamed *et al.* 2021) ML approaches are trained with real-world clinical databases having diverse attributes and external parameters. Different research works are performed with several ML methods for the early detection and classification of diabetes disease. Varma *et al.* (2014) proposed a DT-based classification model for diagnosing

diabetes. Conversely, traditional DT algorithms are deprived by an issue of crisp limits. This work also develops a fuzzy-based computational intelligence method for excluding the sharp edges to increase the classification accuracy. It employs 336 samples for experimentation, which are evaluated using MATLAB R2018b software, and provide 75.8% classification accuracy. Ali *et al.* (2020) applied different kinds of k-NN approaches (Fine, Medium, Weighted, and Cubic) to detect and predict DM. The dataset is constructed according to the guidelines given by the American Diabetes Association. In the learning phase, 4900 data instances are used for the training process of the classification approach to evaluate the output. Subsequently, the remaining 100 data instances are used for testing. The empirical analysis reveals that the Fine k-NN is the optimum technique owing to its precision of categorized data items.

2.4 SVM IN DIABETES CLASSIFICATION

SVM is a non-parametric classifier able to solve classification and regression problems by means of linear and non-linear cost functions. It is a discriminative classification algorithm introduced by Vapnik (1995) to find abnormalities of biological signals due to its superiority and potential to handle the high-dimensional and non-linear database in the domain of the medical sector (Vapnik 1995).

The main concept behind this classification algorithm is to discriminate the unknown testing data into their suitable classes according to the training dataset. The application of SVM in DDS can be pigeonholed as linear and non-linear data. It employs non-linear correlations for varying learning datasets into a higher dimension. Once the training data is modified, it searches for the linear optimal splitting hyperplane. Therefore, in the SVM classification algorithm, the precise margins are made among various classes.

CHAPTER 3

EXSISTING METHOD

A) Data

Data are character, symbol, numbers, picture, video, or audio on which some operation performed on analysis purpose. Data mining uses different source of data and that integrated into data warehousing. Various filed generate large amount of data with unorganized form. All data are not in a same format or same structure. Various types of data that can be mined like:

- Multimedia data consist of audio, video, images etc. used in musical database, digital libraries.
- Transactional data consist of banking data, object database, banking data, inventory data, sales data etc.
- In relational database, data organized in table format used in performing data mining task effectively.
- In data warehousing, flat files are one of the data files to store data in text or binary format that can easily converted into knowledge.

B) Information

Information is collection of useful pattern, association, and relationship between data attributes from the collected data. It is complete, accurate, well organized, valuable, and proper formatted data. Information is meaningful thing that affects the behaviour or used to taking decision. In data mining, meaningful data are known as information.

C) Knowledge

Knowledge will be derived from the past information. Knowledge is a fact or awareness of things which useful for making decision. Based on previous patterns and trends, the knowledge will be built up.

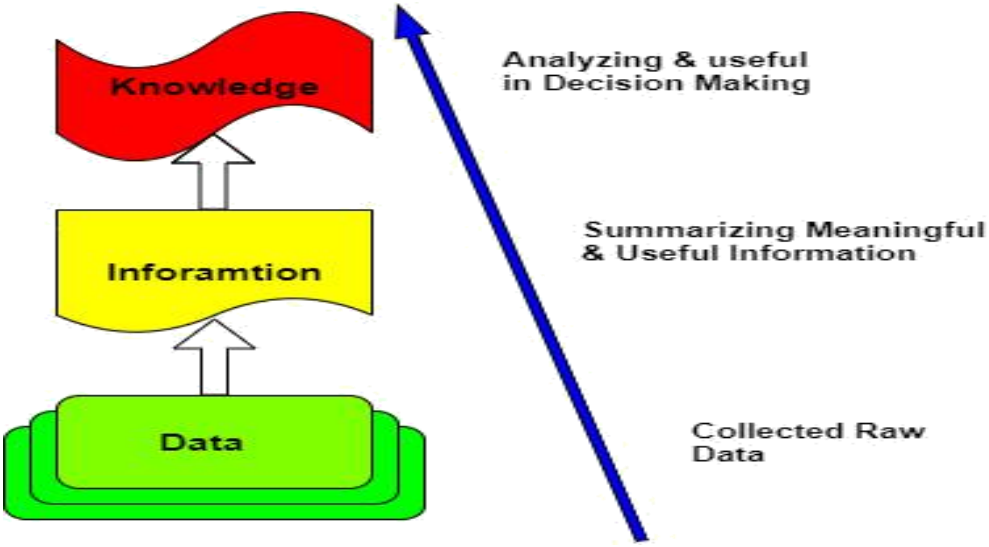

Figure 3.1: Relationship between Data, Information and Knowledge

3.1 EVOLUTION OF DATA MINING

After conducting the long-time of research and the growth in the product, the results of this research are data mining techniques. This concept came under the picture when the organization of data, has been stored in the form of computer data. After that many new technologies developed due to which accessing data, managing data, and updating the data became easy in a possible way.

Data mining was usually created for managing the data and helping in making decision process as it supports.

1. Large amount of data collection.
2. Having powerful and advanced multiprocessor computers.
3. Having various data mining methods and algorithms.

We can categories the four evolutionary steps in data mining technology; those specify the effective answers, organization questions in more accurate way and faster approach. The four evolutionary steps are collection of data, accessing of data, data warehousing and support for decision and now a day's data mining.

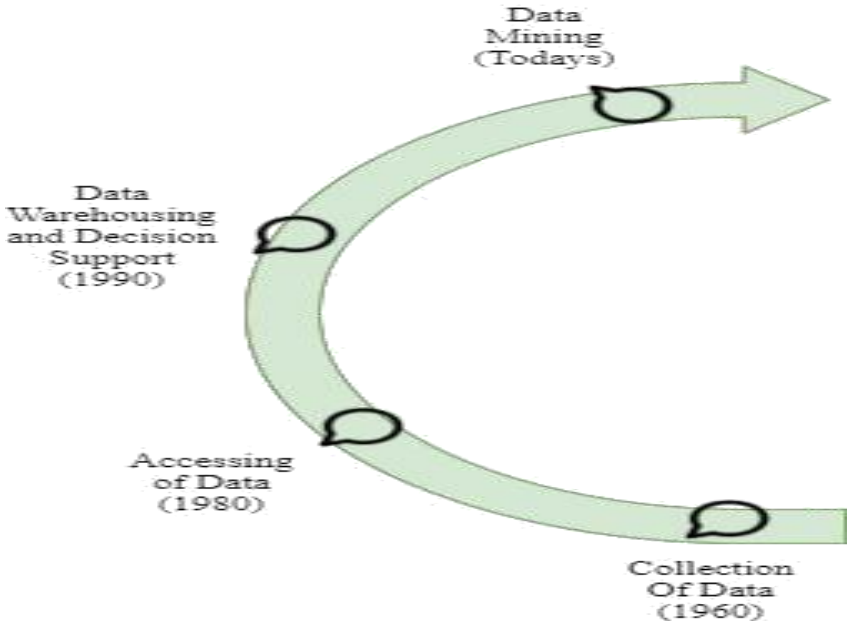

Figure 3.2: Timeline for evolution of data mining

- **Collection of data:**

This concept started in 1960s the question raised at that time was about what were the total sales of all the supermarkets in the past 10 years in South Capricorn region? The commonly used technologies at that time were computers, tapes, disks which was provided by IBM, CDC that usually follows the characteristics of having retrospective and static data delivery.

- **Accessing of data:**

This concept rose in 1980s this generated the questions like what and which were the sales of supermarket in South America in last October? In those times the technology used were RDBMS, SQL, ODBC which was provided by Oracle, Sybase, IBM, Microsoft, Informix. The characteristics followed were retrospective at record level dynamic data delivery. With more storage and relational databases, they use quicker and cheaper machines.

- **Data warehousing and Decision support:**

This can into the picture in 1990s this was generated to solve the questions like what were the sales of supermarket in the South America? Drill down to

America city. OLAP, data warehouses, multidimensional databases were the technologies used at that time which was provided by pilot, co share, arbour, cogon's, micro strategy. That usually has the characteristics of retrospective at multiple level dynamic data delivery. With more storage, they consume quicker and cheaper machine, Multidimensional databases, data centres, as well as remote analytical processing.

- **Data mining:**

This concept is generally used in every aspect which can usually help in solving the questions raised like what happen to South America city sales, in next months? The technologies used are advanced algorithms, massive databases, multiprocessor computers which are been provided by pilot, Lockheed, IBM, SGI, numerous start-ups. This all technologies possess the characteristics of prospective, proactive delivery of information. With more storage, sophisticated computer algorithm makes faster and cheaper as compared to other evolution system.

3.2 DATA MINING ARCHITECTURE

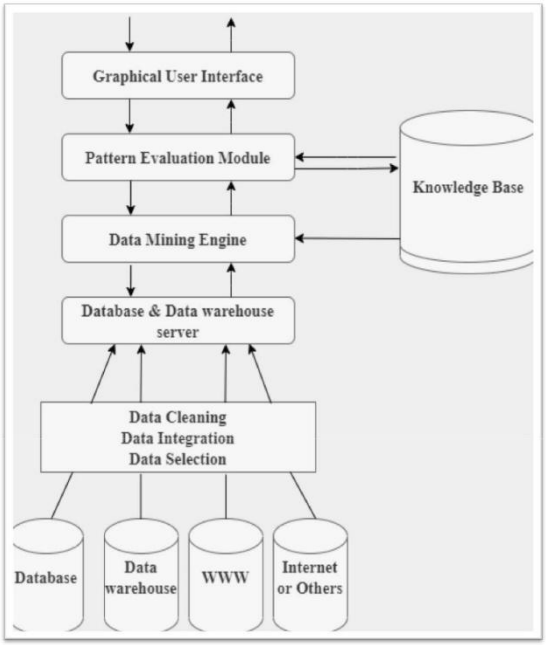

Figure 3.3: Architecture of Data Mining

1. **Data Sources** The initial details of data may in form of plain text, photos, audio, different spreadsheets. The data can be residing anywhere, place in files, spread sheet documents or internet. world wide web (WWW), internet, text files, spreadsheet etc. The collection of data repository is stored in database and data warehouse.

2. **Data Cleaning, Integration of data and collection of data**

   Before data it will be stored in database or data warehouses, the data must be cleaned, integrated as well as selection of data. The data may be obtained from different sources also data should be in different format.

3. **Database & Data warehouse server**

   Selected data is stored in the database server and the data warehouse server after efficient cleaning. The actual informative data, which is ready for proceed stored in database or various server of data warehouses.

4. **Data Mining Engine**

   The data mining engine is a central and critical part of the architecture for data mining. The mining engine contains collection of modules like Association technique, clustering technique, classification technique, characterization, time series analysis, prediction and regression technique. All these modules perform various data mining task easily.

5. **Pattern Evaluation Modules**

   The pattern evaluation module is responsible for measuring of interesting various pattern. The evaluating patterns, threshold value is used. The pattern evaluation module interacts and coordinates with data mining engine for searching of needed and useful pattern which improve better quality of data.

