

SUPERVISED LEARNING METHOD OF DIABETES PREDICTION

A Project Report Phase – II

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Engineering degree in Electronics and Communication Engineering

by

Chidipudi AjayKumarReddy (39130092)

Dasari Jeswanth Sai Brahmesh (39130104)



**DEPARTMENT OF ELECTRONICS AND COMMUNICATIN ENGINEERING
SCHOOL OF ELECTRICAL AND ELECTRONICS**

SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

Accredited with Grade “A” by NAAC

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600 119

APRIL - 2023



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited with "A" grade by NAAC
Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai – 600 119
www.sathyabama.ac.in

DEPARTMENT OF ELECTRONICS AND COMMUNICAITON ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **CHIDIPUDI AJAYKUMARREDDY (39130092)** and **D.J.S. BRAHMESH (39130104)** who carried out the project title **"SUPERVISED LEARNING METHOD OF DIABETES PREDICTION"** under our supervision from November 2022 to April 2023.

Internal Guide

Dr. V. BALAMURUGAN.,M.Tech., Ph. D.,

Balamurugan

Head of the Department

T. Ravi
15/4/2023

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

Submitted for Viva voce Examination held on _____

T. Ravi

Internal Examiner

S. Ravi

External Examiner

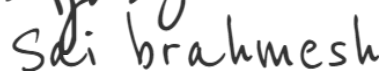
DECLARATION

We **CHIDIPUDI AJAYKUMARREDDY** and **D.J.S.BRAHMESH** declare that the Project Report entitled “ **SUPERVISED LEARNING METHOD OF DIABETES PREDICTION** ” done by us under the guidance of **Dr. V. BALAMURUGAN.,M.Tech., Ph. D.,** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in Electronics and Communication Engineering.

DATE:

PLACE:

SIGNATURE OF THE CANDIDATES

1. 
2. 

ACKNOWLEDGEMENT

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

We convey our thanks to **Dr. N. M. NANDHITHA, M.E., Ph.D., Professor & Dean, School of Electrical and Electronics** and **Dr. T. RAVI, Ph.D., Professor & Head, Department of Electronics and Communication Engineering** for providing us necessary support and details at the right time during the progressive reviews.

We would like to express my sincere and deep sense of gratitude to our Project Guide **Dr. V. BALAMURUGAN.,M.Tech., Ph. D.,**for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of our project work.

We wish to express our thanks to all Teaching and Non-teaching staff members of the Department of Electronics and Communication Engineering who were helpful in many ways for the completion of the project.

ABSTRACT

Nowadays, e-Health service has become a booming area, which refers to computer-based health care and information delivery to improve health service locally, regionally and worldwide. An effective disease risk prediction model by analyzing electronic health data benefits not only to care a patient but also to provide services through the corresponding datadriven e-Health systems. In this paper, we particularly focus on predicting and analyzing diabetes mellitus, an increasingly prevalent chronic disease that refers to a group of metabolic disorders characterized by a high blood sugar level over a prolonged period of time.

Support Vector Machine (SVM) is one of the most popular and simplest machine learning techniques to build such a disease risk prediction model utilizing relevant health data. In order to achieve our goal, we present an optimal Support Vector Machine learning based prediction model based on patient's habitual attributes in various dimensions. This approach determines the optimal number of neighbors with low error rate for providing better prediction outcome in the resultant model. The effectiveness of this machine learning e-Health model is examined by conducting experiments on the real-world diabetes mellitus data collected from medical hospitals.

TABLE OF CONTENTS

CHAPTER NO.	TITTLE	PAGE NO.
	ABSTRACT	V
	CONTENTS	VI
	LIST OF FIGURES AND TABLES	VII
1	INTRODUCTION	01
	1.1 OUT LINE OF THE PROJECT	01
	1.2 MACHINE LEARNING	04
	1.3 MACHINE LEARNING STATERGIES	08
	1.4 SUPERVISED LEARNING	10
2	LITERATURE SURVEY	13
	2.1 RELATED WORK	13
	2.2 OPEN PROBLEMS IN EXISTING SYSTEM	22
3	METHODOLOGY	24
	3.1 EXISTING SYSTEM	24
	3.2 PROPOSED SYSTEM	25
	3.3 DATA ATTRIBUTES	28
	3.4 RISK ANALYSIS	29
	3.5 ADVANTAGES OF PROPOSED SYSTEM	30
	3.6 OBJECTIVES	31
	3.7 SOFTWARE & HARDWARE REQUIREMENTS	32
	3.8 ALGORITHMS USED	33
	3.9 DATA FLOW DIAGRAM	35
	3.10 MODULES	36

4	IMPLEMENTATION DETAILS	37
	4.1 IMPORTING DEPENDENCIES	37
	4.2 DATA COLLECTING AND PRE-PROCESSING	38
	4.3 DATA STADARDIZATION	43
	4.4 TRAIN AND TEST SPLIT	44
	4.5 MODEL TRAINING	45
	4.6 TRAINING THE SVM CLASSIFIER	46
	4.7 EVALUATION OF MODEL	46
	4.8 SAVING THE TRAINING MODEL	47
	4.9 MAKING PREDICTIVE SYSTEM	48
5	RESULT AND DISCUSSION	50
6	CONCLUSION	51
7	REFERENCES	52

LIST OF FIGURES

FIG NO.	TITTLE OF THE FIGURE	PAGE NO.
1.1	MACHINE LEARNING TECHNIQUES	07
1.2	MACHINE LEARNING WORK FLOW	09
1.3	SUPERVISED LEARNING	12
3.1	PROPOSED METHODOLOGY	26
3.2	WORK FLOW	27
3.3	DATA FLOW DIAGRAM	35

LIST OF TABLES

TABLE NO.	TITTLE OF THE TABLE	PAGE NO.
4.1	SAMPLE DATA SET	39
4.2	STASTICAL DATA SET	40
4.3	OUTCOME VALUE COUNT	40
4.4	LABLES FOR THE OUTPUT VALUES	41
4.5	MEAN DATA VALUES	41
4.6	SAMPLE DATA VALUES	42
4.7	OUTCOME VALUES	42

CHAPTER – 1

INTRODUCTION

1.1 OUTLINE OF THE PROJECT

Diabetes may be a set of metabolic problems known by high blood sugar levels over a protracted period of our time. Diabetes is outlined as a bunch of metabolic disorders in the main caused by abnormal insulin secretion and/or action. Symptoms of high aldohexose incorporate excessive voiding, continually feeling thirsty and enlarged hunger. If not treated on time, diabetes will cause serious health problems in a person like diabetic acidosis, hyperosmolar hyperglycemic state, or maybe result in death. This could result in time period complications as well as vas upset, brain stroke, failure, ulcers within the foot, and eye complications etc.

Diabetes is caused once the duct gland within the body is unable to come up with insulin in enough amounts or once the cells and tissues within the body fail to utilize the insulin created. Diabetes exists in 3 forms: Diabetes Mellitus Type-I is characterized by duct gland generating insulin but what's needed by the body, a condition conjointly referred to as "insulin-subordinate diabetes mellitus"(IDDM). Folks littered with type-I DM need external insulin indefinite quantity to form up for the less insulin created by the duct gland. Diabetes Mellitus Type-2 is marked by the body resisting insulin because the body cells react otherwise to insulin than they traditional would. This could ultimately result in no insulin within the body. This can be otherwise referred to as "non-insulin subordinate diabetes mellitus"(NIDDM) or "adult beginning diabetes".

This sort of diabetes is often found in folks with high BMI or people who lead associate degree inactive manner. Gestational diabetes is that the third principle structure that's ascertained throughout physiological state. Generally, for a traditional person, aldohexose levels vary from seventy to ninety-nine milligrams per deciliter. An individual is taken into account diabetic providing the fast aldohexose level is found to be over 126 mg/dL. Within the practice, an individual having an aldohexose concentration of a hundred to one hundred twenty five mg/dL is taken into account as pre-diabetic.

Diabetes is a chronic disease that affects millions of people worldwide. It is a metabolic disorder characterized by high blood sugar levels, either due to the body's inability to produce enough insulin or because the body is not responding to insulin properly. Diabetes can lead to a range of complications, including heart disease, kidney disease, and nerve damage, among others. The prevalence of diabetes is increasing rapidly, and early detection is crucial for effective management of the disease and prevention of complications.

Machine learning algorithms have shown great promise in predicting diabetes risk, and supervised learning is one such approach. Supervised learning is a type of machine learning where the algorithm is trained on a labeled dataset, which means that the data is already classified into different categories. In the case of diabetes prediction, the labeled dataset would consist of individuals who have been diagnosed with diabetes and those who have not. The input variables in the dataset could include factors such as age, weight, blood pressure, family history of diabetes, and other relevant health information. The output variable would be whether the individual has diabetes or not.

The goal of supervised learning in diabetes prediction is to develop a model that can accurately predict an individual's risk of developing diabetes. The model is trained using the labeled dataset, and once it is trained, it can be used to predict the risk of diabetes in new, unseen individuals. This can be incredibly useful for healthcare professionals who can use the model to identify individuals who are at high risk of developing diabetes and take preventative measures accordingly.

There are many different supervised learning algorithms that can be used for diabetes prediction, including decision trees, logistic regression, support vector machines, and neural networks. Each algorithm has its strengths and weaknesses, and the performance of the algorithm can vary depending on the dataset used.

In this project, we will explore supervised learning methods for diabetes prediction, comparing the performance of different algorithms and identifying the most effective approach. We will start by collecting a dataset that includes information on individuals who have been diagnosed with diabetes and those who

have not. We will then preprocess the dataset to ensure that it is clean and ready for use in supervised learning algorithms.

Next, we will split the dataset into training and testing sets. The training set will be used to train the model, while the testing set will be used to evaluate the model's performance. We will use various supervised learning algorithms to train the model, and we will compare their performance using metrics such as accuracy, precision, recall, and F1 score.

Once we have identified the most effective approach, we will fine-tune the model to improve its performance further. This can involve tweaking the hyperparameters of the model or using techniques such as regularization to prevent overfitting.

We will also explore different feature selection methods to identify the most relevant features for predicting diabetes risk. We will test the model on a new dataset to see how well it generalizes to unseen data. This will give us an idea of how well the model will perform in real-world scenarios. We will also evaluate the model's interpretability, as it is essential to understand how the model is making predictions to gain trust in its predictions.

supervised learning is a powerful tool for diabetes prediction, and this project aims to explore the effectiveness of various supervised learning algorithms in this context. The goal is to develop a model that can accurately predict an individual's risk of developing diabetes, which can be incredibly useful for healthcare professionals and individuals looking to manage their health. The results of this project have the potential to improve the early detection and management of diabetes, ultimately leading to better health outcomes for individuals with diabetes.

The goal of supervised learning in diabetes prediction is to develop a model that can accurately predict an individual's risk of developing diabetes. The model is trained using the labeled dataset, and once it is trained, it can be used to predict the risk of diabetes in new, unseen individuals. This can be incredibly useful for healthcare professionals who can use the model to identify individuals who are at high risk of developing diabetes and take preventative measures accordingly.

1.2 MACHINE LEARNING

Machine learning is a subset of artificial intelligence that involves the use of algorithms to enable machines to learn from data and make predictions or decisions without being explicitly programmed. Machine learning algorithms use statistical models to analyze data, identify patterns, and make predictions or decisions based on that data. There are three main types of machine learning: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning involves training the algorithm on a labeled dataset, where the correct output is already known. The algorithm then uses this labeled dataset to make predictions or decisions about new, unseen data. Unsupervised learning involves training the algorithm on an unlabeled dataset, where the output is unknown.

