Diabetes Prediction Using Machine Learning Algorithms

A PROJECT REPORT

Submitted to

Jawaharlal Nehru Technological University Kakinada, Kakinada

in partial fulfillment for the award of the degree of

Bachelor of Technology

in

COMPUTER SCIENCE AND ENGINEERING

Submitted by

S. HARISH (19KN1A05H2)

P. RAKESH (19KN1A05E4)

V. BHARGAV (19KN1A05I1)

Under the esteemed guidance of

Dr. D. SUNEETHA

Professor, CSE Department

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

NRI INSTITUTE OF TECHNOLOGY

Autonomous

(Approved by AICTE, Permanently Affiliated to JNTUK, Kakinada) Accredited by NBA (CSE, ECE & EEE), Accredited by NAAC with 'A' Grade ISO 9001: 2015 Certified Institution

Pothavarappadu (V), (Via) Nunna, Agiripalli (M), Krishna Dist, PIN: 521212, A.P, India.

2019-2023

Diabetes Prediction Using Machine Learning Algorithms

A PROJECT REPORT

Submitted to

Jawaharlal Nehru Technological University Kakinada, Kakinada

in partial fulfillment for the award of the degree of

Bachelor of Technology

in

COMPUTER SCIENCE AND ENGINEERING

Submitted by

S. HARISH (19KN1A05H2)

P. RAKESH (19KN1A05E4)

V. BHARGAV (19KN1A05I1)

Under the esteemed guidance of

Dr. D. SUNEETHA

Professor, CSE Department

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

NRI INSTITUTE OF TECHNOLOGY

Autonomous

(Approved by AICTE, Permanently Affiliated to JNTUK, Kakinada) Accredited by NBA (CSE, ECE & EEE), Accredited by NAAC with 'A' Grade ISO 9001: 2015 Certified Institution

Pothavarappadu (V), (Via) Nunna, Agiripalli (M), Krishna Dist, PIN: 521212, A.P, India.

2019-2023

NRI INSTITUTE OF TECHNOLOGY
(AUTONOMOUS)
Approved by AICTE, New Delhi: Permanently Affiliated to JNTUK, Kakinada
Accredited by NAAC with "A" GRADE, Accredited by NBA (CSE, ECE&EEE)
An ISO 9001:2015 Certified Institution
Pothavarappadu (V), Agiripalli (M), Eluru District, A.P., India, Pin: 521 212
URL: www.nriit.edu.in, email: principal@nriit.edu.in, Mobile: + 91 8333882444

Certificate

This is to certify that the Project entitled **"Diabetes Prediction Using Machine Learning Algorithms"** is a bonafide Work Carried out by **S.Harish(19KN1A05H2), P.Rakesh(19KN1A05E4) and V.Bhargav(19KN1A05I1)** in partial fulfillment for the award of degree of Bachelor of Technology in **Computer Science and Engineering** of **Jawaharlal Nehru Technological University Kakinada, Kakinada** during the year 2019-2023.

Project Guide
(DR.D.SUNEETHA)
Professor, Department of CSE

Head of the Department
(DR. D.SUNEETHA)

EXTERNAL EXAMINER

DECLARATION

We hereby declare that the project report titled "**Diabetes Prediction Using Machine Learning Algorithms"** is a bonafide work carried out in the Department of Computer Science and Engineering, **NRI Institute of Technology, Agiripalli,Vijayawada,** during the academic year 2019-2023, in partial fulfilment for the award of the degree of **Bachelor of Technology** by JNTU Kakinada.

We further declare that this dissertation has not been submitted elsewhere for any Degree.

19KN1A05H2

19KN1A05E4

19KN1A05I1

ACKNOWLEDGEMENT

We take this opportunity to thank all who have rendered their full support to our work. The pleasure, the achievement, the glory, the satisfaction, the reward, the appreciation and the construction of my project cannot be expressed with a few words for their valuable suggestions.

We are grateful to our Project Guide **Dr. D. Suneetha, Professor and Head of the C.S.E Department** for rendering the valuable suggestions and extended his support to complete the project successfully.

We are expressing our heartfelt thanks to **Head of the Department, Dr. D. Suneetha** garu for her continuous guidance for completion of our Project work.

We are extending our sincere thanks to **Dean of the Department, Dr. K. V. Sambasiva Rao** garu for his continuous guidance and support to complete our project successfully.

We are thankful to the **Principal, Dr. C. Naga Bhaskar** garu for his encouragement to complete the Project work.

We are extending our sincere and honest thanks to the **Chairman, Dr. R. Venkata Rao** garu & **Secretary, Sri K. Sridhar** garu for their continuous support in completing the Project work.

Finally, we thank the Administrative Officer, Staff Members, Faculty of Department of CSE, NRI Institute of Technology and our friends, directly or indirectly helped us in the completion of this project.

19KN1A05H2

19KN1A05E4

19KN1A05I1

ABSTRACT:

Diabetes Mellitus is among critical diseases and lots of people are suffering from this disease. Age, obesity, lack of exercise, hereditary diabetes, living style, bad diet, high blood pressure, etc. can cause Diabetes Mellitus. People having diabetes have high risk of diseases like heart disease, kidney disease, stroke, eye problem, nerve damage, etc. Current practice in hospital is to collect required information for diabetes diagnosis through various tests and appropriate treatment is provided based on diagnosis. Big Data Analytics plays an significant role in healthcare industries. Healthcare industries have large volume databases. Using big data analytics one can study huge datasets and find hidden information, hidden patterns to discover knowledge from the data and predict outcomes accordingly. In existing method, the classification and prediction accuracy is not so high. In this paper, we have proposed a diabetes prediction model for better classification of diabetes which includes few external factors responsible for diabetes along with regular factors like Glucose, BMI, Age, Insulin, etc. Classification accuracy is boosted with new dataset compared to existing dataset. Further with imposed a pipeline model for diabetes prediction intended towards improving the accuracy of classification.

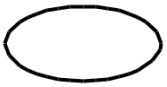
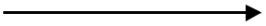
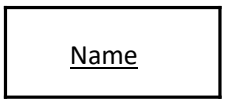
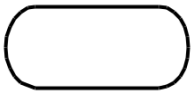
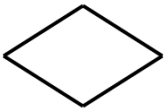
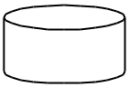
INDEX

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABBREVIATIONS	MEANING
GB	Gigabyte
GHz	Gigahertz
UML	Unified Modeling Language
JVM	Java Virtual Machine
API	Application Program Interface
GUI	Graphical User Interface
IDE	Integrated Development Environment
XML	eXtensible Markup Language
HTML	Hyper Text Markup Language
CSS	Cascading Style Sheets
DFD	Data Flow Diagram
CSV	Comma Separated Values
IDS	Intrusion Detection System
SVM	Support Vector Machine
LS-SVM IDS	Least Square SVM IDS
HFSA	Hybrid Feature Selection Algorithm
DoS	Denial of Service Attack
U2R	User to Root Attack
R2L	Root to Local Attack

LIST OF SYMBOLS

Symbols	Meaning
	Use-case
	Actor
	Transition
	Object
	Activities/States
	Decision Box
	Initial state of workflow
	Final State of work flow
	Dataset(DFD)

CHAPTER-1

INTRODUCTION

1. INTRODUCTION

Healthcare sectors have large volume databases. Such databases may contain structured, semi-structured or unstructured data. Big data analytics is the process which analyses huge data sets and reveals hidden information, hidden patterns to discover knowledge from the given data. Considering the current scenario, in developing countries like India, Diabetic Mellitus (DM) has become a very severe disease. Diabetic Mellitus (DM) is classified as Non-Communicable Disease (NCB) and many people are suffering from it. Around 425 million people suffer from diabetes according to 2017 statistics. Approximately 2-5 million patients every year lose their lives due to diabetes. It is said that by 2045 this will rise to 629 million.[1] Diabetes Mellitus (DM) is classified asType-1 known as Insulin-Dependent Diabetes Mellitus (IDDM). Inability of human's body to generate sufficient insulin is the reason behind this type of DM and hence it is required to inject insulin to a patient. Type-2 also known as Non-Insulin-Dependent Diabetes Mellitus (NIDDM). This type of Diabetes is seen when body cells are not able to use insulin properly.Type-3 Gestational Diabetes, increase in blood sugar level in pregnant woman where diabetes is not detected earlier results in this type of diabetes. DM has long term complications associated with it. Also, there are high risks of various health problems for a diabetic person. A technique called, Predictive Analysis, incorporates a variety of machine learning algorithms, data mining techniques and statistical methods that uses current and past data to find knowledge and predict future events. By applying predictive analysis on healthcare data, significant decisions can be taken and predictions can be made. Predictive analytics can be done using machine learning and regression technique. Predictive analytics aims at diagnosing the disease with best possible accuracy, enhancing patient care, optimizing resources along with improving clinical outcomes.[1] Machine learning is considered to be one of the most important artificial intelligence features supports development of computer systems having the ability to acquire knowledge from past experiences with no need of programming for every case. Machine learning is considered to be a dire need of today's situation in order to eliminate human efforts by supporting automation with minimum flaws. Existing method for diabetes detection is uses lab tests such as fasting blood glucose and oral glucose tolerance. However, this method is time consuming. This paper focuses on building predictive model using machine learning algorithms and data mining techniques for diabetes prediction. The paper is organized as followsSection II-gives literature review of the work done on diabetes prediction earlier and taxonomy.