6. **Graphical User Interface**

   GUI is necessary for use of all the components make user friendly, effective and efficient. The interface helping users to access the system very easily and it can reduce the complexity. GUI cooperates between users and Data mining system.

7. Knowledge Base

The knowledge base aspect is the basis of the entire process of data mining. The module consists of data obtained from the user feedback. The data mining engine interact with Knowledge base components to provide result more accurate and reliable.

8. Understanding application domain:

This an initial process. Here there is understanding process with many decisions like transformation, various algorithm, representation of data etc. The people who use the KDD system will need identify and understand the purpose and the context in which the KDD process is going to play in.

9. Selecting and Creating datasets:

Here which type of data to be used in project with KDD is defined. It involves finding data, necessary information, additional data, integration of all data for KDD into one dataset also including the defined attributes. This all the basic evidence in creation of models.

10.Pre-processing and Cleaning:

Data reliability is improved in this process. This includes data cleaning such as removal of unnecessary data, the handling of missing values, noise removal or other special cases. There are several strategies for this method.

11.Data transformation:

The better generation of knowledge is prepared and produced in data transformation process. Methods such as dimension reduction along with feature selection and extraction are available and sampling is also documented.

12. Selecting the required role for Data Mining:

Here is a selection of which type of data mining, such as classification, regression or clustering, should be applied. This selection depends more on the goals and the previous stages of the KDD process. For data mining, the primary goal is prediction and description.

6. Choosing appropriate Data Mining Algorithm:

This phase includes selection of specific method which can help in searching Various patterns. Data mining algorithm like Decision Tree, Navies Bayes, Neural Network etc. is selection depends on the requirement of the output be performed on data inputs collected.

7. Employing Data Mining algorithm:

Finally, it is possible to apply the Data Mining algorithm. The presented algorithm is employing several times until the satisfied results is produced.

8. Evaluation:

We evaluate the patterns in this process about targets. The effect of pre-processing steps on the outcome of the data mining algorithm is the aspect considered here. This step concentrates on the understandability and usefulness of the model.

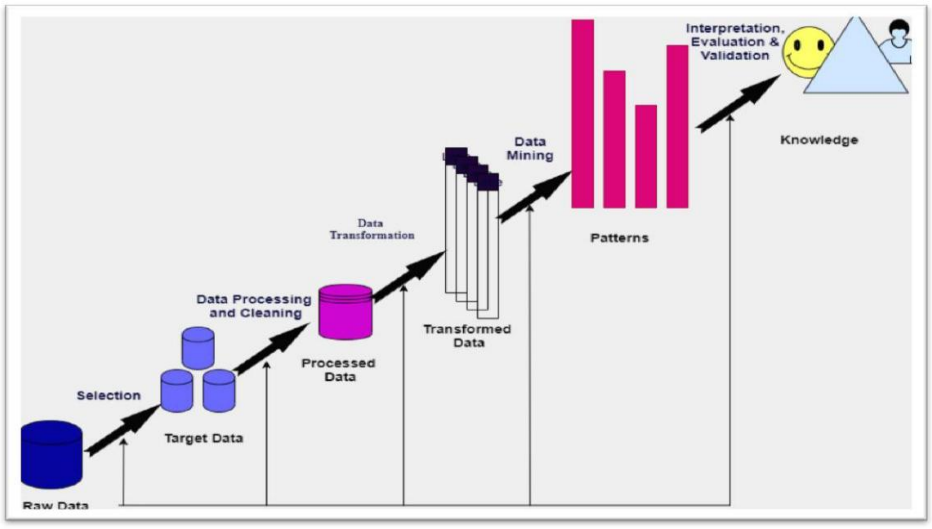

Figure 3.4: Steps involved in Knowledge Discovery Database (KDD) process.

3.3 TAXONOMY OF DATA MINING

As there are multiple techniques involve in data mining which are usually used for various methods, purposes and objectives. In order to understand the variance and variety of techniques, their interconnection and classification between the data and the respective techniques, taxonomy is stared. Are there are two subtype of data mining which are verification-oriented and discovery-oriented.

Here verification-oriented means the user hypothesis is detected by the system and discovery-oriented means process detect certain laws and trends autonomously. Taxonomy helps in distinguishing these two types of data mining.

3.4 DATA MINING TASK

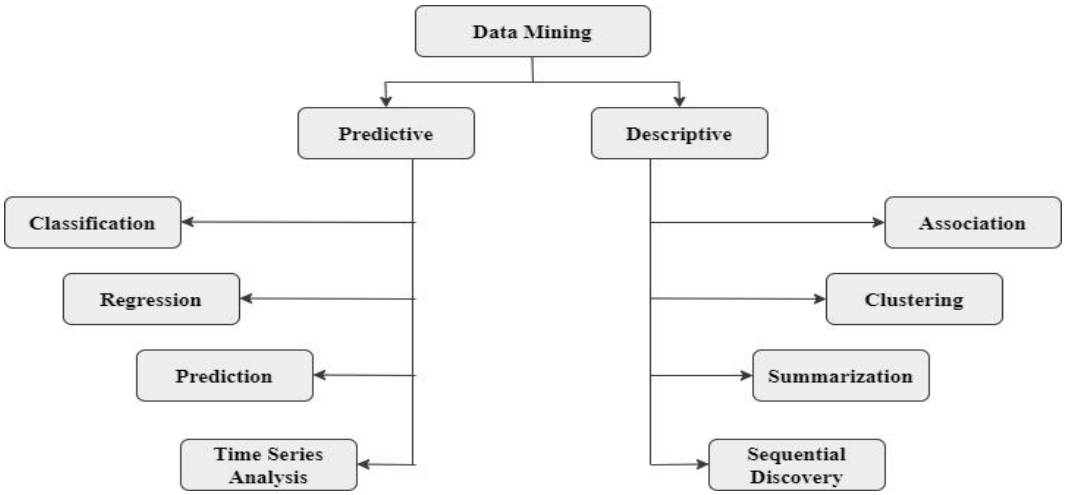

Figure 3.5: Data Mining Task

Since there are various patterns present in a database, the task of data mining is very diverse and distinct. Based on the patterns available various task are classified into Predictive and Descriptive.

The activities of data mining are divided into two groups.

Predictive and Descriptive

Predictive

In order to make a prediction, inference of the present data in turn is performed by predictive data mining tasks. It is the aim of predictive data mining to construct a model that can be used to perform tasks.

The predictive data mining task includes:

- **Classification**

  Classification is a process in which there involves the finding function of predictive learning method. A data object is categorized by this function into one of several predefined groups.

- **Regression**

  Regression is called a task which predicts a number. It is a technique for

data mining that is used to fit into a dataset equation. A function which finds minimal errors to model data is called regression.

- **Prediction**

  With the help of prediction, we can predict missing or unknown values. Prediction is a process done based on some algorithms. It can also be done by query passed on the particular targeted database. Predicting the outcome based on the dataset with the aid of a task or classification of estimation data mining. Guesting the missing data of forecasting the future data is done is prediction. According to the present data set, a model is created in prediction.

Descriptive

Descriptive data mining activities describe the basic features of the data in the database repository. To gain or understand of the system which is analysed by uncovering data patterns and association ship in the larger data sets are the target of the descriptive data mining. The descriptive data mining task includes:

- **Clustering**

Identification of classes or clusters or groups are called as clustering for a group of objects with undefined categories. Clustering of object is done on the basis of similarities. The objects are designated with a particular cluster, once the clusters are decided. To form the class description, general characteristics of the artifacts are summarized in a cluster form. Several groups made based on similarities which help for better understanding.

These is no predefined class during clustering.

- **Summarization**

  The generalization or abstraction of data is called summarization. A general overview is obtained by summarizing a large set of task relevant data. It is possible to obtain various levels of abstraction from summarization analysis. It is important to visualize varying degrees from various viewpoints. It is used to generalize the given data. According to different prospective the data can be summarized.

- **Sequential Discovery**

  The discovery of interesting patterns in the evolution history of the object. Sequence discovery is also known as pattern mining. Sequential patterns can be found out from transactional data set from sequential discovery. The patterns are generated based on time sequence.

3.5 DATA MINING TECHNIQUES

• Statistical Approach

In data mining techniques various statistical tools are used. The Bayesian network model, correction evaluation process, multiple regression analysis and clustering analysis are part of this method. With the help of set of training data statistical methods are built. The statistical approaches are an important technique used in a Bayesian network.

• Machine Learning Approach

Almost all of the times machine learning algorithms use Heuristics type in the searching purpose. How that machine learning approach looks for the suitable classifier model that fits the performance data as a outcome. The most popular machine learning approaches are conceptual modelling, inductive conceptual.

- **Database- Oriented Approach**

Unlike the other two previous kind of methods Database-oriented method don’t look for right and suitable model. Use of development process or repository database heuristics are take advantage of the nature of the data in hands.

1. Attribute-Oriented Induction

Low level data is standardized into definitions of high levels. Using the logical hierarchies, this is achieved.

2. Iterative Database Scanning

A regular element set in or through a transactional database is defined by iterative scanning of the database. From these set of frequent products list, Association rules

are derived.

3. Attributes focusing method

By preferentially inserting attributes into patters, the attributes focusing methods find patterns with unusual probabilities.

3.6 ADVANTAGES OF DATA MINING

As data mining implemented on every aspects of fields like in business, education, medical along with it is also applied on prediction of weather forecasting, medicine, transportation, healthcare, insurance, government and much more. Data mining has common advantage on when used onto industry. Advantage of data mining in various sectors are mentioned bellow.

Disadvantages of Data Mining

Data mining method that helps one person in making decision and that decision process making involves all the features and factors of which mining is involved precisely. While making decision process one can face or one can come across:

Violating user privacy:

It is described facts that data mining technology gathers and collect the relevant information about people using some market-based techniques and some sort of information technology. And this process involves various and serval numbers of factors. As they various means and methods to collect the information, data mining system violates the privacy of user identity and information that leads to matter of safety and security of users..

Supplementary in irrelevant information:

The main working with data mining system that it establishes a relevant space for beneficial and important information. Here the problem comes is that while gathering or extracting the information it collects some overwhelming of information for all and sometimes even gathers some irrelevant contents and topics.

Misuse of information:

As we know and explained that data mining has weak point with respective security and safety measure as they are minimal. As someone can even misuse the information which can raises the issues like information is misuse for personal contents, cause of harm to other, mis conceptual aspects and their own ways of harming. To prevent this issue data mining can sets some sort of time limit and working limit that can reduce the ratio of misuse of information.

Accuracy of data:

As in early period when they need some sort of information about something, they used to seek the help of clients, teachers, books etc. but nowadays the information gets available easily on any topic as data mining provides many technologies and their methods. Here data mining accuracy of information on their own limit of information.

Security:

As there is huge provision of data and huge data is begin collected in data mining systems. There are many possibilities that data integrity is damaged like a kind of data can be hacked by hackers and contents are changed or used related like big companies like Samsung, Sony, etc. this another big issue in data mining. Here there are many cases of hacker stealing the useful information.