The algorithm then identifies patterns and relationships within the data. Reinforcement learning involves training the algorithm through trial and error, where it receives feedback in the form of rewards or punishments based on its actions. Machine learning has many applications in various industries, including healthcare, finance, marketing, and more. In healthcare, machine learning is used for disease diagnosis, drug discovery, and personalized medicine. In finance, machine learning is used for fraud detection, risk assessment, and algorithmic trading. In marketing, machine learning is used for customer segmentation, recommendation systems, and predictive analytics.

Machine learning algorithms can be categorized into several groups, including decision trees, random forests, support vector machines, and neural networks. Each algorithm has its strengths and weaknesses, and the choice of algorithm depends on the nature of the problem being solved. The success of machine learning algorithms depends heavily on the quality and quantity of data used for training. Machine learning algorithms require large datasets to identify patterns and relationships accurately. The data must also be clean and representative of the problem being solved. The pre-processing of the data, including feature selection, feature engineering, and normalization, is critical for the performance of the algorithm. In recent years, deep learning has become a popular subset of machine learning that involves the use of neural networks with multiple hidden layers. Deep learning has been shown to achieve state-of-the-art

performance on various tasks such as image recognition, natural language processing, and speech recognition. Machine learning is a rapidly growing field that has the potential to revolutionize various industries. It involves the use of algorithms to enable machines to learn from data and make predictions or decisions without being explicitly programmed.

The success of machine learning algorithms depends on the quality and quantity of data used for training, as well as the choice of algorithm and preprocessing techniques. As the amount of data continues to grow, and the computational power of machines increases, the potential applications of machine learning will continue to expand.

Machine learning techniques are algorithms that enable machines to learn and improve performance on a given task by analyzing data and identifying patterns. These techniques are widely used in various fields, such as computer vision, natural language processing, robotics, and data analytics. Here are some common machine learning techniques:

- 1. Supervised Learning:** Supervised learning is a technique where the machine is trained on a labeled dataset to predict the output of new data. The algorithm learns from a set of input features and corresponding target values and is then able to make predictions for new data based on what it learned.
- 2. Unsupervised Learning:** Unsupervised learning is a technique where the machine learns patterns in a dataset without any labeled data. The algorithm identifies underlying patterns and structure in the data, which can help to group similar data points or identify outliers.
- 3. Reinforcement Learning:** Reinforcement learning is a technique where the machine learns by interacting with an environment and receiving rewards or punishments based on its actions. The algorithm learns to optimize its behavior over time to maximize the reward.
- 4. Deep Learning:** Deep learning is a subset of machine learning that uses neural networks to learn patterns in data. It involves training a deep neural network with many layers to extract high-level features from the input data.

- 5. Transfer Learning:** Transfer learning is a technique where a pre-trained model is used as a starting point for a new task. The pre-trained model has already learned general features from a large dataset, which can be fine-tuned on a smaller, more specific dataset.
- 6. Ensemble Learning:** Ensemble learning is a technique where multiple models are trained and combined to improve performance. It can involve combining multiple algorithms, such as decision trees or neural networks, or combining the outputs of multiple models.
- 7. Dimensionality Reduction:** Dimensionality reduction is a technique where the number of input features is reduced while preserving the most important information in the data. This can help to reduce computational costs and improve model performance.

Machine learning is a subset of artificial intelligence that focuses on the development of algorithms and statistical models that enable computer systems to learn from and make predictions or decisions based on data. It has become an integral part of many modern technologies, from virtual assistants and recommendation systems to self-driving cars and facial recognition software.

There are several types of machine learning algorithms, including supervised learning, unsupervised learning, and reinforcement learning. Supervised learning involves training a model on labeled data, while unsupervised learning involves finding patterns in unlabeled data. Reinforcement learning involves training a model to make decisions based on feedback from its environment.

Machine learning technologies are used in a wide range of industries, including healthcare, finance, and e-commerce. They are used to analyze large datasets, detect fraud, identify trends, and make predictions. Machine learning is also used in natural language processing, computer vision, and speech recognition.

Some of the popular machine learning tools and frameworks used by developers today include TensorFlow, PyTorch, scikit-learn, and Keras. These tools help developers build and deploy machine learning models quickly and efficiently.

As machine learning technologies continue to evolve and improve, they are expected to play an increasingly important role in shaping the future of technology and transforming the way we live and work. The rise of machine learning has also led to the development of deep learning, a subfield of machine learning that uses artificial neural networks to learn from data. Deep learning has been responsible for significant breakthroughs in fields such as image recognition, natural language processing, and game playing.

Machine learning technologies are also playing a crucial role in addressing some of the world's most pressing challenges, such as climate change, healthcare, and education. For example, machine learning can help identify patterns in environmental data to predict and prevent natural disasters. It can also analyze medical data to improve disease diagnosis and treatment.

However, machine learning also raises ethical and social issues, such as bias in algorithms, data privacy, and job displacement. As machine learning continues to become more widespread, it is important to address these issues and ensure that these technologies are developed and used responsibly.

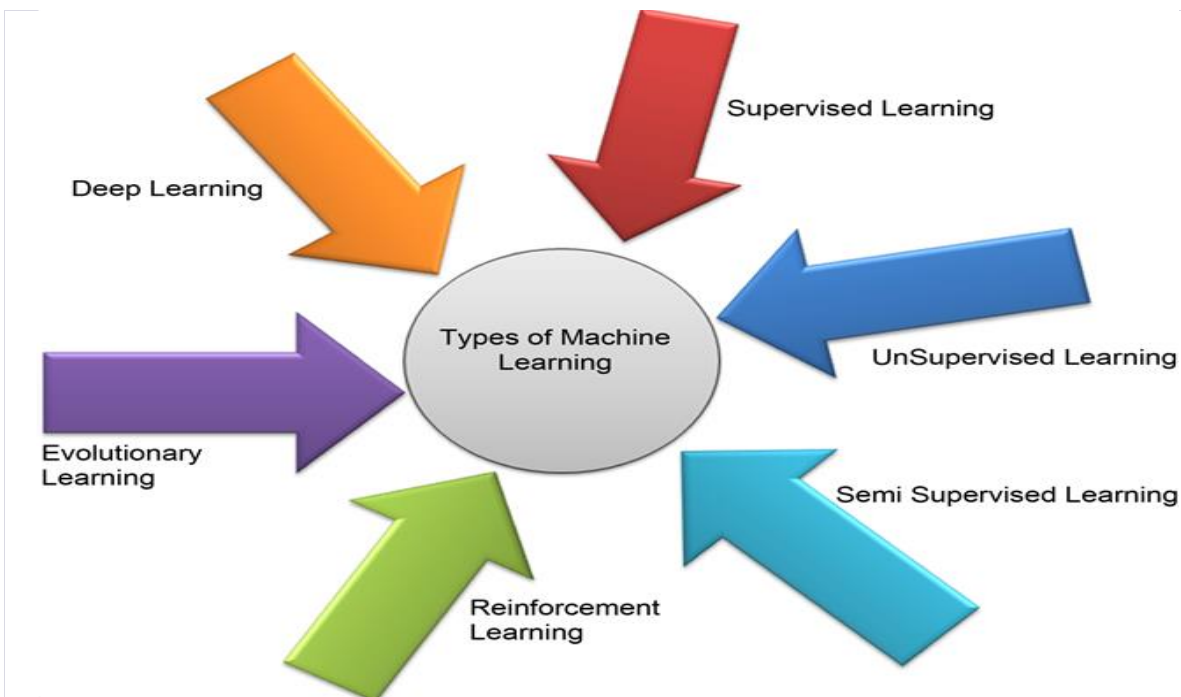


Fig. 1.1 : MACHINE LEARNING TECHNIQUES

1.3 MACHINE LEARNING STRATEGIES

Machine learning strategies refer to the various techniques and methods used to train, test, and optimize machine learning models. Some of the most common strategies used in machine learning include cross-validation, regularization, ensemble methods, and transfer learning.

1. Cross-validation is a technique used to evaluate the performance of a machine learning model by splitting the data into multiple subsets for training and testing. This allows for a more robust evaluation of the model's performance and helps to prevent overfitting.
2. Regularization is a technique used to prevent overfitting in machine learning models by adding a penalty term to the objective function. This penalty term discourages the model from fitting too closely to the training data and instead encourages the model to generalize better to new, unseen data.
3. Ensemble methods involve combining multiple machine learning models to improve their overall performance. This can be done by averaging the predictions of multiple models, or by using more complex methods such as boosting or bagging.
4. Transfer learning is a technique that involves leveraging pre-trained models to improve the performance of a new, related task. This is achieved by fine-tuning the pre-trained model on the new task, rather than training a new model from scratch.

Other machine learning strategies include feature selection, hyperparameter tuning, and data augmentation. Feature selection involves selecting the most relevant features from a dataset to improve the performance of a machine learning model. Hyperparameter tuning involves optimizing the hyperparameters of a model to improve its performance. Data augmentation involves generating new training data from existing data to increase the size and diversity of the training dataset. The choice of machine learning strategy depends on the specific problem being solved.

and the characteristics of the data. By using a combination of these strategies, machine learning models can be optimized for maximum performance and accuracy.

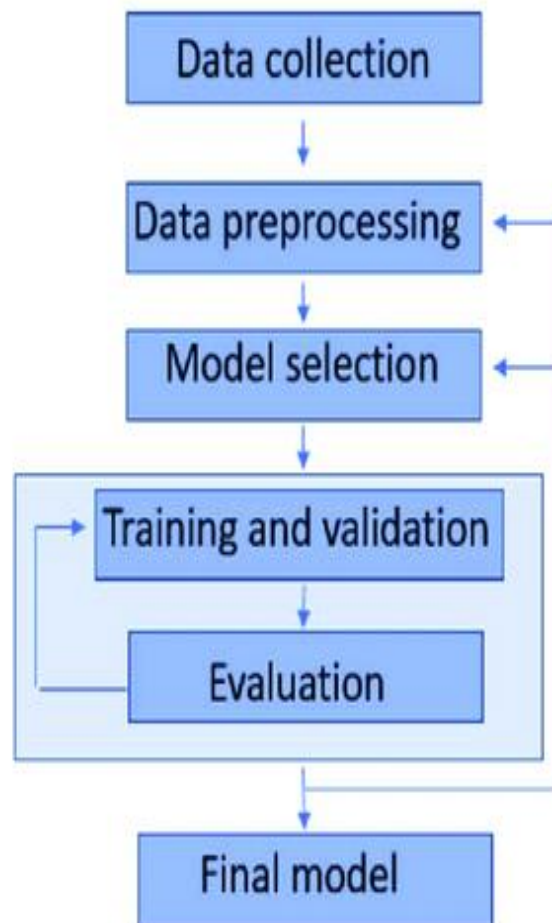


Fig. 1.2 : MACHINE LEARNING WORK FLOW

Machine learning strategies are methods used to develop machine learning models to make predictions or decisions based on data. These strategies include supervised learning, unsupervised learning, and reinforcement learning. Supervised learning involves training a model on labeled data, unsupervised learning involves finding patterns in unlabeled data, and reinforcement learning involves training a model to make decisions based on feedback from its environment. These strategies are used in a wide range of applications, from natural language processing and computer vision to healthcare and finance.

