**CHAPTER-1**

**INTRODUCTION**

**1.1 Machine Learning**

Using a variety of datasets, machine learning generates mathematical models and makes predictions based on real-world or classical data. Without being expressly programmed, a machine can learn from data, perform better based on past experiences, and forecast results **[1]**. When a machine learning system is given unknown data, it uses its predictive models that it has built from classical or real data to forecast the result. Since larger data volumes allow for the creation of better models, the accuracy of the model is calculated using the values upon which it is based. Figure 1's block diagram illustrates how a machine learning technique operates.



**Figure1: working of machine learning**

**1.1.1 Machine Learning Properties:**

In recent years, machine learning has made significant progress. Because of its many benefits—such as the ability to handle a wide range of data types, identify patterns and trends quickly, require no human intervention, and improve over time—machine learning has had a significant impact on a wide range of industries and applications **[2].** Machine learning's primary characteristics are:

* **Predictive modelling:** Predictive modelling is the process of developing models to forecast future events. Applications for predictive modeling include forecasting the likelihood of a loan default, consumer purchasing trends, and illness, among others **[2]**.
* **Automation:** Machine learning can be used to automatically identify patterns in data, requiring less human involvement and allowing for authorized, precise, and effective analysis.
* **Flexibility:** Because ML approaches are built to handle enormous volumes of data, they can be used for large-scale data processing **[3]**.
* **Generalization:** Machine learning techniques can be used to new, unanalyzed data to identify patterns and make helpful predictions about what will happen in the future.
* **Adaptiveness:** These algorithms are designed in such a way so as to to learn and modify continuously. So, they can improve their performance with time, leading to increased precision and efficiency with new data being available to them **[3]**.

**1.1.2. Classification of machine learning techniques**

* **Supervised learning:** It is used to construct predictive models that predict missing values using other values present in the dataset. Examples of learning approaches are decision trees, SVM, the Bayesian method, and artificial neural networks **[4]**.
* **Unsupervised learning:** Unsupervised approaches are employed when inputs are known but the output is unknown. This group includes clustering methods such as K-means clustering, among others **[4].**
* **Semi-supervised learning:** This type of learning can be used to both labeled and unlabeled data. While unlabeled data is used for a wider dispersion of data, labeled datasets are utilized for prediction purposes **[4]**.
* **Reinforcement:** Similar to conventional forms of data analysis, the algorithm in this case gathers data via trial and error before determining which activity yields the most rewards. Reinforcement learning consists of three main parts: the action, the environment, and the agent **[5]**.

**1.1.3 Machine Learning life cycle:**

The ability to automatically learn without explicit programming has been granted to computer systems via machine learning. However, how does a system for machine learning operate. The machine learning life cycle can therefore be used to describe it. A machine learning project can be built efficiently by following a cycle known as the machine learning life cycle. The life cycle's primary goal is to resolve the issue or undertaking **[6]**.

The seven main steps in the machine learning life cycle are listed below:

* Data collection
* Preparation
* Wrangling
* Analysis Information
* Educate the model
* Evaluate the model.
* Implementation



**Fig2: Machine learning life cycle**

**1.1.4. Uses of Machine Learning**

The significance of machine learning is rising as a result of the availability of high-speed Internet, the accessibility and affordability of computing power, and the ever-increasing amounts and variety of data. It is feasible to easily and automatically create models that can accurately and swiftly assess incredibly vast and complicated data sets because to these digital transformation characteristics.

Machine learning can be used in a wide range of applications to save expenses, reduce risks, and enhance overall quality of life. Some examples of these applications include making product and service recommendations, identifying cybersecurity breaches, and enabling self-driving automobiles.

**1.2 Diabetes**

Worldwide, diabetes affects a large number of individuals, and the number of diabetes sufferers is rising quickly. The International Diabetes Federation (IDF) reported that 463 million individuals worldwide were afflicted with the condition in 2019. Additionally, it has been estimated that 700 million adults will suffer from diabetes by 2045, compared to 578 million by 2030. Diabetes can

eventually lead to a number of further issues, including coma, renal and retinal failure, etc., if it is not treated or recognized **[7]**. There may well be numerous reasons, a few of them are age, family history, insulin production, body mass index, stress, pregancy etc. a few of these attributes can be taken as attributes within the dataset, and a think about can be conducted for predicting the disease based on them. These attributes play an imperative part by influencing a person's health **[8]**. In the process of refining our model, we looked at several writing review papers and discovered that traditional methods had previously been used. Hence, in order to produce a show that is unique and state-of-the-art, a modernized dataset is taken into consideration. For example, it may be a sophisticated therapeutic dataset that includes readings of people consuming junk food, regular people taking for exercise, and people going about their daily lives. This typically differs entirely from the traditional method, which uses prediction to take a defined set of numbers and produces inaccurate findings. Consequently, their technique did not yield any effective results. In order to address this problem, machine learning techniques such as decision trees, KNN, Naive Bayes, ANN, Random Forest and others were used to count computerized datasets. Among these algorithms, KNN performed well with our digital dataset, producing better and more effective outcomes.

**1.2.1 Symptoms of diabetes**

* Constantly thirsty
* Foggy vision
* Injuries that refuse to mend
* Infection in the vagina
* Constantly Hungry
* Constantly Tired
* A sexual issue
* Tingling or numbness in the feet or hands
* Continual urination
* Weight loss that is systemic

**1.2.2 Kinds of diabetes**

Type 1, Type 2, and gestational diabetes are the three main forms of diabetes.   
Type 1. diabetes is the most prevalent kind of the disease. Type 1 diabetes is characterized by a weakened immune system and insufficient insulin production by cells, i.e., immune system destruction of insulin-producing beta cells. The body either becomes incapable of producing any insulin at all or can only produce very little of it.

**Autoimmune condition:**

* Type 1. diabetes affects about 10% of people with diabetes.
* Type 2. It indicates that either less insulin is produced or the body is unable to use insulin effectively. Both lifestyle choices and genetics play a role in its cause.
* Sometimes referred to as diabetes resistant to insulin.
* The body cannot effectively use insulin.
* 80% to 90% of cases of diabetes globally are type 2 diabetes, which is the most prevalent kind of the disease.
* Type 3: Gestational diabetes: This type of the disease primarily affects pregnant women and is characterized by rapidly increasing blood sugar levels. Type 2 diabetes develops in about 50% of patients with gestational diabetes.
* Often around the 24th week of pregnancy. Mothers who are ill can also have an impact on their unborn children.
* Certain hormones produced by the placenta inhibit the action of insulin **[9]**.   
  Danger elements:
* The fetus appears larger than usual.
* Women who become pregnant and then develop type 2 diabetes.
* Past pregnancy history of gestational diabetes

**The blood glucose's source:**

* Food source
* Gluconongenesis: Glucose is produced from non-carbohydrate sources like proteins, fatty acids, etc.
* Glycogenolysis: Glucose is produced through the process of glycogenolysis.

**How diabetes affects your heart and vascular system:**

Diabetes can impact the heart and cardiovascular system in a variety of ways. Diabetes can impact these systems in different ways. The first significant one is that diabetes is sometimes referred to as the "silent killer" since elevated blood sugar levels can damage internal veins and arteries without causing symptoms to manifest until the disease improves. Because of this, high blood sugar damages the blood vessel lining, making the arteries somewhat rougher, which makes it easier for plaque from cholesterol to attach to the inner walls and result in early heart disease. And for some people, that can result in earlier heart attacks, which can lead to earlier heart attacks. Additionally, because those arteries are no longer smooth muscles, it may also result in elevated blood pressure. Because of this, it may also have an impact on blood pressure and cause it to increase. The same issue that causes plaque to build up in diabetic heart arteries more quickly and easily also affects the arteries in the legs. As a result, many of our diabetes patients report that walking causes them a lot of pain. And it's typically due to occlusions in the arteries that supply the lower limb **[10].**

**Why is such a model necessary**

* Diabetes is a medical condition that affects how efficiently our body burns food for energy.
* If diabetes is not consistently and carefully treated, blood sugar levels may rise, increasing the risk of serious side effects like stroke and heart attack.
* Diabetes is typically diagnosed every year with a battery of tests, and it requires time.
* If more people were tested at once, it would take longer.

We can automate the procedure to stop this, which will have two advantages:

* Less time will be required.
* It is possible to greatly lower the error rate, or the mistake rate in manually identifying diabetes **[9]**.

**1.2.3 Preventive and Healing Measures**

You can manage Type 2 diabetes even if there is no cure for it. Severe cases necessitate glucagon injections, although mild and moderate cases can be managed on their own **[11]**.  
Here is a list of a few safety measures:

* Drugs and a balanced way of life
* Maintain a healthy weight
* Avoid from tobacco use
* Engage in regular physical activity.
  1. **Problem identification**

It is critical to identify and treat diabetes early in order to prevent more problems from the disease. Numerous issues might arise from diabetes if it is not managed or properly diagnosed by a physician. Prediction models for identifying diabetes and its kind can be created by machine learning. Additionally, it can be utilized to identify diabetes risk factors so that preventative actions can be implemented. Numerous internet data archives, such as KAGGLE, UCI, Mh41ST, and others, have a vast amount of information about diabetes patients. Medical professionals may employ machine learning algorithms for early evaluation and treatment that are based on accessible data. This project primarily aims to provide a comparative analysis of several classification algorithms in order to categorize newly collected data as either diabetes or healthy.

* 1. **Organization of the project**

This research project presents a comparative study of several diabetes patient methods. This project is structured as follows.

* Chapter 2: The diabetic literature is thoroughly examined in this chapter.
* Chapter 3: Explain tools and techniques used in this work.
* Chapter 4: Describes the proposed model and research design.
* Chapter 5: Presents the results and comparative analysis.
* Chapter 6: Conclusion and future work

References

**CHAPTER-2**

**LITERATURE REVIEW**

Various methods have been used to create classification and prediction models for medical datasets. employing data from 402 patients gathered from various places, Tafa et al. **[12]** suggested an ensemble model employing naive Bayes and S VM for diabetes prediction. They reported an accuracy of 97.6%, which is superior than the approaches employed individually.

Sisodia and Sisodia **[13]** applied NB, decision tree, and SVM approaches to the "Pima Indians Diabetes Dataset" and then reported a greater accuracy of 76.3% accuracy utilizing naive bayes.

