ABSTRACT

This study explores the prediction of diabetes using various Machine Learning (ML) algorithms. The project aims to evaluate the performance of these algorithms and its equivalent model in predicting diabetes and to identify the most effective one. The ML algorithms considered in the study include Support Vector Machine, Logistic Regression, K-nearest Neighbor, Naïve Bayes, Random Forest, and Decision Tree. The Diabetes dataset, which contains 768 instances with eight features, is used to train and test the models. Using the above mentioned algorithms for predicting diabetes, the degree of accuracy of the algorithms are also determined. These accuracy rates define the probability of having surety in the diabetes prediction. The most accurate one is considered the most potent one.

TABLE OF CONTENT

CHAPTER 1

INTRODUCTION

1.1 Introduction

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Insulin is a hormone that regulates blood glucose. Hyperglycaemia, also called raised blood glucose or raised blood sugar, is a common effect of uncontrolled diabetes and over time leads to serious damage to many of the body's systems, especially the nerves and blood vessels.

In 2014, 8.5% of adults aged 18 years and older had diabetes. In 2019, diabetes was the direct cause of 1.5 million deaths and 48% of all deaths due to diabetes occurred before the age of 70 years. Another 460,000 kidney disease deaths were caused by diabetes, and raised blood glucose causes around 20% of cardiovascular deaths *(1)*. Between 2000 and 2019, there was a 3% increase in age-standardized mortality rates from diabetes.

1.2 Problem Statement

Diabetes is one of the serious diseases that needs to be cared for in the early stage. Patients might have faced many problems while visiting the hospitals for check-ups like lack of professional doctors. Some hospitals might be using a Decision system but they are not sufficient. Decisions are made on the basis of the doctor rather than information that is stored in the system database. This disease is one of the life changing diseases thus needs to be taken care of very carefully. Hence, to predict the disease and its accuracy easily by using various algorithms, this system might be helpful.

1.3 Objectives

- To predict probabilities for a patient about the chances of having diabetes.
- To provide accuracy rates of all the algorithms used.

1.4 Scope and Limitations

1.4.1 Scope

The implementation of the decision support system is a computer-based system that helps to extract patient’s records. This system is able to predict the chances of having diabetes with some levels of surety. Thus, many algorithms are used to check the records of the patients.

1.4.2 Limitations

- The system cannot refer the patient to a particular doctor as well as hospital.
- The result of this system needs to be shown to a doctor for confirmation.
- The data should be in appropriate format.

CHAPTER 2

LITERATURE REVIEW AND RESEARCH METHODOLOGY

2.1 Literature Review

The developed system called Decision Support in Diabetes Prediction System uses various data mining modeling techniques to discover the relationship between variables in data in the healthcare industry. It can serve as a training tool to train nurses and medical students to diagnose patients with sugar.

The Neural network is splitted into two halves: training data and testing data. There are 8 parameters to be taken care of which are pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree and age. The system identifies the unknown data from comparisons with the trained data, whenever unknown data is fed by the doctors and it finally generates the outcome.

The project is also called Intelligent System that uses data mining techniques namely Support Vector Machine, Logistic Regression, Naive Bayesian algorithm, KN neighbors algorithm, Decision Tree algorithm, and Random Forest algorithm. These algorithms are web-based applications which compare the user values with the trained data set. It thus assists healthcare providers to make intelligent clinical decisions.

The paper has proposed a weighted fuzzy rule based Clinical Decision Support System for the diagnosis of sugar. It obtains the knowledge from the patient's clinical data. A computerized approach for generation of weighted fuzzy rules is developed and then a fuzzy rule based decision support system is generated to accurately determine the chances of having diabetes.