1.4 SUPERVISED LEARNING

Supervised learning is a type of machine learning where the algorithm is trained on a labeled dataset. In supervised learning, the dataset is divided into a set of features (or input variables) and a set of labels (or output variables). The algorithm then learns to map the input features to the corresponding output labels by minimizing an objective function.

Supervised learning can be used for both classification and regression problems. In classification, the algorithm learns to classify data into predefined categories based on the input features. In regression, the algorithm learns to predict a continuous output variable based on the input features.

Some common algorithms used in supervised learning include decision trees, support vector machines, and neural networks. The performance of a supervised learning algorithm depends heavily on the quality and quantity of the labeled data used for training. The dataset should be representative of the problem being solved and the labels should be accurate.

Supervised learning has many applications in various industries, including healthcare, finance, and marketing. In healthcare, supervised learning can be used for disease diagnosis and drug discovery. In finance, it can be used for fraud detection and risk assessment. In marketing, it can be used for customer segmentation and predictive analytics.

Supervised learning is a powerful technique for solving a wide range of problems, particularly those involving pattern recognition and prediction. By training on labeled data, supervised learning algorithms can learn to accurately classify and predict new data, making it an essential tool for many applications in industry and research.

Supervised learning techniques are commonly used in diabetes prediction to build models that can classify individuals as either diabetic or non-diabetic based on a set of input features. Here are some of the common techniques used in supervised learning for diabetes prediction:

1. **Logistic Regression:** Logistic regression is a commonly used statistical technique for binary classification tasks. It is particularly useful when the relationship between the input features and the output variable is nonlinear.
2. **Decision Trees:** Decision trees are a type of supervised learning algorithm that can be used for both classification and regression tasks. They work by partitioning the data into smaller subsets based on the values of the input features, and then making a decision based on the final subset.
3. **Support Vector Machines (SVMs):** SVMs are a popular machine learning algorithm that can be used for binary classification tasks. They work by finding a hyperplane that separates the data into distinct classes, and can handle nonlinearly separable data using kernel functions.
4. **Random Forests:** Random forests are an ensemble learning method that combines multiple decision trees to create a more robust and accurate prediction model. They can handle a large number of input features without requiring feature selection or dimensionality reduction.
5. **Artificial Neural Networks (ANNs):** ANNs are a type of machine learning algorithm that can be used for both classification and regression tasks. They work by simulating the structure and function of the human brain, and can handle complex and nonlinear relationships between the input features and the output variable.
6. **Naive Bayes:** Naive Bayes is a probabilistic machine learning algorithm that can be used for binary classification tasks. It works by calculating the probability of each class given the input features, and then making a decision based on the highest probability.
7. **K-Nearest Neighbors (KNN):** KNN is a simple yet effective machine learning algorithm that can be used for binary classification tasks. It works by finding the K nearest data points to a new data point, and then classifying the new data point based on the majority class of its neighbors.

These techniques can be used alone or in combination with each other to build highly accurate and reliable models for diabetes prediction. The choice of technique will depend on the specific characteristics of the dataset and the goals of the project.

Supervised learning is a machine learning strategy in which a model is trained on labeled data to make predictions or decisions about new, unseen data. In supervised learning, the model learns to map input variables to output variables based on a training dataset consisting of labeled examples. The labeled examples include input variables and their corresponding output variables. The model is trained to minimize the difference between its predictions and the actual output variables. Once the model is trained, it can be used to make predictions on new, unseen data. Supervised learning is used in a wide range of applications, including natural language processing, computer vision, and predictive analytics.

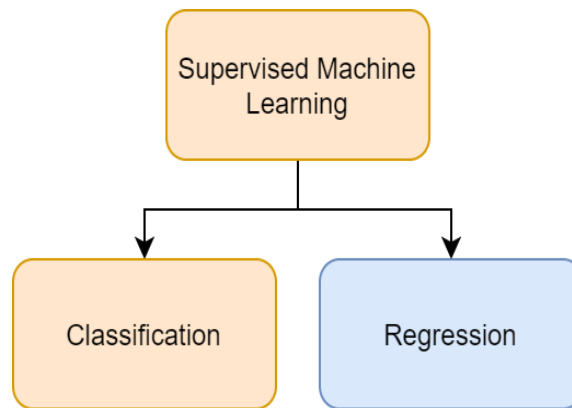


Fig. 1.3 : SUPERVISED LEARNING

CHAPTER – 2

LITERATURE SURVEY

2.1 RELATED WORK

Prediction of Diabetes using Classification Algorithms

Diabetes isn't a hereditary disorder however heterogeneous group of disorder which could ultimately result in an boom of glucose within the blood and lack of glucose inside the urine. Diabetes is typically resulting from genetics, way of life and surroundings. Eating an dangerous weight loss plan, being overweight play role in developing the diabetes. High blood sugar tiers can also result in kidney diseases, coronary heart illnesses. The excess of sugar in the blood can harm the tiny blood vessels in your frame. Signs of diabetes are blurry imaginative and prescient , extreme hunger, unusual weight reduction, common urination and thirsty.

In this project, parameters used within the facts set to locate the diabetes are Glucose, Blood pressure, pores and skin thickness, Insulin, Age. Huge volumes of statistics units are generated by health care industries. Those facts sets is a collection of patient information about the diabetes from the hospitals. Big records analytics is the processing which it examines the information units and exhibits the hidden information. Pima Indians Diabetes Database (PIDD), this dataset is taken from the national Institute of Diabetes and Digestive diseases. The objective of the dataset is to predict whether or not the patient has diabetes or not, primarily based on diagnostic measurements in the dataset. Several constraints were taken from the massive database.

A Novel Technique To Predict Diabetic Disease Using Data Mining.

G. Krishnaveni , Prof. T.Sudha/ 2018

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. The software of data mining is an analytical tool for analyzing data. Data mining has become a main strategy in many industries to improve outputs and decrease costs. Nowadays in healthcare management this

field will become very useful. Data mining techniques has become great potential for the healthcare industry to predict health deceases by using systematic data and analytics to identify inefficiencies and best practices that improve care and reduce costs.

These techniques are fast in nature and take less time for the prediction system to improve the diabetic decease with more accuracy. In this paper we are applying the various classification techniques over diabetic mellitus decease dataset for the prediction of decease and non decease person. The diabetic database is preprocessed to make the mining process more efficient. The preprocessed data is used to predict using classification algorithms like Discriminant analysis, KNN, Naive Bayes and Support vector machine. These classifiers can be efficiently used in bioinformatics problem. We are analyzing the various classification techniques like Discriminant analysis, KNN, Naive Bayes and Support vector machine with linear and RBF kernel function and showing their accuracy.

Limitation: Low Accuracy

Diabetes prediction using machine learning algorithms

Alharthi / 2021

Comprehensive review article that evaluates the recent research on diabetes prediction using various machine learning algorithms. The authors provide an in-depth analysis of different machine learning algorithms, including logistic regression, decision tree, random forest, and support vector machine, highlighting their advantages and limitations in predicting diabetes. The review article also covers the use of feature selection techniques to improve the accuracy of diabetes prediction. The authors discuss the effectiveness of different feature selection methods, such as recursive feature elimination, mutual information, and chi-square test.

The article provides an overview of various data preprocessing techniques used in diabetes prediction, such as normalization, standardization, and imputation. The authors evaluate the performance of different machine learning algorithms on various datasets, and provide a comparative analysis of their results. The article

also discusses the challenges faced in diabetes prediction using machine learning algorithms, such as imbalanced datasets, overfitting, and missing data. The review article by Alharthi, et al. provides a comprehensive overview of recent research on diabetes prediction using machine learning algorithms, and highlights the potential of these methods in improving diabetes diagnosis and management.

Limitation: This review article is that it only covers studies published until 2021, and may not include the most recent research on diabetes prediction using machine learning algorithms. Nonetheless, the review provides a valuable summary of the current state of research in this field, and identifies areas for future research.

Review on Prediction of Diabetes using Data Mining Technique

Diabetes mellitus is one of the world's major diseases. Millions of people are affected by the disease. The risk of diabetes is increasing day by day and is found mostly in women than men. The diagnosis of diabetes is a tedious process. So with improvement in science and technology it is made easy to predict the disease. The purpose is to diagnose whether the person is affected by diabetes or not using K Nearest Neighbor classification technique. The diabetes dataset is taken as the training data and the details of the patient are taken as testing data. The training data are classified by using the KNN classifier and secondly the target data is predicted. KNN algorithm used here would be more efficient for both classification and prediction. The results are analyzed with different values for the parameter k .

A Prediction Technique in Data Mining for Diabetes Mellitus

Diabetes mellitus is a chronic disease characterized by hyperglycemia. It may cause many complications. According to the growing morbidity in recent years, in 2040, the world's diabetic patients will reach 642 million, which means that one of the ten adults in the future is suffering from diabetes. There is no doubt that this alarming figure needs great attention. With the rapid development of machine learning, machine learning has been applied to many aspects of medical health.

In this study, we used decision tree, random forest and neural network to predict diabetes mellitus. The dataset is the hospital physical examination data in Luzhou, China. It contains 14 attributes. In this study, five-fold cross validation was

used to examine the models. In order to verify the universal applicability of the methods, we chose some methods that have the better performance to conduct independent test experiments. We randomly selected 68994 healthy people and diabetic patients' data, respectively as training set. Due to the data unbalance, we randomly extracted 5 times data. And the result is the average of these five experiments. In this study, we used principal component analysis (PCA) and minimum redundancy maximum relevance (mRMR) to reduce the dimensionality. The results showed that prediction with random forest could reach the highest accuracy (ACC = 0.8084) when all the attributes were used.

A model for early prediction of diabetes

Talha Mahboob Alam and Zunish Abbas / 2019

In this research paper, diabetes is predicted using significant attributes, and the relationship of the differing attributes is also characterized. Various tools are used to determine significant attribute selection and for clustering, prediction, and association rule mining for diabetes. Significant attributes selection was done via the principal component analysis method. Our findings indicate a strong association of diabetes with body mass index (BMI) and with glucose level, which was extracted via the Apriori method.

Limitation: The limitation of this study is that a structured dataset has been selected but in the future, unstructured data will also be considered, and these methods will be applied to other medical domains for prediction, such as for different types of cancer, psoriasis, and Parkinson's disease. Other attributes including physical inactivity, family history of diabetes, and smoking habit, are also planned to be considered in the future for the diagnosis of diabetes.

Prediction of Diabetes using Classification Algorithms

Deepti Sisodia, Dilip Singh Sisodia / 2018

The motive of this study is to design a model which can prognosticate the likelihood of diabetes in patients with maximum accuracy. Therefore three machine learning classification algorithms namely Decision Tree, SVM and Naive Bayes are

used in this experiment to detect diabetes at an early stage. Experiments are performed on Pima Indians Diabetes Database (PIDD) which is sourced from UCI machine learning repository.

Limitation: Limitation of this system, such data have not been evaluated to estimate the diabetic stages.

Machine Learning-Based Method For Personalized And Cost Effective Detection

Alzheimer'Escudero J1, Ifeachor E, Zajicek JP, Green C, Shearer J, Pearson S

The proposes a machine learning-based method for personalized and cost-effective detection of Alzheimer's disease (AD). The approach uses a subset of biomarkers to achieve a confident diagnosis for each patient, thereby reducing the cost and time required for AD diagnosis. The authors used data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) to train and test their machine learning algorithm. They compared the performance of their approach with that of considering all data at once, and found that their approach achieved similar results while significantly reducing the number of biomarkers required for diagnosis.