CHAPTER-2

LITERATURE SURVEY

2. LITERATURE SURVEY

2.1 Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop

Authors: Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj

V. Dharwadkar

Abstract: Now days from health care industries large volume of data is generating. It is necessary to collect, store and process this data to discover knowledge from it and utilize it to take significant decisions. Diabetic Mellitus (DM) is from the Non Communicable Diseases (NCD), and lots of people are suffering from it. Now days, for developing countries such as India, DM has become a big health issue. The DM is one of the critical diseases which has long term complications associated with it and also follows with various health problems. With the help of technology, it is necessary to build a system that store and analyze the diabetic data and predict possible risks accordingly. Predictive analysis is a method that integrates various data mining techniques, machine learning algorithms and statistics that use current and past data sets to gain insight and predict future risks. In this work machine learning algorithm in Hadoop MapReduce environment are implemented for Pima Indian diabetes data set to find out missing values in it and to discover patterns from it. This work will be able to predict types of diabetes are widespread, related future risks and according to the risk level of patient the type of treatment can be provided.

2.2 Prediction of Diabetes Based on Personal Lifestyle Indicators

Authors: Ayush Anand and Divya Shakti **Abstract:** Diabetes Mellitus or Diabetes has been portrayed as worse than Cancer and HIV (Human Immunodeficiency Virus). It develops when there are high blood sugar levels over a prolonged period. Recently, it has been

quoted as a risk factor for developing Alzheimer, and a leading cause for blindness & kidney failure. Prevention of the disease is a hot topic for research in the healthcare community. Many techniques have been discovered to find the causes of diabetes and cure it. This research paper is a discussion on establishing a relationship between diabetes risk likely to be developed from a person's daily lifestyle activities such as his/her eating habits, sleeping habits, physical activity along with other indicators like BMI (Body Mass Index), waist circumference etc. Initially, a Chi-Squared Test of Independence was performed followed by application of the CART (Classification and Regression Trees) machine learning algorithm on the data and finally using Cross-Validation, the bias in the results was removed.

2.3 Predictive Analytics in Health Care Using Machine Learning Tools and Techniques

Authors: B. Nithya and Dr. V. Ilango

Abstract: When we have a huge data set on which we would like to perform predictive analysis or pattern recognition, machine learning is the way to go. Machine Learning (ML) is the fastest rising arena in computer science, and health informatics is of extreme challenge. The aim of Machine Learning is to develop algorithms which can learn and progress over time and can be used for predictions. Machine Learning practices are widely used in various fields and primarily health care industry has been benefitted a lot through machine learning prediction techniques. It offers a variety of alerting and risk management decision support tools, targeted at improving patients' safety and healthcare quality. With the need to reduce healthcare costs and the movement towards personalized healthcare, the healthcare industry faces challenges in the essential areas like, electronic record management, data integration, and computer aided diagnoses and disease predictions. Machine Learning offers a wide range of tools, techniques, and frameworks to address these challenges. This paper depicts the study on various prediction techniques and tools for Machine Learning in practice. A glimpse on the applications of Machine Learning in various domains are also discussed here by highlighting on its prominence role in health care industry.

2.4 Diagnosis of Diabetes Using Classification Mining Techniques

Authors: Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly

Abstract: Diabetes has affected over 246 million people worldwide with a majority of them being women. According to the WHO report, by 2025 this number is expected to rise to over 380 million. The disease has been named the fifth deadliest disease in the United States with no imminent cure in sight. With the rise of information technology and its continued advent into the medical and healthcare sector, the cases of diabetes as well as their symptoms are well documented. This paper aims at finding solutions to diagnose the disease by analyzing the patterns found in the data through classification analysis by employing Decision Tree and Naïve Bayes algorithms. The research hopes to propose a quicker and more efficient technique of diagnosing the disease, leading to timely treatment of the patients.

2.5 Software Environment

Python is a high-level, interpreted scripting language developed in the late 1980s by Guido van Rossum at the National Research Institute for Mathematics and Computer Science in the Netherlands. The initial version was published at the alt. Sources newsgroup in 1991, and version 1.0 was released in 1994.

Python 2.0 was released in 2000, and the 2.x versions were the prevalent releases until December 2008. At that time, the development team made the decision to release version 3.0, which contained a few relatively small but significant changes that were not backward compatible with the 2.x versions. Python 2 and 3 are very similar, and some features of Python 3 have been back ported to Python 2. But in general, they remain not quite compatible.

Both Python 2 and 3 have continued to be maintained and developed, with periodic release updates for both. As of this writing, the most recent versions available are 2.7.15 and 3.6.5.

However, an official End of Life date of January 1, 2020 has been established for Python 2, after which time it will no longer be maintained. If you are a newcomer to Python, it is recommended that you focus on Python 3, as this tutorial will do.

Python is still maintained by a core development team at the Institute, and Guido is still in charge, having been given the title of BDFL (Benevolent Dictator For Life) by the Python community. The name Python, by the way, derives not from the snake, but from the British comedy troupe Monty Python's Flying Circus, of which Guido was, and presumably still is, a fan. It is common to find references to Monty Python sketches and movies scattered throughout the Python documentation.

2.6 Why Choose Python

If you're going to write programs, there are literally dozens of commonly used languages to choose from. Why choose Python? Here are some of the features that make Python an appealing choice.

Python is Popular

Python has been growing in popularity over the last few years. The 2018 Stack Overflow Developer Survey ranked Python as the 7th most popular and the number one most wanted technology of the year. World-class software development countries around the globe use Python every single day.

According to research by Dice Python is also one of the hottest skills to have and the most popular programming language in the world based on the Popularity of Programming Language Index.

Due to the popularity and widespread use of Python as a programming language, Python developers are sought after and paid well. If you'd like to dig deeper into Python salary statistics and job opportunities, you can do so here.

Python is interpreted

Many languages are compiled, meaning the source code you create needs to be translated into machine code, the language of your computer's processor, before it can be run. Programs written in an interpreted language are passed straight to an interpreter that runs them directly.

This makes for a quicker development cycle because you just type in your code and run it, without the intermediate compilation step.

One potential downside to interpreted languages is execution speed. Programs that are compiled into the native language of the computer processor tend to run more quickly than interpreted programs. For some applications that are particularly computationally intensive, like graphics processing or intense number crunching, this can be limiting.

In practice, however, for most programs, the difference in execution speed is measured in milliseconds, or seconds at most, and not appreciably noticeable to a human user. The expediency of coding in an interpreted language is typically worth it for most applications.

Python is Free

The Python interpreter is developed under an OSI-approved open-source license, making it free to install, use, and distribute, even for commercial purposes.

A version of the interpreter is available for virtually any platform there is, including all flavors of Unix, Windows, macOS, smart phones and tablets, and probably anything else you ever heard of. A version even exists for the half dozen people remaining who use OS/2.

Python is Portable

Because Python code is interpreted and not compiled into native machine instructions, code written for one platform will work on any other platform that has the Python interpreter installed. (This is true of any interpreted language, not just Python.)

Python is Simple

As programming languages go, Python is relatively uncluttered, and the developers have deliberately kept it that way.

A rough estimate of the complexity of a language can be gleaned from the number of keywords or reserved words in the language. These are words that are reserved for special meaning by the compiler or interpreter because they designate specific built-in functionality of the language.

Python 3 has 33 keywords, and Python 2 has 31. By contrast, C++ has 62, Java has 53, and Visual Basic has more than 120, though these latter examples probably vary somewhat by implementation or dialect.

Python code has a simple and clean structure that is easy to learn and easy to read. In fact, as you will see, the language definition enforces code structure that is easy to read.

But It's Not That Simple For all its syntactical simplicity, Python supports most constructs that would be expected in a very high-level language, including complex dynamic data types, structured and functional programming, and object-oriented programming.

Additionally, a very extensive library of classes and functions is available that provides capability well beyond what is built into the language, such as database manipulation or GUI programming.

Python accomplishes what many programming languages don't: the language itself is simply designed, but it is very versatile in terms of what you can accomplish with it.

Conclusion

This section gave an overview of the **Python** programming language, including:

- A brief history of the development of Python
- Some reasons why you might select Python as your language of choice

Python is a great option, whether you are a beginning programmer looking to learn the basics, an experienced programmer designing a large application, or anywhere in between. The basics of Python are easily grasped, and yet its capabilities are vast. Proceed to the next section to learn how to acquire and install Python on your computer.

Python is an open source **programming** language that was made to be easy-to-read and powerful. A Dutch programmer named Guido van Rossum made Python in 1991. He named it after the television show Monty Python's Flying Circus. Many Python examples and tutorials include jokes from the show.

Python is an interpreted language. Interpreted languages do not need to be compiled to run. A program called an interpreter runs Python code on almost any kind of computer. This means that a programmer can change the code and quickly see the results. This also means Python is slower than a compiled language like C, because it is not running machine code directly.

Python is a good programming language for beginners. It is a high-level language, which means a programmer can focus on what to do instead of how to do it. Writing programs in Python takes less time than in some other languages.

Python drew inspiration from other programming languages like C, C++, Java, Perl, and Lisp.

