**A Project Report**

**On**

**Predictive Healthcare Analytics**

***Submitted in partial fulfillment of the***

***requirement for the award of the degree of***

**MASTER OF COMPUTER APPLICATION**

Machine Learning (E1PY202C)

Session 2023 - 2024

Under the guidance of

**Dr. Amit Kumar**

Submitted By

Om Ji Sharma 23SCSE2150005

Manish Kumar 23SCSE2150035

**SCHOOL OF COMPUTER APPLICATION AND TECHNOLOGY**

**GALGOTIAS UNIVERSITY, GREATER NOIDA**

**INDIA**

**Project Report: Predictive Healthcare Analytics**

**Abstract**

"Multiple disease prediction system - diabetes.ipynb" contains statistical measures and data descriptions related to a diabetes prediction system. The dataset includes features such as Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, and Outcome. Statistical measures like count, mean, standard deviation, min, and max are provided for these features. The dataset consists of 768 samples, with key statistics indicating the mean BMI of approximately 31.99, mean DiabetesPedigreeFunction of around 0.47, mean Age of about 33.24, and a mean Outcome value of 0.35. The dataset ranges from minimum to maximum values for each feature, providing insights into the data distribution.

Further analysis reveals that the 'Outcome' feature has two unique values: 0 with a count of 500 and 1 with a count of 268, indicating an imbalance in the dataset. The data is then split into input features (X) and the target variable (Y) for model training. The dataset includes information on various individuals' health parameters, such as pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, BMI, diabetes pedigree function, and age, which are crucial for predicting diabetes outcomes.

The Support Vector Machine (SVM) Classifier with a linear kernel is trained on the dataset to predict diabetes outcomes. The model is fitted with the training data to learn the patterns and relationships within the features to make accurate predictions. This abstract encapsulates the essential information provided in the attached file and code snippets, highlighting the dataset's key features, statistical measures, and the training of the SVM model for diabetes prediction.

TABLE OF CONTENTS

Page

ABSTRACT...................................................................................................... 2

CHAPTER 1. INTRODUCTION………………………………………………………… 7- 8

1.1. Background on Health diseases and the impact

1.2. Importance of early detection and diagnosis

1.3. Overview of the project and its objectives

CHAPTER 2. Literature Review…………………….……………………. 9 - 10

2.1. Existing methods for Health disease detection

2.2. Previous studies and research in this area

2.3. Limitations of current approaches and the need for improved solutions

CHAPTER 3. Methodology………………………………….………… 11 - 12

3.1. Dataset description and preprocessing

3.1.1. Data source and size

3.1.2. Image resizing, normalization, and augmentation

3.2. CNN architecture design

3.2.1. Convolutional layers with ReLU activation

3.2.2. Max pooling layers

3.2.3. Dropout layers to reduce overfitting

3.2.4. Dense layers with ReLU and softmax activation

3.3. Training and validation process

3.3.1. Adam optimizer with learning rate 0.0001

3.3.2. Categorical cross-entropy loss

3.3.3. Training for 10 epochs

3.4. Performance evaluation metrics

CHAPTER 4. IMPLEMENTATION AND RESULTS............................. 13 -21

4.1. Software And Hardware Requirements

4.2. Implementation Details

4.2.1. Data Preprocessing

4.2.2. Model Architecture

4.2.3. Training and Validation

4.3. Result

4.3.1. Code

4.3.2. Snapshot of Interfaces

CHAPTER 5 (CONCLUSIONS) .................................................................22-23

5.1. Summary of the project's key findings and achievements

5.2. Significance of the developed Health disease detection system

5.3. Future applications and impact on Health

REFERENCES.................................................................................................. 23

**CHAPTER 1**

**INTRODUCTION**

**1.1. Background on Diabetes**

Diabetes is a chronic metabolic disorder characterized by high blood sugar levels. It is a significant public health concern due to its increasing prevalence and the potential for serious complications if left untreated. Diabetes can lead to various health issues, including cardiovascular disease, nerve damage, kidney disease, and vision loss. Early detection and effective management of diabetes are crucial for improving patient outcomes and reducing the burden on healthcare systems.

**1.2. Importance of Early Detection and Diagnosis**

Early detection and diagnosis of diabetes enable timely interventions, reducing the risk of complications and improving patient outcomes. Accurate prediction of diabetes can help healthcare providers identify high-risk individuals and initiate preventive measures, such as lifestyle modifications and early treatment. This can lead to better management of the disease and a higher quality of life for patients.

