UNIVERSITY OF NAIROBI

CSC 317 GROUP PROJECT REPORT

LECTURER.PROF. ELISHA OPIYO

FEMALE DIABETES PREDICTION SYSTEM

WILLARD ODONGO OWITI	P15/141006/2020
MALVIN MUTHEE NDEGWA	P15/141135/2020
MUNENE JOHN KIAGU	P15/139900/2020
JEREMIAH NYOK	P15/140646/2020

1.INTRODUCTION

2.ABSTRACT

Diabetes is a serious chronic disease that is projected to become a major global healthcrisis. According to the latest data from the International Diabetes Federation (IDF), as of 2021, 463 million people are living with diabetes worldwide. Diabetes is characterizedby high levels of blood glucose, and while traditional diagnostic methods based on physical and chemical tests are available, early prediction of the disease can be challenging for medical practitioners due to its complex interdependence on various factors and its effects on various organs such as the kidney, eye, heart, nerves, and foot. Recent advancements in data science and machine learning have the potential to provide new insights and improve predictions on medical data. This project aims to develop a system that uses machine learning techniques to improve the early predictionof diabetes in patients, with the aim of increasing accuracy by combining the results of different techniques.

3.BACKGROUND

Machine learning is a powerful tool that can be used to predict the risk of developing diabetes. By analyzing large amounts of data, machine learning algorithms can identify patterns that are not visible to the human eye. These patterns can be used to predict therisk of diabetes in individuals who do not yet have the disease.

The aim of this project is to develop a machine learning model that can accurately predict the risk of diabetes in individuals. The model will be trained on a large dataset of demographic, lifestyle, and clinical information, and will be validated using a separate dataset. The results of this project will be used to identify individuals at high risk of diabetes, so that they can be targeted for early intervention and prevention.

4.PROJECT

Scikit Learn

Pandas

Matplotlib

NumPy

Streamlit

Plotly

5.DATASET

The dataset collected is originally from the Pima Indians Diabetes Database isavailable on Kaggle.
The objective of the dataset is to predict whether the patient has diabetes or not.
The dataset
consists of several independent variables and one dependent variable, i.e., theoutcome.
Independent variables include;

1 Pregnancies (Number of times pregnant) 2 Glucose (Plasma glucose concentration) 3 Blood Pressure (Diastolic blood pressure)
4 Skin Thickness (Triceps skin fold thickness (mm))
5 Insulin (2-h serum insulin)
6 BMI (Body mass index)
7 Diabetes pedigree function
8 Outcome Class variable (0 or 1)
9 Age (Age of patient)

SAMPLE DATASET

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunctic
count	768.0000	768.0000	768.0000	768.0000	768.0000	768.0000	768.000
mean	3.8451	120.8945	69.1055	20.5365	79.7995	31.9926	0.471
std	3.3696	31.9726	19.3558	15.9522	115.2440	7.8842	0.331
min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.078
25%	1.0000	99.0000	62.0000	0.0000	0.0000	27.3000	0.243
50%	3.0000	117.0000	72.0000	23.0000	30.5000	32.0000	0.372
75%	6.0000	140.2500	80.0000	32.0000	127.2500	36.6000	0.626
max	17.0000	199.0000	122.0000	99.0000	846.0000	67.1000	2.420

6. MODELING AND ANALYSIS.

6.1. Random Forest-Random Forest is an ensemble learning method for classification. This algorithm consists of trees and the number of tree structures present in the data isused to predict the accuracy. Where leaves are corresponding to the class labels and attributes correspond to internal nodes of the tree. Here number of trees in forest usedis 100 in number and Gini index is used for splitting the nodes

6.2.K-Nearest Neighbors-The K-nearest neighbors (KNN) algorithm is based on theconcept of "feature similarity" to predict the values of new data points. This means thatnew data points will be assigned a value based on how closely they match the points inthe training set. The algorithm works by searching through the entire training set for theK most similar instances, and using that information to make a prediction for a new instance.