Python has a very easy-to-read syntax. Some of Python's syntax comes from C, because that is the language that Python was written in. But Python uses whitespace to delimit code: spaces or tabs are used to organize code into groups. This is different from C. In C, there is a semicolon at the end of each line and curly braces ({}) are used to group code. Using whitespace to delimit code makes Python a very easy-to-read language.

Python use [change / change source]

Python is used by hundreds of thousands of programmers and is used in many places. Sometimes only Python code is used for a program, but most of the time it is used to do simple jobs while another programming language is used to do more complicated tasks.

Its standard library is made up of many functions that come with Python when it is installed. On the Internet there are many other libraries available that make it possible for the Python language to do more things. These libraries make it a powerful language; it can do many different things.

Some things that Python is often used for are:

- Web development
- Scientific programming
- Desktop GUIs
- Network programming
- Game programming

CHAPTER-3

SYSTEM ANALYSIS

3. SYSTEM ANALYSIS

3.1 EXISTING SYSTEM:

In existing method, the classification and prediction accuracy is not so high. Healthcare industries have large volume databases. Using big data analytics one can study huge datasets and find hidden information, hidden patterns to discover knowledge from the data and predict outcomes accordingly. Huge number of people is becoming its victim every day and many are unaware if they have it or not. There are two types of diabetes. Diabetes mellitus and diabetes. The test conducted to detect diabetes is physical examination and blood sugar test. More research is required to stop this disease. Machine learning and cloud computing will play a major role in the research related work to detect and timely cure it for the people. Diabetes specially affects the elderly and obese people. Diabetes can cause other variety of health problems like heart attack, kidney failure, high blood pressure and diabetic foot syndrome.

3.2 PROPOSED SYSTEM:

Diabetes can be controlled if it is predicted earlier. To achieve this goal this project work we will do early prediction of Diabetes in a human body or a patient for a higher accuracy through applying, Various Machine Learning Techniques. Machine learning techniques Provide better result for prediction by con- structing models from datasets collected from patients. In this work we will use Machine Learning Classification and ensemble techniques on a dataset to predict diabetes. Which are K-Nearest Neighbor (KNN), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Gradient Boosting (GB) and Random Forest (RF). The accuracy is different for every model when compared to other models. The Project work gives the accurate or higher accuracy model shows that the model is capa- ble of predicting diabetes effectively. Our Result shows that Random Forest achieved higher accuracy compared to other machine learning techniques.

CHAPTER-4

FEASIBILITY STUDY

4. FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

ECONOMICAL FEASIBILITY

TECHNICAL FEASIBILITY

SOCIAL FEASIBILITY

4.1 ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

4.2 TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

4.3 SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to

CHAPTER-5

SYSTEM REQUIREMENT SPECIFICATION

5. SYSTEM REQUIREMENTS

5.1 HARDWARE REQUIREMENTS:

- System : Pentium Dual Core.
- Hard Disk : 320 GB.
- Monitor : 15'' LED
- Input Devices : Keyboard, Mouse
- Ram : 8 GB

5.2 SOFTWARE REQUIREMENTS:

- Operating system : Windows 10
- Coding Language : python
- Tool : PyCharm
- Database : MYSQL
- Server : Flask

CHAPTER-6

SYSTEM DESIGN

6. SYSTEM DESIGN

6.1 SYSTEM ARCHITECTURE:

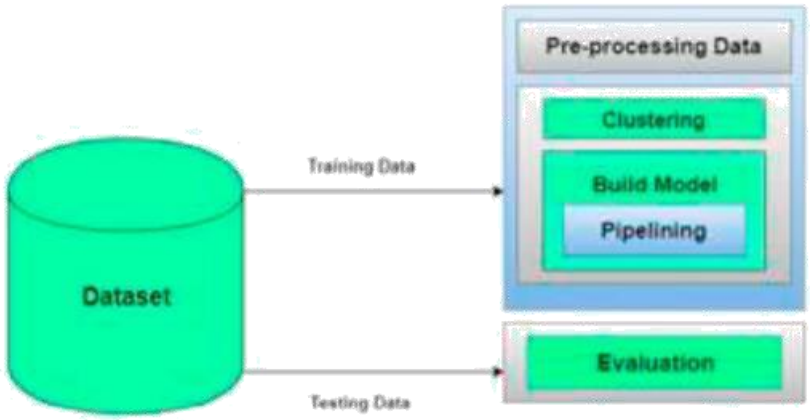


Fig 6.1: System Architecture

6.2 DATA FLOW DIAGRAM:

1. The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.
2. The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.
3. DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.
4. DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.

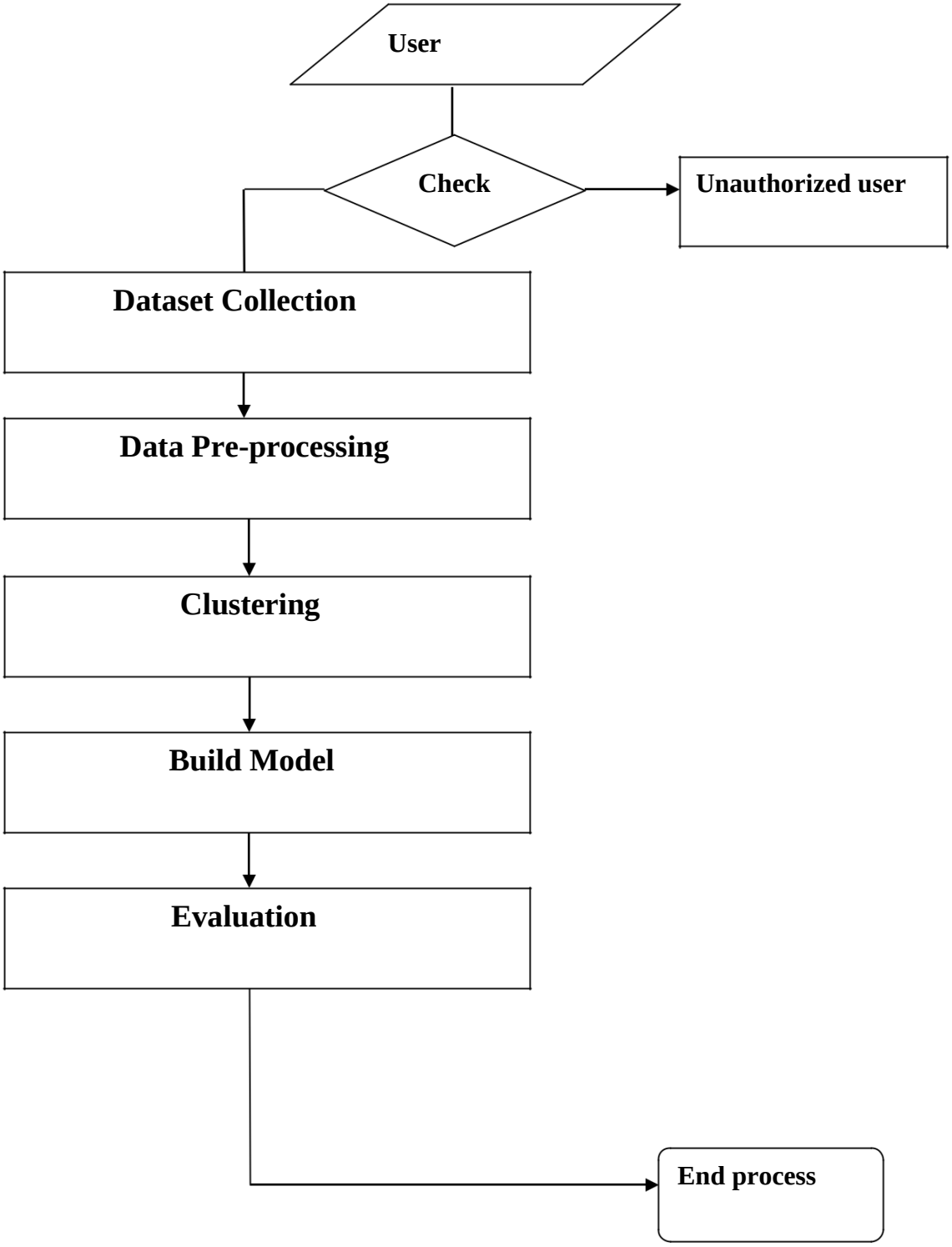


Fig 6.2: Data Flow Diagram

6.3 UML DIAGRAMS:

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

GOALS:

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modeling language.
5. Encourage the growth of OO tools market.
6. Integrate best practices.

USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can bedepicted.