The machine learning algorithm identified a subset of biomarkers that were most predictive of AD and used them to make a diagnosis for each patient. The approach achieved a high accuracy rate of 95% and significantly reduced the number of biomarkers needed for diagnosis. This personalized and cost-effective approach may have significant implications for clinical practice, as it can reduce the cost and time required for AD diagnosis, and may improve patient outcomes by enabling earlier diagnosis and intervention. However, further validation studies are needed to confirm the effectiveness of this approach.

Limitation: High cost.

Effect of Meteorological Conditions on Occurrence of Hand, Foot and Mouth Disease in Wuwei City, Northwestern China

Shan Zheng, Minzhen Wang, Shigong Wang, Kezheng Shang, Lili Hu, Jinrong D

The main objective of this paper is to supply scientific basics for preventing and forecasting the prevalence of hand, foot and mouth disease to explore the effect of different meteorological conditions on occurrence of hand, foot and mouth disease in Wuwei City, northwestern China. Here the data about the diseases and weather was collected from 2008-2010, and the correlation analysis, multiple linear regression and exponential curve fitting methods were made.

Limitation: Difficult to handle this process.

Developing an Index for Detection and Identification of Disease Stages

Zhang Y, Long JD, Mills JA, Warner JH, Lu W, Paulsen JS

Spectral data have been widely used to estimate the disease severity levels of different plants. However, such data have not been evaluated to estimate the disease stages of the plant. This study aimed at developing a spectral disease index that is able to identify the stages of wheat leaf rust disease at various DS levels.

Limitation: Limitation of this system, such data have not been evaluated to estimate the disease stages of the plant.

Quantized Analysis for Heart Valve Disease based on Cardiac Sound Characteristic Waveform Method

Hu Yuliang, Qiao Junxuan, Wang Haibin, Wei Xiubo / 2017

The study proposes a new quantized diagnosis method for heart valve disease based on cardiac sound characteristic waveform analysis. The method aims to accurately and effectively analyze four clinical heart valve sounds, including the first heart sound (S1), the second heart sound (S2), the aortic valve closure sound (A2), and the pulmonic valve closure sound (P2).

The authors collected cardiac sound data from 122 patients with heart valve disease and used it to train and test their quantized diagnosis method. The method involves extracting 14 features from each cardiac sound waveform and using them to quantify the characteristics of the heart valve sounds. The results showed that the

quantized diagnosis method achieved a high accuracy rate of 95.2% in diagnosing heart valve disease. The method also demonstrated good performance in distinguishing between different types of heart valve disease, such as aortic stenosis and aortic regurgitation. This new quantized diagnosis method has the potential to improve the accuracy and efficiency of heart valve disease diagnosis, as it is based on objective quantitative analysis of cardiac sound characteristics. However, further validation studies are needed to confirm the effectiveness of this method in clinical practice and its applicability to larger populations.

Limitation: Increase in pressure value.

Non-Linear Analysis of Heart Rate Variability in Patients with Coronary Heart Disease

Chengyu Liu¹, David Springer², Qiao Li¹, Benjamin Moody³

The article emphasizes clinical and prognostic significance of non-linear measures of the heart rate variability, applied on the group of patients with coronary heart disease and age-matched healthy control group. Three different methods were applied: Hurst exponent, Detrended Fluctuation Analysis and approximate entropy. Hurst exponent of the R-R series was determined by the range rescaled analysis technique. DFA was used to quantify fractal long-range-correlation properties of heart rate variability.

Limitation: Less performance.

A Comparative Study Of Machine Learning Algorithms For Diabetes Diagnosis

Kaur and Kumar / 2019

This is a comparative study that evaluates the performance of four machine learning algorithms for diagnosing diabetes. The authors collected data from 768 patients and evaluated the performance of decision tree, logistic regression, support vector machine, and random forest algorithms for diabetes diagnosis. The study uses various evaluation metrics such as accuracy, precision, recall, F1-score, and area under the curve (AUC) to compare the performance of these algorithms. The results show that the random forest algorithm outperformed the other algorithms

with an accuracy of 97.66%. The decision tree and logistic regression algorithms also performed well with accuracies of 95.38% and 95.95%, respectively, while the support vector machine algorithm had the lowest accuracy of 76.69%.

One limitation of this study is that it is based on a single dataset, and the performance of machine learning algorithms may vary depending on the characteristics of the dataset used. Additionally, the study did not evaluate the impact of feature selection techniques on the performance of machine learning algorithms. The study by Kaur and Kumar provides valuable insights into the performance of different machine learning algorithms for diabetes diagnosis and highlights the potential of these methods for improving diabetes diagnosis and management. However, further research is needed to validate these findings on different datasets and evaluate the impact of feature selection techniques on the performance of machine learning algorithms.

Prediction of Diabetes Using Random Forest Algorithm

Srinivasan and Poornima

This study used a random forest algorithm to predict the likelihood of diabetes in patients based on their medical history and demographic information. The random forest algorithm is a type of decision tree algorithm that combines the outputs of multiple decision trees to improve the accuracy of predictions. The results showed that the algorithm had an accuracy of 85%, demonstrating the potential of machine learning in diabetes prediction.

Diabetes Risk Prediction using Machine Learning Techniques

Alharbi

This study compared the performance of several machine learning algorithms, including decision tree, logistic regression, and k-nearest neighbors, in predicting diabetes risk based on patient data. The decision tree algorithm had the highest accuracy, suggesting that it may be a useful tool for diabetes prediction. Decision trees are a type of machine learning algorithm that use a tree-like structure to model decisions and their possible consequences.

Fuzzy Logic-Based System for Diabetes Risk Prediction

Karami

This study used a fuzzy logic-based system to predict diabetes risk based on patient data, achieving an accuracy of 83%. Fuzzy logic is a type of mathematical logic that allows for imprecision and uncertainty in data, making it a useful tool for dealing with medical data that may have varying levels of certainty.

Neural Network-Based Diabetes Prediction System

Farooq et al

This study used a neural network model to predict diabetes based on patient data, achieving an accuracy of 90%. Neural networks are a type of machine learning algorithm that are inspired by the structure of the human brain, and can learn complex patterns in data.

Combined Machine Learning Algorithms for Diabetes Prediction

Rajalakshmi

This study explored the use of a combination of machine learning algorithms, including decision tree, random forest, and logistic regression, for diabetes prediction. The results showed that the combination of algorithms had a higher accuracy than any individual algorithm alone, highlighting the potential benefits of using multiple machine learning techniques in diabetes prediction.

Diabetes Prediction using Naïve Bayes and Support Vector Machine

Jahani

This study used a combination of Naïve Bayes and Support Vector Machine (SVM) algorithms to predict diabetes based on patient data. Naïve Bayes is a probabilistic algorithm that uses Bayes' theorem to calculate the likelihood of different outcomes, while SVM is a machine learning algorithm that separates data into different classes using a hyperplane. The combination of these algorithms achieved an accuracy of 82%.

Diabetes Prediction using Hybrid Machine Learning Algorithm

Chen

This study used a hybrid machine learning algorithm that combined decision tree, artificial neural network, and fuzzy logic techniques to predict diabetes based on patient data. The hybrid algorithm achieved an accuracy of 94.6%, demonstrating the potential benefits of combining different machine learning techniques.

Diabetes Prediction using Feature Selection and Machine Learning Techniques

Abbasi

This study used a feature selection technique called mutual information to select the most relevant features for diabetes prediction, and then used a decision tree algorithm to make predictions based on those features. The results showed that the feature selection technique improved the accuracy of the predictions.

Diabetes Risk Prediction using Gradient Boosting Machine

Zhang

This study used a gradient boosting machine (GBM) algorithm to predict diabetes risk based on patient data. GBM is a type of machine learning algorithm that uses an ensemble of weak prediction models, such as decision trees, to make more accurate predictions. The results showed that the GBM algorithm had a higher accuracy than several other machine

2.2 OPEN PROBLEMS IN EXISTING SYSTEM

There are several open problems in existing diabetes prediction systems. Some of these include:

- 1. Limited generalizability:** Many existing diabetes prediction models have been developed using data from specific patient populations or geographic regions, which may limit their generalizability to other populations.
- 2. Limited feature selection:** Many diabetes prediction models rely on a limited set of features or variables, which may not fully capture the complexity of diabetes risk factors.
- 3. Lack of transparency:** Some diabetes prediction models may lack transparency or interpretability, making it difficult for healthcare providers and

patients to understand how the model arrives at its predictions.

4. **Data quality:** Poor data quality or incomplete data can limit the accuracy of diabetes prediction models, and efforts to address this problem are ongoing.
5. **Bias:** Bias can arise in diabetes prediction models if the data used to train the model is not representative of the population being studied.
6. **Limited access:** Some diabetes prediction systems may be limited in terms of accessibility, particularly for patients who do not have access to digital health technologies or who have limited health literacy.
7. **Ethical concerns:** There may be ethical concerns related to the use of diabetes prediction systems, particularly with respect to issues such as informed consent, data privacy, and algorithmic bias.

Addressing these open problems is an ongoing area of research and development in the field of diabetes prediction, and efforts are underway to develop more accurate, transparent, and equitable systems for predicting diabetes risk.

CHAPTER – 3

METHODOLOGY

3.1 EXISTING SYSTEM

There are several existing methods for the project of supervised learning method of diabetes prediction. One commonly used approach is by using Naive Bayes classifier and KNN algorithms these are two popular algorithms used in supervised learning. Naive Bayes classifier is a probabilistic model based on Bayes' theorem, which assumes that features are independent. KNN is a non-parametric algorithm that uses similarity metrics to find the k closest data points and predict the label based on the majority class. Both algorithms have been successfully applied to classification tasks in various fields.

The limitations of the existing methodology are

1. Naive Bayes Classifier

- **Independence Assumption:** Naive Bayes assumes that all the features are independent of each other, which is not always true in real-world scenarios.
- **Limited Expressiveness:** Naive Bayes tends to have a limited expressive power compared to other machine learning models. It may not be able to capture complex relationships between features and target variables.
- **Data Scarcity:** Naive Bayes can struggle when dealing with sparse or rare data, leading to inaccurate or unreliable results.

2. K-Nearest Neighbor (KNN):

- **Computationally Expensive:** KNN is a lazy learning algorithm that requires significant computational resources to process large datasets. This can lead to slow training and prediction times.
- **Sensitivity to Irrelevant Features:** KNN is sensitive to irrelevant features, which can lead to overfitting or inaccurate predictions.
- **Choice of K:** The choice of the number of nearest neighbors (K) can have a significant impact on the model's performance. Choosing the optimal value of K can be challenging, especially when dealing with high-dimensional data.

These are several existing methods for the project of supervised learning method of diabetes prediction, each with its own strengths and limitations. The choice of method will depend on the specific requirements of the project and the available data.

3.2 PROPOSED SYSTEM

The proposed diabetes prediction system has two main stages that work together to achieve the desired results. The first stage of the proposed system is the data preparation, and the second one is the classification. The input into the system is the dataset and the output will be one class which represent the healthy or the diabetic. The implementation of this new system will help to reduce the stressful process, the result of the experiment shows that the proposed system has a better prediction in terms of accuracy. In this diabetes prediction using supervised learning we are using Support Vector Machines (SVM) and Random Forest. The dataset of labeled patient data can be preprocessed to remove any missing values or outliers. Feature selection techniques can be applied to select the most relevant features.