2.2 Framework of the model

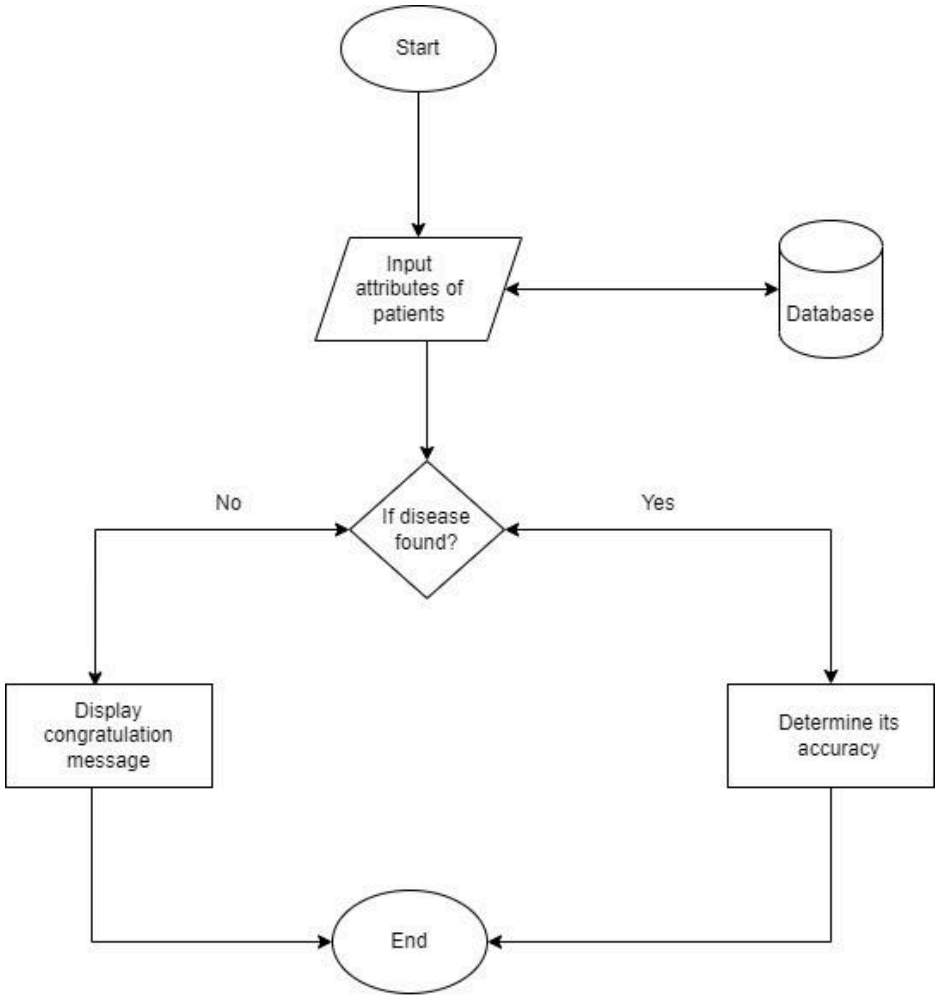
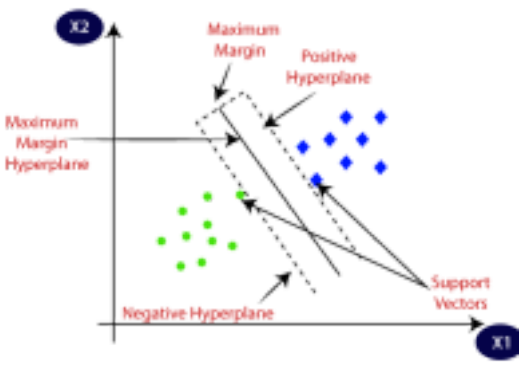


Fig 1: Framework of diabetes prediction system

2.3 Methodology

Following algorithms were used to calculate the probability of having the disease.

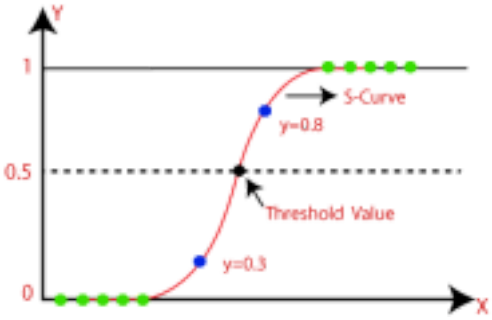
a. Description of Support Vector Machine:



Support Vector Machine (SVM) is a supervised machine learning algorithm that is used for classification and regression analysis. SVM constructs a hyperplane or a set of hyperplanes in a high-dimensional space that can be used for classification or regression. The algorithm aims to maximize the margin between the hyperplane and the nearest data points from each class. The SVM algorithm