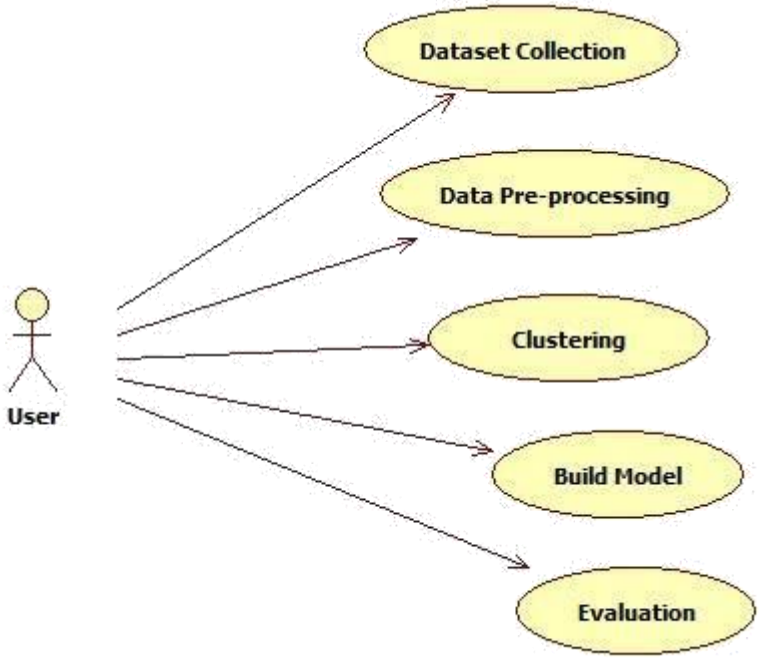


Fig 6.3: Use Case diagram

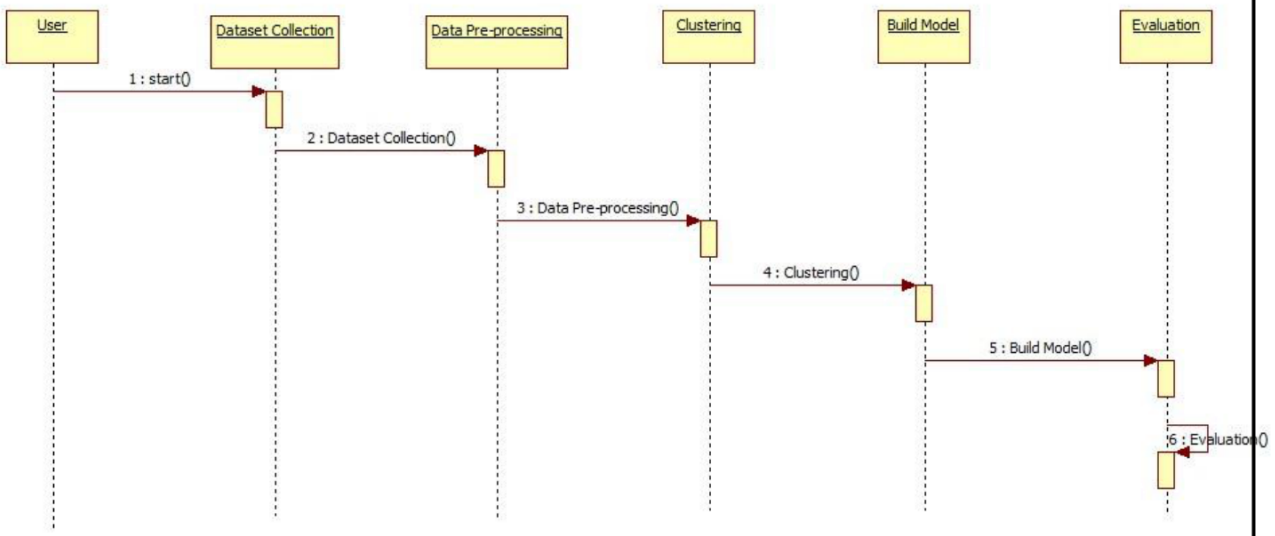


Fig 6.3:Use case Diagram

CLASS DIAGRAM:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

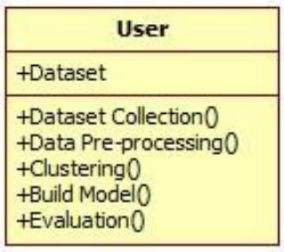


Fig 6.3: Class Diagram

SEQUENCE DIAGRAM:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

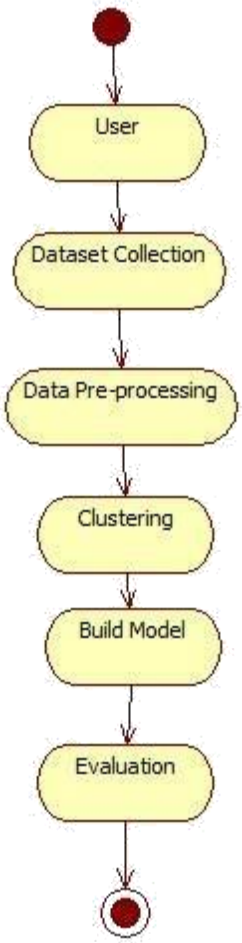


Fig 6.3: Activity Diagram

CHAPTER7

IMPLEMENTATION &CODING

7. IMPLEMENTATION

7.1 MODULES:

- Dataset Collection
- Data Pre-processing
- Clustering
- Build Model
- Evaluation

MODULES DESCRIPTION:

i. **Dataset Collection:** This module includes data collection and understanding the data to study the patterns and trends which helps in prediction and evaluating the results. Dataset description is given belowThis Diabetes dataset contains 800 records and 10 attributes

ii. **Data Pre-processing**: This phase of model handles inconsistent data in order to get more accurate and precise results. This dataset contains missing values. So we imputed missing values for few selected attributes like Glucose level, Blood Pressure, Skin Thickness, BMI and Age because these attributes cannot have values zero. Then we scale the dataset to normalize all values.

iii. **Clustering**: In this phase, we have implemented K-means clustering on the dataset to classify each patient into either a diabetic or non-diabetic class. Before performing K-means clustering, highly correlated attributes were found which were, Glucose and Age. K-means clustering was performed on these two attributes. After implementation of this clustering we got class labels (0 or 1) for each of our record.

iv. **Model Building:** This is most important phase which includes model building for prediction of diabetes. In this we have implemented various machine learning algorithms for diabetes prediction. These algorithms include Support Vector Classifier, Random Forest Classifier, Decision Tree Classifier, Extra Tree Classifier, Ada Boost algorithm, Perceptron, Linear Discriminant Analysis algorithm, Logistic Regression, K-Nearest Neighbour, Gaussian Naïve Bayes, Bagging algorithm,

7.2 SAMPLE CODE

```
# Importing essential libraries
from flask import Flask, render_template, request
import pickle
import pandas as pd
import numpy as np
from sklearn import preprocessing
from sklearn.model_selection import train_test_split
from sklearn. Ensemble
importRandomForestClassifier

# Load the Random Forest CLassifier model
filename = 'diabetes-prediction-rfc-model.pkl'
classifier = pickle.load(open(filename, 'rb'))

#Build the model using 5G-Smart Diabetes
 # SVM Algorithm
import pandas as pd
import numpy as np
from sklearn import preprocessing
from sklearn.model_selection import train_test_split
from sklearn import svm

# Load the dataset
data = pd.read_csv('diabetes_data.csv')
# Select the columns for training
cols = [col for col in data.columns if col not in
['Outcome']]
data = data[cols]

# Create the feature array
```

```
X = data.values

# Create the target array
y = diabetes_data.Outcome

# Normalize the feature array
X = preprocessing.scale(X)

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X,
y, test_size=0.2)

# Create a SVM classifier
clf = svm.SVC(kernel='linear')

# Train the model using the training sets
clf.fit(X_train, y_train)

# Predict the output for the test dataset
y_pred = clf.predict(X_test)

# Model Accuracy
accuracy = clf.score(X_test, y_test)
print("Accuracy:", accuracy)

 # ANN Algorithm
import pandas as pd
import numpy as np
from sklearn import preprocessing
from sklearn.model_selection import train_test_split
from sklearn.neural_network import MLPClassifier
```

```
        # Load the dataset
        data = pd.read_csv('diabetes_data.csv')
        # Select the columns for training
        cols = [col for col in data.columns if col not in
['Outcome']]
        data = data[cols]

        # Create the feature array
        X = data.values

        # Create the target array
        y = diabetes_data.Outcome

        # Normalize the feature array
        X = preprocessing.scale(X)

        # Split the data into training and testing sets
        X_train, X_test, y_train, y_test = train_test_split(X,
y, test_size=0.2)

        # Create an MLP Classifier
        clf = MLPClassifier()

        # Train the model using the training sets
        clf.fit(X_train, y_train)

        # Predict the output for the test dataset
        y_pred = clf.predict(X_test)

        # Model Accuracy
```

```
accuracy = clf.score(X_test, y_test)
print("Accuracy:", accuracy)

# Ensemble Algorithm
import pandas as pd
import numpy as np
from sklearn import preprocessing
from sklearn.model_selection import train_test_split
fromsklearn.ensembleimport
RandomForestClassifier

# Load the dataset
data = pd.read_csv('diabetes_data.csv')
# Select the columns for training
cols = [col for col in data.columns if col not in
['Outcome']]
data = data[cols]

# Create the feature array
X = data.values

# Create the target array
y = diabetes_data.Outcome

# Normalize the feature array
X = preprocessing.scale(X)

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X,
y, test_size=0.2)
```

```
# Create an Ensemble Classifier
clf = RandomForestClassifier()

# Train the model using the training sets
clf.fit(X_train, y_train)

# Predict the output for the test dataset
y_pred = clf.predict(X_test)

# Model Accuracy
accuracy = clf.score(X_test, y_test)
print("Accuracy:", accuracy)
app = Flask(__name__)

@app.route('/')
def home():
        return render_template('index.html')

@app.route('/predict', methods=['POST'])
def predict():
   if request.method == 'POST':
      preg = int(request.form['pregnancies'])
      glucose = int(request.form['glucose'])
      bp = int(request.form['bloodpressure'])
      st = int(request.form['skinthickness'])
      insulin = int(request.form['insulin'])
      bmi = float(request.form['bmi'])
      dpf = float(request.form['dpf'])
      age = int(request.form['age'])

      data = np.array([[preg, glucose, bp, st, insulin,
```

```
bmi, dpf, age]])
        my_prediction = classifier.predict(data)

        return render_template('result.html',
prediction=my_prediction)

    if __name__ == '__main__':
        app.run(debug=True)
```

CHAPTER-8

SYSTEM TESTING

8. SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

TYPES OF TESTS

Unit testing:

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing:

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test:

Functional tests provide systematic demonstrations that functions tested are available as

specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test:

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

White Box Testing:

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

Black Box Testing:

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

8.1 Unit Testing:

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach:

Field testing will be performed manually and functional tests will be written in detail.

