Big Data Analytics Technical Project Proposal

Title:	Pima Indians Diabetes
Author:	Athira Bindhu
Student ID:	L00163585
Supervisor:	Henna Shagufta
Degree:	Master of Science in Big Data Analytics and Artificial Intelligence

Background

In our data set, we are dealing with Pima Indian women whose age is over 21 years which is Pima Indians are believed to be descendents of those who crossed the Bering Strait from Asia to the Americas. The information about Pima Indians Diabetes[6] is obtained from the National Institute of Diabetes and Digestive and Kidney diseases.The main aim of using this data set is to predict whether a patient is having diabetes by using the various measurement given in the dataset.

Dataset

The Pima Indian Diabetes data set is taken from Kaggle data set.The Dataset consist of one csv file having 768 rows and 9 columns.

The Fields in the Dataset are listed below:

Pregnancies

Glucose

BloodPressure

SkinThickness

Insulin

BMI

DiabetesPedigreeFunction

Age

Outcome: The knowledge of whether there is diabetes: 1 or 0

Based on all these datas we are planning to predict the diabetic patients.Our main aim is to reduce this data set to simple form without manual effect to attain an optimal solution as we wish in an effective manner by using my knowledge acquired from the lectures and practicals.

Method

Here I am basically planing to predict diabetic patients from the data set by using ANN (Artificial neural network)[1].we are using matplot python library for visualizing the analysis.[2]

The second method is to predict weather the patient in the given dataset is diabetic or not by implementing certain machine learning models [5]such as SVC(support vector classifier),RandomForest Classifier,GuassianNB,XGB Classifier and Bernoulli NB.These models can be categorised using the accuracy. The related data code of this project will be uploaded to below link.

https://github.com/athirabindhu/pima-indians-diabetes.git

Software and Hardware Requirements

We are planning to do this project in python language by importing some main libraries such as matplot,different classifiers as mentioned above,label encoder and one hot encoder,pandas,sql, kaggle dataset, skompiler etc and visualization of our predictions will be also done in this project.SVC basically used for prediction.Jupyter Notebook interface is using here to develop the notebooks in python.

Goal

[1] By using the ANN method we are planning to predict the patient in the given dataset is diabetic or not.

[2] By comparing the accuracy of the machine learning models we are also planning to find out the number of women who are diabetic above age 25 years.

[3]Comparison of diabetic person based on their accuarcy measurement using machine learning algorithms.

Further improvement will done if needed and will be included in the final project report.

References

[1] Avinash Navlani tutorials] online 2019 hhttps://www.datacamp.com/community/tutorials/neura
network-models-r

[2] Shrish Mohadarkar — October 19, 2021]Implementing Artificial Neural Network in Python from Scratchhttps://www.analyticsvidhya.co

[3] Purdue University]Python AI: How to Build a Neural Network Make Predictionshttps://realpython.com.

[4] sqlite3 — DB-API 2.0 interface for SQLite databases — Python ...https://docs.python.or

[5] online https://builtin.com/data-science/tour-top-10-algorithms-machine-learning-newbies

[6] kaggle-UCI Machine learning 2017https://www.kaggle.com/uciml/pima-indians-diabetes-database