The "Pima Indians Diabetes Dataset" was utilized by Kumari et al. **[14]** to test an ensemble technique consisting of Random Forest, Logistic Regression, and NB. The performance of the ensemble approach was compared to other machine learning techniques, and the ensemble approach proved to execute better.

Gupta et al. **[15]** demonstrated improved support vector machine performance using SVM and Naive Bayes methods on the PIMA India diabetes dataset.

Parveen et al. **[16]** employed two ensemble techniques—bagging using J48 and adaboost using J48—as well as the J48 decision tree on the CPCSSN diabetes dataset. They found that the adaboost ensemble technique performed better in small sample sizes and that bagging performed better in large sample sizes.

Tigga and Garg **[17]** employed logistic regression, K-Nearest Neighbor, SVM, Naive Bayes, decision trees, and random forests to analyze the data they obtained from an online and offline questionnaire as well as the Pima Indian Diabetes Dataset. According to their analysis, random forests yielded the best results for both datasets, with accuracy of 94.1% and 75% for the PIMA dataset, respectively.

Diwani and Sam **[18]** On the Pima Indian Diabetes Dataset, Diwani and Sam [18**]** employed the J48 decision tree and naive bayes. They compared their performance with rule-based approaches and random trees, and found that naive bayes performed best, with an accuracy of 76.3%, followed by J48 with an accuracy of 73.8%. Ahuja et al. **[19]** employed linear discriminate analysis (LDA), which to select the salient features from the Pima Indian Diabetes dataset. They then conducted a comparative analysis of the preprocessed dataset using a number of machine learning algorithms, including SVM, Decision trees, Random Forest, Logistic Regression, and the Multilayer Perceptron, and found that MLP outperformed the other classifiers.

MaIik et al. **[17]** In a study conducted by MaIik et al. **[20]** using data from a German hospital, ten machine learning approaches were compared and it was shown that K-NN, random forest, and decision tree had the best performance.

Soltani and Jafarian **[21]** On the "Pima Indians Diabetes Dataset," Soltani and Jafarian [21] utilized a probabilistic neural network to predict diabetes. They reported accuracy rates of 89.56% and 89.49% on the training and test sets, respectively.

Hasan et al. **[22]** On the Pima Indian diabetes dataset, Hasan et al. [22] performed feature selection using PCA, independent component analysis, and correlation-based approaches. After that, they used k-NN, Decision Trees, Random Forest, Naive Bayes, Multilayer Perceptron’s, Adaboost, and gradient boost classifiers. They found that the combination of Adaboost and gradient boost classifiers outperformed the other methods.

Alehegn et al. **[23]** stated that a variety of machine-learning algorithms should be used in combination to attain the highest level of accuracy. Using Support Vector Machines and Decision Trees, Rathore et al. **[22]** reported an accuracy of 82% with SVM.

Shetty et al. **[24]** Using KNN and Bayesian algorithms, Shetty et al. [24] created an intelligent diabetic prediction system to help doctors predict diabetes.

A web application for predicting diabetes risk was created by Nai-Arun & Moungmai **[25]**. They started by building a dataset using preprocessed data that was gathered from 26 primary care facilities. They then used a variety of data mining approaches, including bagging, logistic regression, random forest, decision trees, naive bayes, and others, to report the highest performance using random forest.

Majji, Ramachandra, and associates **[26]** A website developed by Majji, Ramachandra, and associates [26] provides a person's risk score for diabetes based on a number of variables, including age, ethnicity, body mass index (BMI), smoking status, and steroid use.

Pranto et al. **[27]** An autonomous diabetes prediction method was developed by Pranto et al. [27] using the Pima Indian and Bangladesh hospital datasets. They reported 81.2% and 79.2% accuracy utilizing KNN and decision tree using the Pima Indian dataset for training and testing machine learning algorithms on the other dataset.

Olisah et al. **[28]** In order to predict missing samples, Olisah et al. [28] employed a polynomial regression-based method. They also used hyperparameter tuning on Pima Indian and LMCH Iraqi databases, and they concluded that their Deep Neural Network strategy performed better than alternative methods.

Ramesh et al. **[29]** Using the Pima Indian diabetes dataset, Ramesh et al. [29] employed a variety of preprocessing strategies, including feature scaling, feature selection, and SMOTE. They found that SVM with an RBF kernel produced the highest accuracy.

Singh et al. **[30]** employed MLP, Naive Bayes, and decision tree-based random forest algorithms in conjunction with a correlation-based feature selection technique on the Pima dataset to eliminate unneeded features.

Singh et al. **[31]** employed MLP, Naive Bayes, and decision tree-based random forest algorithms in conjunction with a correlation-based feature selection technique on the Pima dataset to eliminate unneeded features.

Nnamoko N et al. **[32]** Using feature subset selection, Nnamoko N et al. [32] achieved an accuracy of 83% using the ensemble approach on the UCI diabetes dataset.

Joshi et al. **[33]** used ANN networks, logistic regression, and SVM to create an early diabetes prediction model. The authors stated that by employing these methods, their accuracy improved. Naive Bayes, SVM, and decision trees were used in tests conducted by Deepti Sisodia and her cofounder on the PIDD dataset. Among them, Naive Bayes did better.

Nongyao Nai-arun et al. **[34]** used 30,122 patient records with 12 characteristics. They used the Naive Bayes algorithm, decision trees, random forests, ANN, and logistic regression to create online apps. Using random forest, the program achieved an accuracy of 85.55% in classifying patient data into groups with diabetes and those without. In their hybrid model for diabetes diagnosis and prediction, Nahla Barkat and Andrew Bradley **[34]** found that waist circumference, BPDIAS (diastolic blood pressure), and FBS (fasting blood sugar) are the three main risk factors for diabetes prediction. With SQR ex-SVM, they were able to extract rules for the diagnosis of diabetes and attain 94% prediction accuracy, 93% sensitivity, and 94% specificity.

According to Branimir et al. **[35]**, a method was developed to address the two issues of lack of transparency in characteristics and heterogeneity with relation to prior techniques. The PRISMA approach was applied. In addition to tree-based techniques, eighteen different types of model comparison were carried out. The authors came to the conclusion that SVM and KNN are mostly utilized for prediction.

Jyotismita et al. **[36]** In order to overcome the limitations of classification, Jyotismita et al. [36] employed the detection and analysis mechanism of diabetes disease employing six facets: dataset, processing methods, feature extraction, ML identification, and classification and diagnosis of DM. Different supervised, unsupervised, and clustering approaches are compared. There is a great deal of work to be done to increase the effectiveness of the detection of different diabetic illnesses, and different datasets provide distinct obstacles.

The primary focus of Nur et al. **[37]** has been on data preprocessing, which entails the following procedures: feature importance and data augmentation; missing value removal; and data balancing. The algorithms for algorithm classification are LR and RF. Compared to data without preprocessing, the result obtained is 24% higher for recall and 20% higher for precision.

Prajyot Palimkar et.al **[38]** This research also proposes an efficient diabetes prediction model that uses many machine learning methods to improve diabetes prediction accuracy and improve diabetes classification. To predict diabetes at an early stage, several machine learning methods are used, including AdaBoost Classifier, Gaussian Process Classifier, K-Nearest Neighbors, Support Vector Machine, Random Forest Classifier, Logistic Regression, and Gaussian Naïve Bayes. These models' performances are evaluated using the appropriate metrics, such as F-Measure, Error, Accuracy, Precision, and Recall.

Chau,C.y et.al **[39]** Eight separate variables were used in this study to preprocess the data: age, BMI, sebum thickness, insulin level, number of pregnancies, plasma glucose level, diastolic blood pressure, and diabetes pedigree function. This study collected the data for the model performance analysis after training, testing, cross-validation, and comparison. The outcomes demonstrated that while all models produced good results, the two-class decision jungle and two-class enhanced decision tree were the most effective models. As the performance indicator, the area under the curve (AUC) was chosen, and AUC scores of 0.976 and 0.991 were obtained.

**CHAPTER 3**

**TOOLS & TECHNIQUES**

To conduct this study the following tools are used: -

* Python Program Language
* Algorithm for classification
* Dataset

**3.1 Introduction to python**

Introduction to Python Language: Python is a popular high-level, general-purpose programming language. In 1991, Guido van Rossum created the first concept, which was later improved by the Python Software Foundation. Programmers may express concepts in fewer lines of code thanks to its syntax, which was primarily built with code readability in mind **[40]**.

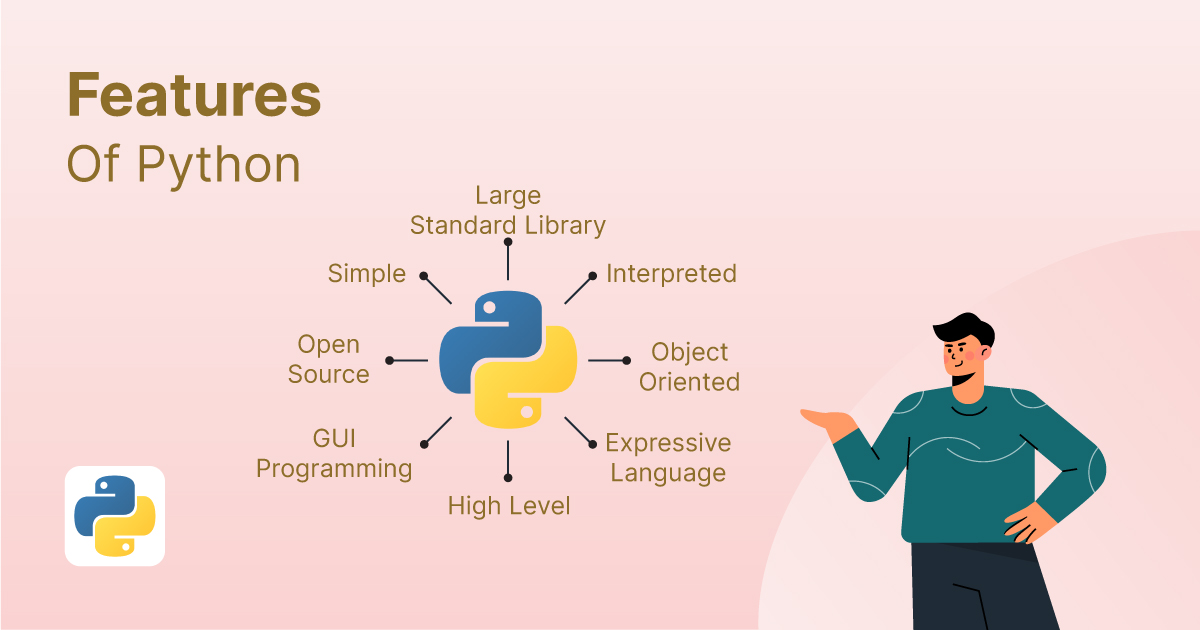
Python is a programming language that facilitates faster work and more effective system integration. Python is an object-oriented, interpreted, high-level programming language.   
The design of Python focuses readability. It has fewer syntactical structures than other languages and typically applies English keywords instead of punctuation.