Test objectives:

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

8.2 Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

8.3 Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

CHAPTER-9

INPUT &OUTPUT DESIGN

9. INPUT DESIGN AND OUTPUT DESIGN

9.1 INPUT DESIGN:

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

OBJECTIVES:

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2.It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

9.2 OUTPUT DESIGN:

A quality output is one, which meets the requirements of the end user and presents the

information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2. Select methods for presenting information.

3. Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

CHAPTER-10

SCREEN SHOTS

10. SCREENSHOTS

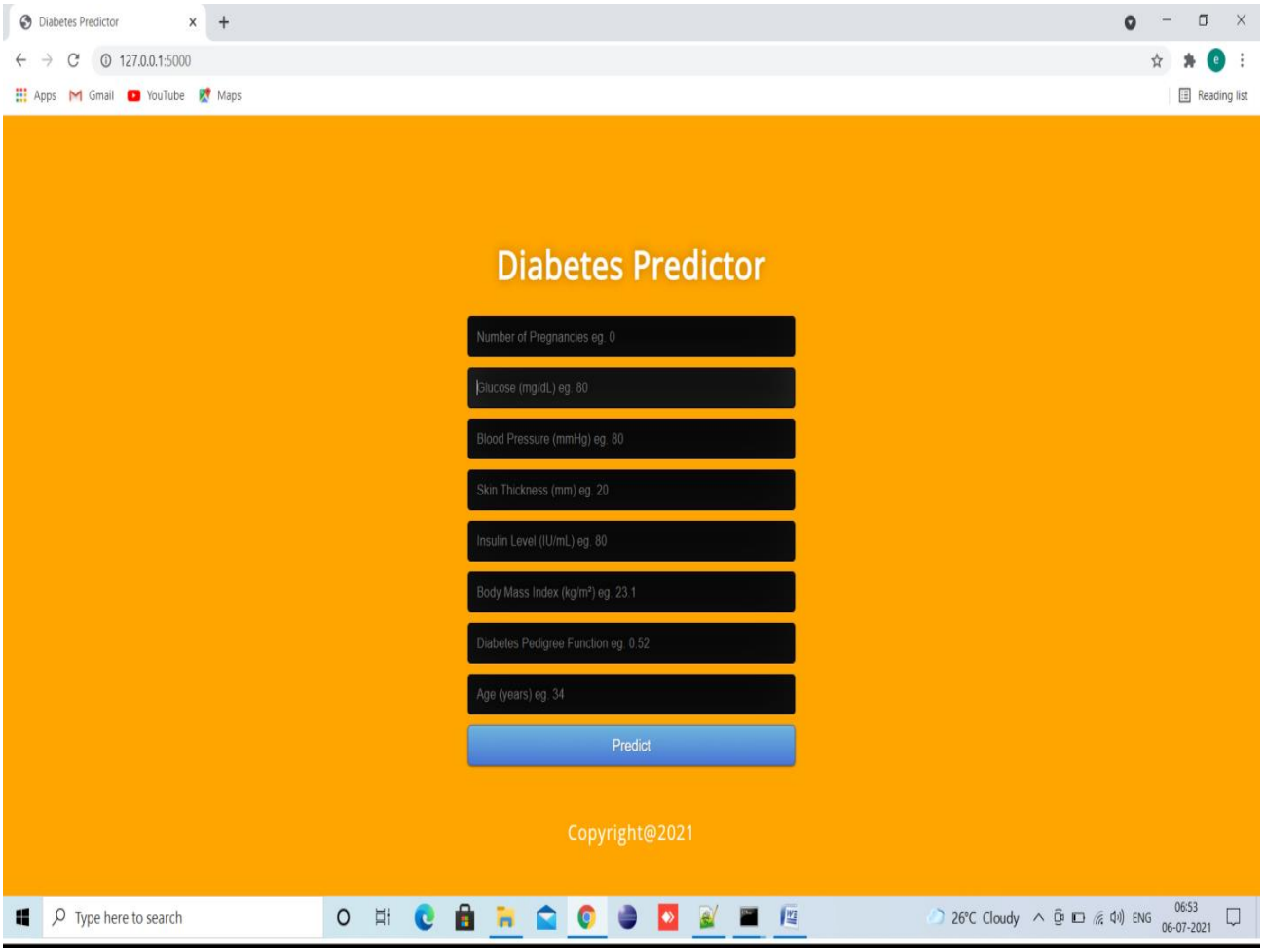
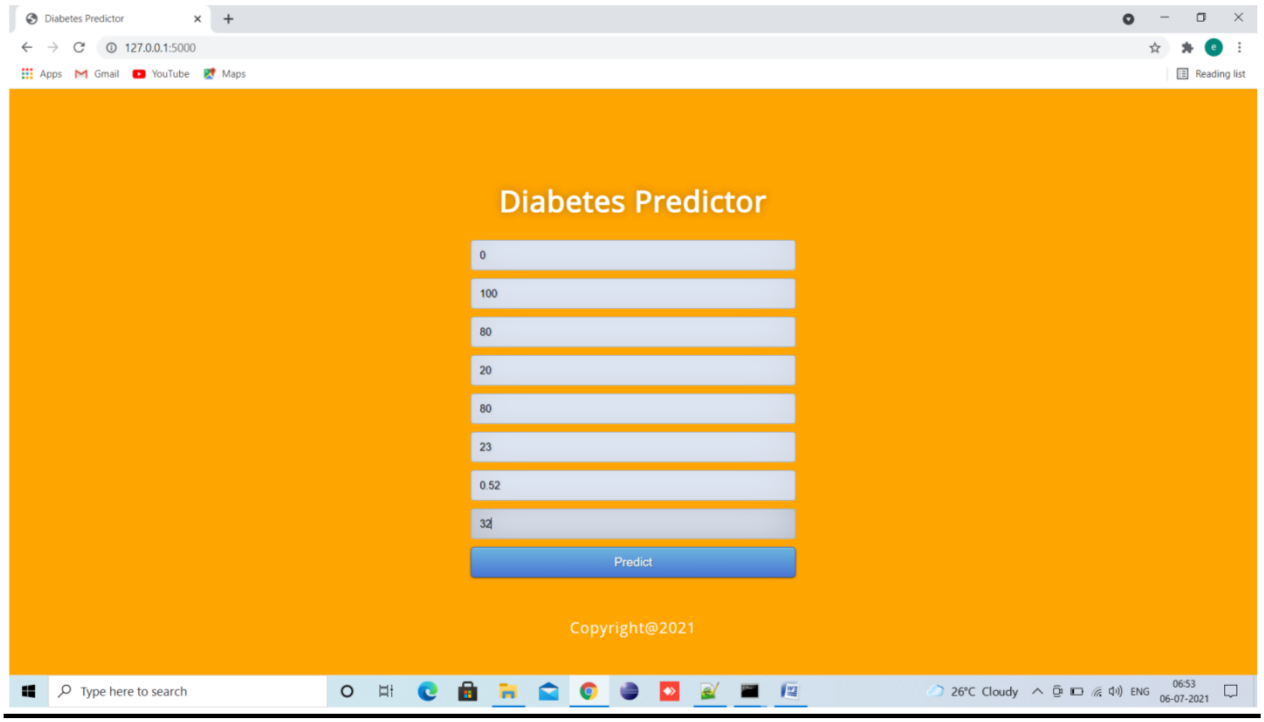
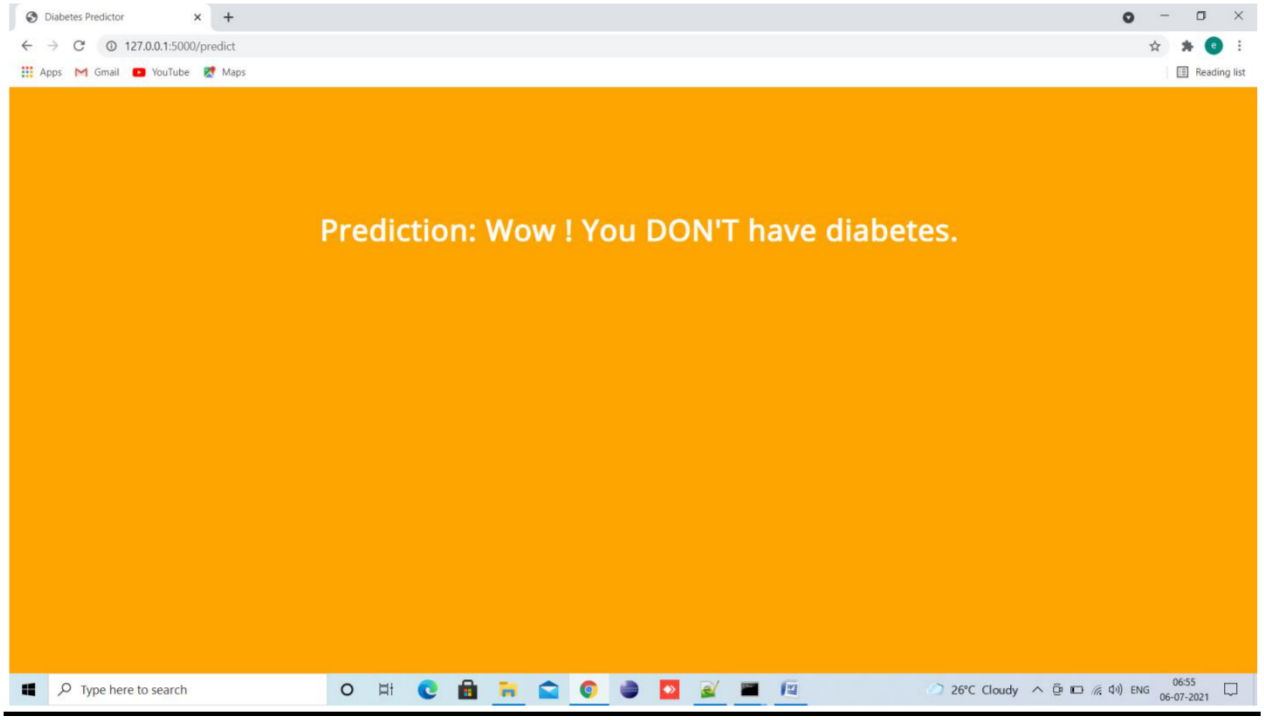