A skilled person for data mining:

As we know data mining tools are very powerful. But still there is need and cause of using a technical skilled person is needed to know the working of it. Even this person is used and knows how to prepare data and then produce the relevant output of it. As there are many variations in patterns and relationship between the data hence a skilled person is needed.

By seeing the above disadvantages, we can conclude that data mining has issues and faces problem in managing and maintaining the privacy, security and misuse of information's. as this all are big problems which are not addressed and resolved.

3.7 CHALLENGES IN DATA MINING

Data Mining system totally depends on repositories to provide the pre-processes data, here the problem arises such as the database appears to be complex, huge, messy and incomplete. Other problem results in inadequate data and the irrelevance of the information to be stored.

Because of the complicated algorithms, data mining is a dynamic method and data is not present at one place. In some section the factor and the issues faced in data mining is categorized into segments like mining framework and user interaction, performance problems and various problem with data types.

1. Methodology of mining and problems with user interaction:

This covers the following categories of problems:

- **Different information forms in repository**:

As various clients are interested in different sorts of important information for implementation and usage as per their requirement. Therefore, data mining wants to understand and handle a wide variety of tasks for information exploration.

- **Requiring collaborative mining information at several abstraction levels**:

As the data mining method needs to be interactive, it allows the search patterns to be attracted and oriented, providing and optimizing the mining query depending on the outcomes provided. Like Amazon sets prime membership example of data mining with its Amazon Check Price Mobile Application.

- **Incorporate of background knowledge**:

The background knowledge is usually used to guide discover patterns and even express the new patterns. Context information is used not just in succinct terms but also at different levels of abstraction to describes the patterns to be identified. Concerning the domain section that integrates the process of knowledge discovery, context knowledge, laws, constraints and other diverse material. These may be used for the assessment of trends and also a reference to discovering interesting patterns.

- **Query language and ad hoc data mining**:

The programming language allowed the client to identify ad hoc mining activities that must be combined with the programming language of the database system. It also helps in optimizing data in efficient and flexible manner. Query languages plays a significant job in searching a flexible manner. In this there should be facility of appropriate sets of data for study, the domain awareness and constraints enforcement in discovering patterns.

The new data patterns are recognized and generated or discovered they need to be fixed in high levels of languages and visualization. These will be easily comprehensible.

- **Managing noisy and inaccurate information**:

  Data cleaning techniques include the handling of messy information and missing items whereas managing data mining. If there is not a data cleaning method, there will be no accuracy of the discovered patterns. We even recognize that the method of extracting data from vast amounts of data is data mining.

- **Pattern evaluation**:

  The different patterns found must be important both as common knowledge expression or lack of creativity. Since the patterns can differ from client to client or person to person. To determine the accuracy and consistency of a pattern based on subjective measures, we need to evaluate techniques.

1. Performance Issues:

It has following performance-related issues:

- **Scalability and efficiency of data mining algorithms**:

  In order to have the productivity to extract information from a large amount of data in the database system, data mining algorithms must be scalable and effective. As the data in today's day continues to multiply day by day, these two factors are especially critical.

- **Algorithms for parallel, distributed and incremental mining:**

  Factors such as huge, enormous libraries, vast sizes of databases, a wide range of data distribution and even the complexity of data mining techniques are driving the development of parallel and distributed data mining algorithms. As they typically split the information into smaller partitions that are further processed in a parallel way. Then the findings from the partitions carried out are combined. Incremental algorithms, without mining the data, update databases.

2. Diverse Data Types Issues:

The following are data types of issues in data mining.

- **Managing relational and complicated types of data:**

  It is not possible to handle data that database contains like complex data items, data objects for graphics, spatial information, temporal data etc. this all types are potential to be handled by one system. Real world data are realty heterogeneous in nature and it can even contain images, audio, natural text etc. It becomes to manage and extract such all sort of data by a single system.

- **Data mining from heterogeneous databases and global systems of information:**

In various sources such as local area network, internet, Wide area network, data is often available. Such information is typically organized, semi-organized or unorganized. As a result of this mining, the information from them presents problems or challenges.

The other faced issues are:

- **Overfitting**: When the constructed model faces or has a greater number of overfitting problem doesn't fit future states. When algorithm comes under overfitting the data, results in generalization. Cross validation, regularization and other sophisticated methods can overcome this problem.

- **Outliers:** This are the data that are not nicely fitted into the derived model. The models that develop the outliers, then the model may not behave well for data that are not outliers.

- **Protection and social issues:** Protection is a main issue in data mining while we collect the data from any resources. It creates problem when used in making the strategic decisions.

3.8 DATA MINING IN HEALTHCARE

In several industries, data mining is used successfully. It also helps to predict the customer relation, many things in banking, calculating the average etc. We also defined data mining is used in education, healthcare, telecom, and much more.

Data mining has much potential to be explore in healthcare sector as by increase the factors in records with respect to healthcare. As in older days the doctors used to maintain the file of patient's information in which they hold patient's data for analysis but this type of handling patient's data in massive format is quite difficult.

System Approach in HealthCare:

There is different approach in data mining with respect to healthcare. For an academic research approach the best approach with data mining is following three system approach.

Following all these three approaches improve real world analysis and initiative with healthcare. But unevenly only few of them follows this three-system approach. The three system comprises such as analytic system, the content system, the development system.

3.9 DATA MINING APPLICATIONS IN MEDICAL SECTOR

The key dimensions in healthcare management are customer relationship, diagnosis and treatment, healthcare resource management, fraud and anomaly detection, hospital management, medical device management, pharmaceutical industry etc.

Diagnosis and Treatment:

Ultrasound images is commonly used in examining which is also known as Sonography, this all supports images-based examination of body that can help in finding tumour, organ swelling, unfortunate conditions in body etc.

Healthcare Resource Management:

Using various model provision in data mining comparison factor in health segment can be resolved like with the help of logistic regression models, we can compare hospitals profile based on risk-adjusted death factor with a limit of 30 days and more with non-cardiac surgery based can be compared easily. Even it can predict disposition in kids present in the urgent care department with the disease of bronchiolitis suing Neural Network system.

This research proposes to address and solve the challenges of enhancing the predicting accuracy of diabetes disease and heart disease in right time. The research tasks are defined as the way by (i) how data mining algorithms with normalisation method used by the medical filed to summarize their outcomes in prediction? (ii) In order to accurately predict the risk of diabetes and heart disease, how specifically does the classification approach help to enhance the prediction model?

The primary aim of this research work is to build a forecasting method effectively by improving the classification algorithm. Using Min Max normalization methods to predict the risk of developing diabetes and heart disease in earlier stages, progress can be accomplished. It also demonstrates that data mining could be used to predict or classify data with great accuracy in healthcare database

CHAPTER 4
PROPOSED METHOD

HYBRID ONLINE PREDICTION USING AN ADAPTIVE PRINCIPAL COMPONENT ANALYSIS AND INCREMENTAL SUPPORT VECTOR MACHINE

This chapter presents a hybrid incremental framework for the timely identification of DM using machine learning techniques. Accurate and the timely identification of DM from clinical datasets are indispensable for the prevention of disease and the choice of correct treatment. Conventional ML-based disease detection models have been originally developed as non-incremental (offline) methods that are learned from a predefined dataset before they can be implemented to solve disease classification problems.

(i) an adaptive PCA algorithm for missing data imputation, attribute selection, and data clustering; and

(ii) an improved incremental SVM for classifying diabetes mellitus in an incremental way. The performance of HOMED is evaluated on the PIDD database using various evaluation measures including accuracy, precision, sensitivity, specificity, PPV, and NPV.

(iii) The extensive empirical analysis proves that HOMED significantly improves the performance of DDS and reduces the processing overhead as compared to the offline classifiers. This model can support clinicians to take appropriate decisions about the disease and therapy.

4.1 HYBRID ONLINE MODEL FOR EARLY DETECTION OF DIABETES DISEASE

Handling DM is a perplexing, expensive, and multifaceted task for care providers and clinicians. There are numerous important data to store about the patients and diseases that supportthe clinicians in making an optimum verdict to increase the life expectancy of people. This fetches the conventional analytical

approaches to stop or criticize the training phase as the training methods become susceptible to the over-fitting problem due to the huge irrelevant and redundant attributes in a high-dimensional database.

One of the extensively used and powerful methods in any disease detection system is the implementation of ML algorithms. Machine learning copes with the growth of cutting-edge techniques that allow machines to learn for making estimates about newly arrived dataset that are expected to derive from the analogous population that created the data used for the detection system.

Furthermore, typically supervised ML approaches often cannot be performed incrementally, and therefore it is essential to recognize all the learning data again to carry out data classification process.

The key goal of this study is to design an efficient and precise indigenous analytic model for identifying DM, nevertheless the availability of several renowned dominant methods, which have already been realized or are in use for the identification of DM. The contributions of this work are: (i) an APCA algorithm is proposed by integrating methods for missing value imputation, data clustering and attribute selection; and (ii) an improved ISVM to detect the DM is devised.

The performance of HOMED is calculated on various evaluation measures including accuracy, precision, specificity, sensitivity, PPV, and NPV. An extensive empirical analysis of PIDD is carried out. The empirical analysis divulges that HOMED significantly upturns the detection performance and reduces processing overhead in terms of the offline detection frameworks. The developed online DDS can help medical experts diagnose diabetes mellitus with superior accuracy.

By considering detection and classification of DM, HOMED employs the results of medical clinical tests as input, mines an a bridged attribute space, and diagnoses diabetes mellitus.

Initially, the input data set is preprocessed for removing the unrelated and redundant data and for missing value imputation.

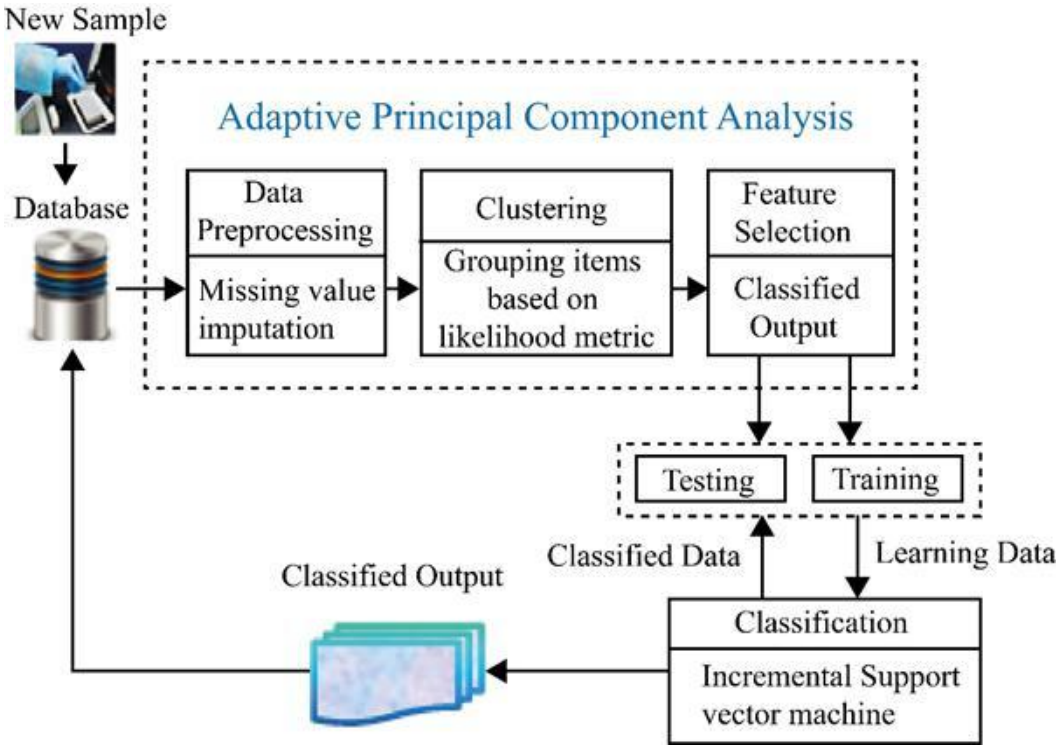

Figure 4.1: System architecture of HOMED.