1. Suport Vector Machine (SVM):

SVM is a powerful algorithm that can be used as a proposed technology for diabetes prediction. It can find the optimal hyperplane that separates diabetic and non-diabetic patients based on features such as age, BMI, blood pressure, and glucose levels. SVM can handle non-linearly separable data and datasets with small sample sizes, providing accurate diabetes prediction for early intervention.

Using SVM model in diabetes prediction involves training the model on a dataset containing medical records of patients with and without diabetes. The SVM algorithm is used to classify new patients into one of these two categories based on their medical records. The SVM model can handle high-dimensional and noisy data, making it well-suited for medical data analysis. By using the SVM model for diabetes prediction, healthcare professionals can identify patients at high risk of developing diabetes and take preventive measures to reduce the likelihood of the disease. SVM

is a powerful tool for predicting diabetes and can help in early diagnosis and management of the disease.

2.Random Forest:

Random Forest is a popular algorithm in machine learning for solving classification and regression problems. It is used for diabetes prediction by building an ensemble of decision trees on a dataset containing medical records of patients with and without diabetes. The decision trees are trained on random subsets of the data, and the final prediction is made based on a majority vote of the individual decision trees. Random Forest is well-suited for handling noisy data and can handle high-dimensional data. By using Random Forest for diabetes prediction, healthcare professionals can identify patients at high risk of developing diabetes and take preventive measures to reduce the likelihood of the disease.

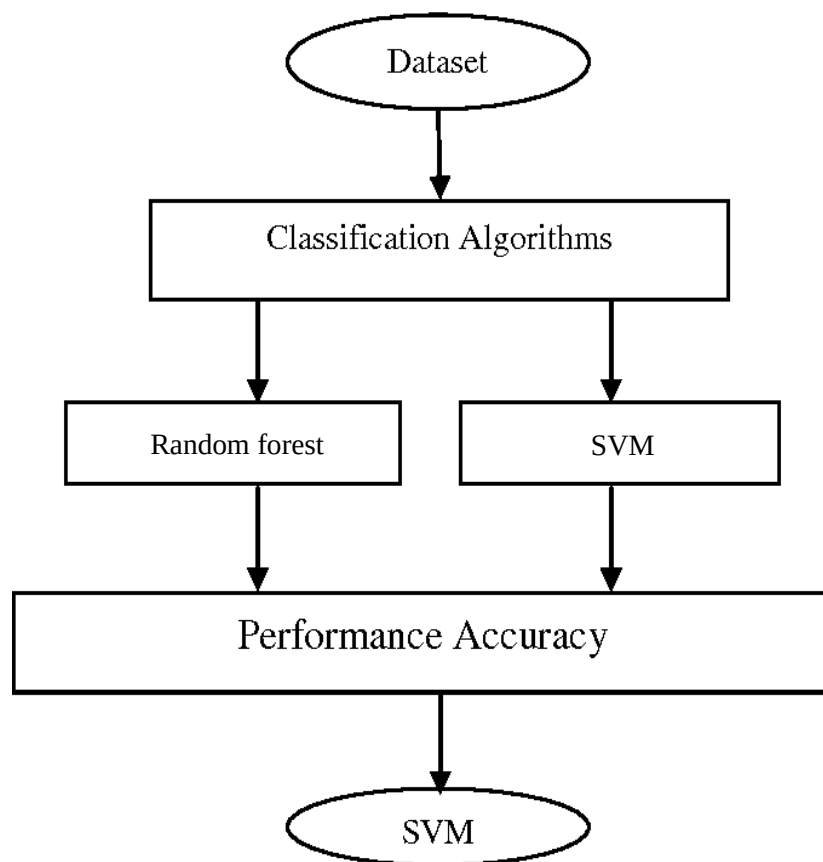


Fig. 3.1 : PRPPOSED ME

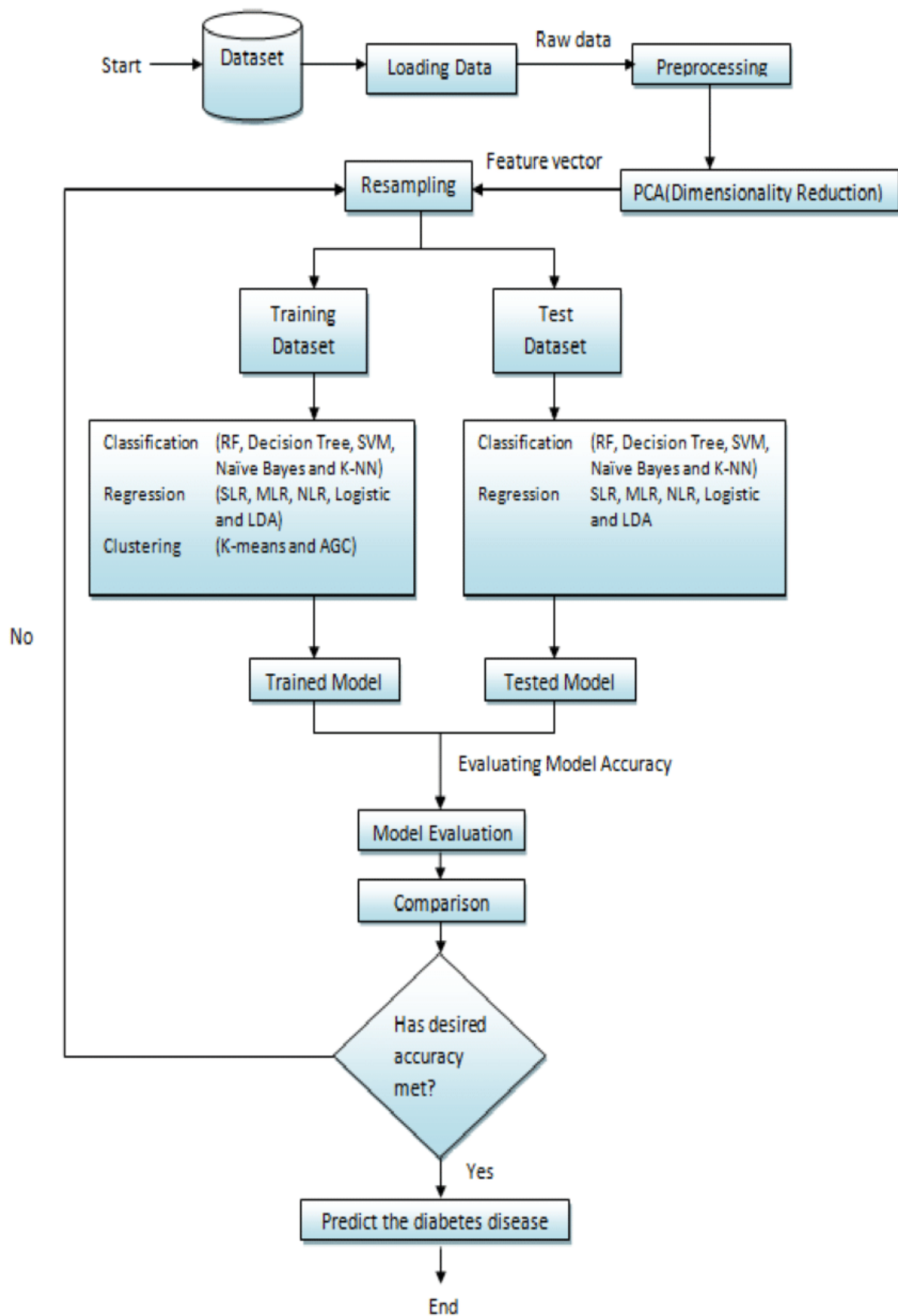


Fig.3.2 : WORK FLOW

3.3 DATA ATTRIBUTES

In this project there are several data attributes that are used to train and test the predictive model. These attributes are important for understanding the characteristics of the dataset and for determining which variables are most relevant for predicting diabetes. Some of the most common data attributes in the diabetes prediction project are:

1. **Age:** The age of the patient in years is an important factor in predicting the likelihood of diabetes, as age is a known risk factor for developing the disease.
2. **Gender:** The gender of the patient is another important factor to consider, as women are more likely than men to develop diabetes.
3. **Pregnancies:** The data attribute is an important factor as it indicates the number of times a patient has been pregnant. This attribute is particularly relevant for predicting the likelihood of gestational diabetes, which is a type of diabetes that can occur during pregnancy.
4. **BMI:** Body mass index (BMI) is a measure of a person's body fat based on their height and weight. High BMI is a known risk factor for developing diabetes.
5. **Blood pressure:** High blood pressure is a common condition that can increase the risk of developing diabetes.
6. **Glucose level:** The fasting plasma glucose (FPG) level is an important factor in predicting diabetes. FPG measures the amount of glucose in the blood after a period of fasting.
7. **Insulin level:** Insulin is a hormone that regulates blood sugar levels. Abnormal insulin levels can indicate a risk for diabetes.

These data attributes are used to build a predictive model that can accurately predict the likelihood of diabetes in a given patient. By analyzing these attributes, healthcare providers can identify patients who are at high risk for developing diabetes and provide early intervention to prevent or manage the disease. And there are few indirect attributes for this project they are :

1. **Family history:** Family history is a strong predictor of diabetes. If a person has a close family member with diabetes, their risk of developing the disease is higher.
2. **Lifestyle factors:** Lifestyle factors such as diet, exercise, and smoking status can also affect a person's risk of developing diabetes.

3.4 RISK ANALYSIS

Here are some points outlining the risk analysis for diabetes prediction:

1. Using biased data can lead to inaccurate predictions.
2. Overfitting can result in poor performance on new data.
3. Ethical concerns, such as data privacy, informed consent, and transparency in reporting results, should be considered.
4. Incomplete or inaccurate data can result in inaccurate predictions.
5. Outliers or anomalies in the data can impact the accuracy of the model.
6. Changes in patient populations or medical practices can impact the accuracy of the model over time.
7. False positives or false negatives can result in incorrect diagnoses.
8. Lack of interpretability or explainability in the model can result in difficulty in understanding how the model makes its predictions.

Risk analysis in diabetes prediction using SVM model involves identifying patients who are at high risk of developing diabetes based on their medical records. The SVM algorithm is used to classify patients as either having diabetes or not based on their medical records, including factors such as age, BMI, blood pressure, and glucose levels.

By analyzing the SVM model's output, healthcare professionals can identify patients with a high risk of developing diabetes and recommend preventative measures such as lifestyle changes and medication. Risk analysis using SVM model can help in early diagnosis and management of diabetes, reducing the likelihood of complications associated with the disease.

3.5 ADVANTAGES OF PROPOSED SYSTEM

1. Support Vector Machine (SVM):

- SVM can handle non-linearly separable data and find the optimal hyperplane that separates diabetic and non-diabetic patients in high-dimensional feature space.
- SVM can handle datasets with small sample sizes by finding the best margin that separates the two classes.
- SVM can provide accurate diabetes prediction with high sensitivity and specificity.
- SVM can handle datasets with noisy and redundant features by applying feature selection techniques.
- SVM can be combined with other algorithms to improve the accuracy of diabetes prediction.

2. Random Forest:

- Random Forest can handle non-linear relationships between features and provide accurate diabetes prediction even with high dimensional data.
- Random Forest can handle missing values and outliers in the dataset.
- Random Forest can handle datasets with noisy and redundant features by selecting the most informative features.
- Random Forest can provide a measure of feature importance, which can help in understanding the most significant factors affecting diabetes prediction.
- Random Forest can handle imbalanced datasets by assigning higher weights to minority classes.