Fig 10:Diabetes Predictor


Diabetes Predictor
127.0.0.1:5000
Apps
Gmail
YouTube
Maps
Reading list
Diabetes Predictor
0
100
80
20
80
23
0.52
32
Predict
Copyright@2021
Type here to search
26°C Cloudy
ENG
06:53
06-07-2021


Diabetes Predictor
127.0.0.1:5000/predict
Apps
Gmail
YouTube
Maps
Reading list
Prediction: Wow ! You DON'T have diabetes.
Type here to search
26°C Cloudy
ENG
06:55
06-07-2021

CHAPTER-11

FUTURE ENHANCEMENT

11. FUTURE ENHANCEMENT

Our future work will focus on integration of other methods into the used model for tuning the parameters of models for better accuracy. Then testing these models with large dataset having minimum or no missing attribute values will reveal more insights and better prediction accuracy.

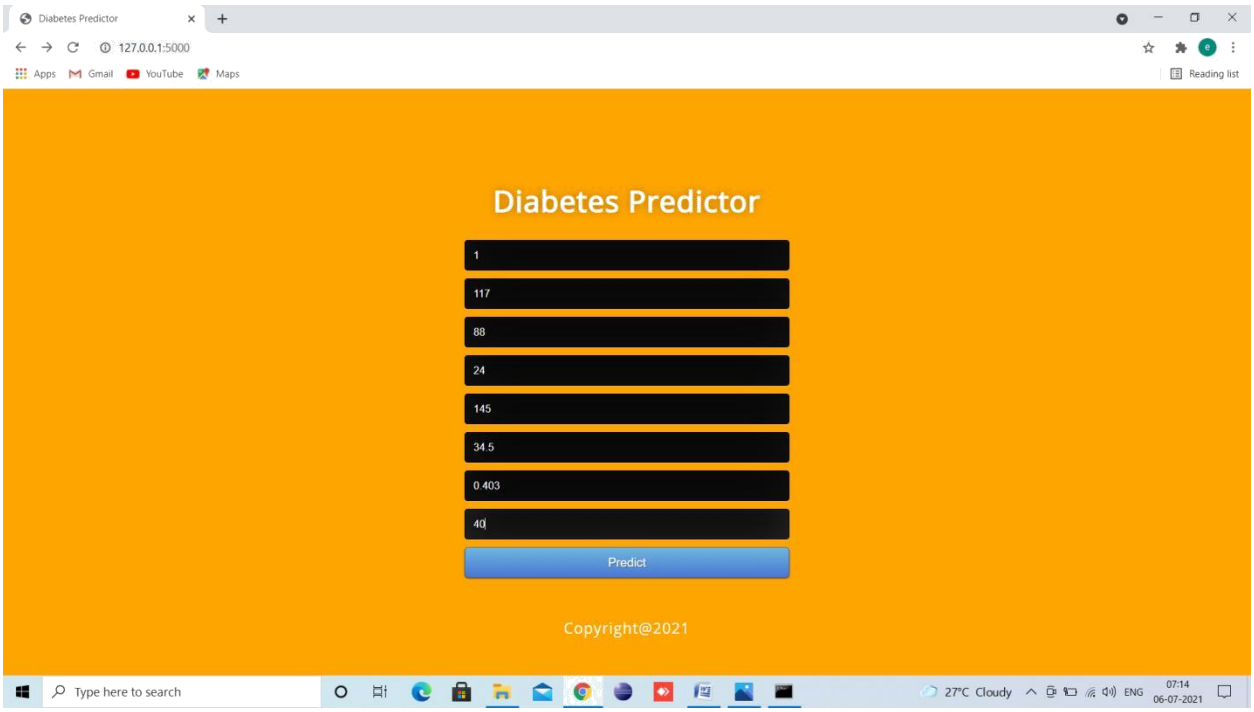


Fig 11: Diabetes Future Enhancement

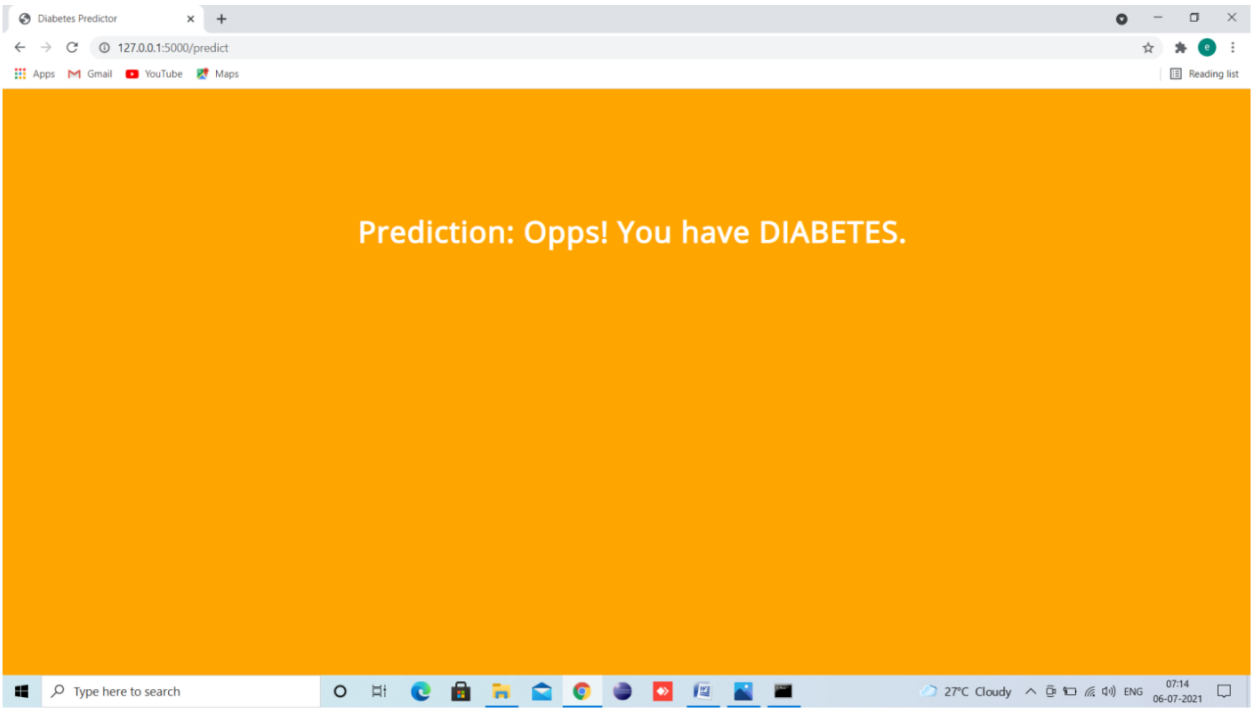


Fig 11.1 Prediction of Diabetes

CHAPTER-12

CONCLUSION

12. CONCLUSION

In this study, various machine learning algorithms are applied on the dataset and the classification has been done using various algorithms of which Logistic Regression gives highest accuracy of 96%. Application of pipeline gave AdaBoost classifier as best model with accuracy of 98.8%. We have seen comparison of machine learning algorithm accuracies with two different datasets. It is clear that the model improves accuracy and precision of diabetes prediction with this dataset compared to existing dataset. Further this work can be extended to find how likely nondiabetic people can have diabetes in next few years.

