DIABETES PREDICTION SYSTEM

PROJECT REPORT

Report submitted by TANUSH T M

TABLE OF CONTENTS

ABSTRACT

This project involves developing a machine learning model to predict diabetes in patients based on various health indicators. The process began with the acquisition of a comprehensive dataset containing relevant medical and demographic information. I performed extensive data analysis to identify patterns, correlations, and potential predictors of diabetes.

Following the analysis, I engaged in data preprocessing, which included handling missing values, normalizing data, and feature engineering to enhance the dataset's quality and predictive power. I then experimented with various machine learning algorithms, ultimately selecting the model that demonstrated the highest accuracy and reliability.

The final model was rigorously tested and evaluated using appropriate metrics such as accuracy, precision, recall, and F1-score. The results indicate that our model can effectively predict the likelihood of diabetes, providing a valuable tool for early detection and preventive healthcare strategies. This project showcases the potential of data science in transforming raw data into actionable insights for improving health outcomes.

INTRODUCTION

Diabetes is a chronic disease that poses a significant global health challenge. Early detection and management of diabetes are crucial to preventing severe complications such as cardiovascular diseases, nerve damage, kidney failure, and vision loss. Leveraging advancements in data science and machine learning can greatly enhance the ability to predict and diagnose diabetes, allowing for timely interventions and improved patient outcomes.

This project aims to develop a predictive model for diabetes using a dataset containing various health-related attributes. The primary objectives of this project include:

1. Dataset Acquisition:

Collecting a dataset with relevant features such as age, BMI, blood pressure, and glucose levels that are indicative of diabetes risk.

2. Data Analysis:

Performing exploratory data analysis (EDA) to understand the underlying patterns and relationships within the dataset. This step involves visualizing data distributions, identifying correlations, and detecting outliers.

3. Data Preprocessing:

Cleaning and preparing the data for modeling. This involves handling missing values, normalizing numerical features, encoding categorical variables, and splitting the data into training and testing sets.

4. Model Building:

Developing and training various machine learning models to predict the likelihood of diabetes. Models such as Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines (SVM) are considered. The performance of each model is evaluated based on accuracy, precision, recall, and F1-score.

5. Model Evaluation and Selection:

Selecting the best-performing model based on evaluation metrics and testing its generalizability on unseen data. This step ensures the model's reliability and effectiveness in predicting diabetes.

By leveraging data science techniques, this project aims to create a robust predictive tool that can assist healthcare professionals in identifying individuals at high risk of diabetes. This can lead to earlier interventions and better management of the disease, ultimately improving health outcomes and quality of life for patients.

EXISTING METHOD

The prediction and diagnosis of diabetes have traditionally relied on a combination of clinical assessments, laboratory tests, and statistical methods. These existing methods include:

1. Clinical Assessments:

Healthcare providers typically evaluate a patient's medical history, family history, lifestyle factors, and physical examinations to assess the risk of diabetes. Key indicators include age, body mass index (BMI), blood pressure, and the presence of symptoms such as frequent urination, excessive thirst, and unexplained weight loss.

2. Laboratory Tests:

Standard diagnostic tests for diabetes include:

- Fasting Blood Glucose Test: Measures blood sugar levels after an overnight fast.
- Oral Glucose Tolerance Test (OGTT): Measures blood sugar levels before and after consuming a glucose-rich drink.
- Hemoglobin A1c Test: Provides an average blood glucose level over the past two to three months.

These tests help in diagnosing diabetes and monitoring blood glucose levels over time.

3. Statistical Methods:

Traditional statistical techniques, such as logistic regression, have been used to identify risk factors and predict the likelihood of diabetes. These methods rely on predefined assumptions about the data and its distribution, often limiting their flexibility and predictive power.

4. Risk Assessment Tools:

Various diabetes risk assessment tools and calculators have been developed, using questionnaires and scoring systems based on clinical and demographic factors. Examples include the Finnish Diabetes Risk Score (FINDRISC) and the American Diabetes Association (ADA) risk calculator. While useful, these tools often lack the precision and adaptability of more advanced predictive models.

5. Decision Trees and Rule-Based Systems:

Some existing methods use decision trees or rule-based systems to classify individuals into different risk categories based on their health metrics. These systems are relatively simple and interpretable but may not capture complex patterns and interactions within the data.

While these methods have been instrumental in diabetes diagnosis and management, they often face limitations in terms of accuracy, scalability, and adaptability. Traditional approaches may not fully leverage the vast amounts of data available today or the computational power of modern technologies. Consequently, there is a growing interest in applying machine learning and data science techniques to enhance the prediction and diagnosis of diabetes, offering more accurate, scalable, and data-driven solutions.