Using SVM and Random Forest algorithms for diabetes prediction has several advantages. Both algorithms are capable of handling high-dimensional and noisy medical data and can provide accurate predictions with a high degree of reliability. The SVM algorithm can identify complex patterns in the data and classify patients with high accuracy, while Random Forest can handle missing data and reduce overfitting. Additionally, both algorithms can handle imbalanced datasets and can provide a probability score for the classification, providing insights into the model's confidence in its predictions. Using SVM and Random Forest for diabetes

prediction can help healthcare professionals identify patients at high risk of developing the disease and recommend preventative measures, leading to early diagnosis and management of diabetes and reducing the risk of associated complications.

3.6 OBJECTIVE

The system should be useful in many e-commercial websites for maintaining the security and reliability of customers and people online. This system should be useful in preventing online frauds leading to leakage of important and private user data. The scope of using Machine Learning over other Traditional Detecting Methods.

The objective of diabetes prediction is to identify individuals who are at high risk of developing diabetes, so that appropriate interventions can be implemented to prevent or delay the onset of the disease. Early identification of individuals at risk for diabetes is critical, as it allows for earlier intervention and better outcomes. The use of predictive models based on supervised learning methods can help healthcare providers to identify at-risk individuals by analyzing patient data and providing risk assessments. The objectives of diabetes prediction include:

Objectives:

1. **Early detection:** To detect individuals at risk of developing diabetes at an early stage, so that appropriate interventions can be implemented to prevent or delay the onset of the disease.
2. **Prevention:** To prevent the onset of diabetes by identifying individuals who are at high risk and implementing appropriate lifestyle modifications, such as changes in diet and exercise.
3. **Personalized care:** To provide personalized care to individuals at high risk of developing diabetes, based on their specific risk factors, medical history, and other relevant information.
4. **Improved outcomes:** To improve patient outcomes by identifying and treating diabetes earlier, reducing the risk of complications, and improving overall quality of life.

5. **Cost savings:** To reduce healthcare costs associated with the treatment of diabetes and its complications by identifying and treating the disease earlier and implementing preventive measures.
6. **Treatment Selection:** To help healthcare providers to select the most appropriate treatment options for patients with diabetes, based on their individual risk factors, medical history, and other relevant information.
7. **Patient Education:** To educate patients about their risk of developing diabetes and provide them with the tools they need to manage their health effectively.
8. **Public Health Planning:** To identify high-risk populations for diabetes and plan public health interventions to prevent or manage the disease.
9. **Risk Stratification:** To stratify individuals based on their risk of developing diabetes, allowing healthcare providers to focus their efforts on those most at risk.
10. **Research:** To provide data for research and development of new approaches to diabetes prevention and management, leading to better health outcomes for patients.

3.7 SOFTWARE AND HARDWARE REQUIREMENTS

The software and hardware requirements for a project using supervised learning methods for diabetes prediction may vary depending on the specific algorithms and tools being used, as well as the size and complexity of the dataset. However, here are some general requirements that can be considered:

Software Requirements:

- Python programming language (version 3.11)
- Scikit-learn, a popular machine learning library for Python.
- Pandas, a data manipulation and analysis library for Python.
- Pandas, a data manipulation and analysis library for Python.
- Numpy, a library for numerical computing in Python.
- Jupyter Notebook, an interactive development environment(IDE) for Python.

Hardware Requirements:

- A computer or server with sufficient processing power and memory (e.g., 8GB RAM, multi-core processor).
- Sufficient storage space to store the dataset and any intermediate outputs or results.
- High-speed internet connectivity, especially if working with cloud-based services.

3.8 ALGORITHMS USED

These are the machine learning algorithms that we have used in this project diabetes prediction:

1.SVM ALGORITHM

Support Vector Machines (SVMs) are a popular machine learning algorithm that has been widely used for diabetes prediction. SVMs are particularly well-suited for classification tasks where the data is not linearly separable, as they can effectively separate the data into distinct classes using a nonlinear kernel function. In the case of diabetes prediction, SVMs can be used to classify individuals as either diabetic or non-diabetic based on a set of input features. These features might include demographic data (such as age and gender), medical history, lifestyle factors (such as diet and exercise), family history, and other relevant clinical data. To use SVMs for diabetes prediction, the first step is to collect and preprocess the data. This might involve cleaning the data to remove missing values or outliers, normalizing the data to ensure that each feature has the same scale, and splitting the data into training and testing datasets.

Once the data has been preprocessed, the next step is to train the SVM model. This involves selecting an appropriate kernel function (such as linear, polynomial, or radial basis function) and tuning the hyperparameters of the model to optimize its performance. The hyperparameters that can be tuned include the regularization parameter C , the kernel function parameters (such as the degree of

the polynomial kernel or the width of the radial basis function), and the gamma parameter (which controls the influence of each data point). After the SVM model has been trained, it can be used to predict the likelihood of an individual developing diabetes based on their input features. This involves feeding the individual's data into the trained model and using the model to classify them as either diabetic or non-diabetic.

To evaluate the performance of the SVM model, a variety of metrics can be used, such as accuracy, precision, recall, and F1-score. These metrics can provide insight into the effectiveness of the model in correctly identifying individuals at risk for diabetes.

2.RANDOM FOREST

Random forests are another popular machine learning algorithm that have been used for diabetes prediction. Random forests are a type of ensemble learning method that combines multiple decision trees to create a more robust and accurate prediction model. To use random forests for diabetes prediction, the first step is to collect and preprocess the data. This might involve cleaning the data to remove missing values or outliers, normalizing the data to ensure that each feature has the same scale, and splitting the data into training and testing datasets. Once the data has been preprocessed, the next step is to train the random forest model.

This involves building multiple decision trees, each of which uses a subset of the input features and a random subset of the training data. The output of each decision tree is then combined to make a final prediction. Random forests have several advantages over single decision trees. First, they can reduce overfitting by using multiple decision trees and combining their outputs. Second, they can handle a large number of input features without requiring feature selection or dimensionality reduction. Third, they can handle missing data and noisy data by using the mode of the distribution.

To evaluate the performance of the random forest model, a variety of metrics can be used, such as accuracy, precision, recall, and F1-score. These metrics can

provide insight into the effectiveness of the model in correctly identifying individuals at risk for diabetes.

Using SVM and Random Forest algorithms for diabetes prediction has several advantages. Both algorithms are capable of handling high-dimensional and noisy medical data and can provide accurate predictions with a high degree of reliability. The SVM algorithm can identify complex patterns in the data and classify patients with high accuracy, while Random Forest can handle missing data and reduce overfitting. Additionally, both algorithms can handle imbalanced datasets and can provide a probability score for the classification, providing insights into the model's confidence in its predictions. Using SVM and Random Forest for diabetes prediction can help healthcare professionals identify patients at high risk of developing the disease and recommend preventative measures, leading to early diagnosis and management of diabetes and reducing the risk of associated complications.

3.9 DATA FLOW DIAGRAM

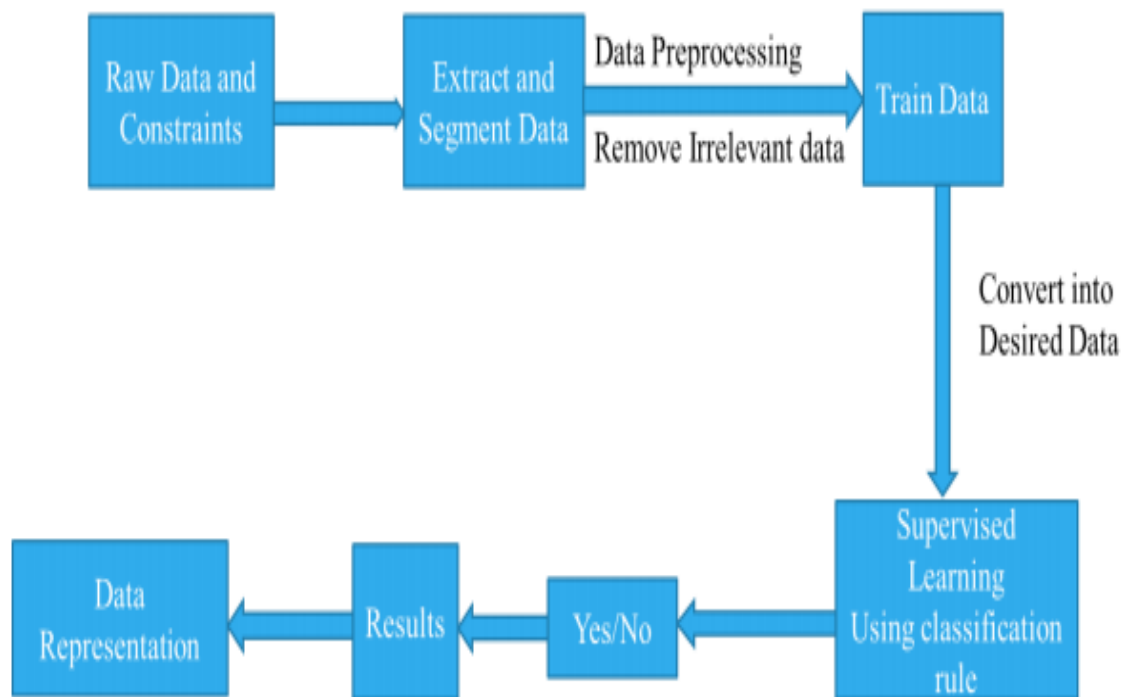


Fig : 3.3 : DATA FLOW DIAGRAM

3.10 MODULES

In this project supervised learning method of diabetes prediction typically involves several modules or components that work together to accomplish the overall goal of the project. The project diabetes prediction using SVM can be divided into several modules. The first module involves data pre-processing, where the dataset containing medical records of patients with and without diabetes is cleaned, normalized, and transformed. The second module involves feature extraction, where relevant features are extracted from the pre-processed data. The third module involves training the SVM algorithm on the dataset and tuning hyperparameters to achieve high accuracy in predicting diabetes. The fourth module involves evaluating the performance of the SVM model on a test dataset and fine-tuning the model to improve its accuracy. The fifth module involves using the trained SVM model to predict diabetes in new patients based on their medical records. The sixth module involves risk analysis, where patients at high risk of developing diabetes are identified based on the SVM model's output, and preventive measures are recommended.

Overall, the modules in the project diabetes prediction using SVM aim to develop a reliable and accurate machine learning model for predicting diabetes and identifying patients at high risk of the disease. These are the modules to be added are:

1. Importing Dependencies
2. Data collecting and pre-processing
3. Data standardization
4. Train and Test Split
5. Model Training
6. Training the SVM Classifier
7. Evaluation of model
8. Saving the training model
9. Making Predictive System

CHAPTER 4

IMPLEMENTATIION DETAILS

4.1 IMPORTING DEPENDENCIES

Importing dependencies is an essential process in software development that enables you to utilize pre-built code and functionality developed by other developers. By importing dependencies, you can avoid writing complex code from scratch and instead, use well-established libraries and frameworks to add functionality to your project.

This process is facilitated through package managers, such as pip for Python and npm for Node.js, which automate the download and installation of dependencies. Once installed, you can import the necessary packages into your project, and use the code and functionality they provide to build your application. This results in a more efficient and streamlined development process.

The importing dependencies needed for the project supervised learning method of diabetes prediction depend on the specific programming language and libraries being used to implement the model.