Since the clinical datasets are unremittingly collected from the new samples, it is beneficial to incrementally appraise the previous framework of DM detection by focusing only on new input data to reduce the temporal overhead related to the detection process.

This proposed HOMED model employs an APCA algorithm and an improved ISVM for classifying diabetes mellitus in an incremental way. The proposed APCA helps DDS models eliminate unsuitable attributes, hence minimizing the learning cost, time, and also increasing the enactment of the DDS. In this work, the APCA exploits the Gaussian mixture model to achieve data clustering, and the expectation-maximization technique is used to compute the model variables.

4.2 CLASSIFICATION OF ALGORITHMS

K-Nearest Neighbor

K-Nearest Neighbour (KNN) is a simplest supervised learning classification algorithm for categorizing the object. It used in multiple application like data mining, pattern recognition, online marketing, cluster analysis, machine learning for correct prediction and estimation.

- KNN algorithm predict class label of any dataset. By observing majority rates of neighbours, object is easily classified. Assigned class for specific object is decides by positive integer value of k.

- Implementation of algorithm done is parallel manner. The reason behind that, every datapoint, algorithm check continuously training table for K nearest neighbour. Every data point is independent that‟s way the process executed parallel way.

The number of samples of learning contained in n-dimensional storage. The new test samples arrive, classifier search K training samples in n-dimensional space. The closeness calculated using Euclidean distance between two data points.

K= {k1, k2, k3,.., kn} and V={v1, v2, v3,……., vn}represent attributes values. The Euclidean distance is

$$d(k, v) = \sqrt{\sum_{i=1}^{n} (v_i - k_i)^2}$$

The first step of K-Nearest Neighbour algorithm is to check new instance with available cases using Euclidean distance and classified with help of k value. Thus, the KNN algorithm is dependent on k value for better prediction.

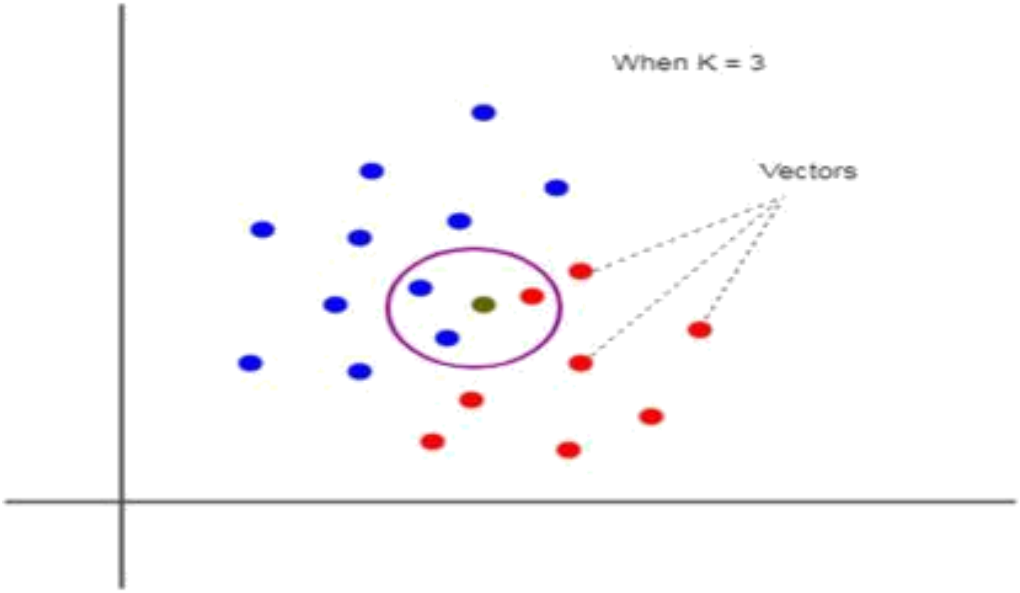

Figure 4.2: Graphical representation of working of KNN.

K Nearest Neighbour Algorithm

The following steps shows how KNN algorithm works:

1. Loading a dataset for training and testing.
2. Choose the parameter k. The value of k is the closest point of knowledge and should be an integer value.
3. For every point of test data, iterate the following process for predicted class.

    I. Compute the distance between data from the test and each of the training data by using either Euclidean method, Hamming distance and Manhattan method. Most probability Euclidean distance method are used for calculating the distance.

    II. Calculated distance sort in ascending order according to the getting value from Euclidean distance.

    III. The it selects first k value of having smallest distance in order to sorted array.

    IV. It will assign an object or data point to majority of class.

4. End

Advantages of KNN:

- Effective and easy for implementation when huge amount of training data available.

- Robust for instance, classes in dataset have noisy data which is not linearly separable.

Disadvantages of KNN:

- Expensive because need to compute distance for all instance available in training samples.
- Does not more effective with high dimensions and large dataset. Hence it shows poor accuracy.

Support Vector Machine

A supervised algorithm for machine learning used for classification and regression purposes is the Support Vector Machine (SVM). The aim of SVM is to find the suitable margine called hyperplane between two categories. The distance from the hyperplane to the data point should be far as possible for better generalization method. The algorithms draw hyperplane that divides positive and negative dataset samples. Face detection, Bioinformatics, classification of image is some application of SVM. The input of labeled sample data, draw a line in between two group.

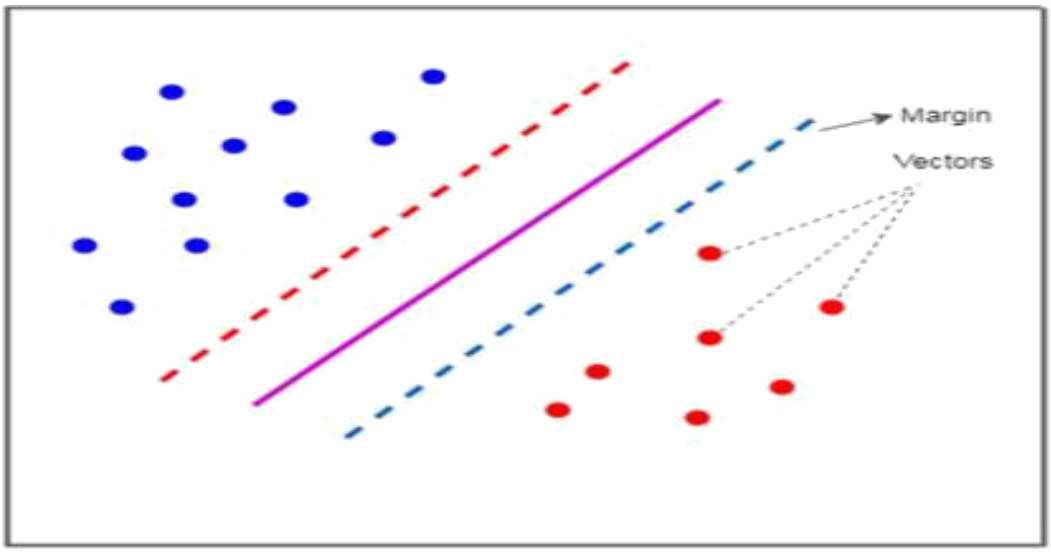

Figure 4.3: Graphical representation of working of SVM

Advantages of SVM:

1. SVM work effective algorithm when there is no idea on the dataset.

2. The SVM is also compatible with unstructured data and semi structured.
3. SVM is much more effective in high dimensional data space.

Disadvantages of SVM:

1. For the large data set, algorithm does not perform very well. It required long time for the classification.
2. Selecting type of kernel is difficult task in SVM.
3. Difficult to handle final classification model.

Random Forest

It is an important ensemble-supervised classification algorithm for learning which executes during the training process by building multiple decision trees. To finding prediction of value and probability estimation, Random forest method widely used. The opinion of the bulk of the trees is selected as the final choice by the random forest. Each tree in the forest treated separately according to value of random vector and equally allotted to all the trees in the forest that form a grouping of tree predictors in random forest algorithm.

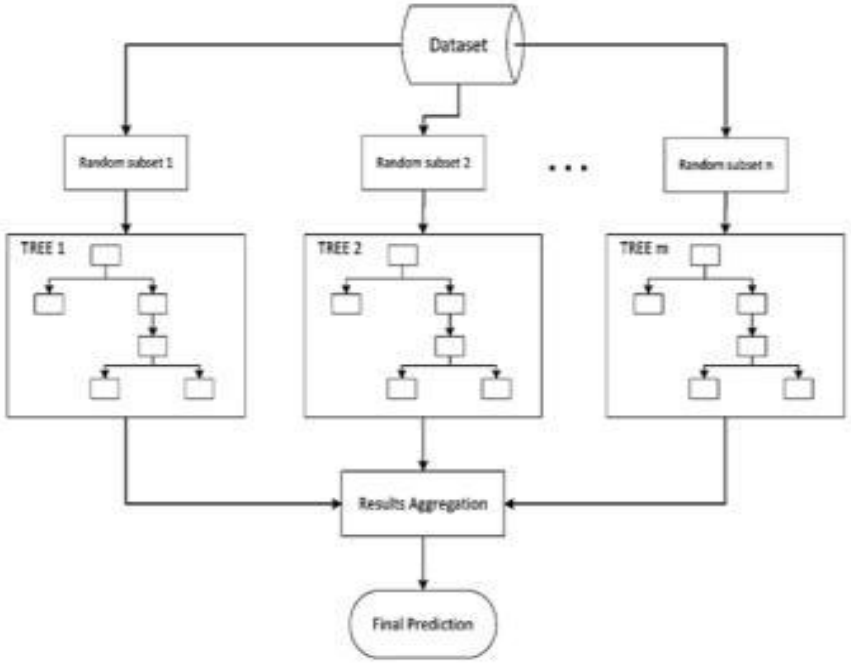

Figure 4.4: Graphical representation of working of Random Forest

Random Forest Algorithm:

1. Select random sample „r‟ from the given training dataset „d‟, whereas r<<d.
2. Make decision trees with help of data points.
3. From sample r, compute the node „n‟ to determine the best split.
4. Repeating step number 1 to 3 until it reached to final node of the tree.
5. For new data point, determine the all decision tree and put data point to the majority vote wins by tree.