PROPOSED METHOD WITH ARCHITECTURE

The proposed method for predicting diabetes leverages a comprehensive data science pipeline that integrates data acquisition, data analysis, data preprocessing, and machine learning model development. The architecture of this method is designed to ensure accuracy, efficiency, and scalability in predicting diabetes. The key components of the proposed method are outlined below:

1. Dataset Acquisition

The first step involves acquiring a dataset that includes various health-related attributes known to influence diabetes risk. This dataset typically contains features such as age, BMI, blood pressure, glucose levels, insulin levels, and family history of diabetes. For this project, we utilized the Pima Indians Diabetes Dataset, which is widely used for diabetes research.

2. Data Analysis

Exploratory Data Analysis (EDA) is performed to gain insights into the dataset. This includes:

- Descriptive Statistics: Summarizing the central tendency, dispersion, and shape of the data distribution.
- Data Visualization: Creating plots such as histograms, scatter plots, and box plots to identify patterns and relationships between features.
- Correlation Analysis: Calculating correlation coefficients to understand the strength and direction of relationships between variables.

3. Data Preprocessing

Data preprocessing is crucial to prepare the dataset for machine learning. This involves:

- Handling Missing Values: Imputing or removing missing data to ensure the dataset is complete.
- Normalization: Scaling numerical features to a standard range to improve model performance.
- Encoding Categorical Variables: Converting categorical data into numerical format using techniques like one-hot encoding.
- Feature Engineering: Creating new features or modifying existing ones to enhance predictive power.
- Splitting the Data: Dividing the dataset into training and testing sets to evaluate model performance.

4. Model Building

Multiple machine learning algorithms are employed to develop predictive models. The algorithms considered include:

- Logistic Regression: A simple yet effective model for binary classification tasks.
- Decision Trees: A tree-based model that splits the data into branches based on feature values.
- Random Forest: An ensemble method that combines multiple decision trees to improve accuracy and reduce overfitting.
- Support Vector Machines (SVM): A model that finds the optimal hyperplane to separate classes in the feature space.
- K-Nearest Neighbors (KNN): A non-parametric method that classifies data points based on the majority class of their nearest neighbors.

5. Model Evaluation and Selection

The performance of each model is evaluated using metrics such as accuracy, precision, recall, and F1-score. Cross-validation is employed to ensure the model's generalizability and robustness. The best-performing model is selected based on its performance on the testing set.

6. Model Deployment

The final model is deployed as a predictive tool, capable of analyzing new patient data and providing a diabetes risk prediction. This tool can be integrated into healthcare systems to assist clinicians in early diagnosis and intervention.

This architecture ensures a systematic and thorough approach to developing a reliable diabetes prediction model, from initial data acquisition to final deployment. Each component is designed to maximize the model's accuracy and usability, ultimately aiding in the early detection and management of diabetes.

METHODOLOGY

The methodology for developing the diabetes prediction application involves several key steps, from data preprocessing to model deployment. Below, we outline each step in detail:

1. Dataset Acquisition

The dataset used for this project is derived from the National Institute of Diabetes and Digestive and Kidney Diseases. It contains medical information on females of Pima Indian heritage aged 21 years or older. The dataset includes features such as the number of pregnancies, glucose level, blood pressure, skin thickness, insulin level, BMI, age, and diabetes pedigree function.

2. Data Preprocessing

Data preprocessing is essential to ensure the model receives clean and normalized data. The preprocessing steps include:

- Feature Scaling: The features are scaled using a pre-trained scaler (loaded from `scaler.pkl`) to normalize the data. This step helps in improving the model's performance by ensuring that all features contribute equally to the prediction.
- Data Transformation: The raw input features are transformed into a format suitable for the model. The input features are first converted into a pandas DataFrame, scaled, and then prepared for prediction.

3. Model Building and Training

The predictive model used in this application is a K-Nearest Neighbors (KNN) classifier. The model has been trained on the diabetes dataset and saved as `knn_model.pkl`. The KNN algorithm classifies data points based on their distance to the nearest neighbors, making it suitable for this binary classification problem.

