CHAPTER 1

INTRODUCTION

Diabetes is a chronic disease which is caused due to increase level of blood sugar. Various methods are available for diagnosing diabetes. However early prediction of diabetes is a quite challenging task. Diabetes is a fast growing disease among the people even among the youngsters. This disease is caused due to increase level of blood sugar.

Types of Diabetes:

* Type 1 diabetes means that immune system is compromised and the cells fail to produce insulin in sufficient amounts. There are no eloquent studies that prove the causes of type 1 diabetes and there are currently no known methods of prevention.
* Type 2 diabetes means that the cells produce a low quantity of insulin or the body can’t use the insulin correctly. This is the most common type of diabetes, thus affecting 90% of persons diagnosed with diabetes. It is caused by both genetics factors and the manner of living.

Machine Learning is an emerging scientific field in data science dealing with the ways in which machine learns from the experience. The aim of this project is to design a system which performs early prediction on patient data. In this project we will use diabetes CSV file which is in kaggle.

This Dataset contains nine features in that eight are independent features and one is outcome. Independent Features are

* Pregnancies
* Glucose
* Blood Pressure
* Insulin
* Skin thickness
* BMI
* DiabetesPedigreeFunction
* Age

And dependent feature is Outcome. Here we will train several Machine Learning algorithms to predict Outcome features. We will choose algorithm which gives better results.

MACHINE LEARNING

Machine learning is a field in computer science aiming to imitate the human learning process. Machine Learning is a branch of Artificial Intelligence where computer learns from the data (past experiences) and makes future prediction. It finds the pattern in data, based on pattern it predicts for unseen data.

Machine Learning classified into 3 types:

1. Supervised Machine Learning
2. Unsupervised Machine Learning
3. Re-Inforcement Machine Learning

Supervised Machine Learning

Supervised Learning algorithms learn from labeled input data, where labeled input data associated with output responses that consist of numeric values or string labels. Based on Output Variable Supervised Machine Learning classified into two types:

1. Regression:

If output variable is continuous then these types of problems can be considered as regression type.

Example: Predicting salary of person based on past experience of person.

1. Classification:

If output variable is categorical then these types of problems can be considered as classification type.

Example: Predicting whether patient is having cancer or not.

Unsupervised Machine Learning

If Machine Learning learns from unlabeled data, where data is not associated with target response.

Algorithms:

1. Clustering Algorithms
2. Association Clustering

Reinforcement Machine Learning

Reinforcement learning differs from supervised learning in a way that in supervised learning the training data has the answer key with it so the model is trained with the correct answer itself whereas in reinforcement learning, there is no answer but the reinforcement agent decides what to do to perform the given task.

This problem is classification problem. Here we will use Logistic Regression, Random Forest, Gradient Boosting, XG Boost, KNN algorithms and we will compare which will give best.

CHAPTER 2

PROBLEM STATEMENT

Diabetes is a chronic disease which is caused due to increase level of blood sugar. Various methods are available for diagnosing diabetes. However early prediction of diabetes is a quite challenging task. Diabetes is a fast growing disease among the people even among the youngsters. This disease is caused due to increase level of blood sugar.

Types of Diabetes:

* Type 1 diabetes means that immune system is compromised and the cells fail to produce insulin in sufficient amounts. There are no eloquent studies that prove the causes of type 1 diabetes and there are currently no known methods of prevention.
* Type 2 diabetes means that the cells produce a low quantity of insulin or the body can’t use the insulin correctly. This is the most common type of diabetes, thus affecting 90% of persons diagnosed with diabetes. It is caused by both genetics factors and the manner of living.

Machine Learning is an emerging scientific field in data science dealing with the ways in which machine learns from the experience. The aim of this project is to design a system which performs early prediction on patient data.

CHAPTER 3

DATASET

The dataset collected is originally from the Pima Indians Diabetes Database is available on Kaggle. Dataset contains several independent variables and one dependent variable.

Independent features are Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, DiabetesPedigreeFunction and Age. Dependent feature is Outcome.

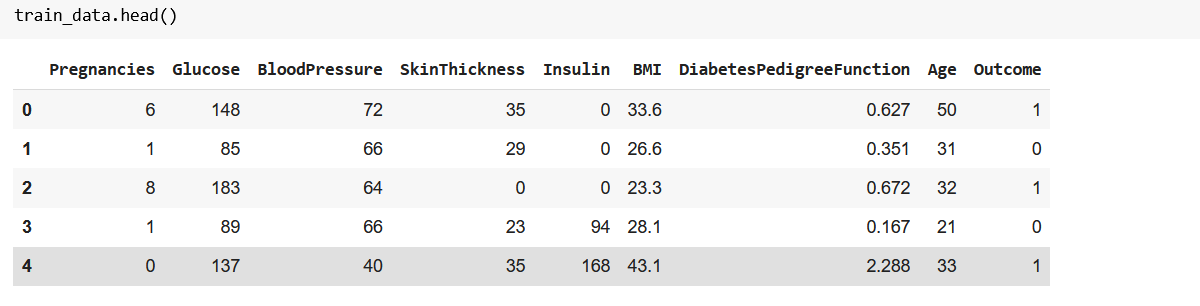


FIG 3.1: Dataset

This dataset has 768 observations and nine columns. Outcome is dependent features. Here we need to predict outcome column. All columns are numerical columns