Advantages of Random Forest:

- Use of multiple trees that overcome the over fitting problem.
- It can maintain accuracy while missing value in dataset.

Disadvantages of Random Forest:

- Not much effective in real time prediction.

Decision Tree

Decision tree is the supervised algorithms of learning being used execute the task of classification and regression. The decision tree algorithm is also used for classification problems. It presents a structure-like tree in which each internal node will perform as a test feature of a dataset., branches act like a decision rules and every leaf represent final outcome of a tree.

Every decision tree is classified into two category such as decision node and leaf node. Decision node take decision when having more than one branch or decision node having multiple leaf node. The leaf node represents final outcome of the decision node which cannot be further split. The split of decision node totally depends on the dataset.

Terminology:

1. Root Node: The root features of the dataset which can split into multiple subsets to take a decision.
2. Leaf Node: Final outcome of a tree which cannot further split.
3. Pruning: It is technique to removing unnecessary branches of the tree.
4. Parent Node: The root node called parent node.
5. Child Node: The sub node of the parent node is called child node.

Algorithm:

- Tree starts from the root node, contains entire dataset present in root node.
- Discover suitable attributes with help of attribute selection approach.
- Split the root node, into sub leaf node having good attributes value.
- Make decision tree using attributes selection approach has best attribute on each split.
- Continue this process until reach to final outcome where leaf node cannot be further divided.

Attribute Selection Approach:

To selecting best attributes from the dataset attributes selection approach is used, which take decision to split the dataset into leaf node.

Entropy:

It measures impurity or randomness of particular attributes or features in the data.

The entropy can be calculated as

Entropy(S) = -P(Positive Probability)$\log_2$(Positive Probability) - P(Negative Probability) $\log_2$ (Negative Probability) Where, S represent total number of samples in data

Advantages of Decision Tree:

- Simplicity: For decision-making, the classifier is user - friendly and easy to understand.
- Beneficial for decision making problem.
- Perform more efficiently on numerical value as well as categorical data.

- Data cleaning process is not taking ling time as compared to other classifiers.

Disadvantages of Decision Tree:

- Overfitting problem increases complexity of algorithm.
- More number of attributes in dataset, making model more complex.
- Selecting proper parameter is necessary and important task.

Logistic Regression

Logistic Regression is a classification algorithm for supervised machine learning, essentially used to accumulate independent variables and estimate the performance of the categorical dependent variable using an independent variable. The outcome of logistic regression algorithm will give probabilistic value which is either categorical value or discrete value such as outcome may be either "yes" or "no", either "0" or "1", either "true" or "false" etc.

$\text{Log}[\ ] = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + \ldots.. + a_nx_n$

Types of Logistic Regression:

1.Binomial

The binomial logistic regression gives two different types of outcome that is "Win" or "loss" , "0" or "1", "True" or "False", "Pass" or "Fail", "Yes" or "No" etc.

2.Multinomial

The multinomial logistic regression gives 3 or more different possible types of unordered outcome such as "Mango" "Banana" "Orange", "Disease X" "Disease Y" "Disease Z", or "Cat" "Goat" "Cow" etc.

3.Ordinal

The Ordinal logistic regression gives 3 or more different possible types of ordered outcome such as "very good" "good" "bad", "high", "medium" "low" etc.

Naïve Bayes

Naïve Bayes is the work basis of the Bayes theorem on supervised machine learning classification algorithm. The application work on text or document

classification on dataset. Naïve bayes algorithm make machine learning model stronger and give quick prediction result. The classifier used probability of an attribute for prediction of data.

(|)()
(|)
( )

Where,

P(F|G) represent probability of even F, given G has occurred.

P(G|F) represent probability of even G, given F has occurred.

P(F) represent probability of event F

P(G) represent probability of event G [111]

The Naïve bayes classifier used in data classification purpose, sentiment analysis, real time prediction and credit scoring.

Types of Naïve Bayes:

1. Gaussian

The predictors value of each feature is continuous in nature and value are not discrete, the assume value of sampled is from normal distribution or gaussian distribution.

2. Multinomial

The classified model used feature frequency of words for classification problem. This type of naïve bayes classifier used when available data is multinomial distributed.

3. Bernoulli

This classifier is similar to multinomial classifier. The difference is that predictor is Bernoulli distribution are Boolean variable. The word is either present in document or not using independent Boolean variable that give Yes or No value.

Advantages of Naïve Bayes:

- Fast machine learning classifier to predicting result and easy to implementation.

- Text classification task perform very well using naïve bayes classifier.
- Naïve bayes performs on binary classification task as well as multi class classification.
- For train a model, naïve bayes algorithm required small amount of dataset, so no longer time is required to train a model.
- Naïve bayes classifier perform best when predictor value is true as compared to another classifier.

Disadvantages of Naïve Bayes:

- Difficult to learn the relationship between different features because classifier assume all available feature in dataset are independent in nature.
- The model assume zero probability when particular variable is not visible in training data. That is different situation for predicting data.

4.3 DATA ACQUISITION

The collection of different data from various source is called data acquisition. The data acquisition is essential and equally important task in data mining. A data can contains image, records, various entities, elements, points etc. The collecting data has some attributes.

- **Nominal Attributes:** The dataset contains set of distinct data that represent same category of unique value with no any particular order in values. For example, colour attributes having value Red, Black, White. Fruit attributes having values Mango, Banana, Apple.
- **Ordinal Attributes:** The value present in ordinal attributes having some meaningful sequence between them. For example, Grade attributes having value A, B, C, D. The Range attributes having value High, Medium, Low.
- **Binary Attributes:** The Binary attributes having only two values either Yes or No, True or False etc.

- **Ratio based Attributes:** This attribute is type of quantitative attributes which value is measure in numeric value for example kelvin temperature.

4.4 DATACOLLECTION

Data is an integral part of the process of data mining. The vast volume of data helps with data mining techniques for information retrieval, pattern detection and health disease diagnosis. For analysis and better prediction more data is required. Large and proper data gives greatest prediction rate and it increase accuracy rate.

4.5 IMPLEMENTATION MODEL

The objective of the study is to boost the accuracy of supervised learning algorithms such as K Nearest Neighbour, Support Vector Machine and Random Forest using the Min Max techniques on diabetes dataset. The KNN, SVM and Random Forest these three algorithms are used in research work.

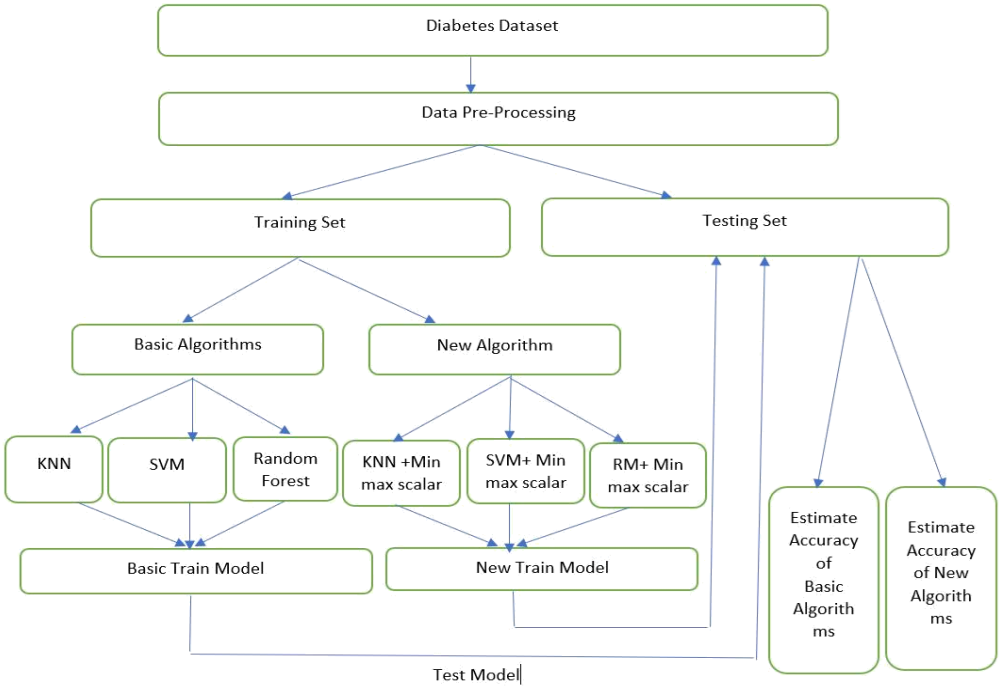

Figure 4.5: Implementation for diabetes model

One Hot Encoding

- Machine learning algorithms cannot work with categorical data directly.
- Categorical data must be converted to numbers.
- This applies when you are working with a sequence classification type problem and plan on using deep learning methods such as Long Short-Term Memory recurrent neural networks.
- A one hot encoding is a representation of categorical variables as binary vectors.
- This first requires that the categorical values be mapped to integer values.
- Then, each integer value is represented as a binary vector that is all zero values except the index of the integer, which is marked with a 1.

Worked Example of a One Hot Encoding

Let's make this concrete with a worked example.

Assume we have a sequence of labels with the values 'red' and 'green'.

We can assign 'red' an integer value of 0 and 'green' the integer value of 1. As long as we always assign these numbers to these labels, this is called an integer encoding. Consistency is important so that we can invert the encoding later and get labels back from integer values, such as in the case of making a prediction.

Next, we can create a binary vector to represent each integer value. The vector will have a length of 2 for the 2 possible integer values.

The 'red' label encoded as a 0 will be represented with a binary vector [1, 0] where the zeroth index is marked with a value of 1. In turn, the 'green' label encoded as a 1 will be represented with a binary vector [0, 1] where the first index is marked with a value of 1.

If we had the sequence:

```
1        'red', 'red', 'green'
```

We could represent it with the integer encoding:

```
1       0, 0, 1
```

And the one hot encoding of:

```
1       [1, 0]
2       [1, 0]
3       [0, 1]
```

A one hot encoding allows the representation of categorical data to be more expressive.We could use an integer encoding directly, rescaled where needed. This may work for problems where there is a natural ordinal relationship between the categories, and in turn the integer values, such as labels for temperature ‘cold’, warm’, and ‘hot’.

One Hot Encode with scikit-learn

In this example, we will assume the case where you have an output sequence of the following 3 labels:

```
1       "cold"
2       "warm"
3       "hot"
```

An example sequence of 10 time steps may be:

```
1  cold, cold, warm, cold, hot, hot, warm, cold, warm, hot
```

This would first require an integer encoding, such as 1, 2, 3. This would be followed by a one hot encoding of integers to a binary vector with 3 values, such as [1, 0, 0].The sequence provides at least one example of every possible value in the sequence. Therefore we can use automatic methods to define the mapping of labels to integers and integers to binary vectors.

One Hot Encode with Keras

You may have a sequence that is already integer encoded.