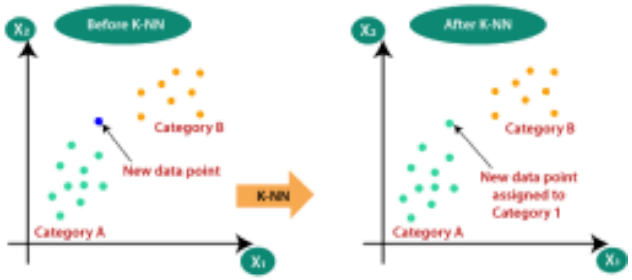
can handle linearly separable and non-linearly separable datasets by using different types of kernels such as linear, polynomial, radial basis function, and sigmoid kernels. SVM has been widely used in various fields including image classification, text classification, bioinformatics, and finance.

b. Description of Logistic Regression:



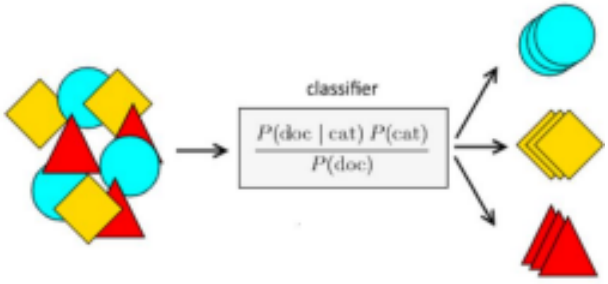
Logistic Regression is a statistical method used for binary classification problems, where the goal is to predict the probability of an event occurring or not occurring. It is a type of regression analysis where the dependent variable is binary, meaning it can only take on two possible values. The logistic regression model uses a logistic function to estimate the probability of the event occurring based on one or more independent variables. The model calculates the odds ratio, which is the ratio of the odds of the event occurring in one group compared to the odds of the event occurring in another group. The logistic regression model is widely used in various fields, including healthcare, marketing, and social sciences, among others, to predict the probability of an outcome based on a set of predictor variables.

c. Description of K-nearest neighbor:



K-nearest neighbor (KNN) is a supervised machine learning algorithm used for classification and regression tasks. It is a non-parametric and lazy learning algorithm, which means that it doesn't make any assumptions about the underlying data distribution and it doesn't require any training to make predictions. Instead, KNN uses the distance between the input data and the labeled data points in the training dataset to make predictions. The KNN algorithm calculates the distance between the input data and all the labeled data points in the training dataset and selects the K nearest neighbors to the input data point. The output of the KNN algorithm is the mode (for classification) or the mean (for regression) of the K-nearest neighbors labels. The value of K is a hyperparameter that needs to be tuned based on the data and the problem at hand.

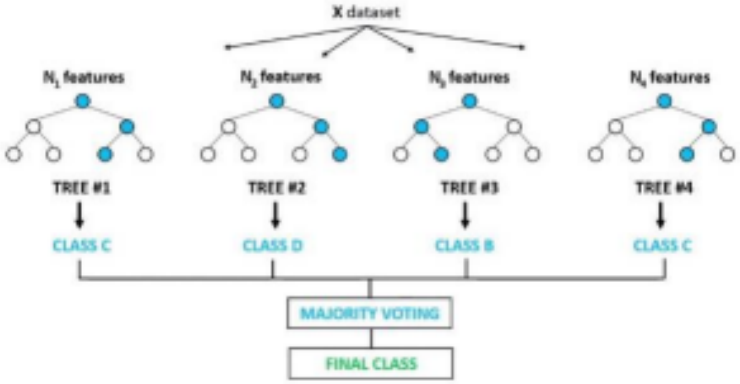
d. Description of Naïve Bayes:



Naïve Bayes is a popular probabilistic algorithm used for classification problems in machine learning. It is based on Bayes' theorem, which describes the probability of a hypothesis given evidence. Naïve Bayes assumes that the features are conditionally independent of each other, which means that the presence or absence of a particular feature does not affect the likelihood of the presence or absence of other features. This makes the algorithm fast and efficient, especially for large datasets. Naïve Bayes works by calculating the probability of each class given the input features and selecting the class with the highest probability as the output.