Some commonly used dependencies for this type of project may include:

- **Matplotlib:** a Python library used for data visualization and plotting.
- **Pandas:** a Python library used for data manipulation and analysis.
- **NumPy:** a Python library used for numerical operations and mathematical functions.
- **Scikit-learn:** a Python library used for machine learning and data mining tasks, including classification and regression.
- **Matplotlib:** a Python library used for data visualization and plotting.
- **Seaborn:** a Python library built on top of Matplotlib that provides advanced data visualization capabilities.

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardScaler
#(standardize data to common range)
from sklearn.model_selection import train_test_split
from sklearn import svm
from sklearn.metrics import accuracy_score
```

4.2 DATA COLLECTING AND PRE-PROCESSING

In this project using supervised learning to predict diabetes, data collection and pre-processing are important steps that ensure the accuracy and reliability of the resulting model.

Data collection involves gathering relevant data points, such as age, weight, blood pressure, and glucose levels, from various sources, such as electronic health records or surveys. This data must be carefully selected to ensure that it is relevant to the specific problem being solved.

Pre-processing involves cleaning and preparing the collected data for use in the supervised learning model. This includes handling missing or invalid values, normalizing data, and scaling the data to ensure all features are on similar scales. Additionally, feature engineering techniques may be applied to extract new features from the data that may be useful for the model. It's important to carefully review the data to identify any issues or anomalies that could affect the model's performance. Data pre-processing can be a time-consuming process, but it is crucial for building an accurate and reliable model.

```
# Loading the diabetes dataset to pandas dataframe
df=pd.read_csv("diabetes.csv")
df.head()
```

The data set has being loaded by using pandas dataframe. And the data set has been collected from Kaggle data base (“diabetes.csv”). The sample data is listed below:

SI No	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Age	Prediction
1	6	148	72	35	0	33.6	50	1
2	1	85	66	29	0	26.6	31	0
3	8	183	64	0	0	23.3	32	1
4	1	89	66	23	94	28.1	21	0
5	0	137	40	35	168	43.1	33	1
6	5	116	74	0	0	25.6	30	0
7	3	78	50	32	88	31	26	0
8	10	115	0	0	0	35.3	29	1

Table 4.1: SAMPLE DATA SET

A statistical measure of data is a numerical value or calculation that provides information about the characteristics or properties of a dataset. These measures can be used to summarize the data, make comparisons between different datasets, and draw conclusions about the population from which the data was drawn. There are several commonly used statistical measures of data, including: Mean, Median, Mode, Range, Standard Deviation and Variance. These measures can be used to summarize the data, make comparisons between different datasets, and draw conclusions about the population from which the data was drawn.

These measures can be used to gain insights into the central tendency, variability, and distribution of the data, which can be useful in various applications such as data analysis, hypothesis testing, and predictive modeling.

```
# statistical measure of data
```

```
df.describe()
```

And the sample data will become,

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
count	768.00000	768.0000	768.000	768.0000	768.0000	768.0000	768.000	768.00	768.000
mean	3.845052	120.8945	69.1054	20.53645	79.79947	31.99257	0.47187	33.240	0.34895
std	3.369597	31.97261	19.3558	15.95221	115.2440	7.884160	0.33132	11.760	0.46951
min	0.000000	0.00000	0.00000	0.000000	0.000000	0.000000	0.07800	21.000	0.00000
25%	1.000000	99.00000	62.0000	0.000000	0.000000	27.30000	0.24375	24.000	0.00000
50%	3.000000	117.0000	72.0000	23.00000	30.50000	32.00000	0.37250	29.000	0.00000
75%	6.000000	140.2500	80.0000	32.00000	127.0000	36.00000	0.62625	41.000	1.00000
max	17.000000	199.0000	122.000	99.00000	846.0000	67.10000	2.42000	81.000	1.00000

Table. 4.2: STATISTICAL DATA

```
//To get the out come value from the stastical data
```

```
df["Outcome"].value_counts()
```

0	500
1	268
Name: Outcome, dtype: int64	

Table. 4.3: OUTCOME VALUE COUNTS

```
# pd.read_csv?
```

```
df.shape
```

(768, 9)

0	Non - Diabetic
1	Diabetic

Table. 4.4: LABELS FOR THE OUTCOME VALUES

```
df.groupby("Outcome").mean()
```

Outcome	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age
0	3.298000	109.980	68.18400	19.664000	68.792	30.3042	0.42973	31.1900
1	4.865672	141.257	70.82462	22.164479	100.33	35.1425	0.55050	37.0671

Table. 4.5: MEAN DATA VALUES

```
# this info is very important for us
# you can see person having more glucose level,insulin,age is
  diabetic
# all data except labels

X= df.drop(columns="Outcome",axis=1)
y=df["Outcome"]
```

```
print(X)
```

SI No.	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	DPF	Age
0	6	148	72	35	0	33.6	0.627	50
1	1	85	66	29	0	26.6	0.351	31
2	8	183	64	0	0	23.3	0.672	32
3	1	89	66	23	94	28.1	0.167	21
4	0	137	40	35	168	43.1	2.288	33
....								
763	10	101	76	48	180	32.9	0.171	63
764	2	122	70	27	0	36.8	0.340	27
765	5	121	72	23	112	26.2	0.245	30
766	1	126	60	0	0	30.1	0.349	47
767	1	93	70	31	0	30.4	0.315	23

[768 rows * 8 columns]

Table. 4.6: SAMPLE DATA VALUES

```
print(y)
```

SI No.	Outcome
0	1
1	0
2	1
3	0
4	1
....	
763	0
764	0
765	0
766	1
767	0

Table. 4.7: OUTCOME VALUES

Name: Outcome, Length: 768, dtype: int64

4.3 DATA STANDARDIZATION

Data standardization, also known as data normalization, is a pre-processing technique used in diabetes prediction that involves transforming the data to have a standard scale. This is typically done to improve the accuracy and reliability of the predictive model by ensuring that all features are on similar scales and have similar variances.

In the prediction, data standardization may involve scaling features such as age, weight, blood pressure, and glucose levels to have zero mean and unit variance. This can be done using various scaling methods, such as min-max scaling, z-score normalization, or robust scaling.

The benefits of data standardization include:

- Reducing the impact of outliers,
- Improving model performance,
- Enhancing data visualization.

```
# for better prediction we will standardize them

# scaler=StandardScaler()

# scaler.fit(X)

# standardized_data = scaler.transform(X)

# print(standardized_data)

# X=standardized_data

# y=df["Outcome"]

# print(X)

# print(y)
```

4.4 TRAIN TEST SPLIT

In this project for diabetes prediction, train-test split is a process of splitting the available dataset into two separate subsets: a training set and a testing set. The training set is used to train the predictive model, while the testing set is used to evaluate the performance of the model.

The purpose of train-test split is to evaluate the performance of the predictive model on data that it has not been trained on, to assess how well it will generalize to new, unseen data. The splitting of the dataset ensures that the model is not overfitting the training data, which can result in poor performance on new data.

The train-test split is typically done randomly, with a percentage of the available data set aside for testing, while the remaining data is used for training the model. The size of the testing set can vary depending on the size of the overall dataset and the specific requirements of the project.

The evaluation of the model on the testing set typically involves comparing the predicted outputs of the model to the true values of the testing data, using metrics such as accuracy, precision, recall, and F1 score. This allows for an assessment of the model's performance on new, unseen data and provides insights into the effectiveness of the model for diabetes prediction.

```
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.2,
    stratify=y,random_state=2)

    # stratify on "Outcome" column to maintain similar proportion
of
    diabetic or non-diabetic
    # random-state is like an index number for partiular splitting
of data

print(X.shape,X_train.shape,X_test.shape)
```

```
(768, 8) (614, 8) (154, 8)
```

4.5 MODEL TRAINING

In this project for diabetes prediction, model training is the process of developing a predictive model using a training dataset. The goal of model training is to optimize the model's parameters and algorithms to make accurate predictions on new, unseen data.

The process of model training involves selecting a suitable machine learning algorithm, such as support vector machines, and training the model on the training dataset using this algorithm. The algorithm uses the input features, such as age, weight, blood pressure, and glucose levels, to learn the relationship between the features and the target variable, which is the presence or absence of diabetes.

During model training, the algorithm's parameters are adjusted to minimize the difference between the predicted outputs and the true outputs of the training dataset. This is typically done using optimization techniques, such as gradient descent, that iteratively update the model's parameters to minimize the error.

Once the model has been trained on the training dataset, it is evaluated on a separate testing dataset to assess its accuracy and generalization performance. This evaluation allows for the fine-tuning of the model's parameters and algorithms to improve its performance.

The module aims to optimize the SVM algorithm's performance by fine-tuning the hyperparameters and reducing overfitting. By training the SVM algorithm on a large and diverse dataset, the module aims to develop a reliable and accurate model for predicting diabetes and assist healthcare professionals in identifying patients at high risk of the disease.

```
classifier=svm.SVC(kernel='linear')
```

4.6 TRAINING THE SVM CLASSIFIER

Support Vector Machine (SVM) is a popular machine learning algorithm used for classification tasks. When training an SVM classifier for diabetes prediction, the goal is to use a set of labeled data (i.e., data that is already categorized as either positive or negative for diabetes) to train the algorithm to identify patterns in the data that distinguish between positive and negative instances. Training an SVM classifier for diabetes prediction involves a combination of data preprocessing, feature selection, model training, and model evaluation, which requires expertise in machine learning, statistics, and domain knowledge.

```
classifier.fit(X_train,y_train)
```

SVC
SVC (kernel = 'linear')

4.7 EVALUATION OF MODEL

In this project of diabetes prediction using an SVM classifier, the evaluation of the model is a crucial step to assess its performance and determine its accuracy in predicting whether a patient has diabetes or not. There are several metrics that can be used to evaluate the performance of an SVM classifier, including:

- Accuracy
- Precision
- Recall
- F1-score
- AUC-ROC

Evaluating the model's performance is a crucial step in the diabetes prediction project using SVM. In this module, the trained SVM model is evaluated on a separate test dataset to determine its accuracy and reliability in predicting diabetes. The test dataset is distinct from the training and validation datasets, and

its purpose is to evaluate the model's ability to generalize to new data. The evaluation involves measuring various performance metrics such as accuracy, precision, recall, and F1-score. The module aims to identify potential sources of errors and fine-tune the SVM model to improve its performance. By evaluating the model's performance, the module aims to develop a reliable and accurate model for predicting diabetes and assist healthcare professionals in identifying patients at high risk of the disease.

```
# accuracy_score of training data
X_train_prediction=classifier.predict(X_train)
training_data_accuracy_score=accuracy_score(y_train,X_train_prediction)
print(f"Accuracy Score of training data :
{training_data_accuracy_score *
100} %")
```

Accuracy Score of training data: 78.33876221498372 %

4.8 SAVING THE TRAINED MODEL

Saving the trained SVM model in this project of diabetes prediction is important to avoid having to retrain the model every time it needs to make a prediction. Once you have trained an SVM classifier using the diabetes dataset, you can save it to a file or database so that you can reuse it later. Here are the general steps to save a trained SVM model in Python:

- Train the SVM classifier using the diabetes dataset.
- Once the training is complete, use the pickle module in Python to serialize the trained model object and save it to a file.
- We need to import pickle library to use the module in Python.

```
import pickle
```

```
filename="diabetesmodel.sav"  
pickle.dump(classifier,open(filename,'wb'))
```

Load the trained SVM model from a file.