You could work with the integers directly, after some scaling. Alternately, you can one hot encode the integers directly. This is important to consider if the integers do not have a real ordinal relationship and are really just placeholders for labels.

The Keras library offers a function called to_categorical() that you can use to one hot encode integer data.
In this example, we have 4 integer values [0, 1, 2, 3] and we have the input sequence of the following 10 numbers:

```
1    data = [1, 3, 2, 0, 3, 2, 2, 1, 0, 1]
```

A complete example of this function is listed below.

```
1    from numpy import array
2    from numpy import argmax
3    from keras.utils import to_categorical
4    # define example
5    data = [1, 3, 2, 0, 3, 2, 2, 1, 0, 1]
6    data = array(data)
7    print(data)
8    # one hot encode
9    encoded = to_categorical(data)
10   print(encoded)
11   # invert encoding
12   inverted = argmax(encoded[0])
13   print(inverted)
```

Running the example first defines and prints the input sequence.

The integers are then encoded as binary vectors and printed. We can see that the first integer value 1 is encoded as [0, 1, 0, 0] just like we would expect.

We then invert the encoding by using the NumPy argmax() function on the first value in the sequence that returns the expected value 1 for the first integer.

```
1    [1 3 2 0 3 2 2 1 0 1]
2
3    [[ 0.  1.  0.  0.]
4     [ 0.  0.  0.  1.]
5     [ 0.  0.  1.  0.]
6     [ 1.  0.  0.  0.]
7     [ 0.  0.  0.  1.]
8     [ 0.  0.  1.  0.]
9     [ 0.  0.  1.  0.]
10    [ 0.  1.  0.  0.]
11    [ 1.  0.  0.  0.]
12    [ 0.  1.  0.  0.]]
13
14   1
```

CHAPTER 5

RESULTS

DIABETES PREDICTION USING ML

Sample data set

```
df.head()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Figure 5.1: Sample data set

Information about the data

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Figure 5.2: Information about the data

Description about the data

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Figure 5.3: Description about the data

Histogram of the age

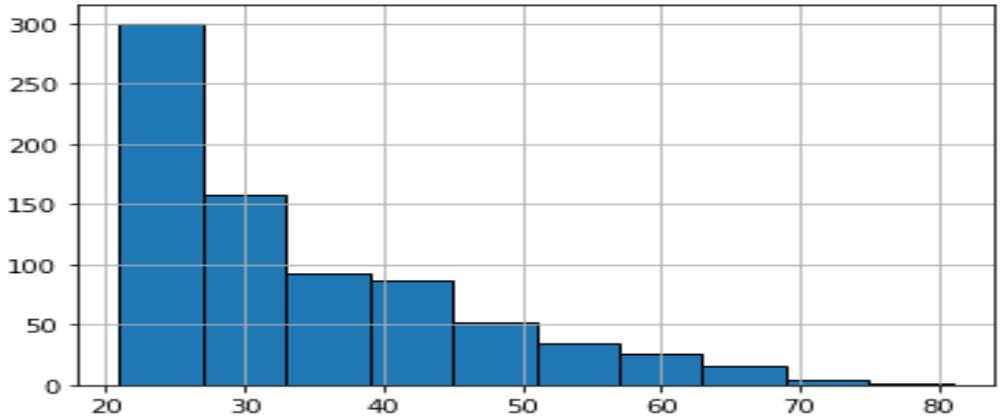

Figure 5.4: Histogram of the age

Histogram of different variables

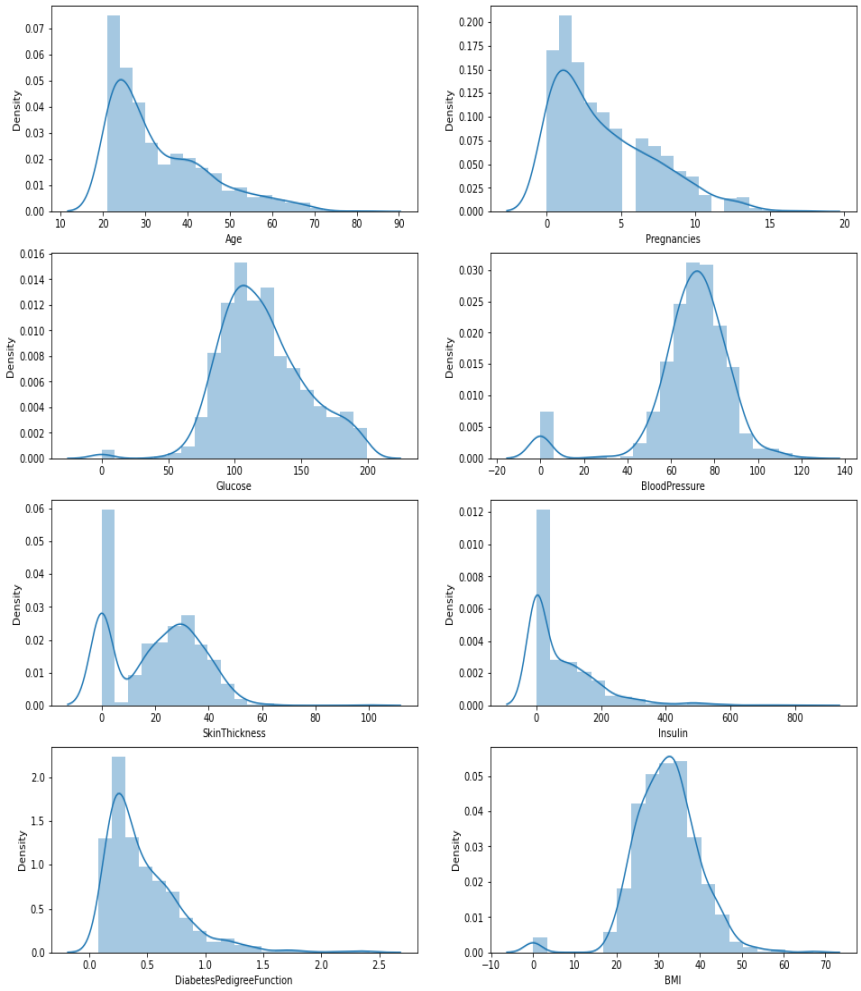

Figure 5.4: Histogram of different variables

Pie chart of the diabetic patients in the database

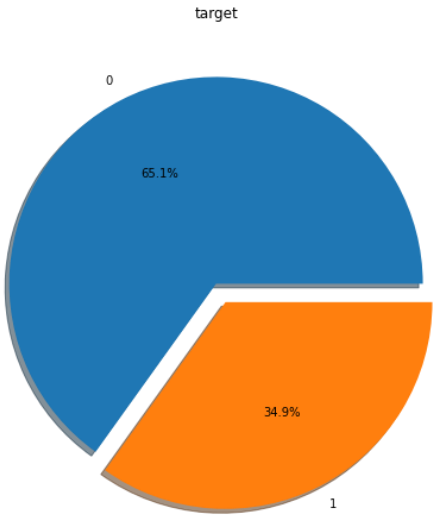

Figure 5.5: Pie Chart

Correlation between the features of the dataset

- Access to the correlation of the data set was provided. What kind of relationship is examined between the variables.
- If the correlation value is> 0, there is a positive correlation. While the value of one variable increases, the value of the other variable also increases.
- Correlation = 0 means no correlation.
- If the correlation is <0, there is a negative correlation. While one variable increases, the other variable decreases.

Correlation matrix

```
df.corr()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.000000	0.129459	0.141282	-0.081672	-0.073535	0.017683	-0.033523	0.544341	0.221898
Glucose	0.129459	1.000000	0.152590	0.057328	0.331357	0.221071	0.137337	0.263514	0.466581
BloodPressure	0.141282	0.152590	1.000000	0.207371	0.088933	0.281805	0.041265	0.239528	0.065068
SkinThickness	-0.081672	0.057328	0.207371	1.000000	0.436783	0.392573	0.183928	-0.113970	0.074752
Insulin	-0.073535	0.331357	0.088933	0.436783	1.000000	0.197859	0.185071	-0.042163	0.130548
BMI	0.017683	0.221071	0.281805	0.392573	0.197859	1.000000	0.140647	0.036242	0.292695
DiabetesPedigreeFunction	-0.033523	0.137337	0.041265	0.183928	0.185071	0.140647	1.000000	0.033561	0.173844
Age	0.544341	0.263514	0.239528	-0.113970	-0.042163	0.036242	0.033561	1.000000	0.238356
Outcome	0.221898	0.466581	0.065068	0.074752	0.130548	0.292695	0.173844	0.238356	1.000000

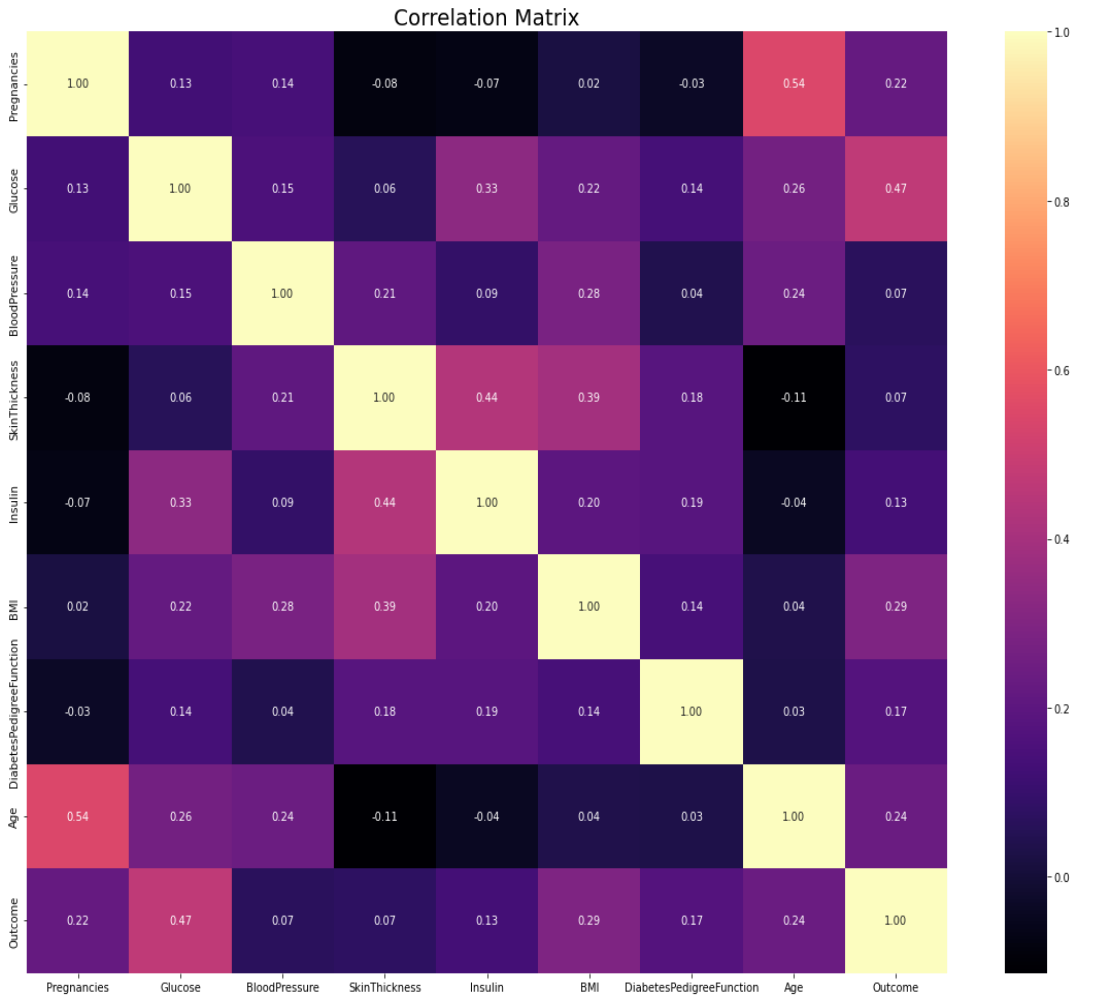

Figure 5.6: Correlation Matrix

Sample outcome which consists of the data missing

•

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148.0	72.0	35.0	NaN	33.6	0.627	50	1
1	1	85.0	66.0	29.0	NaN	26.6	0.351	31	0
2	8	183.0	64.0	NaN	NaN	23.3	0.672	32	1
3	1	89.0	66.0	23.0	94.0	28.1	0.167	21	0
4	0	137.0	40.0	35.0	168.0	43.1	2.288	33	1