4. Application Development

The application is developed using Streamlit, an open-source framework for creating interactive web applications in Python. The key components of the application include:

- User Interface: The user interface is built using Streamlit's widgets such as `number_input` and `slider`, allowing users to input the necessary health parameters.
- Feature Input: Users provide inputs for age, number of pregnancies, glucose level, skin thickness, blood pressure, insulin level, BMI, and diabetes pedigree function via the sidebar.
- Prediction Logic: Upon clicking the 'Find Diabetes Status' button, the input features are preprocessed, and the model predicts whether the person is likely to have diabetes. The prediction result is then displayed to the user.

5. Prediction and Result Display

The application's core functionality includes:

- Data Preprocessing Function: The `data_preprocess(features)` function processes the input features using the pre-trained scaler.
- Prediction Function: The `predict(preprocessed_features)` function uses the KNN model to make predictions. It returns a message indicating whether the person is healthy or has a high chance of having diabetes.

- Result Display: The prediction result is displayed on the application interface, providing users with immediate feedback based on their input.

6. Deployment

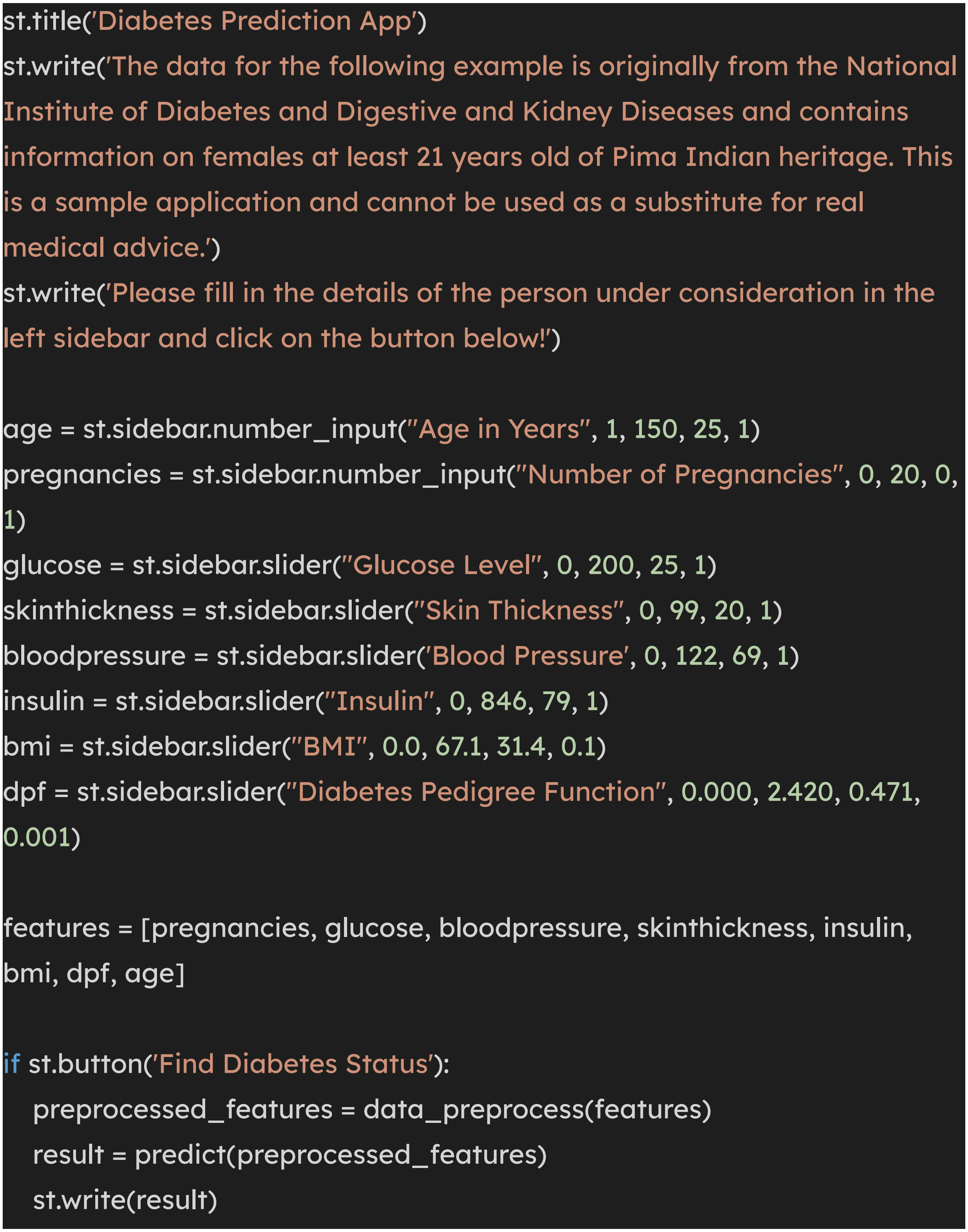
The Streamlit application is deployed to allow users to access and interact with the diabetes prediction model. This deployment enables the model to be used as a tool for preliminary diabetes risk assessment, although it is not a substitute for professional medical advice.