* **Python is Interpreted:** The interpreter processes Python at runtime. It is not necessary for you to compile your software before running it. This reminds me of PERL and PHP.
* **Python is Interactive:** You can write programs by just interacting with the interpreter while seated at a Python prompt.
* **Python is an object-oriented:** Programming language that supports the encapsulation of code within objects.

**3.1.1 Features of Python include:**

* Python's structure is straightforward, with minimal keywords, making it easy to learn.
* Specified syntax. This supports the student's rapid learning of a new language.
* Readable − Python code is more readable and visually appealing.
* Simple to maintain – The source code of Python is relatively simple to maintain.
* A large standard library—the majority of Python's library is cross-platform compatible and highly portable on UNIX, Windows, and Macintosh.
* Interactive Mode: Python comes with an interactive mode that lets you test and debug small portions of code interactively.
* Portable: Python runs with the same interface across a broad range of hardware devices.
* Extendable: The Python interpreter can have low-level modules added to it.
* Programmers can enhance or modify these modules to make their tools more effective.
* Databases: All of the major commercial databases have interfaces available in Python.
* GUI Programming: Python allows the creation and conversion of GUI applications to a variety of system calls, libraries, and Windows systems, including Macintosh, Windows MFC, and Unix's X Window system.

**Scalable:** Python offers larger projects more organization and assistance than shell programming.



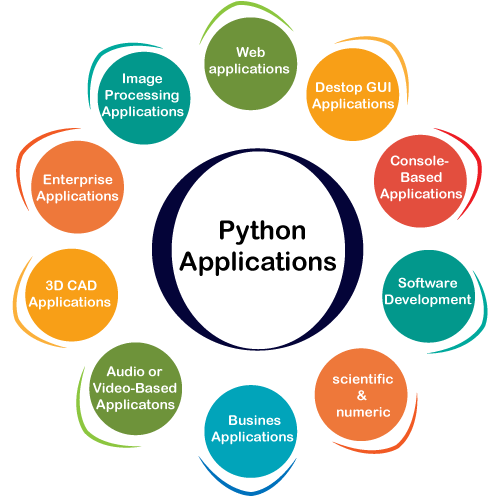
**Fig:3Features of python**

**3.1.2 Applications utilizing Python**

* Python has been effectively integrated as a scripting language into several software applications.
* GNU Debugger display sophisticated structures, such C++ containers, using Python as a nice printer.
* Python is frequently used for jobs involving natural language processing, but it has also been utilized in artificial intelligence.

**3.1.3 Programs:**

* Desktop GUI apps.
* Applications for image processing, games, scientific and computational applications, and graphic design.
* Applications and frameworks for the web.
* Applications for businesses and enterprises
* System Operating
* Access to Education Databases
* Language learning
* Making prototypes
* Program Development



**Fig :4 Applications of python**

**3.1.4 Python Syntax in Relation to Other Languages**

* Python was created with readability in mind and shares several characteristics with the English language, although with a mathematical bent.
* Python finishes a command with a new line, unlike other programming languages that frequently utilize parentheses or semicolons.
* Python defines scope—such as the scope of loops, functions, and classes—by indentation with whitespace. Curly brackets are frequently used in other computer languages for this reason.

**3.1.5 Benefits of Python**

A number of Python's benefits help explain why the language is so well-liked and often

used.

* **Readability:** Python's simple and straightforward syntax facilitates code readability, which makes it simpler to create, comprehend, and update.
* **Large library**: Python's vast library and framework ecosystem offers solutions for many different jobs, cutting down on the amount of time and effort needed to construct a program.
* **Cross-platform compatibility:** Python is compatible with a variety of operating systems, including Windows, macOS, and Linux, which guarantees the portability and interoperability of code.
* **Huge community:** Python boasts a large and vibrant community of developers who build tools, frameworks, libraries, and other ecosystem components while also offering learning resources and support.
* Python's adaptability makes it a useful tool for both developers and academics, as it can be applied to a wide range of tasks, including web development, data analysis, and artificial intelligence.

**3.1.6 Installation of python**

Installing the Python interpreter from the official website (python.org) is the first step in getting started with Python. After installation, you can use an integrated development environment (IDE) like PyCharm or Jupyter Notebook, or a text editor, to begin creating and running Python code. To further aid in learning Python, a plethora of online resources are accessible, such as documentation, tutorials, and online courses. The official Python manual, "Automate the Boring Stuff with Python" by Al Sweigart, and online learning environments like Coursera, Udemy, and Codecademy are a few suggested resources.

**3.2 Algorithm for Classification**The process of classifying data into a group according to a set of rules or similarities is called classification. Although datasets for classification problems contain many features, not all of them are helpful for classification. Performance is decreased by duplicate and irrelevant features. These characteristics could be regarded as loud. In order to accomplish this, features that accurately characterize a certain class must be extracted and chosen. Since the instances are provided with known class labels, classification is also known as supervised learning. An instance is represented by a collection of features, which might be continuous or categorical. Classification is the process of building the model from the training set. The resulting model is then used to predict the class label of the testing instances or to predict the class of new observation.

The classification techniques used in the present work are Decision tree, Navie bayes, Random Forest, ANN, KNN.

**3.2.1 Decision Tree:**

In addition, decision trees can be utilized as supervised learning techniques by segment A decision tree is a supervised learning technique that predicts fresh input by dividing the space of outcomes into many regions **[41]**. Regression and classification are two uses for it **[42]**. Two methods of handling continuous data have been developed: the ID3 approach (Quinlan, 1986) and the C4.5 method (Quinlan, 1993). The decision tree is a table shaped like a tree with links to nodes that are available. Every node is categorized as either a single leaf node or a branch node that is followed by several nodes. Decision trees are used in a strategic manner to solve problems by breaking down complicated issues into several smaller ones and solving the smaller issues by continually applying the data mining technique to find training examples of different types of knowledge classification by building decision trees. Developing a small-scale, very precise decision tree is the key component of the decision tree concept. The Decision Tree Method has numerous advantages. The non-parameter method's extreme flexibility and lack of consideration for data distribution are apparent at first. The opposite side is in good health **[43]**.

Below diagram explains the general structure of a decision tree



**Fig: 5 Decision Tree structure**

**How the decision tree algorithm operates**

* According to S, start the tree at the root node, which has the entire dataset.
* Use the Attribute Selection Measure (ASM) to determine which attribute in the dataset is the best.
* Separate the S into subsets that include potential values for the best characteristics.
* Create the decision tree node that has the best feature. Using the subsets of the dataset generated in above.
* Involves repeatedly creating new decision trees. Proceed in this manner until the nodes reach a point where more classification is not possible, at which point the last node is referred to as a leaf node.

**3.2.2 Naive Bayes:**

Naive Bayes classifier is a very popular and classic machine learning model for classification problems. Typically, the model used a series of features to categorize the element. The concept of Naive Bayes Classifier is simple and straightforward which is to categorize element via using the conditional probability of naïve Bayes theorem **[42]**. A probabilistic machine learning technique based on the probability-theorized Bayes theorem is the Naïve Bayes classification method. It performs better than other classifiers given its simplicity, making it one of the finest. Finding the posterior probability, P(c/x), from P(c), P(x), and P(x/c) is made possible by the Bayes theorem. When a predictor (x) has an impact on a class (c), the Naive Bayes classifier assumes that this influence is independent of the values of other predictors. Class conditional independence is the term for this kind of presumption **[43]**.

|  |  |  |
| --- | --- | --- |
|  |  |  |

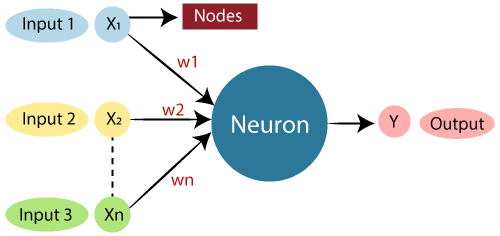
* The posterior probability of the class (target) given the predictor (attribute) is denoted by P(c/x).
* The prior probability of class is denoted by P(c).
* The likelihood, or probability of the predictor given class, is represented by P(x/c).
* P(x) represents the predictor's prior probability.

Why Use the Naïve Bayes Method

* Easy going.
* Excellent performance and scalability.
* Flexible enough to perform nearly any type of categorization task.
* Amazing quick and effective.
* Frequently yields positive outcomes.
* Documents don't need a lot of ram.
* Not as overly fitted and more general.
* Effective even for a limited training set.
* "Resistant" to the qualities of no significance.

Ideal for most data types, including nominal and numerical properties. The naive Bayes technique is also known as independent Bayes, simple Bayes, and idiot's Bayes. This approach is significant since it requires no complex iterative parameter estimate algorithms and is incredibly simple to implement. This implies that it might be used with ease to massive data sets. Because of its simplicity of interpretation, even inexperienced users of classifier technology can easily understand it. Lastly, it frequently performs shockingly well: it might not modelling the conditional probability distribution P(C|D), where C ranges over classes and D over descriptions, in some language, of objects to be categorized, is a common step in probabilistic approaches to classification **[44]**.

**3.2.3 Artificial Neural Network (ANN):** A neural network (ANN)The biological neural networks that give rise to the structure of the human brain are the source of the phrase "Artificial Neural Network". like that of the human brain. The foundation of deep learning techniques, neural networks, sometimes referred to as Artificial Neural Networks (ANNs), are a subset of machine learning **[45]**and have been used to numerous applications **[46–48]**. The input, output, and one or more hidden layers comprise the node layer of an artificial neural network. Every node, or artificial neuron, has a weight and threshold and is connected to other nodes. Any node whose output exceeds the designated threshold value becomes active and begins transmitting data to the network's upper layer. If not, no data is transferred to the network's subsequent layer. In addition, the neural networks have the ability to learn from the training data to improve the accuracy of the outcome **[41]**.