Figure 5.7: Sample outcome

Number of missing values in each feature

```
Pregnancies                 0
Glucose                     5
BloodPressure              35
SkinThickness             227
Insulin                   374
BMI                        11
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

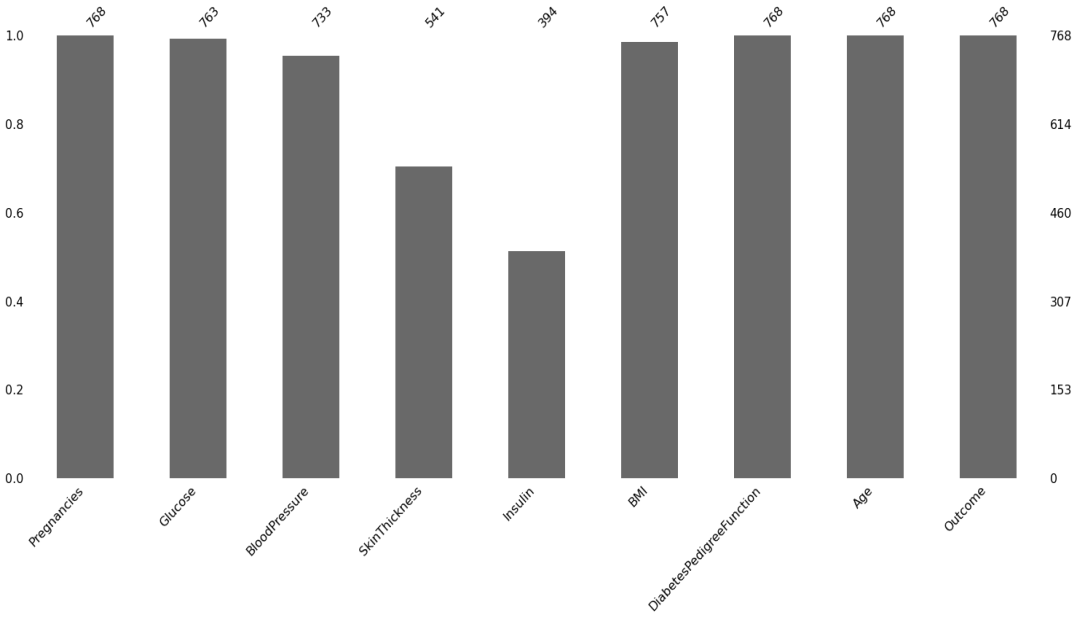

Figure 5.8: Number of Missing Values in Each feature

After data cleaning by median operation

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148.0	72.0	35.0	169.5	33.6	0.627	50	1
1	1	85.0	66.0	29.0	102.5	26.6	0.351	31	0
2	8	183.0	64.0	32.0	169.5	23.3	0.672	32	1
3	1	89.0	66.0	23.0	94.0	28.1	0.167	21	0
4	0	137.0	40.0	35.0	168.0	43.1	2.288	33	1

```
df.isnull().sum()
```

```
Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
```

Figure 5.9: Dataset After Cleaning by Median operation

According to BMI, some ranges were determined, and categorical variables were assigned.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome	NewBMI
0	6	148.0	72.0	35.0	169.5	33.6	0.627	50	1	Obesity 1
1	1	85.0	66.0	29.0	102.5	26.6	0.351	31	0	Overweight
2	8	183.0	64.0	32.0	169.5	23.3	0.672	32	1	Normal
3	1	89.0	66.0	23.0	94.0	28.1	0.167	21	0	Overweight
4	0	137.0	40.0	35.0	168.0	43.1	2.288	33	1	Obesity 3

Figure 5.10: Assignment of Ranges and categorial variables

A categorical variable creation process is performed according to the insulin value.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome	NewBMI	NewInsulinScore
0	6	148.0	72.0	35.0	169.5	33.6	0.627	50	1	Obesity 1	Abnormal
1	1	85.0	66.0	29.0	102.5	26.6	0.351	31	0	Overweight	Normal
2	8	183.0	64.0	32.0	169.5	23.3	0.672	32	1	Normal	Abnormal
3	1	89.0	66.0	23.0	94.0	28.1	0.167	21	0	Overweight	Normal
4	0	137.0	40.0	35.0	168.0	43.1	2.288	33	1	Obesity 3	Abnormal

Figure 5.11: Creation of categorial variables

Some intervals were determined according to the glucose variable, and these were assigned categorical variables.

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome	NewBMI	NewInsulinScore	NewGlucose
6	148.0	72.0	35.0	169.5	33.6	0.627	50	1	Obesity 1	Abnormal	Secret
1	85.0	66.0	29.0	102.5	26.6	0.351	31	0	Overweight	Normal	Normal
8	183.0	64.0	32.0	169.5	23.3	0.672	32	1	Normal	Abnormal	Secret
1	89.0	66.0	23.0	94.0	28.1	0.167	21	0	Overweight	Normal	Normal
0	137.0	40.0	35.0	168.0	43.1	2.288	33	1	Obesity 3	Abnormal	Secret

Figure 5.12: Determination of Intervals

Use one hot coding for better results.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome	NewBMI_Obesity 1	NewBMI_Obesity 2	NewBMI_Obesity 3	NewBMI_Overweight
0	6	148.0	72.0	35.0	169.5	33.6	0.627	50	1	1	0	0	0
1	1	85.0	66.0	29.0	102.5	26.6	0.351	31	0	0	0	0	1
2	8	183.0	64.0	32.0	169.5	23.3	0.672	32	1	0	0	0	0
3	1	89.0	66.0	23.0	94.0	28.1	0.167	21	0	0	0	0	1
4	0	137.0	40.0	35.0	168.0	43.1	2.288	33	1	0	0	1	0

```
categorical_df = df[['NewBMI_Obesity 1','NewBMI_Obesity 2', 'NewBMI_Obesity 3', 'NewBMI_Overweight','NewBMI_Underweight',
                     'NewInsulinScore_Normal','NewGlucose_Low','NewGlucose_Normal', 'NewGlucose_Overweight', 'NewGlucose_Secret']]

categorical_df.head()
```

	NewBMI_Obesity 1	NewBMI_Obesity 2	NewBMI_Obesity 3	NewBMI_Overweight	NewBMI_Underweight	NewInsulinScore_Normal	NewGlucose_Low	NewGlucose_Normal	NewGlucose_Overweight	NewGlucose_Se
0	1	0	0	0	0	0	0	0	0	
1	0	0	0	1	0	1	0	1	0	
2	0	0	0	0	0	0	0	0	0	
3	0	0	0	1	0	1	0	1	0	
4	0	0	1	0	0	0	0	0	0	

Figure 5.13: Using Hot coding Method

The dataset after performing the normalization.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	0.6	0.775	0.000	1.000000	1.000000	0.177778	0.669707	1.235294
1	-0.4	-0.800	-0.375	0.142857	0.000000	-0.600000	-0.049511	0.117647
2	1.0	1.650	-0.500	0.571429	1.000000	-0.966667	0.786971	0.176471
3	-0.4	-0.700	-0.375	-0.714286	-0.126866	-0.433333	-0.528990	-0.470588
4	-0.6	0.500	-2.000	1.000000	0.977612	1.233333	4.998046	0.235294

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	NewBMI_Obesity 1	NewBMI_Obesity 2	NewBMI_Obesity 3	NewBMI_Overweight	NewBMI_
0	0.6	0.775	0.000	1.000000	1.000000	0.177778	0.669707	1.235294	1	0	0	0	
1	-0.4	-0.800	-0.375	0.142857	0.000000	-0.600000	-0.049511	0.117647	0	0	0	1	
2	1.0	1.650	-0.500	0.571429	1.000000	-0.966667	0.786971	0.176471	0	0	0	0	
3	-0.4	-0.700	-0.375	-0.714286	-0.126866	-0.433333	-0.528990	-0.470588	0	0	0	1	
4	-0.6	0.500	-2.000	1.000000	0.977612	1.233333	4.998046	0.235294	0	0	1	0	

Figure 5.14: Dataset after performing the Normalization

Algorithms used their average accuracy, and the standard deviation is mentioned within the brackets.