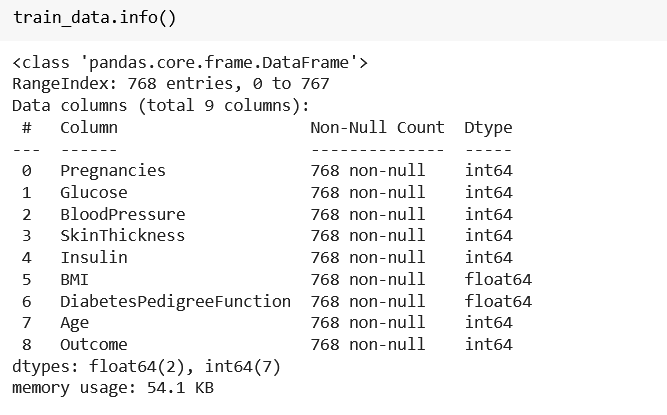


FIG 3.2: Dataset Description

CHAPTER 4

SOFTWARE REQUIREMENTS SPECIFICATION

4.1 INTRODUCTION:

Software Requirement Specification is the beginning point of the software developing activity. As system grew more complex it became evident that the goal of the entire system cannot be easily understood. Hence the need for the requirement phase arises. The Software Requirements Specification is one of the means of translating the ideas of the minds of clients (the input) into format document (the output of the requirement phase).

4.2 PURPOSE:

The purpose of this project is to predict the disease for given patient information. Using Machine Learning algorithms we can able to detect the diabetes.

4.3 SCOPE:

This project is to predict whether patient is having diabetes or not.

4.4 OBJECTIVE:

The objective of this project is to implement algorithms which can able to predict whether given information of patient is having diabetes or not.

4.5 REQUIREMENTS:

4.5.1 HARDWARE REQUIREMENTS

• Processor - Pentium –III

• Speed - 1.1 GHz

• RAM - 256 MB (min)

• Hard Disk - 260 GB

4.5.2 SOFTWARE REQUIREMENTS

* Operating System – Windows10/98/2000/XP
* Web Framework : Flask
* Programming Language: Python.
* Libraries: Sklearn

CHAPTER 5

SYSTEM IMPLEMENTATION

5.1 Modules

5.1.1 Data Preprocessing

* + 1. Model Training

5.1.1 Data Preprocessing:

There are several steps in Data preprocessing like Null values identification, Imbalanced Dataset treatment and etc.

Null Values Identification: There are no null values.

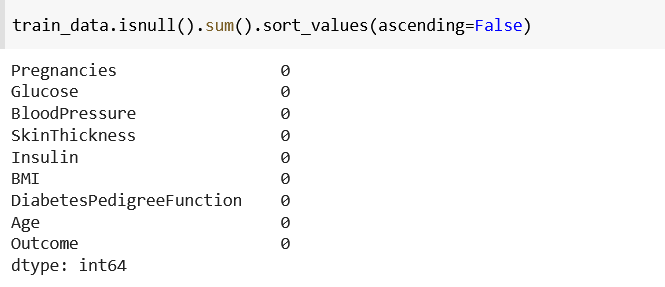


FIG 5.1: Null values Treatment

Imbalanced Dataset Treatment: This dataset is imbalanced; we will use Sklearn to treat imbalanced dataset.

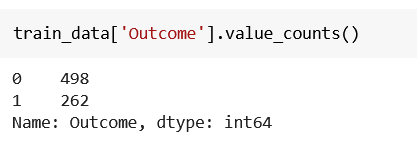


FIG 5.2: Imbalanced Dataset

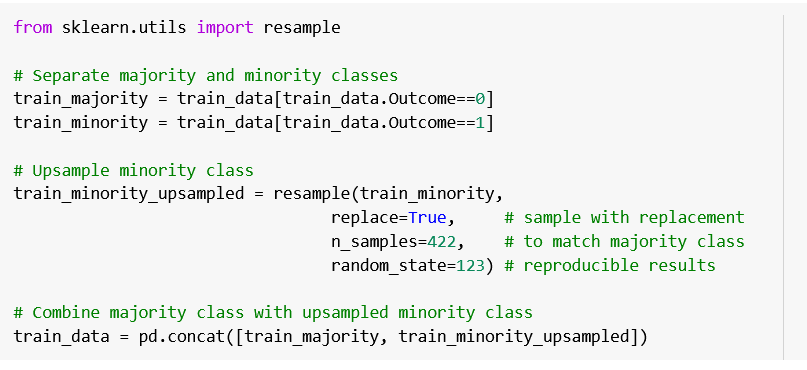


FIG 5.3: Treating Imbalanced Dataset

5.1.2 Model Training

We will trail various Machine Learning algorithms and we will compare all with the results and we will select the best one.

