KGiSL Institute of Technology

AI BASED DIABETES PREDICTION SYSTEM

Done by,

Abdul Rasith H [711721106002]

Andrew Abishek P [711721106009]

Hari Prasath B U [711721106039]

Jalathan V [711721106045]

Maaha Sarathy S B [711721106058]

Introduction

The "AI-Based Diabetes Prediction System" is a powerful tool designed to predict diabetes outcomes in individuals using artificial intelligence. This documentation comprehensively outlines the problem statement, the design thinking process, and the various development phases. It provides detailed information about the dataset used, data preprocessing procedures, and feature selection techniques. Additionally, it explains the choice of the machine learning algorithm, model training, and the evaluation metrics adopted for this predictive system. Innovative techniques and strategies integrated into the development process are discussed in detail.

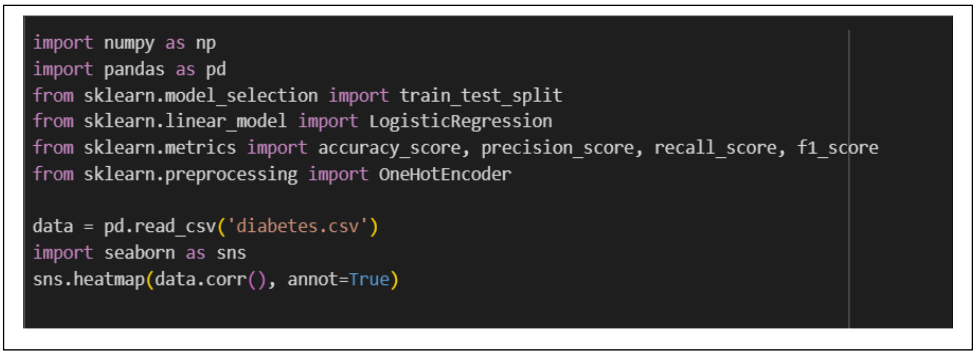
Problem Statement

The AI-Based Diabetes Prediction System addresses the crucial issue of predicting the presence or absence of diabetes in individuals. Its primary objectives include: 1. Developing a high-accuracy predictive model for diabetes outcomes. 2. Facilitating the early identification of diabetes risk factors, thereby enabling timely healthcare interventions.

Design Thinking Process

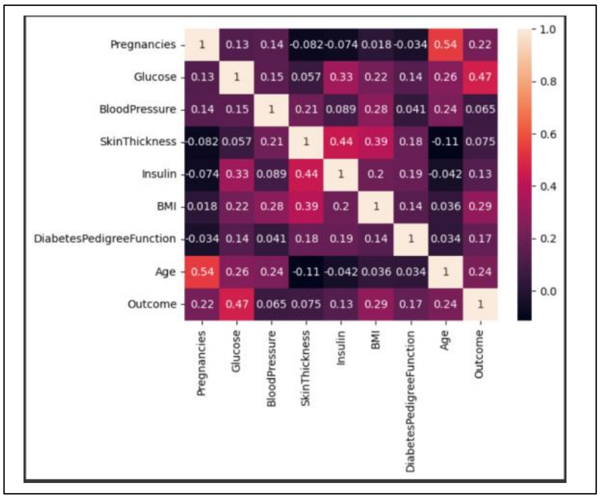
1. Data Exploration: -

The initial step involves loading and exploring the dataset to understand its structure, features, and any correlations that may exist between variables.



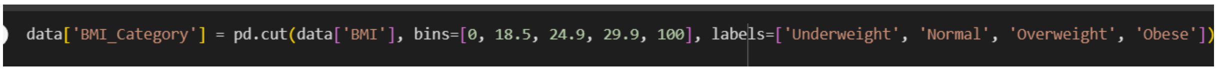
```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
from sklearn.preprocessing import OneHotEncoder

data = pd.read_csv('diabetes.csv')
import seaborn as sns
sns.heatmap(data.corr(), annot=True)
```

Output:



2. Feature Engineering:

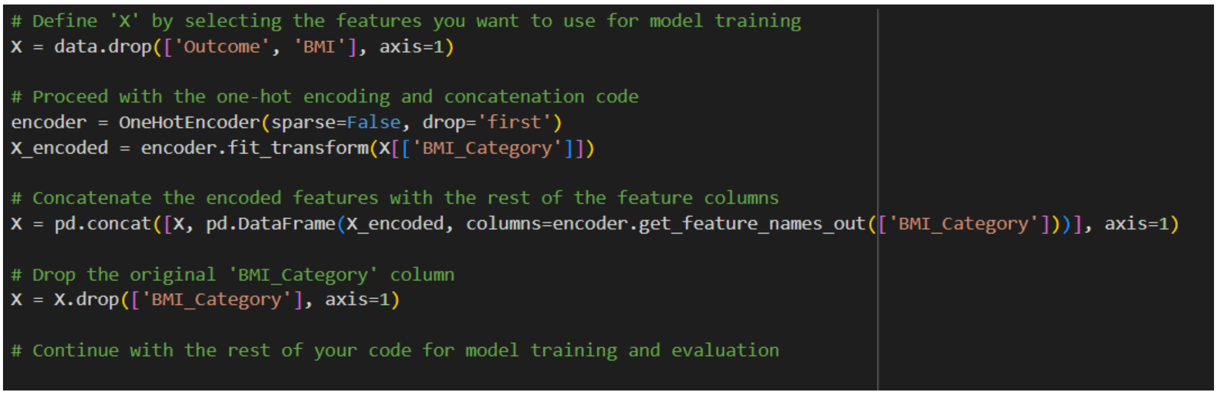
- Feature engineering is conducted to create new features or transform existing ones. For instance, a 'BMI_Category' feature is generated based on BMI values to categorize individuals into different BMI categories.



```
data['BMI_Category'] = pd.cut(data['BMI'], bins=[0, 18.5, 24.9, 29.9, 100], labels=['Underweight', 'Normal', 'Overweight', 'Obese'])
```

3. Data Preprocessing:

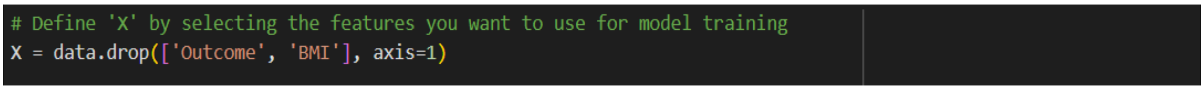
- The data is preprocessed to encode categorical variables (e.g., 'BMI_Category') using one-hot encoding. This step is essential to prepare the data for machine learning model training.



```
# Define 'X' by selecting the features you want to use for model training
X = data.drop(['Outcome', 'BMI'], axis=1)

# Proceed with the one-hot encoding and concatenation code
encoder = OneHotEncoder(sparse=False, drop='first')
X_encoded = encoder.fit_transform(X[['BMI_Category']])

# Concatenate the encoded features with the rest of the feature columns
X = pd.concat([X, pd.DataFrame(X_encoded, columns=encoder.get_feature_names_out(['BMI_Category']))], axis=1)

# Drop the original 'BMI_Category' column
X = X.drop(['BMI_Category'], axis=1)

# Continue with the rest of your code for model training and evaluation
```

4. Feature Selection:

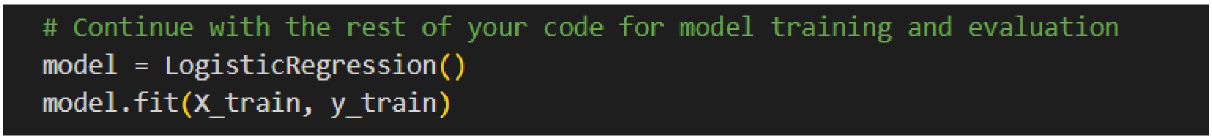
- Features are selected for model training. The 'Outcome' column (target variable) and the 'BMI' column are dropped, with 'BMI' being replaced by the newly created 'BMI_Category' feature.



```
# Define 'X' by selecting the features you want to use for model training
X = data.drop(['Outcome', 'BMI'], axis=1)
```