**Fig 6: Artificial neural network**

**3.2.4 KNN:** The K-Nearest Neighbor (KNN) method is typically employed to address classification problems in business settings, but it may also be used to handle regression and classification difficulties. Its primary benefits are its quick computation time and ease of translation **[49]**. KNN determined using a function of distance. Equation gives the Euclidean distance between two points, x and y.

Euclidean distance=

Examining the data set yields the value of k, the positive integer. Another method for determining an appropriate k value in retrospect is cross-validation, which involves validating the k value using a separate set of data. In this case, the values (k = 1, 3, and 5) yield good results at k = 5. This suggests that when the k value increases, the accuracy of the output will increase. The ideal k value will typically fall between three and ten **[50]**. Among the categorization algorithms is the K-Nearest Neighbor algorithm. Compared to other data mining approaches, it is the easiest and most straight forward. This method uses a similarity measure to categorize new possessions **[52]**. Positive integer numbers are always assigned to the value of k. The training data are saved in this algorithm based on the closest or neighbors, the test data prediction is finished.

Step/phase I: Find out how many neighbors there are in your immediate vicinity, k.   
Phase II. Calculate the separation between the training samples and the instance.

stage/phase III: The training samples' degree of remoteness is sorted, and the neighbor who is nearest to you based on the shortest distance is identified in this stage.

Phase IV: All the classes from all the training data are obtained in this step.

Step/phase V: As the query instance's prediction value, use the bulk of the class of nearest neighbors **[51]**. Random Forest **[42]** It can be applied to both classification and regression. One supervised learning technique that can be used to forecast and categorize data is the Random Forest method. High dimensionality and huge data sets are easily handled by RF. Random selection is used to choose the samples **[51]**.

Algorithm for random forests **[50]** The two parts of this algorithm are

* Bagging trees
* From picking trees to a random forest.

Every tree is grown in this manner:

* Select N cases at random, using replacement, from the original data if the training set has N cases. The training set for the tree's growth will be this sample.
* The node is split using the optimal split if there are M input variables, and a random number of attributes are chosen. Throughout the growth of the forest, the value of M remains unchanged.

**3.3 Dataset and Parameters**

In this dataset 9 features, including glucose, pregnancies, skin thickness, insulin, BMI, diabetes pedigree function, and outcome, are included in the dataset, which has 768 entries. The original source of the diabetes data set <https://www.kaggle.com/johndasilva/diabetes>. Diabetes dataset of two thousand instances. The goal is to determine whether or not the patient has diabetes by using the measurements.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Pregnancies** | **Glucose** | **Blood**  **Pressure** | **Skin**  **Thickness** | **Insulin** | **BMI** | **Diabetes**  **PedigreeFun.** | **Age** | **Outcome** |
| 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 1 |
| 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1 |
| 5 | 116 | 74 | 0 | 0 | 25.6 | 0.201 | 30 | 0 |
| 3 | 78 | 50 | 32 | 88 | 31 | 0.248 | 26 | 1 |
| 10 | 115 | 0 | 0 | 0 | 35.3 | 0.134 | 29 | 0 |
| 2 | 197 | 70 | 45 | 543 | 30.5 | 0.158 | 53 | 1 |
| 8 | 125 | 96 | 0 | 0 | 0 | 0.232 | 54 | 1 |
| 4 | 110 | 92 | 0 | 0 | 37.6 | 0.191 | 30 | 0 |
| 10 | 168 | 74 | 0 | 0 | 38 | 0.537 | 34 | 1 |
| 10 | 139 | 80 | 0 | 0 | 27.1 | 1.441 | 57 | 0 |
| 1 | 189 | 60 | 23 | 846 | 30.1 | 0.398 | 59 | 1 |
| 5 | 166 | 72 | 19 | 175 | 25.8 | 0.587 | 51 | 1 |
| 7 | 100 | 0 | 0 | 0 | 30 | 0.484 | 32 | 1 |
| 0 | 118 | 84 | 47 | 230 | 45.8 | 0.551 | 31 | 1 |
| 7 | 107 | 74 | 0 | 0 | 29.6 | 0.254 | 31 | 1 |
| 1 | 103 | 30 | 38 | 83 | 43.3 | 0.183 | 33 | 0 |
| 1 | 115 | 70 | 30 | 96 | 34.6 | 0.529 | 32 | 1 |
| 3 | 126 | 88 | 41 | 235 | 39.3 | 0.704 | 27 | 0 |
| 8 | 99 | 84 | 0 | 0 | 35.4 | 0.388 | 50 | 0 |
| 7 | 196 | 90 | 0 | 0 | 39.8 | 0.451 | 41 | 1 |
| 9 | 119 | 80 | 35 | 0 | 29 | 0.263 | 29 | 1 |

* There are 2000 data points in the diabetes data collection, each with 9 attributes.
* The feature we will forecast is called "Outcome," where 0 denotes no diabetes and 1 denotes diabetes.

**CHAPTER 4**

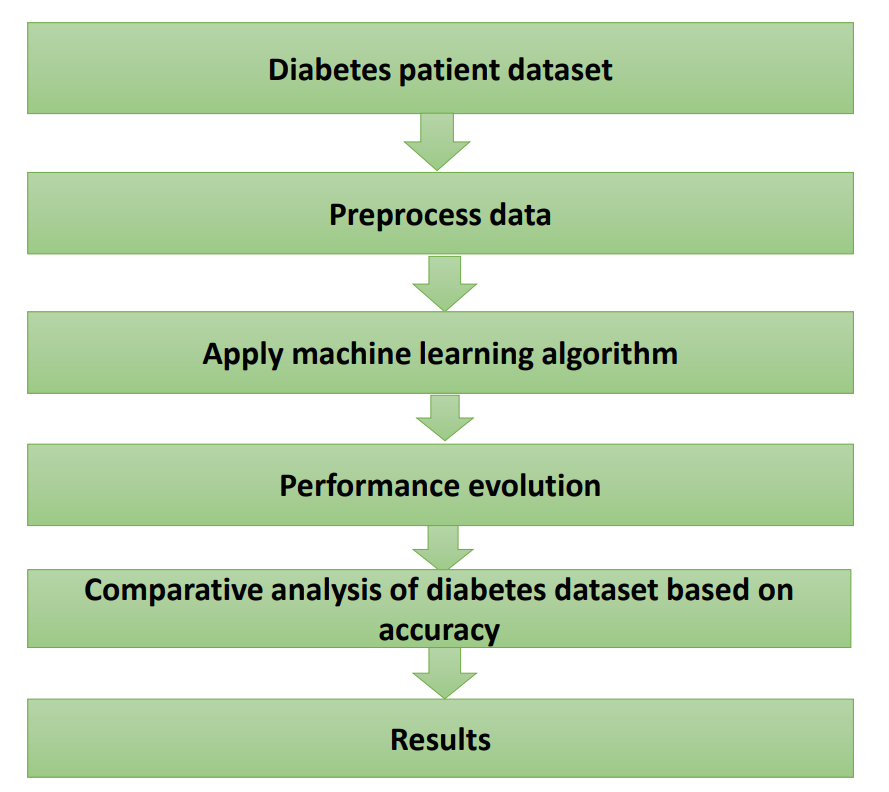
**RESEARCH OBJECTIVES AND METHODOLOGY**

**4.1 Research objectives**

The aim of this study is to conduct a thorough literature evaluation on machine learning methods with a focus on diabetes prediction. In addition, the project intends to perform comparison studies of machine learning models and analyze machine learning models using pertinent assessment metrics.

1. To study machine learning techniques for diabetes prediction.
2. To collect public dataset for diabetes prediction.
3. To compare different machine algorithm to analyse the performance and find out the best one.

**4.2 Methodology**

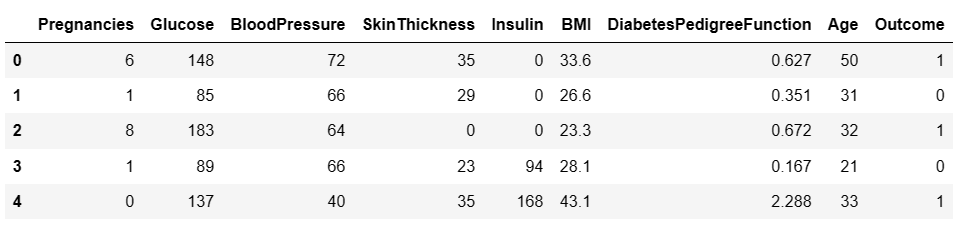


**Fig4.2 Methodology diagram**

Step diabetes patient dataset.

* Step Preprocess data
* Step Apply machine learning algorithm
* Step Performance evolution
* Step Comparative analysis of diabetes dataset based on accuracy
* Step Result

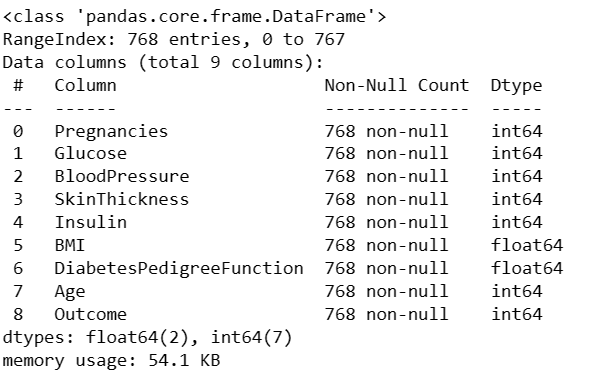
**4.2.1 Description of the Dataset**   
The original source of the diabetes data set <https://www.kaggle.com/johndasilva/diabetes>. Diabetes dataset of two thousand instances. The goal is to determine whether or not the patient has diabetes by using the measurements.