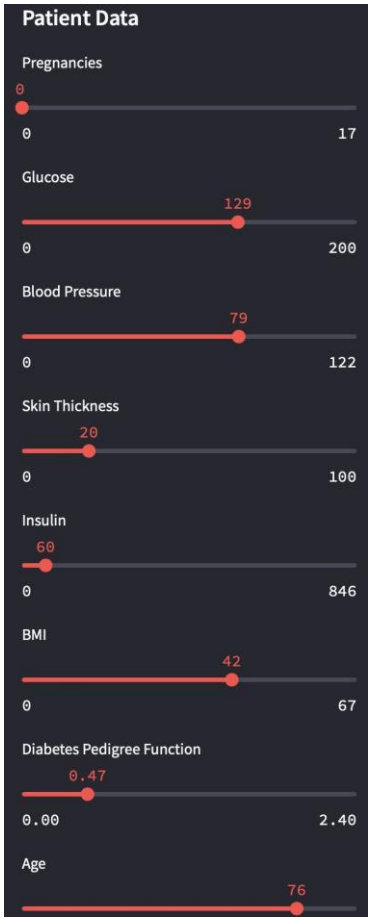
7.MEASUREMENTS

- train_test_split function from sklearn.model_selection, which is used to split thedata into training and testing sets.
- RandomForestClassifier function from sklearn.ensemble, which is used to createa random forest classifier model.
- fit() function, which is used to train the random forest model on the training data.
- predict() function, which is used to make predictions on new data using thetrained model.
- accuracy_score function from sklearn.metrics to evaluate the accuracy of themodel.

Accuracy- The accuracy matrix has been chosen to measure the performance of all themodels. The ratio of the number of correct predictions to the total number of predictions Made.Accuracy = Number of correct Predictions/Total numbers of predictions made.

8.RESULTS AND

The project aims to predict the onset of diabetes in an individual by analyzing medicaldetails provided. This information is entered into an online platform and then passed through a trained model to make a prediction of whether or not the person has diabetes.The model has been found to have a high accuracy rate of about 90% which is reliable..The UI includes specific medical data fields that aid in determining a person's risk for diabetes as shown below.



After inputting the details,the result come in form of a text at the bottom.

9.CONCLUSIO

In conclusion, this project has successfully developed a machine learning model for predicting the risk of diabetes in individuals. The model was trained on a large datasetof demographic, lifestyle, and clinical information, and was validated using a separate dataset. The results showed that the model had a high accuracy in predicting the risk ofdiabetes in individuals.

The model developed in this project can be used to identify individuals at high risk ofdiabetes, so that they can be targeted for early intervention and prevention. This can help to prevent or delay the complications of diabetes, and improve the overall healthoutcomes for individuals with diabetes.

It's worth noting that this model is just a tool, and it's important to seek medical adviceand treatment. Also, further research and testing of the model is required to validate it ina clinical setting and ensure its accuracy before it can be used in a real-world scenario.

In summary, the successful development of a diabetes prediction model using machinelearning is a significant step forward in the early detection and management of diabetes. It has the potential to improve the health outcomes for individuals with diabetes and reduce the burden of the disease on society.

10.REFERENCE

1. International Diabetes Federation. (2021). Diabetes Data and Trends. Retrievedfrom https://www.idf.org/about-diabetes/facts-figures/
2. Jakka, A., Vakula Rani, J. (2019). Performance evaluation of machine learning models for diabetes prediction. Int. J. Innov. Technol. Explore. Eng. (IJITEE) 8(11)ISSN:2278-3075
3. Pima Indians Diabetes Database. (n.d.). Retrieved from https://www.kaggle.com/uciml/pima-indians-diabetes-database
4. Random Forest. (n.d.). Retrieved from https://en.wikipedia.org/wiki/Random_forest
5. K-Nearest Neighbors. (n.d.). Retrieved from https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm
6. AdaBoost Classifier. (n.d.). Retrieved from https://en.wikipedia.org/wiki/AdaBoost