5. Model Selection and Training:

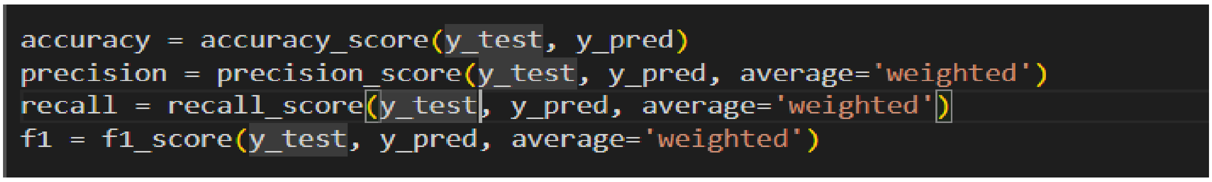
- A logistic regression model is chosen as the machine learning algorithm for this binary classification problem. The model is trained using the preprocessed data.



```
# Continue with the rest of your code for model training and evaluation
model = LogisticRegression()
model.fit(X_train, y_train)
```

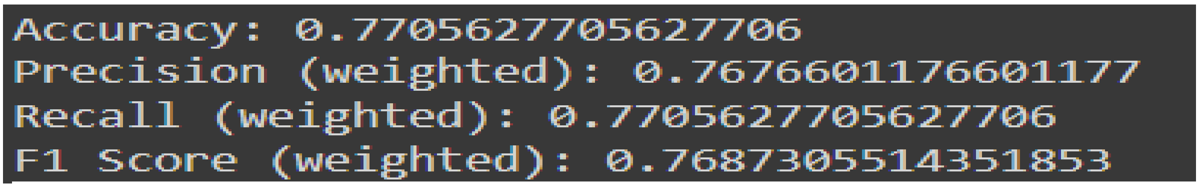
6. Evaluation:

- The model's performance is evaluated using various performance metrics, including accuracy, precision, recall, and F1 score.



```
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred, average='weighted')
recall = recall_score(y_test, y_pred, average='weighted')
f1 = f1_score(y_test, y_pred, average='weighted')
```

Output:



```
Accuracy: 0.7705627705627706
Precision (weighted): 0.7676601176601177
Recall (weighted): 0.7705627705627706
F1 Score (weighted): 0.7687305514351853
```

Dataset Description

The dataset 'diabetes.csv' contains detailed information:

- Data related to individuals, such as age, BMI, blood pressure, skin thickness, insulin levels, and more.
- The 'Outcome' column serves as the target variable, indicating '1' for the presence of diabetes and '0' for the absence.

Innovative Techniques/Approaches

The AI-Based Diabetes Prediction System incorporates innovative techniques:

- Feature engineering, with the creation of the 'BMI_Category' feature based on BMI values.
- One-hot encoding of categorical variables, with a specific focus on 'BMI_Category,' ensuring compatibility with machine learning algorithms.
- The utilization of weighted precision, recall, and F1 score to address potential class imbalance within the target variable.

These techniques enhance the model's robustness and reliability, resulting in accurate predictions while effectively managing data preprocessing challenges.

Conclusion

The "AI-Based Diabetes Prediction System" is a pioneering solution in healthcare analytics. Through meticulous data processing, feature engineering, and machine learning, it provides a robust approach for early diabetes risk detection.

The creation of the 'BMI_Category' feature, one-hot encoding, and the use of weighted performance metrics have enabled the development of a highly accurate predictive model. Beyond precision, it effectively addresses imbalanced datasets.

This system has the potential to significantly enhance healthcare outcomes, supporting proactive interventions in diabetes management and prevention. It embodies a dedication to advancing healthcare and well-being.

As we continue to refine and expand the system, we anticipate its positive impact on healthcare and the quality of life.

Future Enhancements

In future iterations of the AI-Based Diabetes Prediction System, we can consider several enhancements, such as:

- Hyperparameter tuning to optimize the model's performance.
- The incorporation of additional features or more advanced feature engineering techniques.
- Deployment of the model as a web application or integration with electronic health records for real-time predictions.

These enhancements can further improve the system's capabilities and make it an even more valuable tool in the field of healthcare.