➔ There are 2000 data points in the diabetes data collection, each with 9 attributes.   
➔ The feature we will forecast is called "Outcome," where 0 denotes no diabetes and 1 denotes diabetes.

|  |  |
| --- | --- |
| **Attribute No.** | **Attribute Name** |
| 1 | Pregnancies |
| 2 | Glucose |
| 3 | Blood pressure |
| 4 | Skin thickness |
| 5 | Insulin |
| 6 | BMI |
| 7 | Diabetes pedigree fun. |
| 8 | Age |
| 9 | Outcome |

**Table-1: Attribute of patient’s dataset**



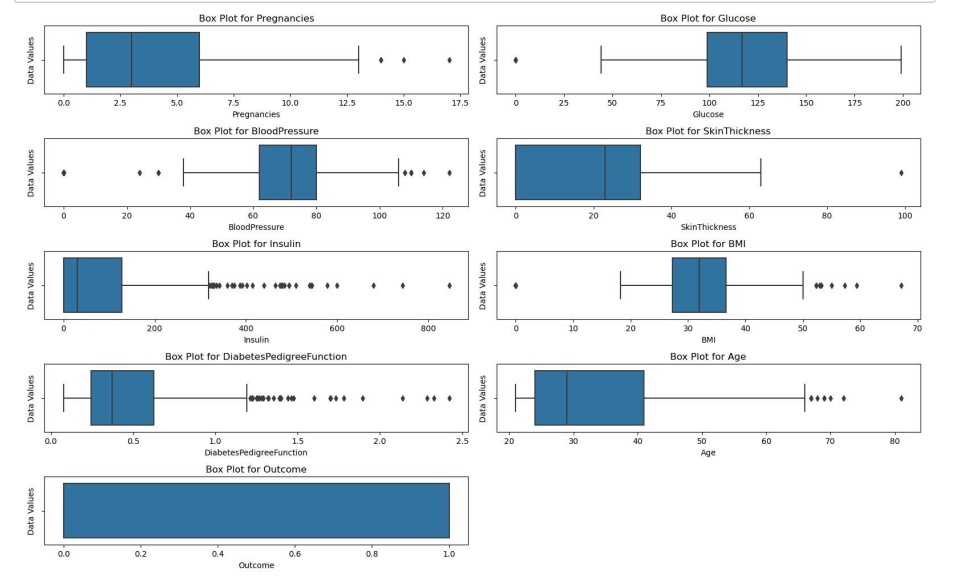
There is no null values in dataset.

**4.2.2 Preprocessing Data**

* Handling missing value
* Find Duplicate value

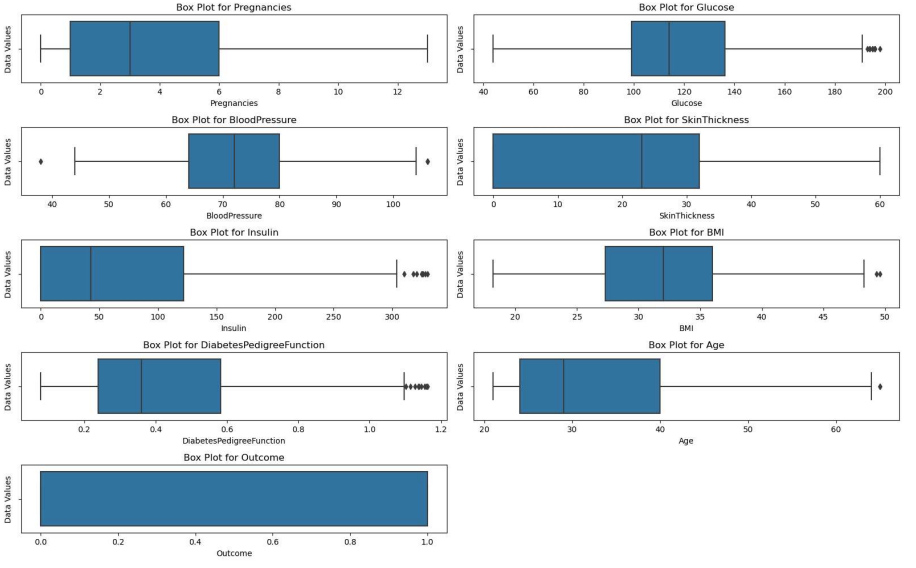
From the above we can see there is no missing value/null value.

Outlier detection using box plot



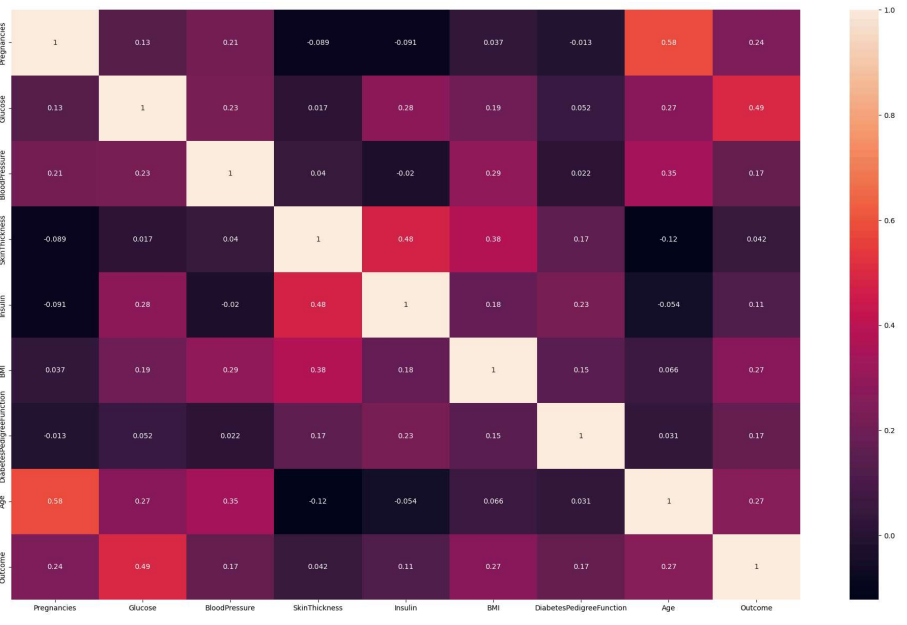
**Fig:7 Outlier Detection**

**Remove outlier**



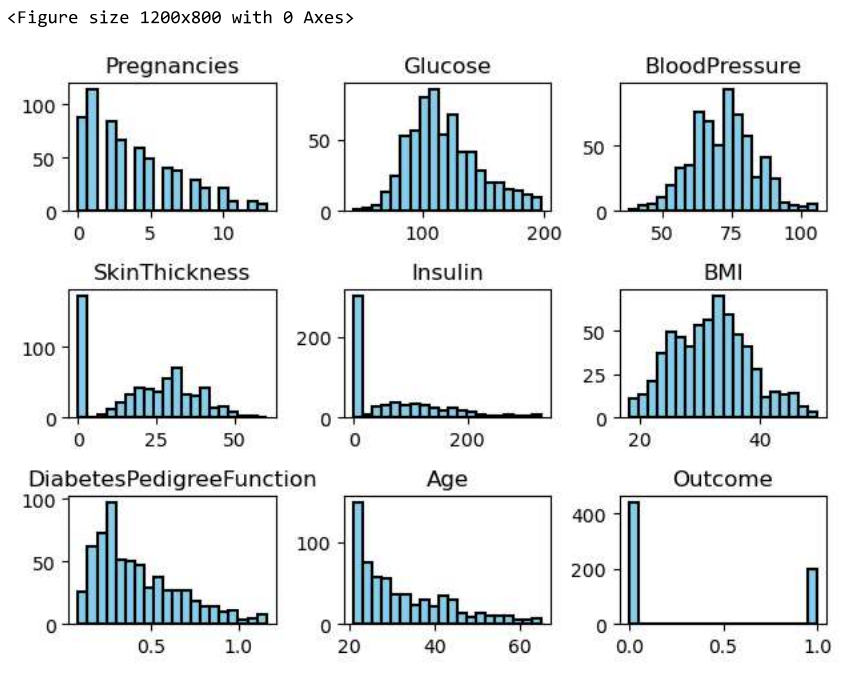
**Fig: 8 Remove outlier**

**Co-relation matrix**



**Fig 9: Co-relation matrix**

**Plot histogram**

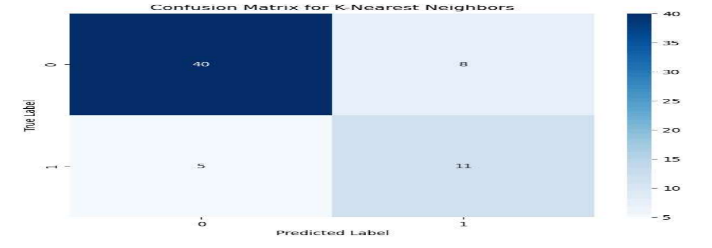


**Fig:10 Plot histogram**

**4.2.3 Used algorithm**

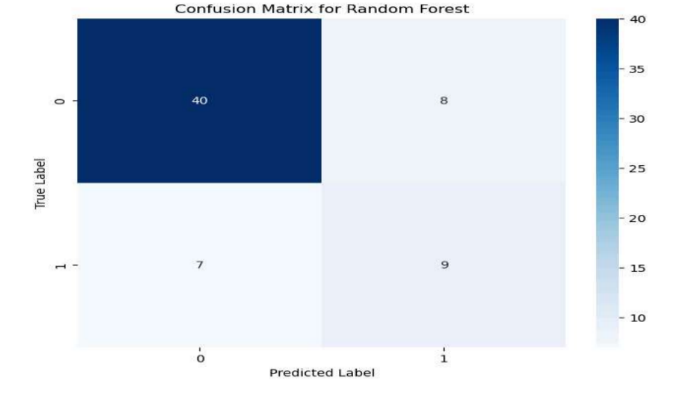
Five different types of algorithms are used in this which is KNN, Random Forest, Decision Tree, Navie bayes, ANN.

**KNN:** The KNN algorithm is a supervised learning classifier that operates in a non-parametric manner. It employs proximity to classify or anticipate how a single data point will be grouped. This is a widely used and straightforward classifier for regression and classification in modern machine learning.



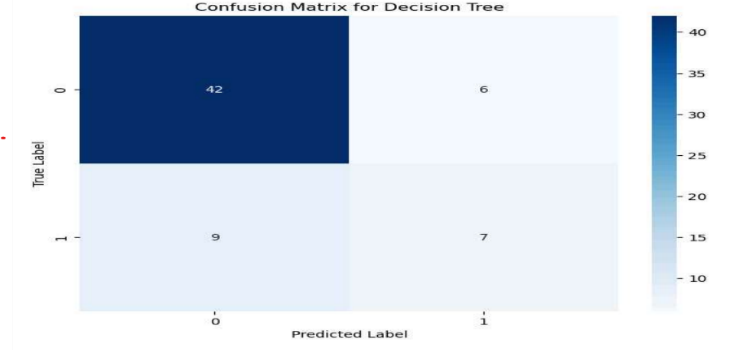
**Fig: 11 Confusion matrix of KNN**

**Random Forest:** One of the well-liked and flexible algorithms utilized in ensemble technique is RF. In terms of improving performance and prediction accuracy, it is the most well-liked machine learning method under the hybrid model approach. Large data sets and high dimensionality can be handled with ease with RF. Random selection is used to choose the samples.