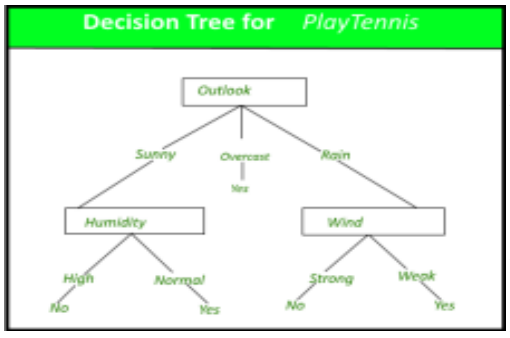
e. Description of Random Forest:

Random Forest is a machine learning algorithm that can be used for both classification and regression tasks. It is an ensemble method that combines multiple decision trees to make more accurate predictions. The algorithm creates a set of decision trees on randomly selected subsets of the training data, and each tree votes for the final prediction. Random Forest helps to reduce overfitting as compared to a single decision tree, as it uses a combination of different trees with different random subsets of features and data points. It also provides feature importance, which can help in feature selection and interpretability of the model.



f. Description of Decision Tree:

A Decision Tree is a popular supervised learning algorithm in machine learning used for classification and regression tasks. It is a decision support tool that uses a tree-like model of decisions and their possible consequences. The algorithm builds a tree-like model based on the training data by making a series of decisions to split the data into subsets, with the goal of maximizing the information gain at each split. Each internal node of the tree represents a decision based on a feature value, while each leaf node represents a class label or a regression value. Decision Trees are easy to interpret, fast to train and can handle both categorical and continuous features.



CHAPTER 3

SYSTEM DESIGN

3.1 Requirement Collection

The system requirement specification of the project consists of functional and nonfunctional requirements.

3.1.1 Functional Requirements

These are statements of services that the system should provide. First, the user must provide the attributes and symptoms the patients have faced then the system predicts whether the disease is present or not. Further process is explained by the use case diagram.The user interface is simple to use.

a. Input data: The user can input different symptoms and attributes.
b. View result: The user can view the result as per the prediction done by the system.

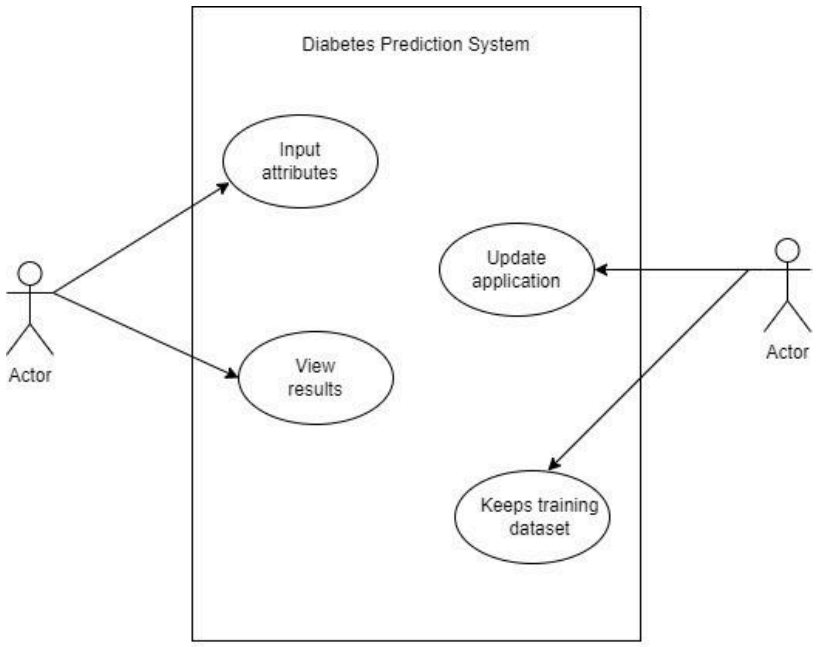


Fig 2: Use case diagram

3.1.2 Non-Functional Requirements

Some of the non-functional requirements of Diabetes Prediction System are summarized as following:

- Availability: The system can be accessed by the authorized person and anywhere having a PC.
- Maintainability: The system is highly maintainable.
- Scalability: The system can be enhanced for other diseases in future.

- Reliability: The reliability of the system is defined by the overall accuracy of the system.
- Ease of Use: The system has a user friendly interface so that any user can use the system without facing any difficulties.