CHAPTER-13
BIBLOGRAPHY

13. BIBLIOGRAPHY

[1] A. Belle, R. Thiagarajan, S. M. R. Soroushmehr, F. Navidi, D. A. Beard, and K. Najarian, "Big Data Analytics in Healthcare," Hindawi Publ. Corp., vol. 2015, pp. 1–16, 2015.

[2] J. Andreu-Perez, C. C. Y. Poon, R. D. Merrifield, S. T. C. Wong, and G.-Z. Yang, "Big Data for Health," IEEE J. Biomed. Heal. Informatics, vol. 19, no. 4, pp. 1193–1208, 2015

[3] E. Ahmed et al., "The role of big data analytics in Internet of Things," Comput. Networks, vol. 129, no. December, pp. 459–471, 2017

[4] "The big-data revolution in US health care: Accelerating value and innovation | McKinsey & Company." [Online]. Available: https://www.mckinsey.com/industries/healthcare-systems-andservices/our-insights/the-big-data-revolution-in-us-health-care. [Accessed: 12-May-2018].

[5] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease Prediction by Machine Learning over Big Data from Healthcare Communities," IEEE Access, vol. 5, no. c, pp. 8869–8879, 2017.

[6] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, "Machine learning on big data: Opportunities and challenges," Neurocomputing, vol. 237, pp. 350–361, May 2017.

[7] J. B. Heaton, N. G. Polson, and J. H. Witte, "Deep learning for finance: deep portfolios,"

[8] K. Lin, M. Chen, J. Deng, M. M. Hassan, and G. Fortino, "Enhanced Fingerprinting and Trajectory Prediction for IoT Localization in Smart Buildings," IEEE Trans. Autom. Sci. Eng., vol. 13, no. 3, pp. 1294–1307, Jul. 2016.

[9] K. Lin, J. Luo, L. Hu, M. S. Hossain, and A. Ghoneim, "Localization Based on Social Big Data Analysis in the Vehicular Networks," IEEE Trans. Ind. Informatics, vol. 13, no. 4, pp.

[10] P. A. Chiarelli, J. S. Hauptman, and S. R. Browd, "Machine Learning and the Prediction of Hydrocephalus," JAMA Pediatr., vol. 172, no. 2, p. 116, Feb. 2018.

[11] A. Jindal, A. Dua, N. Kumar, A. K. Das, A. V. Vasilakos, and J. J. P. C. Rodrigues, "Providing Healthcare-as-a-Service Using Fuzzy Rule-Based Big Data Analytics in Cloud Computing," IEEE J. Biomed. Heal. Informatics, pp. 1–1, 2018.

[12] N. M. S. kumar, T. Eswari, P. Sampath, and S. Lavanya, "Predictive Methodology for Diabetic Data Analysis in Big Data," Procedia Comput. Sci., vol. 50, pp. 203–208, Jan. 2015.

[13] J. Zheng and A. Dagnino, "An initial study of predictive machine learning analytics on

large volumes of historical data for power system applications," in 2014 IEEE International Conference on Big Data (Big Data), 2014, pp. 952–959.

[14] International Journal of Advanced Computer and Mathematical Sciences. Bi Publication-BioIT Journals, 2010.

[15] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," IEEE Access, vol. 5, pp. 8869–8879, 2017.

[16] R. A. Taylor et al., "Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big DataDriven, Machine Learning Approach," Acad. Emerg. Med., vol. 23, no. 3, pp. 269–278, Mar. 2016

[17] S. Das and A. Thakral, "Predictive analysis of dengue and malaria," in 2016 International Conference on Computing, Communication and Automation (ICCCA), 2016, pp. 172–176.

[18] M. S. Simi, K. S. Nayaki, M. Parameswaran, and S. Sivadasan, "Exploring female infertility using predictive analytic," in 2017 IEEE Global Humanitarian Technology Conference (GHTC), 2017, pp. 1–6.

[19] R. Lafta, J. Zhang, X. Tao, Y. Li, and V. S. Tseng, "An Intelligent Recommender System Based on Short-Term Risk Prediction for Heart Disease Patients," in 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015, pp. 102–105.

[20] S. T. Prasad, S. Sangavi, A. Deepa, F. Sairabanu, and R. Ragasudha, "Diabetic data analysis in big data with predictive method," in 2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET), 2017, pp

CHAPTER-14

PUBLICATIONS

International Journal for Modern Trends in Science and Technology, 9(02): 164-168, 2023

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0902030
Available online at: http://www.ijmtst.com/vol9issue02.html

Diabetes Prediction Using Machine Learning Algorithms

Dr.D.Suneetha [1] | S.Harish [2] | P.Rakesh [2] | V.Bhargav [2]

[1]Professor & HOD, Department of CSE, NRI Institute of Technology, India
[2]B.Tech Student, Department of CSE, NRI Institute of Technology, India

To Cite this Article
Dr.D.Suneetha, S.Harish, P.Rakesh, V.Bhargav. Research on Diabetes Prediction Using Machine Learning Algorithms International Journal for Modern Trends in Science and Technology 2023, 9(02), pp. 164-168. https://doi.org/10.46501/IJMTST0902030

Article Info
Received: 30 December 2022; Accepted: 01 February 2023; Published: 04 February 2023

ABSTRACT

Many individuals struggle with diabetes mellitus, making it a serious health problem. Diabetes Mellitus may be brought on by many different things, including becoming older, being overweight, not getting enough exercise, having a family history of diabetes, leading an unhealthy lifestyle, eating poorly, having high blood pressure, etc. Diabetics are at increased risk for a wide range of health issues, including cardiovascular disease, renal failure, stroke, vision and nerve problems, and more. Hospitals now use a battery of tests to determine a patient's diabetes type and, after a diagnosis has been made, patients get care tailored to their specific condition. The healthcare industry benefits greatly from the use of big data analytics. The healthcare industry uses massive database systems. Using big data analytics, one may examine massive datasets in order to unearth previously unknown facts and trends in order to draw conclusions and make predictions. The accuracy of present methods for categorization and prediction is low. In this study, we offer a model for the prediction of diabetes that incorporates traditional parameters like glucose, body mass index, age, insulin levels, etc., as well as a few external factors responsible for diabetes. When compared to the original dataset, the new one improves classification precision. In addition, we imposed a pipeline model for diabetes prediction with the goal of elevating the classification precision.
KEY WORDS: Big data clouds, Diabetes System

1. INTRODUCTION

Databases with enormous volumes are common in the healthcare industry. These databases could include data that is organized, semi-structured, or even completely unstructured. The process of analyzing large data sets, often known as "big data analytics," aims to unearth previously unknown facts and trends in order to derive new insights from the data that has been provided. In light of the present situation, the condition known as diabetes mellitus (DM) has developed into a highly serious illness in developing nations like India. The condition known as diabetes mellitus (DM) is an example of a non-communicable disease (NCD), and it affects a significant number of individuals. According to the estimates from 2017, there are over 425 million individuals living with diabetes. Diabetes is the leading cause of death worldwide, taking the lives of around 2-5 million people per year. It is projected that by the year 2045, this number would have increased to 629 million. [1] Type-1 diabetes mellitus, often known as insulin-dependent diabetes mellitus, is a subtype of

diabetes mellitus (DM) (IDDM). Diabetes mellitus type 1 is characterized by the body's inability to produce an adequate amount of insulin, which is why patients with this form of the disease must get insulin injections. Diabetes mellitus type 2, sometimes referred to as non-insulin-dependent diabetes (NIDDM). Diabetes mellitus type 1 is characterized by the inability of body cells to make appropriate utilization of insulin. Diabetes type 3, often known as gestational diabetes, is caused by an increase in the amount of sugar in a pregnant woman's blood when the condition is not recognized as diabetes at an earlier stage. Diabetes Mellitus is accompanied by a number of long-term consequences. A person with diabetes is also at a significantly increased risk for a variety of health complications. A method known as Predictive Analysis makes use of data from the present as well as the past in order to gain information and make forecasts about the future. This approach is comprised of a number of different machine learning algorithms, data mining techniques, and statistical methodologies. The use of predictive analysis to data pertaining to healthcare enables crucial judgments to be made as well as predictions to be made. Machine learning and regression approach are two methods that may be used to perform predictive analytics. The goal of predictive analytics is to improve clinical outcomes while simultaneously maximizing resources, boosting patient care, and detecting diseases with the highest possible degree of precision. [1] Machine learning is widely regarded as one of the most essential characteristics of artificial intelligence. It enables the construction of computer systems that can learn from their own experiences without requiring explicit programming for each scenario. It is generally agreed that machine learning is an absolute need in the current climate in order to eradicate the need for human labor and to facilitate automation with a minimum of errors. Laboratory testing, such as blood glucose levels while fasting and oral glucose tolerance, are now the standard way for diagnosing diabetes. On the other hand, this approach requires a lot of time. Building a predictive model for diabetes using machine learning algorithms and data mining approaches is the primary emphasis of this article. The paper is structured in the following manner: In Section II, a literature overview of the prior work done on diabetes prediction is presented, as well as a classification system for machine learning algorithms. The rationale for working on this issue is presented in Section III. In Section IV, a potential model for diabetes prediction is presented and addressed. The findings of the experiment are presented in Section V, which is then followed by a Conclusion and References.

