CHAPTER 1

INTRODUCTION

1.1 Introduction

- Python works on different platforms (Windows, Mac, Linux, Raspberry Pi, etc).
- Python has a simple syntax similar to the English language.
- Python has syntax that allows developers to write programs with fewer lines than some other programming languages.
- Python runs on an interpreter system, meaning that code can be executed as soon as it is written. This means that prototyping can be very quick.
- Python can be treated in a procedural way, an object-orientated way or a functional way.

Artificial Intelligence is an approach to make a computer, a robot, or a product to think how smart human think. AI is a study of how human brain think, learn, decide and work, when it tries to solve problems. And finally this study outputs intelligent software systems. The aim of AI is to improve computer functions which are related to human knowledge, for example, reasoning, learning, and problem-solving. The intelligence is intangible. It is composed of reasoning, learning, problem solving, perception and linguistic intelligence.

Machine learning gives computers the ability to learn without being explicitly programmed (Arthur Samuel, 1959). It is a subfield of computer science.

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. Machine learning algorithms are used in a wide variety of applications, such as email filtering, and computer vision, where it is in feasible to develop an algorithm of specific instructions for performing the task. Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning.

Diabetes is a chronic condition prevalent among more than 25 million people across the world, affecting people of all ages. It can be said as a condition of the body where it cannot produce enough insulin to break down the sugar or it cannot use the insulin produced by the body. It can also be considered as a “slow poison”, which does not show its entire effects immediately but destroys the body step by step, rather slowly. Despite major advances in science and technology, diabetes continues to be a chronic disease, with a thirty-day readmission rate of around 20%, as compared to an average of 12% for the rest of the diseases. Additionally, readmissions cost hospitals a fair amount of money, so the end goal is to identify and reduce the possibility of a readmission. Prevention of patient readmission has been given a greater importance due to large cost involvement.

As the health care system moves toward value-based care, CMS has created many programs to improve the quality of care of patients. One of these programs is called the Hospital Readmission Reduction Program, which reduces reimbursement to hospitals with above average readmissions. For those hospitals which are currently penalized under this program, one solution is to create interventions to provide additional assistance to patients with increased risk of readmission. But how do we identify these patients? We can use predictive modeling from data science to help prioritize patients.

One patient population that is at increased risk of hospitalization and readmission is that of diabetes. Diabetes is a medical condition that affects approximately 1 in 10 patients in the United States. According to etal, patients with diabetes have almost double the chance of being hospitalized than the general population. Therefore, I will focus on predicting hospital readmission for patients with diabetes.

1.2 OBJECTIVE OF RESEARCH

The primary objective of this project is to predict whether the patient will be readmitted to the hospital or not. We have used different classification methods for this purpose. Detailed explanation on the same is present in the below sections. Additionally, we created models to predict, 1) the time a patient is likely to spend in the hospital based on the preliminary diagnoses

This is a classification problem. To predict whether a patient will be readmitted or not, we created six different classifiers, namely, Support Vector Machines, Generalized logistic regression, Artificial Neural Networks, Random Forest Classifier, Naïve Bayes Classifier, Decision Trees.

Feature selection for these classes was done by conducting Correlation Analysis, and eliminating features with class imbalance.

Readmission rate was found to be high among the Caucasians. Test result was a good predictor of readmission (Normal results implies less chance of readmission).

1.3 PROBLEM STATEMENT

Diabetic patients experience poorer self-management ability and a higher frequency of hospital readmissions when compared to their Caucasian counterparts as a result of patient education deficits, underutilization of racially conducive learning styles, and minimal attention devoted to evaluating patient perception of diabetes mellitus.

Diabetes linked to an increased likelihood for a number of serious, sometimes life threatening compilations. Proper diabetes management can reduce the risk of compilations and improve the quality of life. It is widely known that diabetes patients who are educated on disease management techniques are more effective in maintaining proper diet, blood pressure, and blood glucose. They are also more likely to make healthy lifestyle adjustments that results in positive changes, such as ideal body weight, they suffer fewer compilations and they require emergency medical care.