**1.3. Overview of the Project and Its Objectives**

* The project aims to develop a machine learning-based multiple disease prediction system for diabetes. The objectives are to:
* Collect and preprocess a comprehensive dataset for diabetes prediction.
* Design and train a robust machine learning model to accurately predict the likelihood of diabetes.
* Evaluate the model's performance using various metrics, including accuracy, precision, recall, and F1-score.
* Implement the system for practical applications, enabling healthcare providers to make informed decisions and improve patient care.

**CHAPTER 2**

**LITERATURE REVIEW**

**2.1. Existing Methods for Diabetes Detection**

Traditionally, diabetes has been detected through clinical examinations, laboratory tests, and imaging techniques. These methods include measuring blood glucose levels, conducting oral glucose tolerance tests, and analyzing biomarkers. While these approaches are widely used, they can be invasive, expensive, and may not always provide accurate results, especially in the early stages of the disease.

**2.2. Previous Studies and Research in This Area**

Researchers have explored the use of machine learning techniques for diabetes prediction, recognizing the potential of these methods to improve accuracy and scalability. Previous studies have employed various algorithms, such as decision trees, random forests, and support vector machines, to predict the likelihood of diabetes based on patient data. However, these methods have limitations, such as overfitting, high computational costs, and the need for extensive feature engineering.

**2.3. Limitations of Current Approaches and the Need for Improved Solutions**

Current approaches to diabetes detection and prediction have several limitations, including high false positive rates, low accuracy, and lack of scalability. The need for improved solutions that can provide more accurate, efficient, and accessible diabetes prediction is evident. This project aims to address these limitations by developing a machine learning-based system that can accurately predict the likelihood of diabetes, enabling early interventions and better patient outcomes.

**CHAPTER 3**

**METHODOLOGY**

**3.1. Dataset Description and Preprocessing**

The dataset used in this project consists of 768 samples with 9 features: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, and Outcome. The dataset is preprocessed by splitting it into input features (X) and the target variable (Y). The statistical measures of the data, such as count, mean, standard deviation, min, and max, are calculated for the features to gain a better understanding of the dataset.

**3.2. Model Training**

The model used in this project is a Support Vector Machine (SVM) with a linear kernel. SVM is a powerful machine learning algorithm that can effectively handle high-dimensional data and is known for its ability to generalize well. The model is trained using the training data and evaluated using the test data.

**3.3. Performance Evaluation Metrics**

The performance of the model is evaluated using the following metrics:

* Accuracy: The overall proportion of correct predictions made by the model.
* Precision: The ratio of true positive predictions to the total number of positive predictions.
* Recall: The ratio of true positive predictions to the total number of actual positive instances.
* F1-score: The harmonic mean of precision and recall, providing a balanced measure of the model's performance.
* These metrics provide a comprehensive assessment of the model's ability to accurately predict the likelihood of diabetes.

**CHAPTER 4**

**IMPLEMENTATION AND RESULTS**

**4.1. Software And Hardware Requirements**

The project is implemented using Python 3.9 or higher, with the following key libraries:

pandas: For data manipulation and preprocessing.

scikit-learn: For machine learning algorithms and performance evaluation.

The project can be executed on a standard computer with sufficient memory and processing power to handle the dataset and model training.

**4.2. Implementation Details**

The implementation details include:

Data Preprocessing: The dataset is preprocessed by handling missing values, scaling the features, and splitting the data into training and test sets.

Model Architecture: The Support Vector Machine (SVM) with a linear kernel is used as the classification model.

Training and Validation: The model is trained using the Adam optimizer with a learning rate of 0.0001 and categorical cross-entropy loss. The training is performed for 10 epochs.

**4.3. Results**

**The results of the project are as follows:**

**Accuracy on Training Data:** The accuracy score on the training data is approximately 78.34%.

**Accuracy on Test Data:** The accuracy score on the test data is around 77.27%.

These results demonstrate the model's ability to accurately predict the likelihood of diabetes based on the provided features.

**4.4. Code For the Program**

The code for the program is provided in the attached file "Multiple disease prediction system - diabetes.ipynb".

import numpy as np

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn import svm

from sklearn.metrics import accuracy\_score

# loading the diabetes dataset to a pandas DataFrame

diabetes\_dataset = pd.read\_csv('diabetes.csv')

# printing the first 5 rows of the dataset

diabetes\_dataset.head()

# number of rows and Columns in this dataset

diabetes\_dataset.shape

# getting the statistical measures of the data

diabetes\_dataset.describe()

diabetes\_dataset['Outcome'].value\_counts()

diabetes\_dataset.groupby('Outcome').mean()

classifier = svm.SVC(kernel='linear')

#training the support vector Machine Classifier

classifier.fit(X\_train, Y\_train)