```
LR: 0.848684 (0.036866)
KNN: 0.840789 (0.023866)
CART: 0.857895 (0.024826)
RF: 0.881579 (0.026316)
SVM: 0.853947 (0.036488)
XGB: 0.890789 (0.020427)
LightGBM: 0.882895 (0.026610)
```

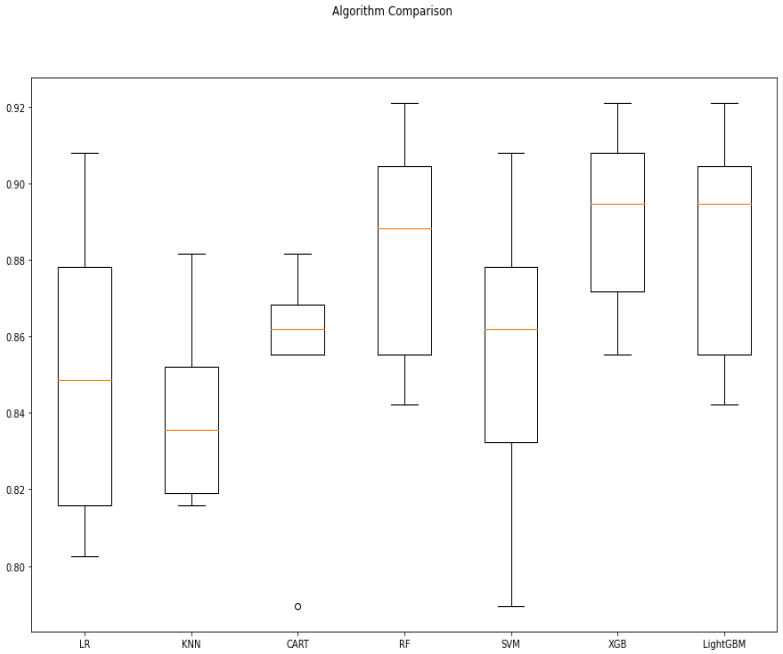

Figure 5.15: Representation of Algorithms

Random Forest

The best features found after multiple iterations

```
'max_depth': 8,
'max_features': 7,
'min_samples_split': 2,
'n_estimators': 500
 Accuracy was   0.8921052631578947
```

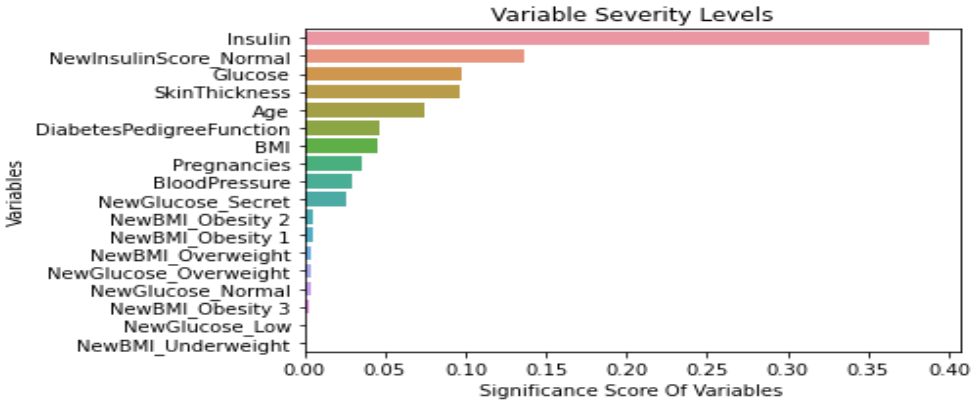

Figure 5.16: Graphical representation of random forest

LGBM Classifier

```
Accuracy 0.8960526315789474
```

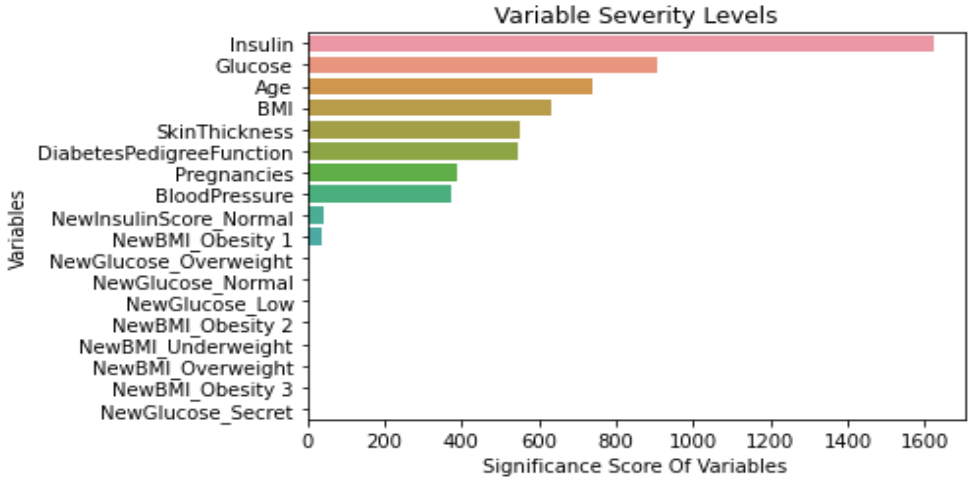

Figure 5.17: Graphical representation of LGBM Classifier

XGBOOST Model

```
Accuracy 0.9013157894736843
```

Figure 5.18: Graphical representation of XGBOOST

Comparison of Algorithms

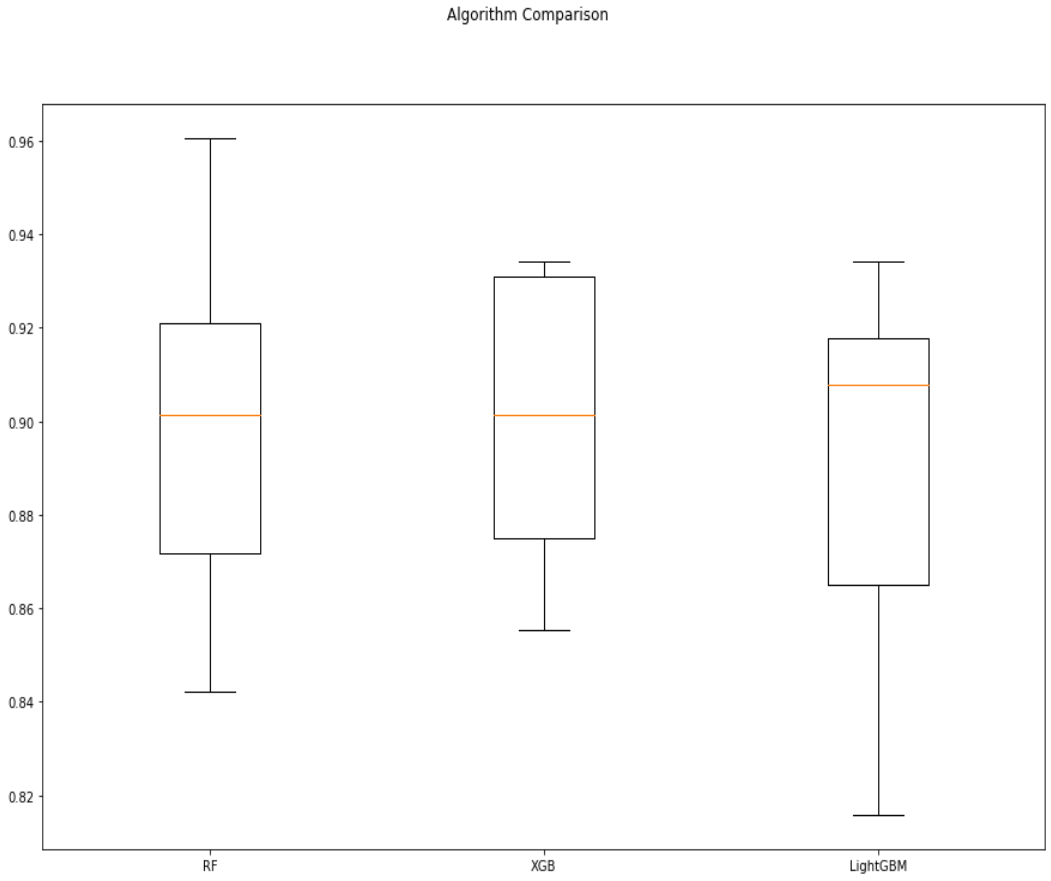

Figure 5.19: Comparison of Algorithms

CHAPTER 6

CONCLUSION

To design and implement Diabetes Prediction Using Machine Learning Methods and Performance Analysis of that methods and it has been achieved successfully. Machine learning has been proven successful over the conventional methods and the results presented in the previous section proves it. Various machine learning algorithms are applied on the data after various stages of cleaning operations performed on the data. The proposed approach uses various classification and ensemble learning method in which SVM, Knn, Random Forest, Decision Tree, Logistic Regression and One hot encoding are used. The machine learning models better performance when the data is subjected to cleaning and normalization.

CHAPTER 7

REFERENCES

[1] Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importance's for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.

[2] K.Vijiya Kumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest]Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Computation Automation and Networking, 2019.

[3] Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019.

[4] Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int. Journal of Engineering Research and Application, Vol. 8, Issue 1, (Part -II) January 2018, pp.-09-13

[5] Nonso Nnamoko, Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC), 2018.

[6] Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabetes Disease Prediction Using Data Mining ".International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017.

[7] Nahla B., Andrew et al,"Intelligible support vector machines for diagnosis of diabetes mellitus. Information Technology in Biomedicine", IEEE Transactions. 14, (July. 2010), 1114-20.

[8] A.K., Dewangan, and P., Agrawal, "Classification of Diabetes Mellitus Using Machine Learning Techniques," International Journal of Engineering and Applied Sciences, vol. 2, 2015

[9] Ayon, SI, Islam, MM & Hossain, MR 2020, 'Coronary artery heart disease prediction: A comparative study of computational intelligence techniques', IETE, Journal of Research, pp. 1-20.

[10] Azhar, M & Thomas, PA 2020, 'Heart disease prediction based on an optimal feature selection method using autoencoder', International Journal of Scientific Research in Science and Technology, vol. 7, no. 4, pp. 25-38.

[11] Babič, F, Olejár, J, Vantová, Z & Paralič, J 2017, 'Predictive and descriptive analysis for heart disease diagnosis', in 2017 Federated Conferences on Computer Science and Information Systems (Fedcsis), pp. 155-163.

[12] Babu, SB, Suneetha, A, Babu, GC, Kumar, YJN & Karuna, G 2018, 'Medical disease prediction using grey wolf optimization and auto encoder based recurrent neural network', Periodicals of Engineering and Natural Sciences (PEN), vol. 6, pp. 229-240.

[13] Beyene, C & Kamat, P 2018, 'Survey on prediction and analysis the occurrence of heart disease using data mining techniques', International Journal of Pure and Applied Mathematics, vol. 118, pp. 165-174.

[14] Bharti, R, Khamparia, A, Shabaz, M, Dhiman, G, Pande, S & Singh, P 2021, 'Prediction of heart disease using a combination of machine learning and deep learning', Computational Intelligence and Neuroscience, vol. 2021.

[15] Buchan, K, Filannino, M & Uzuner, Ö 2017, 'Automatic prediction of coronary artery disease from clinical narratives', Journal of Biomedical Informatics, vol. 72, pp. 23-32, 2017.

[16] Burse, K, Kirar, VPS, Burse, A & Burse, R (Eds.) 2019, 'Various preprocessing methods for neural network-based heart disease prediction', Smart Innovations in Communication and Computational Sciences, Springer, Singapore.

[17] Chantar, H, Mafarja, M, Alsawalqah, H, Heidari, AA, Aljarah, I & Faris, H 2020, 'Feature selection using binary grey wolf optimizer with elite-based crossover for Arabic text classification', Neural Computing and Applications, vol. 32, no. 16, pp. 12201-12220.

[18] Aljumah, A.A., Ahamad, M.G., Siddiqui, M.K., Application of data mining: Diabetes health care in young and old patients. Journal of King

Saud University - Computer and Information Sciences 25, 127–136.

[19] Nesreen Samer El Jerjawi, and Samy S. Abu-Naser* "Diabetes Prediction Using Artificial Neural Network". Department of Information Technology, Faculty of Engineering Information Technology.

[20] Srikar Sistla, Predicting Diabetes using SVM Implemented by Machine Learning,International Journal of Soft Computing and Engineering (IJSCE), 2022.