1.4 **INDUSTRY PROFILE**

Hospital readmission is a high-priority health care quality measure and target for cost reduction. Despite board interest is readmission, relatively the little research has focused on patients with diabetes. The burden of diabetes among hospitalized patients, however, is substantial, growing and costly, and readmissions contribute a significant portion of this burden. Reducing readmission rates of diabetic patients has the potential to greatly reduce health care cost while simultaneously improving care. Risk factors for readmission in the population include lower socioeconomic status, racial minority, comorbidity burden, public insurance, emergent admission, and a history of recent prior hospitalization. Hospitalized patients with diabetes may be at higher risk of readmission then those without diabetes. Potential ways to reduce readmission risk are inpatient, education, specialty care, better discharge instructions, coordinate of care, and post discharge supports. More studies are needed to test the effect of these interventions on the readmission rates of patient with diabetes.

CHAPTER 2

REVIEW OF LITERATURE

Health care Organizations continue to develop optimal solutions in chronic care models. Using such chronic illness management models to develop personalized interventions in hospitals has yielded better patient outcomes, decreased length of stay, and less rates of readmission in a 30 day period.

This study investigated the relationship between unnecessary hospital readmissions of diabetes patients and the key aspects related to diabetes: patient demographics, lifestyle components, biomarkers, and disease management. The concept of the study is based in the interconnected dynamic relationship of diabetes care aspects with the predictability of emergency treatment or unplanned admissions into a hospital and the economic impact of suboptimal disease management of diabetes. The purpose of this study was to evaluate the contribution of patient demographics, (e.g., age, gender, ethnicity), lifestyle components biomarkers (e.g., blood glucose, blood pressure) and disease management aspects (e.g., physician specialty or not and patient participation in a diabetes education program) to predicting hospitalization readmissions for patients with diabetes. This review of the literature will first provide background and context, including definitions of the two types of diabetes, an overview of prevalence of the disease, and an explanation of the economic impact of diabetes. It will present the demographic groups of patients with diabetes and their lifestyle components. The review of patient characteristics and lifestyle will consider age, gender, ethnicity, and patient controlled components that are known to directly affect diabetes such as smoking status and social support. 9 Essential biomarkers will be explained. Biomarkers such as blood glucose level and blood pressure level are indicators of the level of success of disease management efforts. The types of diabetes education available to assist in patient self-management are described, and the relationship between the physician specialty and avoidable readmission is also examined. Examination of these factors may help predict future acute episodes of diabetes that lead to recidivism hospital admissions. The economic impact of suboptimal hospital care and self-care practices resulting in recidivistic readmissions will be detailed

CHAPTER 3

DATA COLLECTION

Prediction models for the recidivistic admissions of patients with diabetes were generated by logistic regression (LR), Support vector machine (SVM), Random Forest Classifier and K nearest neighbors (KNN). A comparison of these models was conducted to determine which method produced the best predictors. To assess the likelihood of recidivistic diabetes admissions, a predictive equation was developed using data from 389 cases. An examination of medical records was assessed for 11 predictors common to this data set. A number of variables were considered as possible predictors for recidivistic admissions of patients with diabetes. Logistic regression, SVC, KNN and Random Forest Classifier models were used to assess the correlation of the variables to the time to the next visit in days. These included dichotomous variables: pregnancies, glucose, Insulin, Skin thickness, BMI, Diabetics. Categorical variables included age and blood glucose.

CHAPTER 4

METHODOLOGY

4.1 Exploratory Data Analysis

Bike buyer prediction project works under classification model. A classification model tries to draw some conclusion from the input values given for training. It will predict the class labels/categories for the new data.Classification can be performed on structured or unstructured data. Classification is a technique where we categorize data into a given number of classes. The main goal of a classification problem is to identify the category to which a new data will fall under.

Prediction models for the bike buyers were generated by logistic regression (LR), Support vector machine (SVM), Random Forest Classifier and K nearest neighbors (KNN),Decision Tree Algorithm. A comparison of these models was conducted to determine which method produced the best accuracy. Accuracy is only really useful when there are an even distribution of values in a data set.

The good news for us is in our data set they are nearly perfectly even. To assess the likelihood of bike buyer prediction, a predictive equation was developed using data from 700 cases. Decision Trees are type of Supervised Machine Learning(that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter.The tree can be explained by two entities,namely decision nodes and leaves.The leaves are the decisions or the final outcomes.