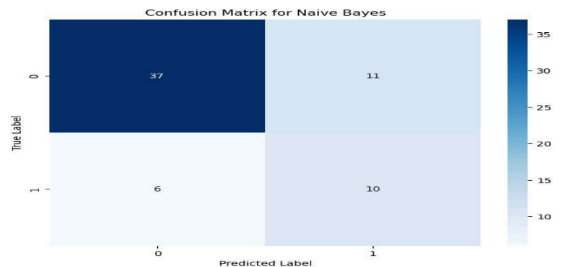
**Fig 12 Confusion matrix of Random Forest:**

**Decision Tree:** One of the most effective supervised learning techniques for both regression and classification problems is the decision tree. With each internal node signifying a test on an attribute, each branch representing the test's outcome, and each leaf node (terminal node) holding a class name, it constructs a tree structure like a flowchart.



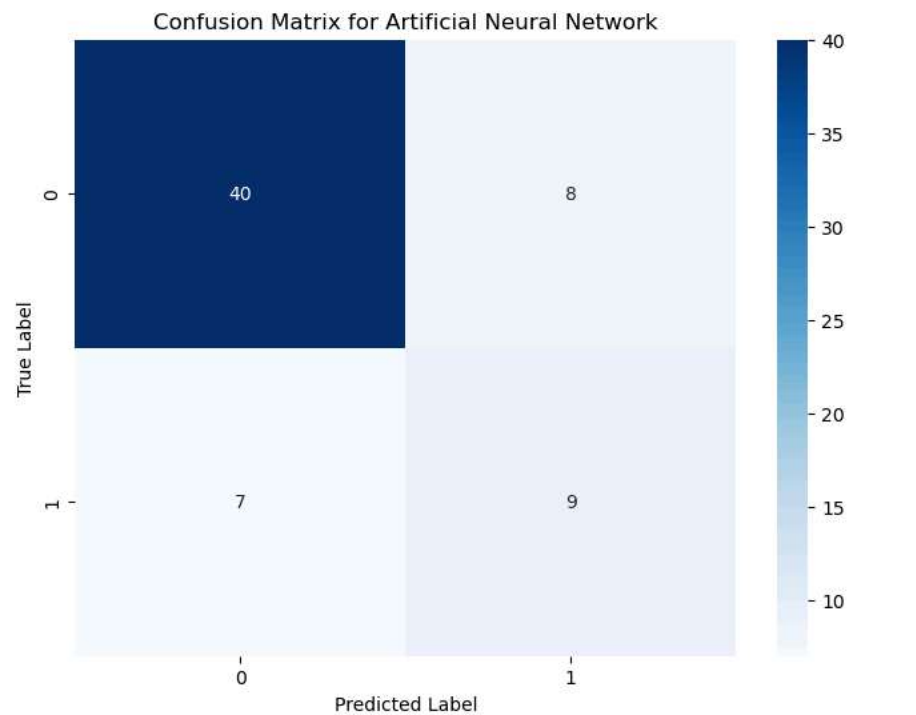
**Fig 13 Confusion matrix of Decision Tree**

**Navie Bayes:** Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle.



**Fig 14: Confusion matrix of Navie bayes:**

**ANN:** ANNs are composed of artificial neurons which are conceptually derived from biological neurons. Each artificial neuron has inputs and produces a single output which can be sent to multiple other neurons. The inputs can be the feature values of a sample of external data, such as images or documents, or they can be the outputs of other neurons. The outputs of the final of the neural net accomplish the task, such as recognizing an object in an image.



**Fig15: Confusion matrix of ANN**

**4.2.4 Performance Evaluation**

Performance matrices are essential for evaluating trained models and learning in classification. The confusion matrix's content is used to analyze the metrics listed below while analyzing the outcomes.The table known as the confusion matrix provides an explanation of how the model performed on test data for values that were known to be true. These are some example confusion measures and a sample matrix for binary classification, which is the work being done here.

Confusion Matrix

|  |  |  |
| --- | --- | --- |
|  | Actual Positive | Actual Negative |
| Predictive Positive | TP | FP |
| Predictive Negative | FP | TN |

**Table 2 Confusion matrix**

TP: The values that both match the positive prediction and the actual value.   
TN: The value that is both negatively expected and negative in reality.   
FP: Positive expected but negative actual results.   
FN: Positive outcomes that were predicted to be negative.

**Accuracy:** The most widely used and straightforward metric to determine a classifier's performance is accuracy. It can be simply described as a model's degree of correct predictions. The trained classifier's accuracy is evaluated using total correctness, which is the total number of occurrences that the trained classifier correctly predicts when tested with unseen data. The ratio of accurate predictions to the total number of cases assessed is generally measured by the accuracy matrix.

Accuracy = TP+TN / TP+TP+FP+TN+FN

Or

Accuracy = TP+TN/ Total No. Sample

**Accuracy has the following benefits:**

Suitable for issues with multiple classes and labels.   
Simple to score   
Simple enough for humans to understand.   
Simple to calculate.   
Humans can understand it easily.

**Limitations of Accuracy:**

One of the primary limits of accuracy is that it generates values that are less distinguishing and less discriminable. As a result, it results in decreased accuracy and discriminating ability when choosing and identifying the best classifier.

**Precision**: Precision is the ratio/relationship between the True Positives and all the points that are classified as Positives.

F-Measures metric represents the harmonic mean between recall and precision values. Precision is used to measure the positive patterns that are correctly predicted from the total predicted patterns in a positive class. Recall is used to measure the fraction of positive patterns that are correctly classified.

**Precision = TP/TP+TN+FP+FN**

Precision: Precision can be defined as the ratio of correctly identified positive examples to the total number of positive examples that the system has labeled.   
Recall: Recall is a metric that quantifies how well our model is able to recognize True Positives. Another name for it is True positive rate.

Recall = TP/TP+FN

**Recall:** Recall is the ratio of successfully classified positive examples to total positive examples in the dataset. Combining the above to get the F-measure.  
**F1 score**: The symphonic mean of recall and precision is represented by the value of f1-score-score.

F1 = 2 \* precision \* recall

With poor performance in terms of recall or either precision, the F-measure is more helpful than the accuracy. Another term for the F-measure is Fl score.

**CHAPTER 5**

**RESULTS AND DISCUSSION**

The PIMA diabetes data is used in this .768 records with two classes—Pregnancies, Glucose, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, and Outcome make up PIMA Dataset.   
A comparison is conducted between different methods. A variety of performance measurements are used to analyze performance.

We apply five machine learning algorithms: KNN, Naive Bayes, Decision Tree, Random Forest, and Artificial Neural Network (ANN). **The resulting Confusing Matrix is displayed in Table 3**

|  |  |
| --- | --- |
| Algorithm | Confusion Matrix |
| Random Forest | |  |  | | --- | --- | | 40 | 8 | | 7 | 9 | |
| Naïve Bayes | |  |  | | --- | --- | | 37 | 11 | | 6 | 10 | |
| KNN | |  |  | | --- | --- | | 40 | 8 | | 5 | 11 | |
| ANN | |  |  | | --- | --- | | 40 | 8 | | 7 | 9 | |
| Decision Tree | |  |  | | --- | --- | | 42 | 6 | | 9 | 7 | |

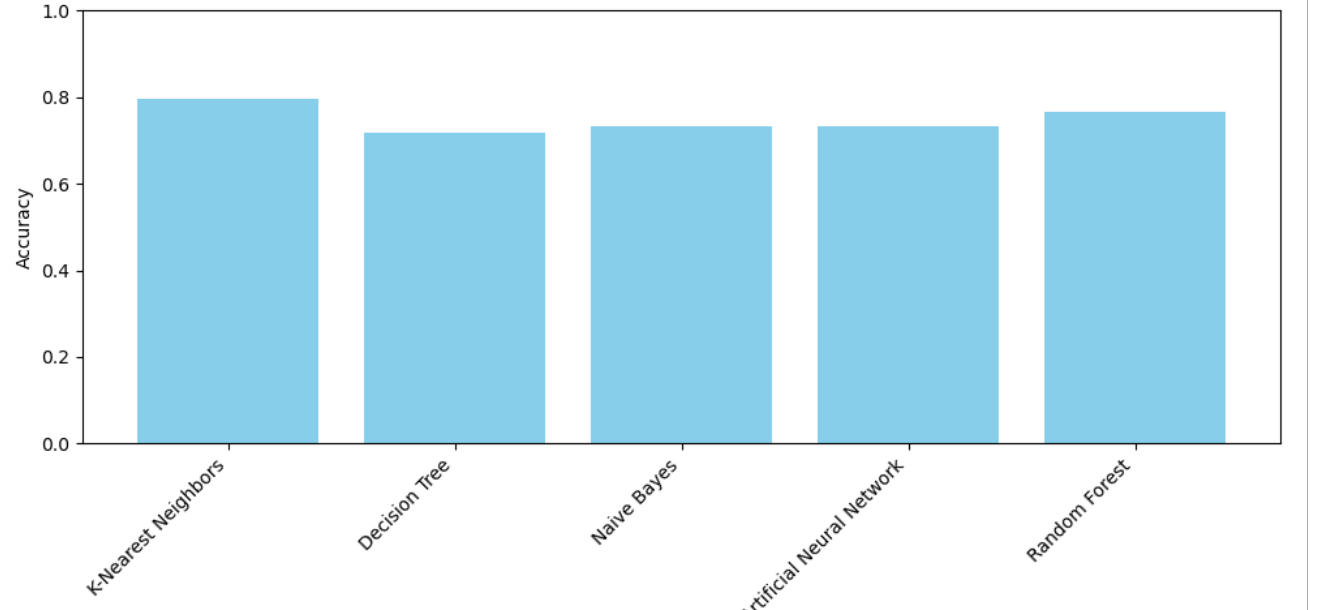
**Table 3 Confusion matrix**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 Score |
| K-Nearest Neighbors | 0.796875 | 0.811404 | 0.796875 | 0.802304 |
| Decision Tree | 0.71875 | 0.71875 | 0.71875 | 0.71875 |
| Naive Bayes | 0.734375 | 0.764396 | 0.734375 | 0.745025 |
| Artificial Neural Network | 0.734375 | 0.739987 | 0.734375 | 0.737002 |
| Random Forest | 0.765625 | 0.770651 | 0.765625 | 0.767943 |