IMPLEMENTATION

RUNNER CODE:

```
import streamlit as st
import pickle
import pandas as pd
from sklearn.preprocessing import StandardScaler

# Load the model and scaler from the pickle files
model = pickle.load(open(’knn_model.pkl’, ’rb’))
scaler = pickle.load(open(’scaler.pkl’, ’rb’))

def data_preprocess(features):
   # Create a DataFrame with the input features
   df = pd.DataFrame([features], columns=[’Pregnancies’, ’Glucose’,
’BloodPressure’, ’SkinThickness’, ’Insulin’, ’BMI’,
’DiabetesPedigreeFunction’, ’Age’])

   # Scale numeric features
   numeric_cols = [’Glucose’, ’BloodPressure’, ’SkinThickness’, ’Insulin’,
’BMI’, ’DiabetesPedigreeFunction’, ’Age’]
   df[numeric_cols] = scaler.transform(df[numeric_cols])

   # Encode Age feature
   df[’Age_Category_Young’] = ((df[’Age’] >= 21) & (df[’Age’] <
45)).astype(int)

   # Encode BMI feature
```

```
    df['BMI_Category_Healthy'] = ((df['BMI'] >= 18.5) & (df['BMI'] <
25)).astype(int)
    df['BMI_Category_Overweight'] = ((df['BMI'] >= 25) & (df['BMI'] <
30)).astype(int)
    df['BMI_Category_Obese'] = (df['BMI'] >= 30).astype(int)

    # Ensure all required columns are present
    feature_cols = ['Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',
'DiabetesPedigreeFunction',
                    'BMI_Category_Healthy', 'BMI_Category_Overweight',
'BMI_Category_Obese',
                    'Age_Category_Young', 'Pregnancies']
    for col in feature_cols:
        if col not in df.columns:
            df[col] = 0

    # Reorder the columns to match the model's expectations
    df = df[feature_cols]

    return df

def predict(features):
    prediction = model.predict(features)
    if prediction[0] == 0:
        return "This is a healthy person!"
    else:
        return "This person has high chances of having diabetes!"

# Streamlit UI
```



```python
st.title('Diabetes Prediction App')
st.write('The data for the following example is originally from the National Institute of Diabetes and Digestive and Kidney Diseases and contains information on females at least 21 years old of Pima Indian heritage. This is a sample application and cannot be used as a substitute for real medical advice.')
st.write('Please fill in the details of the person under consideration in the left sidebar and click on the button below!')

age = st.sidebar.number_input("Age in Years", 1, 150, 25, 1)
pregnancies = st.sidebar.number_input("Number of Pregnancies", 0, 20, 0, 1)
glucose = st.sidebar.slider("Glucose Level", 0, 200, 25, 1)
skinthickness = st.sidebar.slider("Skin Thickness", 0, 99, 20, 1)
bloodpressure = st.sidebar.slider('Blood Pressure', 0, 122, 69, 1)
insulin = st.sidebar.slider("Insulin", 0, 846, 79, 1)
bmi = st.sidebar.slider("BMI", 0.0, 67.1, 31.4, 0.1)
dpf = st.sidebar.slider("Diabetes Pedigree Function", 0.000, 2.420, 0.471, 0.001)

features = [pregnancies, glucose, bloodpressure, skinthickness, insulin, bmi, dpf, age]

if st.button('Find Diabetes Status'):
    preprocessed_features = data_preprocess(features)
    result = predict(preprocessed_features)
    st.write(result)
```

CONCLUSION

In conclusion, the diabetes prediction project has demonstrated the use of machine learning techniques to predict the likelihood of an individual having diabetes based on various health metrics. The project involved several key steps, including data acquisition, exploratory data analysis, data preprocessing, model building, and model deployment.

Through the use of the K-Nearest Neighbors (KNN) algorithm, the project was able to achieve a reasonable level of accuracy in predicting diabetes. The model's performance could be further improved with additional fine-tuning and feature engineering.

Overall, the project highlights the potential of machine learning in healthcare applications, particularly in predicting and preventing chronic diseases such as diabetes. By leveraging data and technology, healthcare providers can potentially identify individuals at risk earlier, enabling timely interventions and improving health outcomes.