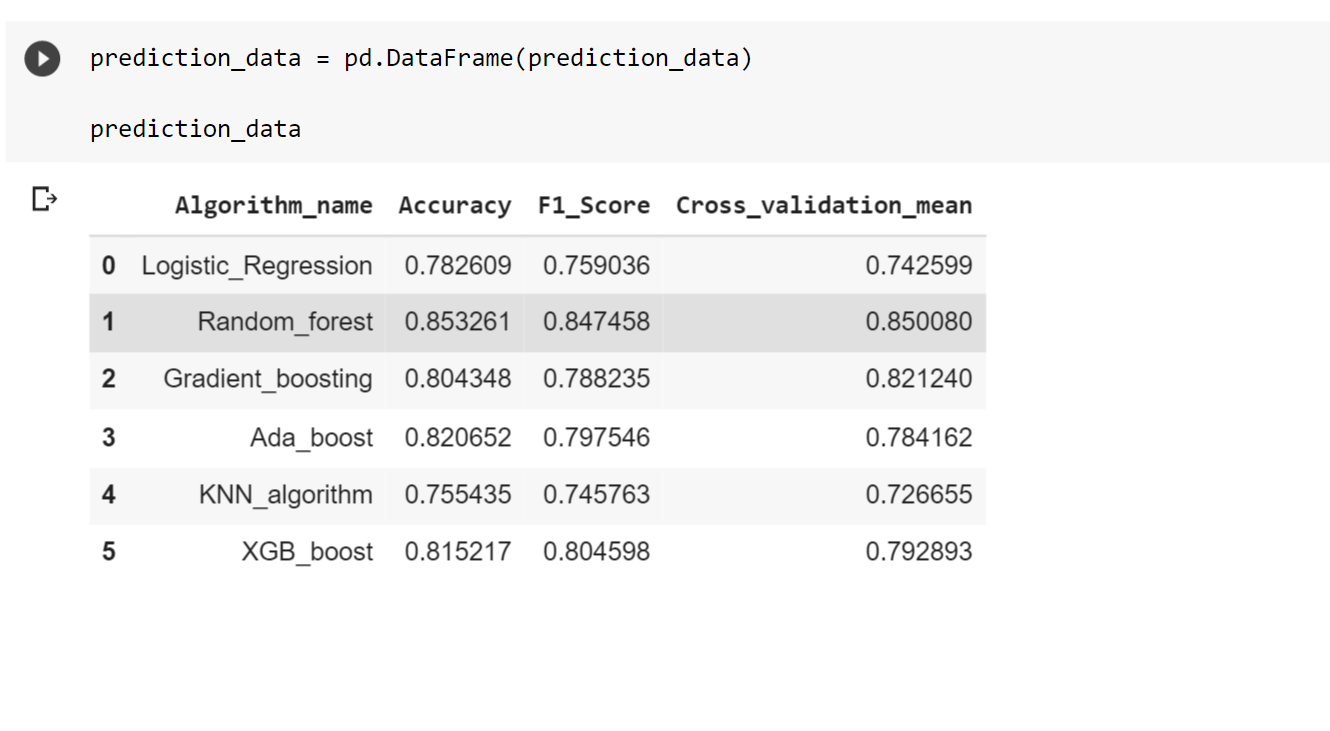


FIG 5.4: Training results

CHAPTER 6

TESTING

6.1 SOFTWARE TESTING

GENERAL

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, subassemblies, 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.

DEVELOPING METHODOLOGIES

The test process is initiated by developing a comprehensive plan to test the general functionality and special features on a variety of platform combinations. Strict quality control procedures are used. The process verifies that the application meets the requirements specified in the system requirements document and is bug free. The following are the considerations used to develop the framework from developing the testing methodologies.

6.2 Types of Tests

6.2.1 Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program input produces 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 structural testing, that relies on knowledge of its construction and is invasive configuration.

6.2.2 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.

6.2.3 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.

6.2.4 Performance Test

The Performance test ensures that the output is produced within the time limits, and the time taken by the system for compiling, giving response to the users and request being send to the system for to retrieve the results.

6.2.5 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.

CHAPTER 7

SNAPSHOT