3.2 System Design

3.2.1 Process Design

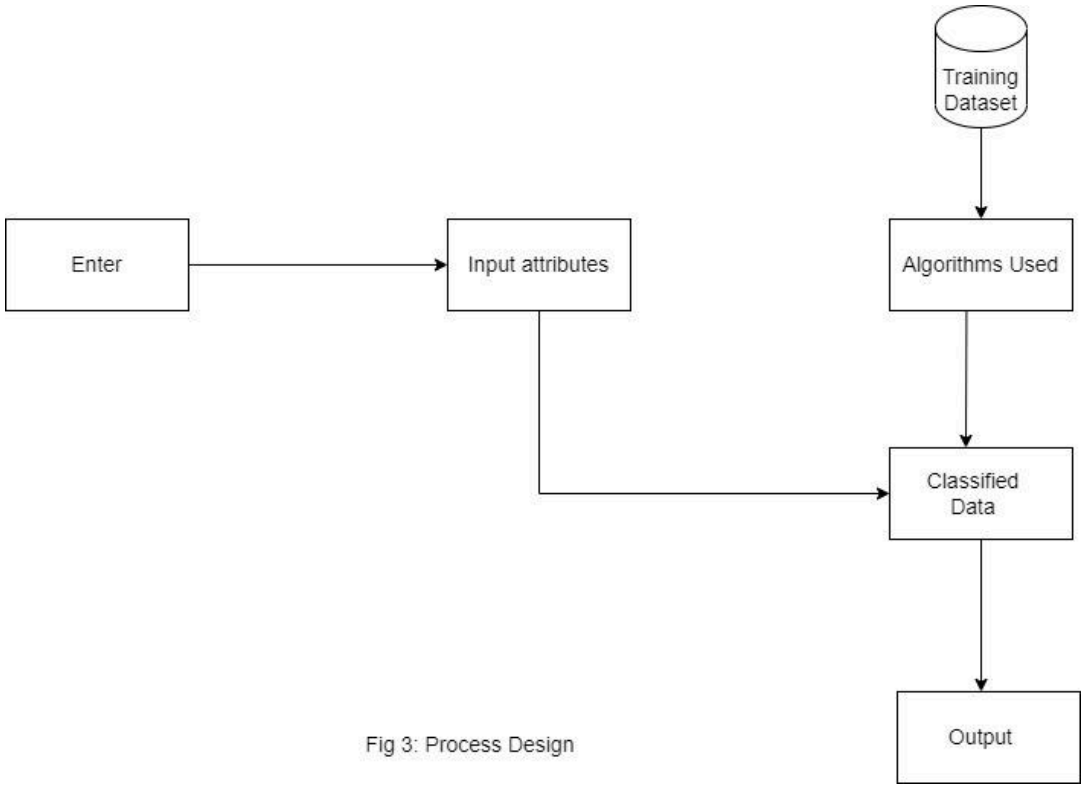


Fig 3: Process Design

3.4 Structuring System Requirement

3.4.1 Process Modelling

- Level 0 DFD

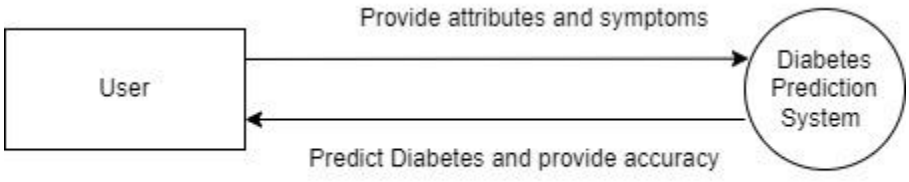


Fig 4: Level 0 DFD

CHAPTER 4

IMPLEMENTATION

Description of imported libraries and their implementation:

i. NumPy:

NumPy, Numerical Python, is a Python library used for working with arrays. It also has functions for working in the domain of linear algebra, Fourier transform, and matrices. NumPy arrays are faster and more convenient to work with than Python lists. NumPy arrays are a grid of values, all of the same type, and are indexed by a tuple of non-negative integers. NumPy provides many functions to create arrays, such as arrange, linspace, and random. It also provides a wide range of mathematical operations that can be performed on arrays, such as addition, subtraction, multiplication, and division. NumPy is widely used in scientific computing, data analysis, and machine learning.