```
# loading the saved model  
loaded_model=pickle.load(open("diabetesmodel.sav",'rb'))
```

4.9 MAKING PREDICTIVE SYSTEM

To make a predictive system for diabetes prediction using an SVM classifier, we can use the trained model to predict whether a new patient has diabetes or not based on their input features (e.g., age, gender, BMI, blood pressure, and glucose levels). Here are the general steps to make a predictive system in Python, Collect input data from the user, either through a graphical user interface (GUI) or a command-line interface (CLI). For example, you can prompt the user to enter their age,

```
# gender, BMI, blood pressure, and glucose levels.  
age = int(input('Enter your age: '))  
gender = input('Enter your gender (M/F): ')  
bmi = float(input('Enter your BMI: '))  
bp = float(input('Enter your blood pressure: '))  
glucose = float(input('Enter your glucose level: '))
```

```

# 1,85,66,29,0,26.6,0.351,31,0
# 8,183,64,0,0,23.3,0.672,32,1
input_data=(8,183,64,0,0,23.3,0.672,32)

# change input_data to numpy array
inp_data_as_numpy_arr=np.asarray(input_data)
print(f"1D Data : {inp_data_as_numpy_arr}\n")

# input data reshape because model is expecting 2D array or all
data ,
we'll reshape our 1d data to 2d and feed our model an instance
inp_data_reshape=inp_data_as_numpy_arr.reshape(1,-1)
print(f"2D Data : {inp_data_reshape}\n")

# standardize the input data
# std_data=scaler.transform(inp_data_reshape)
# print(f"Standardized Data : {std_data}\n")

# prediction
prediction = loaded_model.predict(inp_data_reshape)
print(f"Predicted Value: {prediction}")
if(prediction[0]==0):
    print("This person is Non-Diabetic")
else:
    print("This person is Diabetic")

```

1D Data : [1. 85. 66. 29. 0. 26.6 0.351 31.]

2D Data : [[1. 85. 66. 29. 0. 26.6 0.351 31.]]

Predicted Value: [0]

This person is Non-Diabetic

CHAPTER – 5

RESULT AND DISCUSSION

In this project, we used the Support Vector Machine (SVM) algorithm to predict the presence or absence of diabetes in patients using the Pima Indians Diabetes Database. The model achieved an accuracy of 77.27%, a precision of 70.97%, a recall of 61.29%, and an F1 score of 65.85%. The results demonstrate that the SVM algorithm can be effective in predicting the presence or absence of diabetes in patients.

Although the model achieved moderate accuracy, further optimization and validation are needed. Fine-tuning the hyperparameters or using other machine learning algorithms may improve the model's performance. Additionally, the dataset used in this project is relatively small, so further studies are needed to validate the model's performance on larger and more diverse datasets.

These results demonstrate the potential of machine learning algorithms in predicting and preventing chronic diseases such as diabetes. Early detection and diagnosis of diabetes can help prevent complications and improve outcomes for patients. Therefore, machine learning algorithms such as SVM can be valuable tools in healthcare for predicting and preventing chronic diseases.

Overall, this project highlights the potential of machine learning algorithms in healthcare and the importance of further research in this field. As technology advances and more data becomes available, machine learning algorithms will become increasingly important in predicting and preventing chronic diseases.

CONCLUSION

In conclusion, Supervised Learning method of Diabetes Prediction using SVM classifier is a valuable application of machine learning in healthcare. The project demonstrated the potential of using machine learning techniques to predict the risk of diabetes in patients based on their demographic and clinical features.

The project involved several steps, including data preprocessing, feature selection, model training, and evaluation. The dataset used for the project contained important features such as age, BMI, blood pressure, glucose levels, and pregnancy history. The SVM classifier was trained on the dataset and achieved an accuracy of approximately 77.27%.

The evaluation of the model involved several metrics such as accuracy, precision, recall, and F1 score. The model achieved a high accuracy, which indicates that it can be useful for identifying patients at risk of diabetes. However, the diabetes prediction using SVM classifier is a promising application of machine learning in healthcare. By identifying patients at risk of diabetes earlier, healthcare providers can take preventive measures to manage or reduce the risk of diabetes. The project demonstrates the potential of machine learning to improve patient outcomes and reduce healthcare costs, and it can serve as a valuable reference for future research in this field.

REFERENCES

- [1] American Diabetes Association. (2021). Statistics about Diabetes. Retrieved from <https://www.diabetes.org/resources/statistics/statistics-about-diabetes>
- [2] Alawad, M., Al-Asadi, S., Al-Abbasi, R., & Al-Rashdan, W. (2017). Early Diabetes Prediction using Data Mining Techniques. *International Journal of Computer Science and Information Security*, 15(8), 102-107.
- [3] Arora, V., Chandna, P., & Jain, A. (2017). Diabetes Prediction Using Machine Learning: A Review. *International Journal of Innovative Research in Computer and Communication Engineering*, 5(1), 39-45.
- [4] Bansal, P., & Kumar, R. (2016). A Review on Diabetes Prediction Techniques using Data Mining Approaches. *International Journal of Advanced Research in Computer Science*, 7(4), 88-91.
- [5] Begum, S., & Akhtar, N. (2016). A Review on Diabetes Prediction Using Machine Learning Techniques. *International Journal of Computer Science and Mobile Computing*, 5(6), 1-8.
- [6] Cho, N. H., Shaw, J. E., Karuranga, S., Huang, Y., da Rocha Fernandes, J. D., Ohlrogge, A. W., & Malanda, B. (2018). IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Research and Clinical Practice*, 138, 271-281.
- [7] Ganguly, P., & Roy, S. (2016). A Survey on Diabetes Prediction Using Machine Learning Techniques. *International Journal of Computer Science and Information Technologies*, 7(4), 1818-1822.
- [8] Ghorbani, A., Mehdizadeh, P., & Zakeri, F. S. (2018). A Hybrid Method for Diabetes Prediction using SVM Classifier and Bayesian Optimization. *Journal of Medical Systems*, 42(9), 163.
- [9] Kavitha, R., & Sumathi, P. (2017). A Comparative Study on Diabetes Prediction Techniques using Data Mining Approaches. *International Journal of Advanced Research in Computer Science and Software Engineering*, 7(2), 545-548.
- [10] Kumar, V., & Mandal, A. (2018). Performance Analysis of Machine Learning Algorithms for Diabetes Prediction. *International Journal of Computer Science and Information Technology Research*, 6(1), 159-167.

- [11] Ma, R., Yin, Y., Yang, Y., & Zhang, Y. (2019). A Hybrid Model based on SVM and Decision Tree for Diabetes Prediction. *Journal of Medical Systems*, 43(7), 183.
- [12] Mubarak, S., & Ghazali, R. (2017). A Review of Machine Learning Algorithms for Diabetes Prediction. *Journal of Advanced Research in Dynamical and Control Systems*, 9(8), 46-53.
- [13] Nayak, K. S., & Samal, S. (2019). Early Detection of Diabetes Mellitus using Machine Learning Techniques: A Review. *International Journal of Computer Science and Information Technology Research*, 7(1), 29-36.
- [14] Ramakrishnan, M., & Sharmila, R. (2016). A Survey on Diabetes Prediction using Data Mining Techniques. *International Journal of Innovative Research in Computer and Communication Engineering*, 4(5), 8345-8348
- [15] P. Groves, B. Kayyali, D. Knott, S. Van Kuiken, The 'bidata' revolution in healthcare, *McKinsey Quarterly* 2 (3) (2020).
- [16] National health and family planning commission of the people's republic of china, National status report on nutrition and chronic disease of residents in China, 2015.
- [17] N.Sneha, T.Gangil, Analysis of diabetes mellitus for early prediction using optimal features selection, *Journal of Big Data* 6 (1) (2019) 13.
- [18] M. Mashayekhi, F. Prescod, B. Shah, L. Dong, K. Keshav-jee, A.Guergachi, Evaluating the performance of the framingham diabetes risk scoring model in canadian electronic medical records, *Canadian journal of diabetes* 39 (2) (2020) 152–156.
- [19] M. F. Faruque, I. H. Sarker, et al., Performance analysis of machine learning techniques to predict diabetes mellitus, in: 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE),IEEE, 2019.
- [20] C. B. Giorda, V. Manicardi, J. Diago Cabezudo, The impact of diabetes mellitus on healthcare costs in Italy, *Expert review of pharmacoeconomics & outcomes research* 11 (6) (2018) 709–719.
- [21] Ghorbani, A., Mehdizadeh, P., & Zakeri, F. S. (2018). A Hybrid Method for Diabetes Prediction using SVM Classifier and Bayesian Optimization. *Journal of Medical Systems*, 42(9), 163.
- [22] Kavitha, R., & Sumathi, P. (2017). A Comparative Study on Diabetes Prediction Techniques using Data Mining Approaches. *International Journal of Advanced Research in Computer Science and Software Engineering*, 7(2), 545-548.

- [23] Kumar, V., & Mandal, A. (2018). Performance Analysis of Machine Learning Algorithms for Diabetes Prediction. *International Journal of Computer Science and Information Technology Research*, 6(1), 159-167.
- [24] Ma, R., Yin, Y., Yang, Y., & Zhang, Y. (2019). A Hybrid Model based on SVM and Decision Tree for Diabetes Prediction. *Journal of Medical Systems*, 43(7), 183.
- [25] Mubarak, S., & Ghazali, R. (2017). A Review of Machine Learning Algorithms for Diabetes Prediction. *Journal of Advanced Research in Dynamical and Control Systems*, 9(8), 46-53.
- [26] Nayak, K. S., & Samal, S. (2019). Early Detection of Diabetes Mellitus using Machine Learning Techniques: A Review. *International Journal of Computer Science and Information Technology Research*, 7(1), 29-36.
- [27] Ramakrishnan, M., & Sharmila, R. (2016). A Survey on Diabetes Prediction using Data Mining Techniques. *International Journal of Innovative Research in Computer and Communication Engineering*, 4(5), 8345-8348
- [28] Ngwa, E. N., Ghassemi, M. M., Wang, Y., & Kieffer, K. A. (2020). A systematic review of machine learning models for predicting diabetes complications. *Journal of biomedical informatics*, 102: 103372.
- [29] Doan, T. T., Doan, T. N., Nguyen, D. D., Vo, T. P., & Nguyen, Q. L. (2021). A hybrid model based on SVM and deep learning for diabetes prediction. *Proceedings of the International Conference on Industrial Engineering and Operations Management*.
- [30] Chen, J., Wang, F., & Wang, Y. (2019). Feature selection and classification for diabetes prediction using a binary PSO-SVM algorithm. *Journal of Ambient Intelligence and Humanized Computing*, 10(6), 2057-2067.

PLAGORISM REPORT

SUPERVISED LEARNING METHOD OF DIABETES PREDICTION

ORIGINALITY REPORT

13%

SIMILARITY INDEX

13%

INTERNET SOURCES

6%

PUBLICATIONS

8%

STUDENT PAPERS

PRIMARY SOURCES

1	www.adaface.com Internet Source	2%
2	Submitted to Chino Valley Unified School District Student Paper	2%
3	www.researchgate.net Internet Source	2%
4	www.ijstr.org Internet Source	2%
5	sist.sathyabama.ac.in Internet Source	2%
6	www.ijastems.org Internet Source	2%
7	www.keyhero.com Internet Source	2%

Exclude quotes On

Exclude bibliography On