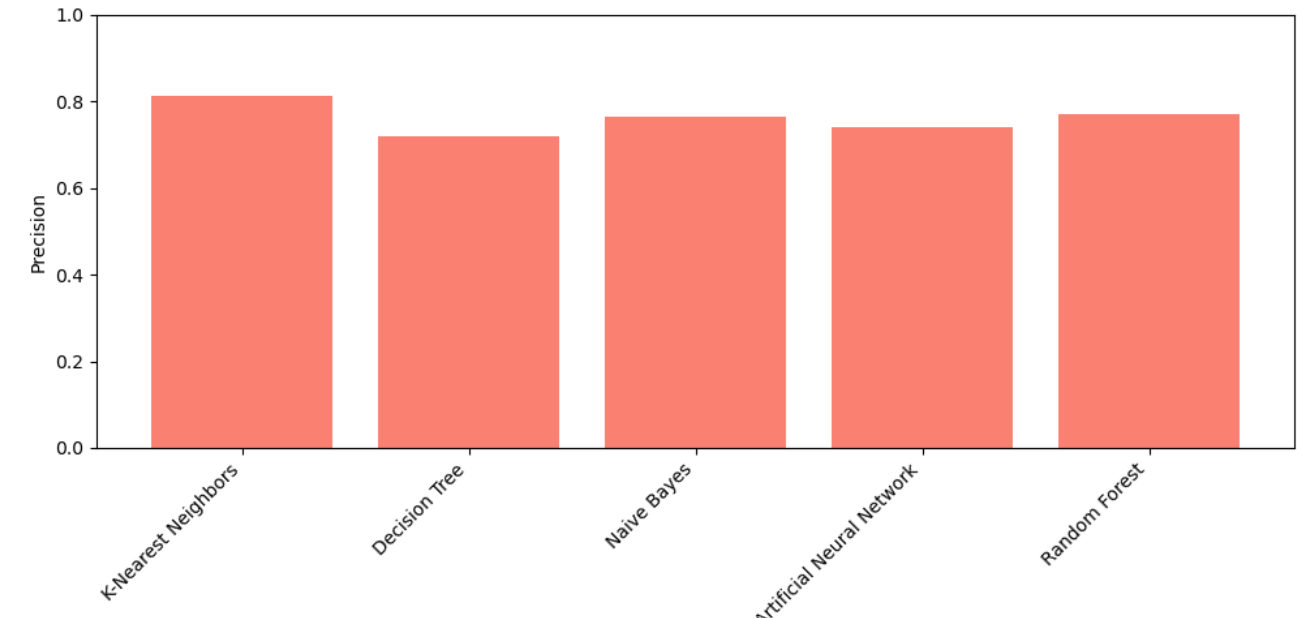
**Table 4 Performance metrics of Different Models**

The accuracy, precision, recall, and F1 score for each of the five machine learning models are displayed in Table 4; it is clear from this table that KNN offers the highest accuracy when compared to the other models, as well as the highest precision, recall, and F1 score. Consequently, when compared to other models, KNN is the best.

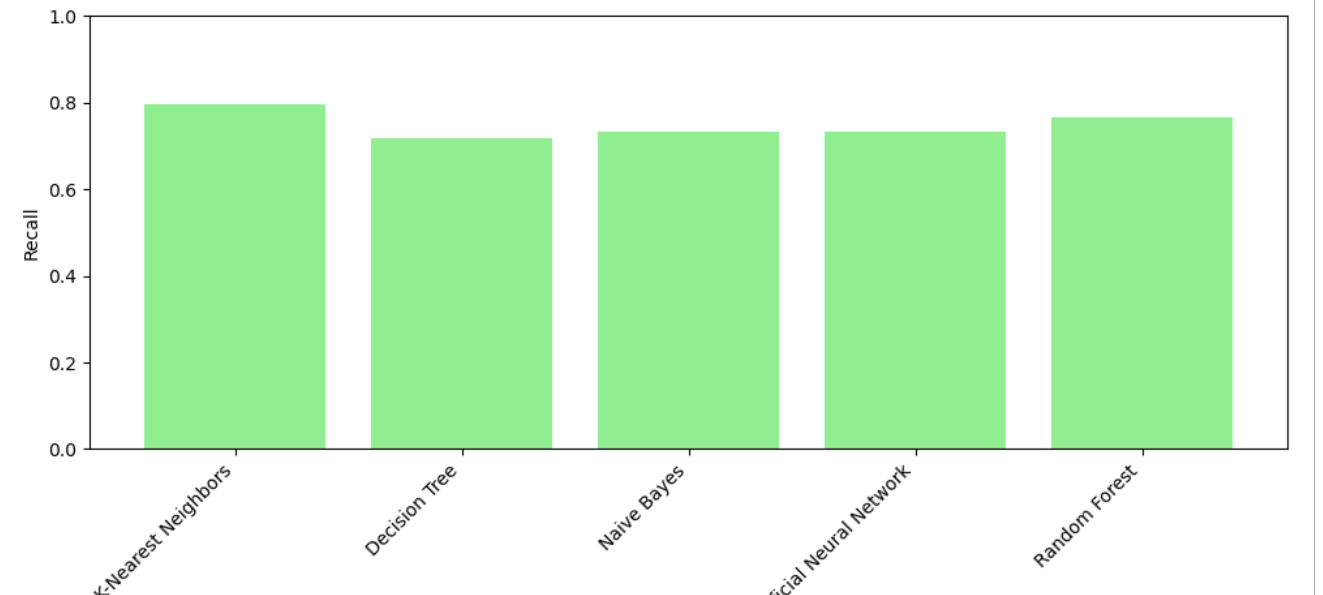
**5.1 Graphical representation**



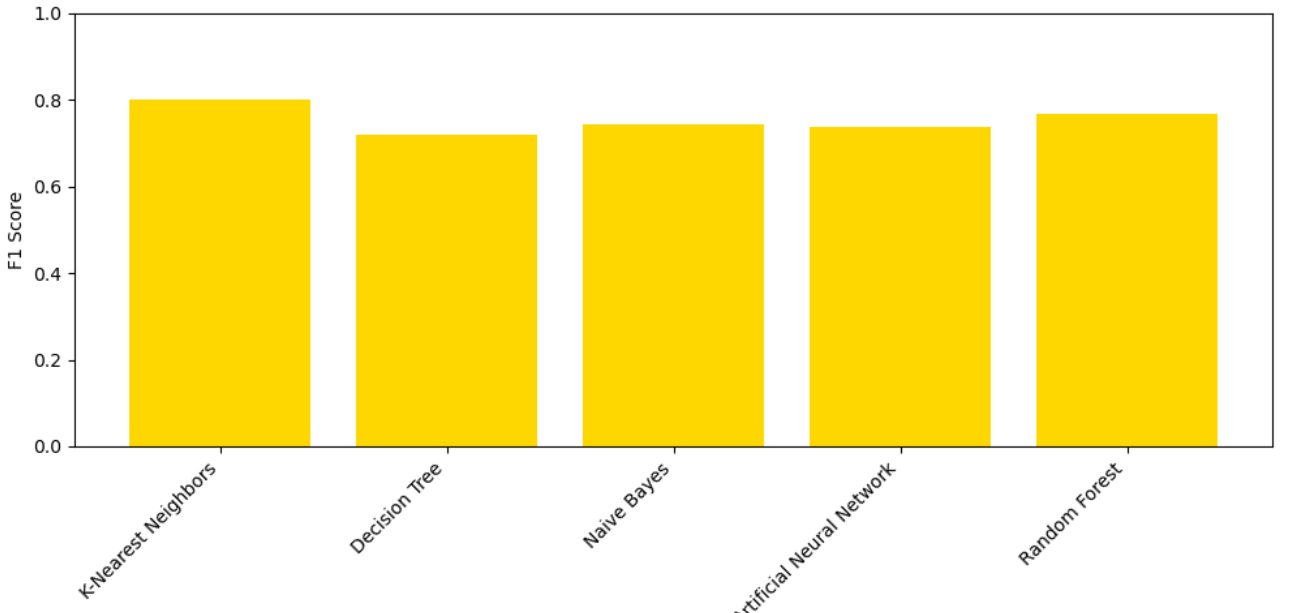
**Fig 16: Comparison Result of accuracy**

****

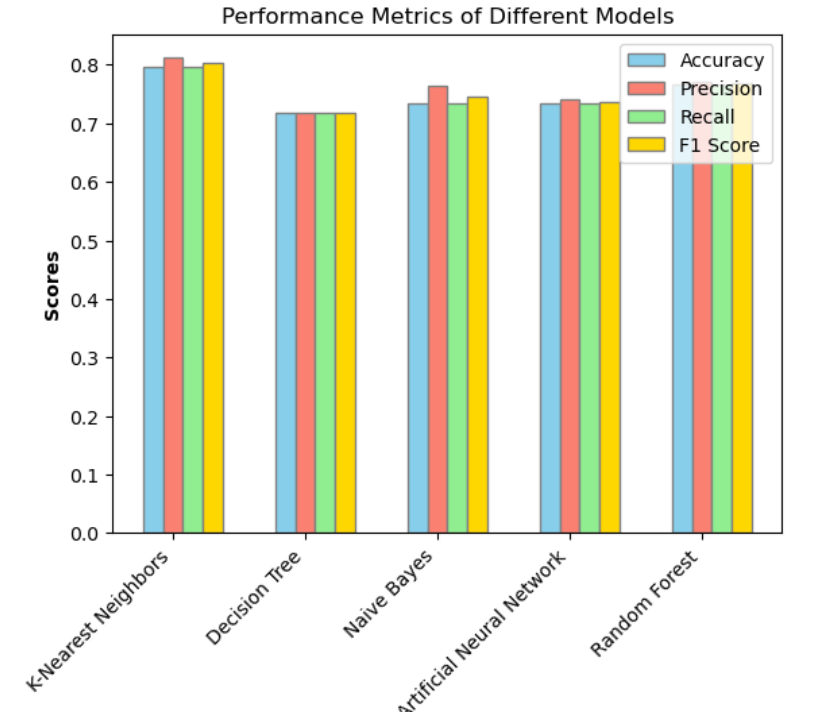
**Fig17: Comparison Result of precision**



**Fig:18 Comparison result of recall**



**Fig19: Comparison result of f1-score**



**Fig 20: Comparison of all performance measure**

**CHAPTER 6**

**CONCLUSION AND FUTURE SCOPE**

A significant portion of the population has diabetes. One of the main concerns is accurately and quickly predicting diabetes. Comparing the different machine learning algorithms' capacity to predict diabetes is the main objective of study. The methods that have been used include Naive Bayes, ANN, KNN, and Decision Trees. The outcomes of KNN are assessed based on f-measure, recall, and precision. Based on the outcome, KNN is the most accurate predictor, and after that Random Forest perform better on the provided dataset than nave bayes, decision trees, and artificial neural networks.