Imported functions:

- numpy.asarray(*a*, *dtype=None*, *order=None*, *, *like=None*)
  - Convert the input to an array.

ii. Pandas:

Pandas is an open-source data analysis and manipulation library for Python. It provides easy-to use data structures and data analysis tools for handling structured data. The two primary data structures in Pandas are Series (1-dimensional) and DataFrame (2-dimensional). Pandas can load data from a variety of file formats, including CSV, Excel, SQL databases, and more. It also provides powerful tools for filtering, grouping, and transforming data, as well as handling missing or null values. With Pandas, it is easy to perform statistical analysis, merge and join datasets, and visualize data using built-in plotting functions. Pandas is widely used in data science and machine learning applications.

Imported functions:

- pandas.read_csv
  - Read a comma-separated values (csv) file into DataFrame.
  - Also supports optionally iterating or breaking of the file into chunks.

- pandas.DataFrame.head
  - Return the first *n* rows.
    - This function returns the first *n* rows for the object based on position. It is useful for

quickly testing if your object has the right type of data in it.

• pandas.DataFrame.value_counts
  - Return a Series containing counts of unique rows in the DataFrame.

• pandas.DataFrame.describe
  - Generate descriptive statistics.

• pandas.DataFrame.groupby
  - Group DataFrame using a mapper or by a Series of columns.

• pandas.DataFrame.mean
  - Return the mean of the values over the requested axis.

• pandas.DataFrame.drop
  - Drop specified labels from rows or columns.

• pandas.DataFrame.values
  - Return a Numpy representation of the DataFrame.

• pandas.DataFrame.shape
  - Return a tuple representing the dimensionality of the DataFrame.

iii. Scikit-learn:

Scikit-learn, also known as sklearn, is a popular machine learning library for Python. It provides tools for data mining, data analysis, and machine learning. Sklearn is built on top of other popular libraries such as NumPy, SciPy, and matplotlib. It provides a wide range of algorithms for various machine learning tasks such as classification, regression, clustering, and dimensionality reduction. Sklearn also includes various preprocessing and feature selection techniques to prepare data for machine learning models. It is widely used in academia and industry for building machine learning models and data analysis.

Imported functions:

• sklearn.preprocessing.StandardScaler
  - Standardize features by removing the mean and scaling to unit variance. •

sklearn.model_selection.train_test_split

- Split arrays or matrices into random train and test subsets.

• sklearn.metrics.accuracy_score

- Accuracy classification score.

• sklearn.svm.SVC

- C-Support Vector Classification.
- The implementation is based on libsvm.
- The fit time scales at least quadratically with the number of samples and may be impractical beyond tens of thousands of samples.

• sklearn.linear_model.LogisticRegression

- Logistic Regression (aka logit, MaxEnt) classifier.
- This class implements regularized logistic regression using the 'liblinear' library, 'newton-cg', 'sag', 'saga' and 'lbfgs' solvers.

• sklearn.neighbors.KNeighborsClassifier

- Classifier implementing the k-nearest neighbors vote.

• sklearn.naive_bayes.GaussianNB

- Gaussian Naive Bayes (GaussianNB).

• sklearn.tree.DecisionTreeClassifier

- A decision tree classifier.

• sklearn.ensemble.RandomForestClassifier

- A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

CHAPTER 5

RESULT AND ANALYSIS

The main objective of this project is to provide instant guidance about the current health condition of any patients. The system is a desktop-based application which takes the necessary inputs from the patients and then determines the probability of occurrence of diabetes. At first, the patient must identify the different attributes such as his or her high blood pressure, low blood pressure, pregnancy status,etc and on the basis of that attribute this system will predict the necessary results.

Support Vector Model, Logistic Regression, Naive Bayesian Classification Algorithm, KN Neighbors Algorithm, Decision Tree Algorithm and Random Forest Algorithm are used to identify the probability. All these algorithms have their own methods and techniques to predict the outcome. Thus, an accurate prediction can be made through this.

Furthermore, an accuracy score is also generated to determine which of these algorithms is the efficient one among all. By analyzing the accuracy of all the algorithms, it can be concluded that Logistic Regression has the highest accuracy (77.92%) thus, it is considered the most efficient one.

CHAPTER 6

CONCLUSION

In recent years, machine learning in medicine has gained in interest by the scientific and research community. Diabetes is a world's growing disease so it needs continuous self-management and control in order to prevent complications and prevent the occurrence of any life-threatening events.

In this project, the data (768) is collected from Kaggle. All the data are entered in the datasets and these datasets are trained and classified using many algorithms. When the user provides valid input, the system is able to predict the relatable disease with highest probability from the given data set.