There are two main types of Decision Trees:

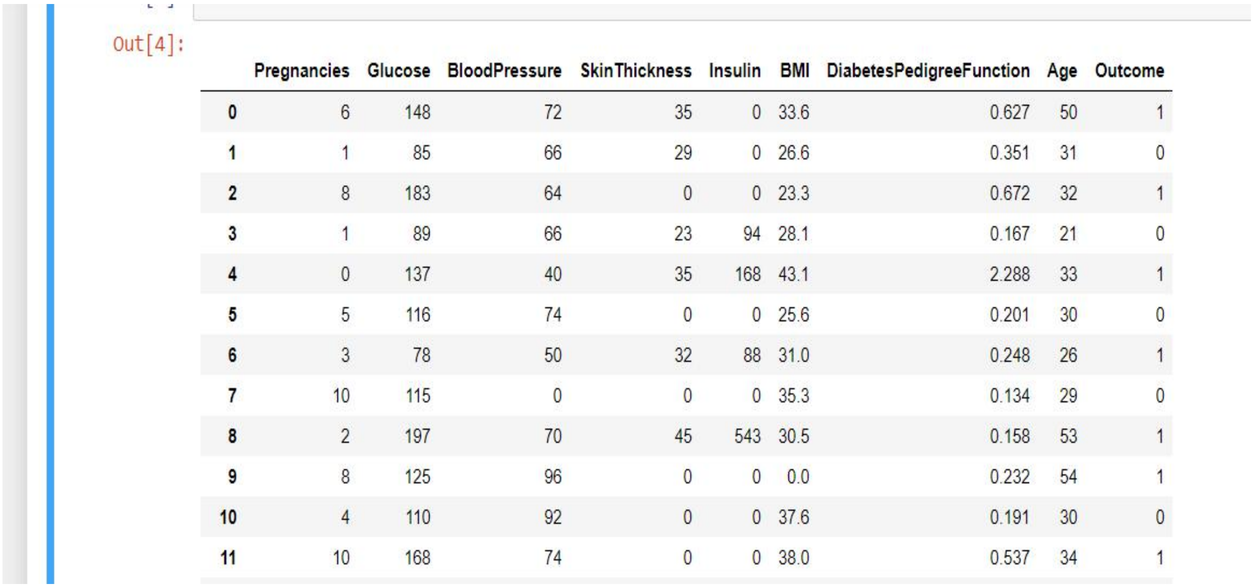
1. Classification trees(Yes/No types)

What we've seen above is an example of classification tree, where the outcome was a variable like 'fit' or 'unfit'. Here the decision variable is Categorical.

2.Regression trees (Continuous data types)

Here the decision or the outcome variable is Continuous, e.g. a number like 123.

4.1.1 Figures and tables



Out[4]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31.0	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0.0	0.232	54	1
10	4	110	92	0	0	37.6	0.191	30	0
11	10	168	74	0	0	38.0	0.537	34	1