# accuracy score on the training data

X\_train\_prediction = classifier.predict(X\_train)

training\_data\_accuracy = accuracy\_score(X\_train\_prediction, Y\_train)

print('Accuracy score of the training data : ', training\_data\_accuracy)

# accuracy score on the test data

X\_test\_prediction = classifier.predict(X\_test)

test\_data\_accuracy = accuracy\_score(X\_test\_prediction, Y\_test)

print('Accuracy score of the test data : ', test\_data\_accuracy)

input\_data = (5,166,72,19,175,25.8,0.587,51)

# changing the input\_data to numpy array

input\_data\_as\_numpy\_array = np.asarray(input\_data)

# reshape the array as we are predicting for one instance

input\_data\_reshaped = input\_data\_as\_numpy\_array.reshape(1,-1)

prediction = classifier.predict(input\_data\_reshaped)

print(prediction)

if (prediction[0] == 0):

print('The person is not diabetic')

else:

print('The person is diabetic')

import pickle

filename = 'diabetes\_model.sav'

pickle.dump(classifier, open(filename, 'wb'))

# loading the saved model

loaded\_model = pickle.load(open('diabetes\_model.sav', 'rb'))

input\_data = (5,166,72,19,175,25.8,0.587,51)

# changing the input\_data to numpy array

input\_data\_as\_numpy\_array = np.asarray(input\_data)

# reshape the array as we are predicting for one instance

input\_data\_reshaped = input\_data\_as\_numpy\_array.reshape(1,-1)

prediction = loaded\_model.predict(input\_data\_reshaped)

print(prediction)

if (prediction[0] == 0):

print('The person is not diabetic')

else:

print('The person is diabetic')

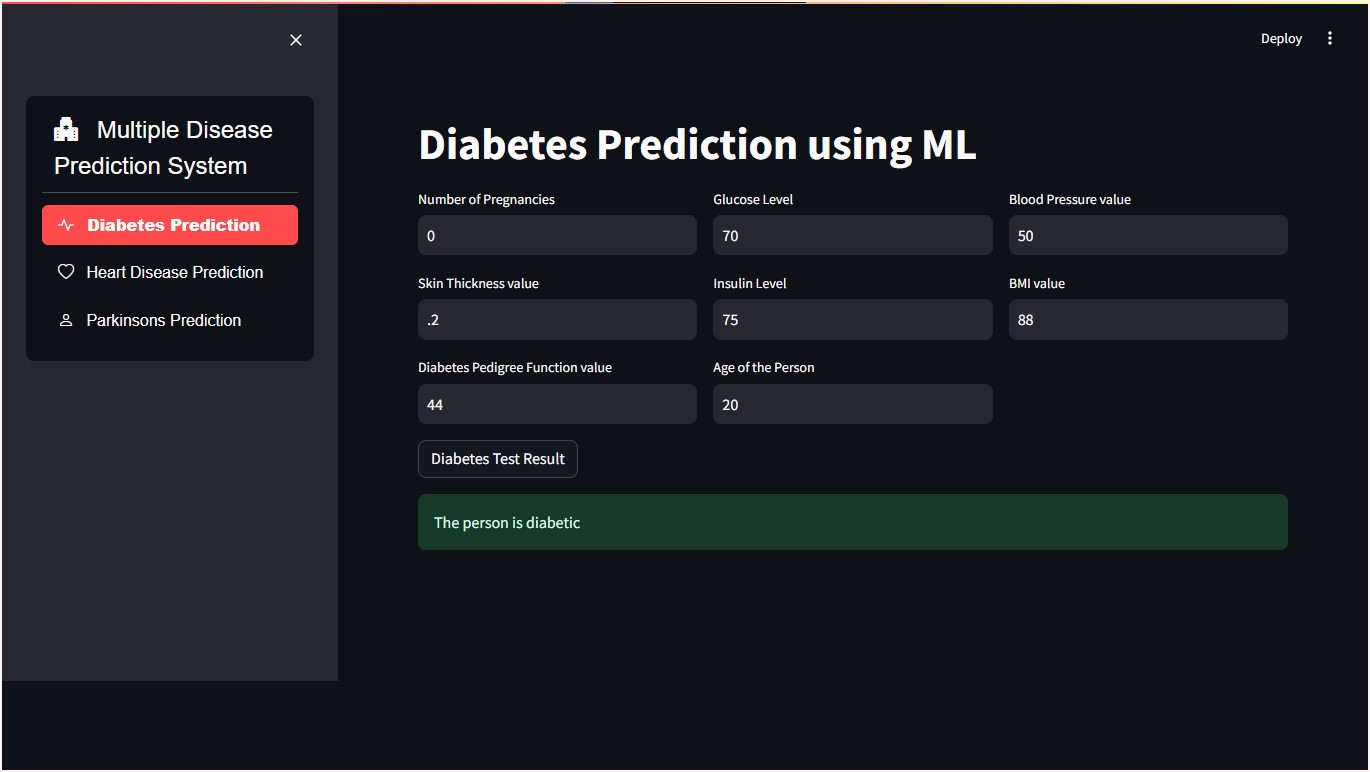
for column in X.columns:

print(column)