In the future, additional diseases may be predicted or diagnosed using the system that was created and the machine learning classification algorithms that were employed. The approach can be expanded upon and enhanced to automate the examination of diabetes, using more machine learning algorithms.

**References:**

[1].Https://www.javatpoint.com/basic-concepts-in-machine-lcarning.g [2].Https://www.geeksforgeeks.org/machine-learning-with-python.

[3].Https://en.wikipedia.org/wiki/Machine learning.

[4]. B. Nithya and Dr. V. Ilango," Predictive Analytics in Health Care Using Machine Learning Tools and Techniques", International Conference on Intelligent Computing and Control Systems, 978-1-5386-2745-7,2017.

[5].https://www.simplilearn.com/tutorials/machine-learning-tutorial/what-is-machine-learning

[ 6]. <https://www.javatpoint.com/machine-learning-life-cycle>.

[7]. Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar," Predictive Analysis

of Diabetic Patient Data Using Machine Learning and Hadoop", International Conference on

[8]. Saravanan, Renuka, and Pothula Sujatha. "A state of art techniques on machine learning algorithms: a perspective of supervised learning approaches in data classification." In 2018 Second international conference on intelligent computing and control systems (ICICCS), pp. 945-949. IEEE, 2018.

[9]. S.Ruba Arockia Archana and Dr.M.S.Thanabal. "Optimization Algorithms for Feature Selection in Classification: A Survey", International Journal of Innovative Research in Computer and Communication Engineering, ISSN: 2320-9798, vol. 4, Issue 2, pp. 1124-1127, February 2016.

[10]. Https://www.cdc.gov/diabetes/basics/diabetes.html.

[11]. Gupta, S., Verma, H. K., & Bhardwaj, D. (2021). Classification of diabetes using Naive Bayes and support vector machine as a technique. In Operations Management and Systems Engineering: Select Proceedings of CPIE 2019 (pp. 365-376). Springer Singapore.

[12]. Tafa, Z., Pervetica, N., & Karahoda, B. (2015, June). An intelligent system for diabetes prediction. In 2015 4th Mediterranean Conference on Embedded Computing (MECO) (pp. 378-382). IEEE.

[13]. Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia computer science, 132, 1578- 1585.

[14]. Kumari, S., Kumar, D., & Mittal, M. (2021). An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. International Journal of Cognitive Computing in Engineering, 2, 40-46.

[15]. Gupta, S., Verma, H. K., & Bhardwaj, D. (2021). Classification of diabetes using Naive Bayes and support vector machine as a technique. In Operations Management and Systems Engineering: Select Proceedings of CPIE 2019 (pp. 365-376). Springer Singapore.

[16]. Gupta, S., Verma, H. K., & Bhardwaj, D. (2021). Classification of diabetes using Naive Bayes and support vector machine as a technique. In Operations Management and Systems Engineering: Select Proceedings of CPIE 2019 (pp. 365-376). Springer Singapore.

[17].Tigga.N P. , & Garg, S. (2020). Prediction of type 2 diabetes using machine learning classification methods. Procedia Computer Science, 167, 706-716.

[18]. Diwani, S. A., & Sam, A. (2014). Diabetes forecasting using supervised learning techniques. Adv Computer Science an Int J, 3, 10-18.

[19]. Ahuja, R., Sharma, S. C., & Ali, M. (2019). A diabetic disease prediction model based on classification algorithms. Annals of Emerging Technologies in Computing (AETiC), Print ISSN, 2516-0281.

[20]. Malik, S., Harous, S., & El-Sayed, H. (2020, September). Comparative analysis of machine learning algorithms for early prediction of diabetes mellitus in women. In *International Symposium on Modelling and Implementation of Complex Systems* (pp. 95-106). Cham: Springer International Publishing.

[21]. Soltani, Z., & Jafarian, A. (2016). A new artificial neural networks approach for diagnosing diabetes disease type II. *International Journal of Advanced Computer Science and Applications*, *7*(6).

[22]. Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access*, *8*, 76516-76531.

[23]. Alehegn, M., Joshi, R., & Alehegn, M. (2017). Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach. *International Research Journal of Engineering and Technology*, *4*(10), 426-436.

[24]. Rathore, A., Chauhan, S., & Gujral, S. (2017). Detecting and Predicting Diabetes Using Supervised Learning: An Approach towards Better Healthcare for Women. *International Journal of Advanced Research in Computer Science*, *8*(5).

[25]. Shetty, D., Rit, K., Shaikh, S., & Patil, N. (2017, March). Diabetes disease prediction using data mining. In *2017 international conference on innovations in information, embedded and communication systems (ICIIECS)* (pp. 1-5). IEEE.

[26]. Nai-Arun, N., & Moungmai, R. (2015). Comparison of classifiers for the risk of diabetes prediction. *Procedia Computer Science*, *69*, 132-142.

[27]. GMRIT, R. (2018). Type 2 Diabetes Classification and Prediction Using Risk Score. *International Journal of Pure and Applied Mathematics*, *119*(15), 1099-1111.

[28]. Pranto, B., Mehnaz, S. M., Mahid, E. B., Sadman, I. M., Rahman, A., & Momen, S. (2020). Evaluating machine learning methods for predicting diabetes among female patients in Bangladesh. *Information*, *11*(8), 374.

[29]. Olisah, C. C., Smith, L., & Smith, M. (2022). Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. *Computer Methods and Programs in Biomedicine*, *220*, 106773.

[30]. Ramesh, J., Aburukba, R., & Sagahyroon, A. (2021). A remote healthcare monitoring framework for diabetes prediction using machine learning. *Healthcare Technology Letters*, *8*(3), 45-57.

[31]. Singh, D. A. A. G., E. Jebamalar Leavline, and B. Shanawaz Baig. "Diabetes prediction using medical data." *Journal of Computational Intelligence in Bioinformatics* 10.1 (2017): 1-8.

[32]. Nnamoko, N., Hussain, A., & England, D. (2018, July). Predicting diabetes onset: an ensemble supervised learning approach. In *2018 IEEE Congress on evolutionary computation (CEC)* (pp. 1-7). IEEE.

[33]. Joshi TN. Prof. Pramila M. Chawan," Diabetes Prediction Using Machine Learning Techniques". Int. Journal of Engineering Research and Application. 2018;8(1):2278-0181.

[34]. Nai-Arun, N., & Moungmai, R. (2015). Comparison of classifiers for the risk of diabetes prediction. *Procedia Computer Science*, *69*, 132-142.

[35]. Branimir et al. proposed a system to overcome the following two challenges: heterogeneity regarding previous techniques and lack of transparency in features. It used the PRISMA methodology. 18 varieties of model comparison along with algorithms based on trees were performed. The authors concluded that KNN and SVM are majorly used for prediction.

[36]. Chaki, Jyotismita, et al. "Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review." *Journal of King Saud University-Computer and Information Sciences* 34.6 (2022): 3204-3225.

[37]. A. Nur Ghaniaviyanto Ramadhan and R. Ade, "Preprocessing handling to enhance detection of type 2 diabetes mellitus based on random forest," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 12, 2021.

[38]. Advanced Computing and Intelligent Technologies, 2022, Volume 218 N ISBN : 978-981-16-2163-Prajyot Palimkar, Rabindra Nath Shaw, Ankush Ghosh

[39]. Chou, C. Y., Hsu, D. Y., & Chou, C. H. (2023). Predicting the onset of diabetes with machine learning methods. *Journal of Personalized Medicine*, *13*(3), 406.

[40]. <https://www.geeksforgeeks.org/introduction-to-python/>.

[41] Simou, Chen. (2023). Multiple Machine Learning Algorithms for Spam Mail Detection.Highlights in Science, Engineering and Technology, doi: 10.54097/hset.v39i.6698.

[42] Chaudhary, K., Yadav, J., & Mallick, B. (2012). A review of fraud detection techniques: Credit card. *International Journal of Computer Applications*, *45*(1), 39-44.

[43]. <https://www.saedsayad.com/naive_bayesian.htm>

[44] Redjeki, S., Guntara, M., & Anggoro, P. (2015). Naive bayes classifier algorithm approach for mapping poor families potential. *Int. J. Adv. Res. Artif. Intell*, *4*(12), 29-33.

[45]Meunier, Laurent, et al. A dynamical system perspective for Lipschitz neural networks. International

[46] Q. Yu et al. Semantic segmentation of intracranial hemorrhages in head CT scans. In 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS) (pp. 112-115). IEEE.2019.

[47]Guo, Jianyuan, et al. Cmt: Convolutional neural networks meet vision transformers. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[48] IBM Cloud Education. Neural Networks. Available at https://www.ibm.com/cloud/learn/neural-networks,2020.  
[49] Tigga, N. P., & Garg, S. (2020). Prediction of type 2 diabetes using machine learning classification methods. *Procedia Computer Science*, *167*, 706-716.

[50] Kandhasamy, J. P., & Balamurali, S. J. P. C. S. (2015). Performance analysis of classifier models to predict diabetes mellitus. *Procedia Computer Science*, *47*, 45-51.

[51] Alehegn, M., Joshi, R., & Alehegn, M. (2017). Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach. *International Research Journal of Engineering and Technology*, *4*(10), 426-436.

[52] Saravananathan, K., & Velmurugan, T. (2016). Analyzing Diabetic Data using Classification Algorithms in Data Mining. Indian Journal of Science and Technology, 9(43). [19] Ramzan, M. (2016, August).

|  |
| --- |
|  |

|  |
| --- |
|  |
|  |
|  |
|  |  |
|  |  |