Fig:1.1

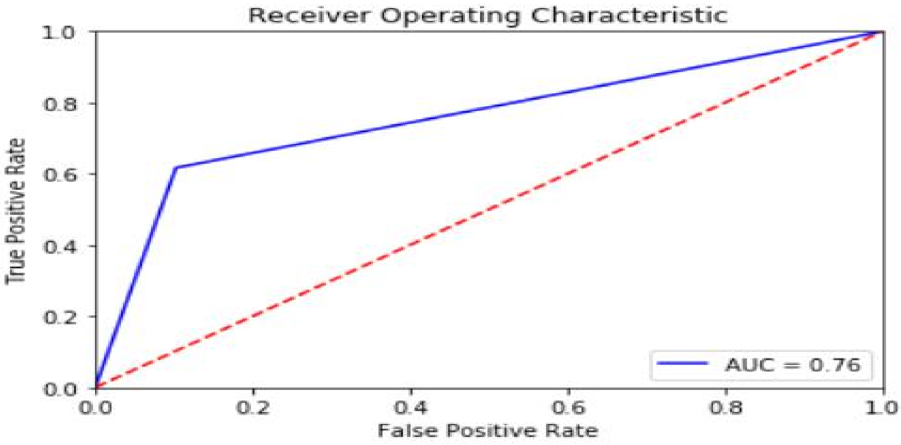


Fig:1.2

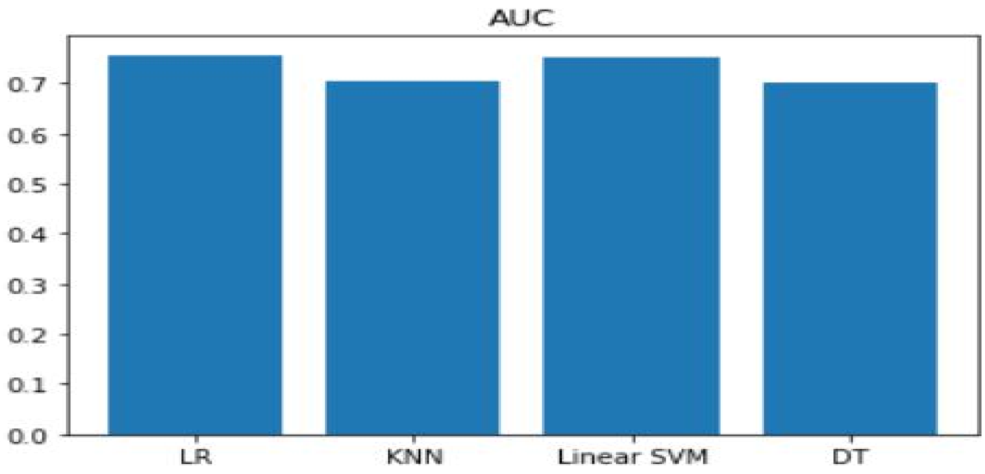
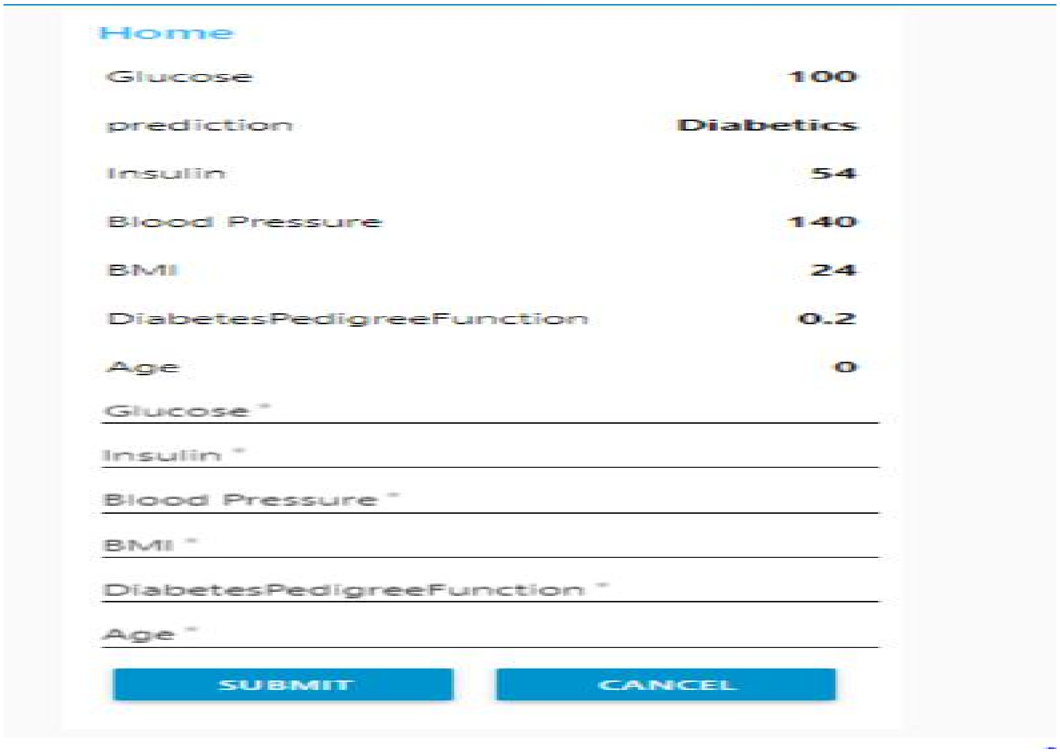


Fig:1.3



Glucose	100
prediction	Diabetics
Insulin	54
Blood Pressure	140
BMI	24
DiabetesPedigreeFunction	0.2
Age	0

Glucose *

Insulin *

Blood Pressure *

BMI *

DiabetesPedigreeFunction *

Age *

Fig:1.4

4.2 Statistical techniques and visualization

NUMPY

NumPy stands for 'Numerical Python' or 'Numeric Python'. It is an open source module of Python which provides fast mathematical computation on arrays and matrices. Since, arrays and matrices are an essential part of the Machine Learning ecosystem, NumPy along with Machine Learning modules like Scikit-learn, Pandas, Matplotlib, TensorFlow, etc. complete the Python Machine Learning Ecosystem.