2. LITERATURE SURVEY

2.1A Predictive Analysis of Data Collected from Diabetic Patients Employing Machine Learning and Hadoop

Authors: Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar

Abstract: A significant amount of data is being produced these days by many health care companies. It is required to gather, store, and interpret this data in order to draw conclusions and make important choices based on what can be learned from it. A significant number of individuals are afflicted with the non-communicable disease known as diabetes mellitus (DM), which belongs to the NCD category. Diabetes mellitus has evolved into a significant public health problem in recent years for emerging nations like India. Diabetes mellitus is one of the most dangerous illnesses since it may cause difficulties over a lengthy period of time and it can also lead to a variety of other health issues. It is essential to create a system that, with the assistance of modern technology, can save diabetes data, do analysis on that data, and identify probable dangers in accordance with those findings. The term "predictive analysis" refers to a strategy that combines a number of data mining techniques, machine learning algorithms, and statistical methods. Predictive analysis makes use of both present and historical data sets in order to acquire insight and anticipate potential dangers. In this work, a machine learning method is constructed for the Pima Indian diabetes data set using the Hadoop MapReduce environment. The goals of this work are to identify missing values within the data set and to detect patterns from the data. Because of this study, it will be possible to forecast which varieties of diabetes are most common, the associated dangers for the future, and the sorts of treatments that may be administered to patients based on their individual risk levels.

2.2 Diabetes Risk Assessment Based on Individual Lifestyle Factors

Ayush Anand and Divya Shakti are the authors of this piece.

Abstract:

Diabetes Mellitus, which is more often referred to as diabetes, has been presented as being more severe than both cancer and HIV (Human Immunodeficiency Virus). It manifests as when elevated quantities of sugar are present in the blood for an extended length of time. In recent times, it has been suggested that it may be a role in the development of Alzheimer's disease, in addition to being a primary cause of blindness and renal failure. Within the medical research community, the issue of most interest right now is the disease's avoidance. There are a variety of approaches that have been developed to investigate the reasons for and treat diabetes. This research paper is a discussion on establishing a relationship between diabetes risk likely to be developed from a person's daily lifestyle activities such as his or her eating habits, sleeping habits, physical activity along with other indicators like BMI (Body Mass Index), waist circumference, etc. This paper is a discussion on establishing a relationship between diabetes risk likely to be developed from a person's daily lifestyle activities such as his or her eating habits, sleeping habits, physical activity along with First, a Chi-Squared Test of Independence was carried out, then the data were subjected to the machine learning method known as CART (Classification and Regression Trees), and lastly, the bias in the findings was eradicated by means of cross-validation.

2.3 Applications of Predictive Analytics in the Health Care Industry Utilizing the Resources and Methods of Machine Learning

B. Nithya and Dr. V. Ilango are the authors of this piece. Machine learning is the way to go when we have a massive data collection on which we would want to do predictive analysis or pattern identification. This is because machine learning can learn from its own mistakes. The field of computer science that is seeing the most rapid growth right now is machine learning (ML), and the field of health informatics is very difficult. The purpose of machine learning is to create algorithms that are capable of learning and improving themselves over the course of time, as well as being able to make accurate predictions. The principles of machine learning are extensively applied in a variety of industries, but the health care business in particular has benefited greatly from the use of machine learning prediction techniques. It provides a wide range of alerting and risk management decision support tools, all of which are geared at enhancing the safety of patients and the quality of healthcare. The healthcare business is facing problems in important areas such as electronic record administration, data integration, computer-aided diagnosis, and illness prediction as a result of the desire to lower healthcare costs and the push towards individualized treatment. The field of machine learning provides a broad variety of tools, strategies, and frameworks that may be used to overcome these issues. This article presents the results of a research on several prediction approaches and tools that may be used in the context of machine learning. This article also provides a brief overview of the applications of machine learning in a variety of fields while putting an emphasis on the prominent role it plays in the health care business.

2.4 Diagnosis of Diabetes Through the Utilization of Different Classification Mining Methods Aiswarya Iyer, S. Jeyalatha, and Ronak Sumbaly are the authors of this piece.

Abstract: Over 246 million people throughout the globe are living with diabetes, with women making up the vast majority of those afflicted. According to the data from the WHO, it is anticipated that by the year 2025, this number would have increased to more than 380 million. The illness has been ranked as the fifth most lethal disease in the United States, and there is currently no sign of an impending cure. Cases of diabetes, in addition to the symptoms that accompany it, have been meticulously recorded thanks to the emergence of information technology and its ongoing introduction into the medical and healthcare industries. Through the use of Decision Tree and Naive Bayes algorithms, the purpose of this study is to identify solutions that can be used to diagnose the illness. These solutions will be identified by doing an analysis of the patterns that can be found in the data. The purpose of the study is to investigate the possibility of developing a method that is both speedier and more effective in detecting the illness, which might ultimately result in patients receiving treatment at an earlier stage.

3.PROPOSED SYSTEM

Diabetes is more easily managed if it is diagnosed at an earlier stage. In order to accomplish this mission, the work that we are doing for this project entails making an early prediction of diabetes in a human body or in a patient by using a number of different machine learning

techniques. This will ensure that the aim is met. Machine learning approaches To get more accurate results when making predictions, construct models using patient data that has been gathered. In this study, we will attempt to forecast cases of diabetes by using Machine Learning Classification and ensemble methods to a dataset. Which of the following algorithms are: K-Nearest Neighbor (KNN), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Gradient Boosting (GB), and Random Forest? (RF). When compared to the accuracy of other models, each model's own accuracy varies. The work done on the project provides an accurate or better accuracy model, which demonstrates that the model is capable of making an accurate prediction of diabetes. According to the results of our analysis, Random Forest attained a greater level of accuracy when compared to other machine learning strategies.

4. RESULTS

Figure 1: Predictor

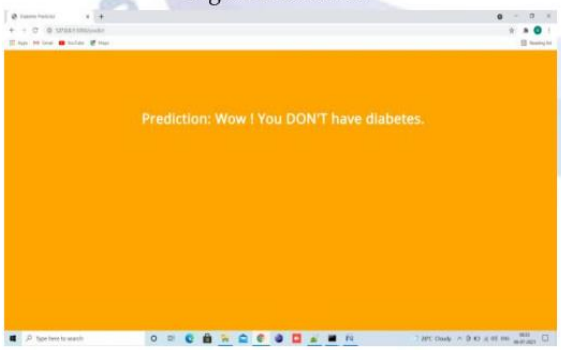


Figure 2: Results of the proposed algorithm

5. CONCLUSION

The dataset in this research was subjected to a number of different machine learning algorithms, and the classification was carried out using a number of different techniques as well, with the Logistic Regression algorithm providing the greatest accuracy at 96%. Following application of the pipeline, the AdaBoost classifier was determined to be the most accurate model. The accuracy of machine learning algorithms was evaluated using two separate datasets, and the results were compared. When compared to the previous dataset, it is abundantly obvious that the model enhances both the accuracy and precision of the diabetes prediction using this dataset. This approach may also be expanded to determine the likelihood that persons who do not currently have diabetes will get diabetes in the next years.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] A. Belle, R. Thiagarajan, S. M. R. Soroushmehr, F. Navidi, D. A. Beard, and K. Najarian, "Big Data Analytics in Healthcare," Hindawi Publ. Corp., vol. 2015, pp. 1–16, 2015.

[2] J. Andreu-Perez, C. C. Y. Poon, R. D. Merrifield, S. T. C. Wong, and G.-Z. Yang, "Big Data for Health," IEEE J. Biomed. Heal. Informatics, vol. 19, no. 4, pp. 1193–1208, 2015

[3] E. Ahmed et al., "The role of big data analytics in Internet of Things," Comput. Networks, vol. 129, no. December, pp. 459–471, 2017

[4] "The big-data revolution in US health care: Accelerating value and innovation | McKinsey & Company." [Online]. Available: https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/the-big-data-revolution-in-us-health-care. [Accessed: 12-May-2018].

[5] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease Prediction by Machine Learning over Big Data from Healthcare Communities," IEEE Access, vol. 5, no. c, pp. 8869–8879, 2017.

[6] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, "Machine learning on big data: Opportunities and challenges," Neurocomputing, vol. 237, pp. 350–361, May 2017.

[7] J. B. Heaton, N. G. Polson, and J. H. Witte, "Deep learning for finance: deep portfolios," Appl. Stoch. Model. Bus. Ind., vol. 33, no. 1, pp. 3–12, Jan. 2017.

[8] K. Lin, M. Chen, J. Deng, M. M. Hassan, and G. Fortino, "Enhanced Fingerprinting and Trajectory Prediction for IoT Localization in Smart Buildings," IEEE Trans. Autom. Sci. Eng., vol. 13, no. 3, pp. 1294–1307, Jul. 2016.

[9] B. Srinivasa Rao D. Vijaya Kumar and K. Kiran Kumar, "Power quality improvement using Cuckoo search based

multilevel facts controller", Journal of Engg. Research Vol.10 No. (4A) pp. 252-261, DOI: 10.36909/jer.10895.
[10] P. A. Chiarelli, J. S. Hauptman, and S. R. Browd, "Machine Learning and the Prediction of Hydrocephalus," JAMA Pediatr., vol. 172, no. 2, p. 116, Feb. 2018.