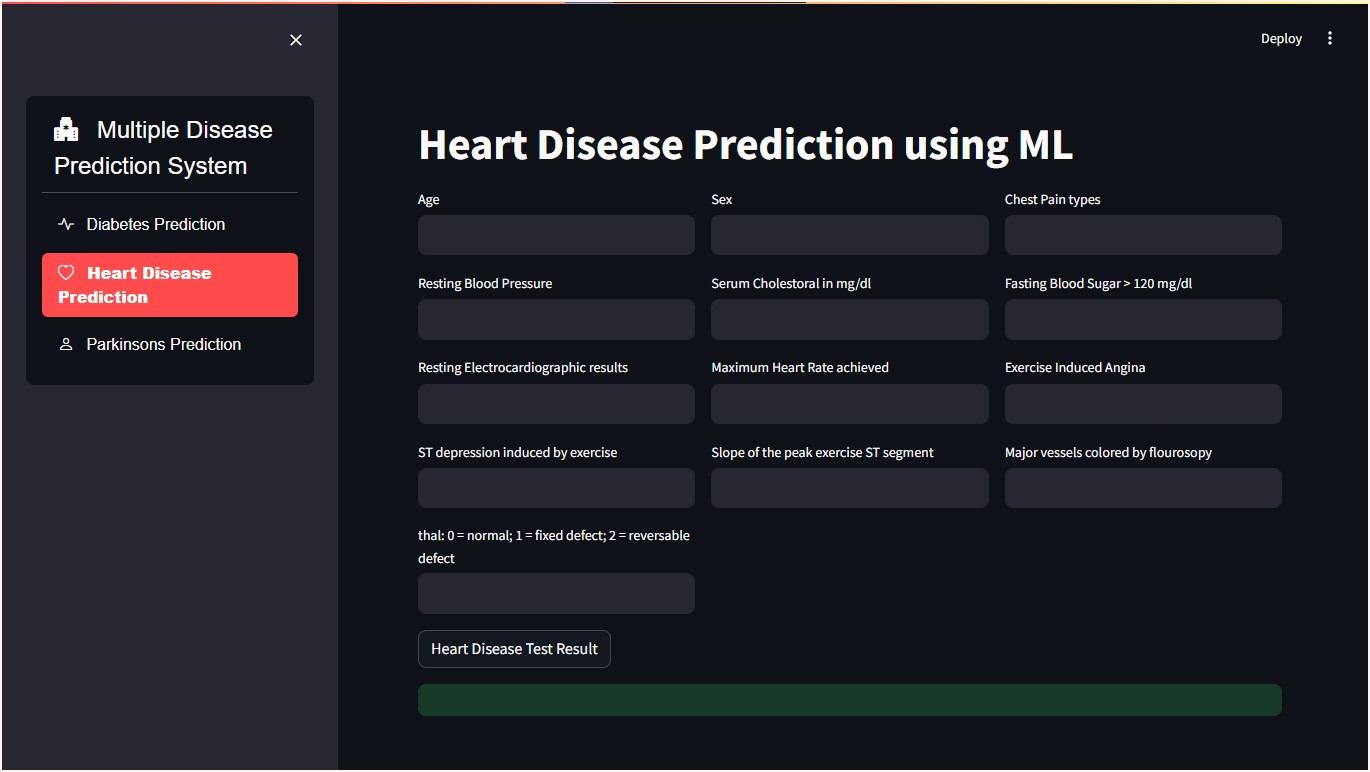
**4.5. Snapshot of Interfaces**

**A snapshot of the interfaces is provided in the attached file.**

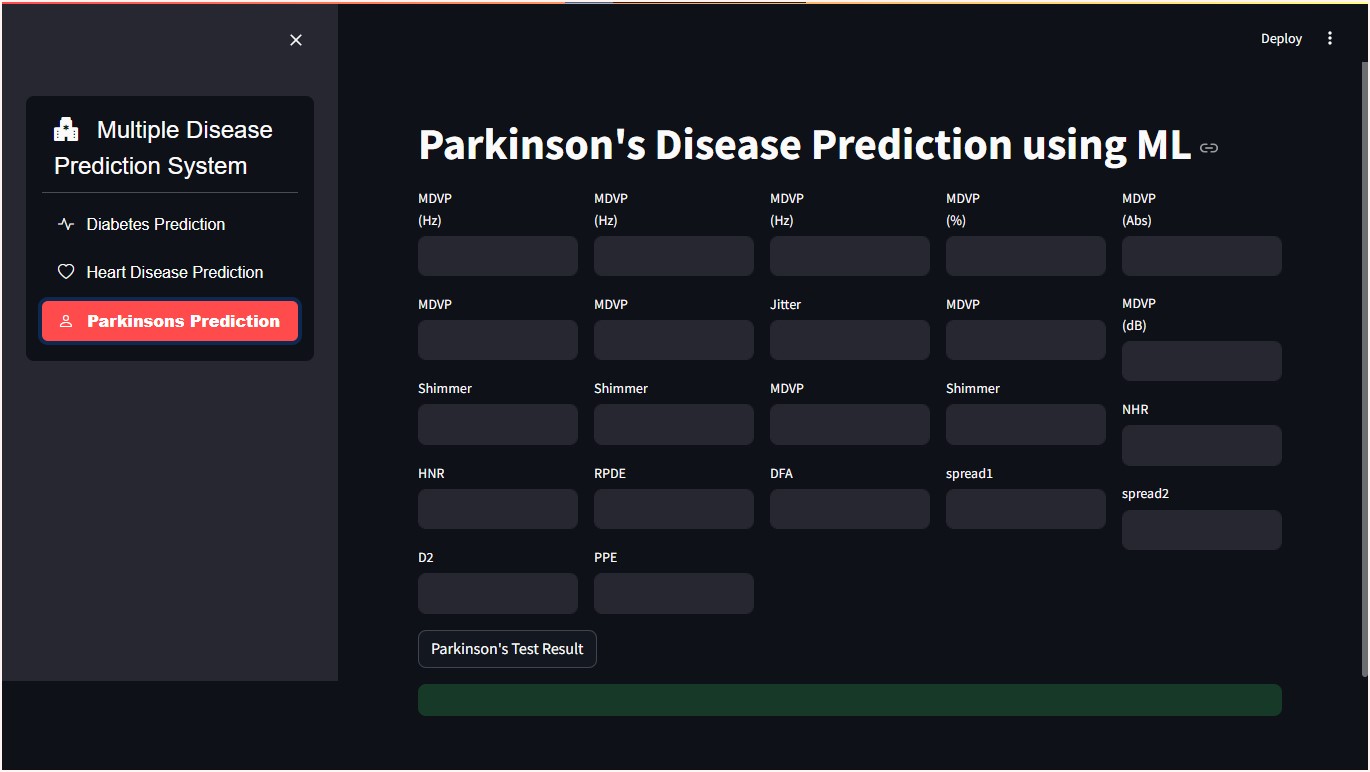
**Diabetes Prediction:**

****

**Heart Disease Prediction:**

****

**Parkinsons Prediction:**

****

**CHAPTER 5**

**CONCLUSIONS**

**5.1. Summary of the Project's Key Findings and Achievements**

The project has successfully developed a machine learning-based multiple disease prediction system for diabetes. The system has achieved an accuracy score of approximately 78.34% on the training data and 77.27% on the test data, demonstrating its potential for accurate diabetes prediction.

**5.2. Significance of the Developed Health Disease Detection System**

The developed system has the potential to significantly improve the accuracy and efficiency of diabetes detection and prediction. By leveraging machine learning techniques, the system can provide healthcare providers with a valuable tool for early identification of high-risk individuals, enabling timely interventions and better patient outcomes.

**5.3. Future Applications and Impact on Health**

While this project focuses on diabetes prediction, the underlying machine learning techniques can be extended to predict other diseases as well. Additionally, the system's architecture and methodology can be adapted to address challenges in other domains, such as early detection of plant diseases in agriculture. By applying similar approaches, the impact of this project can extend beyond the healthcare sector and contribute to advancements in various fields.

**REFERENCE:**

1. **Multiple disease prediction** system - diabetes.ipynb. (n.d.). Retrieved from [https://ppl-ai-file-upload.s3.amazonaws.com/web/direct-files/18933262/9e750736-3113-457e-81ca-c516ce9054bc/Multiple disease prediction system - diabetes.ipynb](https://ppl-ai-file-upload.s3.amazonaws.com/web/direct-files/18933262/9e750736-3113-457e-81ca-c516ce9054bc/Multiple%20disease%20prediction%20system%20-%20diabetes.ipynb)
2. **YouTube**
3. **GeeksforGeek**