NumPy provides the essential multi-dimensional array-oriented computing functionalities designed for high-level mathematical functions and scientific computation. Numpy can be imported into the notebook using import numpy as np. NumPy's main object is the homogeneous multidimensional array. It is a table with same type elements, i.e, integers or string or characters (homogeneous), usually integers. In NumPy, dimensions are called axes. The number of axes is called the rank.

PANDAS

Similar to NumPy, Pandas is one of the most widely used python libraries in data science. It provides high-performance, easy to use structures and data analysis tools. Unlike NumPy library which provides objects for multi-dimensional arrays, Pandas provides in-memory 2d table object called Dataframe. It is like a spreadsheet with column names and row labels.

Hence, with 2d tables, pandas is capable of providing many additional functionalities like creating pivot tables, computing columns based on other columns and plotting graphs. Pandas can be imported into Python using import pandas as pd. New columns and rows can be easily added to the dataframe. In addition to the basic functionalities, pandas dataframe can be sorted by a particular column. Dataframes can also be easily exported and imported from CSV, Excel, JSON, HTML and SQL database.

MATPLOTLIB:

Matplotlib is a 2d plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments. Matplotlib can be used in Python scripts, Python and IPython shell, Jupyter Notebook, web application servers and GUI toolkits. Matplotlib.pyplot is a collection of functions that make matplotlib work like MATLAB. Majority of plotting commands in pyplot have MATLAB analogs with similar arguments.

On the X array below we saying... include all items in the array from 1,2,4,5,6,7. On the y array below we are saying... just use the column in the array mapped to the **8rd row**. The **outcome** column. We are using group by to view the distribution of values in our **outcome** column. Recall that this column is our **target variable**. It's that thing we are trying to predict.

4.3 Data modeling and visualization:

Imported libraries are numpy,pandas,matplotlib. Numpy stands for 'Numerical Python' or 'Numeric Python'. It is an open source module of Python which provides fast mathematical computation on arrays and matrices. Pandas is one of the most widely used python libraries in data science. It provides high-performance, easy to use structures and data analysis tools. Matplotlib is a 2d plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments. Matplotlib can be used in Python scripts, Python and IPython shell, Jupyter Notebook, web application servers and GUI toolkits.

A library is essentially a collection of modules that can be called and used. A lot of the things in the programming world do not need to be written explicitly ever time they are required. There are functions for them, which can simply be invoked. This is a list for most popular Python libraries for Data Science. A lot of datasets come in CSV formats. We will need to locate the directory of the CSV file at first (it's more efficient to keep the dataset in the same directory as your program) and read it using a method called *read_csv* which can be found in the library. entire line of data but what if you are unknowingly removing crucial information? Of course we would not want to do that. One of the most common idea to handle the problem is to take a mean of all the values of the same column and have it to replace the missing data. Sometimes our data is in qualitative form, that is we have texts as our data. We can find categories in text form. Now it gets complicated for machines to understand texts and process them, rather than numbers, since the models are based on mathematical equations and calculations. Therefore, we have to encode the categorical data.

Now we need to split our dataset into two sets—a Training set and a Test set. We will train our machine learning models on our training set, i.e our machine learning models will try to understand any correlations in our training set and then we will test the models on our test set to check how accurately it can predict. A general rule of the thumb is to allocate 80% of the dataset to training set and the remaining 20% to test set. For this task, we will import *test_train_split* from *model_selection* library of scikit. The final step of data preprocessing is to apply the very important feature scaling. It is a method used to standardize the range of independent variables or features of data.

CHAPTER 5

FINDINGS AND SOLUTION

In this project we provide guidance to diabetic patients whether they are re-admit or not. When the diabetes level is high then they visit to hospital for checkup. By using SVC we predict the accuracy value.

.

CHAPTER 6

CONCLUSION

Through this project, we created a machine learning model that is able to predict the patients with diabetes with highest risk of being readmitted within 30 days. The best model was a gradient boosting classifier with optimized hyper parameters. The model was able to catch 58% of the readmissions and is about 1.5 times better than just randomly picking patients. Overall, I believe many healthcare data scientists are working on predictive models for hospital readmission.The goal of the project was to identify the factors that affect the readmission of diabetic patients into hospitals. Finally, analysis on the dataset revealed that Caucasian women over the age of 60 are more susceptible to being readmitted. Also, we discovered the duration of being admitted and medication administered play an important role, as patients who stayed between 4 and 8 days and received diabetes medication are more likely to be readmitted.