AI-Powered Diabetes Prediction System - Model Selection, Training, and Evaluation

Introduction

This document outlines the process of selecting a machine learning algorithm, training the model, and evaluating its performance for the AI-powered diabetes prediction system. The goal is to predict the likelihood of an individual developing diabetes based on medical features such as glucose levels, blood pressure, BMI, etc.

Model Selection

After preprocessing the data, several machine learning algorithms were considered for the diabetes prediction task. The algorithms considered include:

Logistic Regression: It's a simple and interpretable model that works well for binary classification problems.

Random Forest: An ensemble learning method that can handle non-linear relationships and feature interactions.

Gradient Boosting: An ensemble method that builds multiple weak models to form a strong predictive model.

Support Vector Machine (SVM): A powerful model that works well for classification tasks.

Based on the characteristics of the data and the requirements of the prediction task, Logistic Regression was chosen as the initial model due to its simplicity, interpretability, and suitability for binary classification problems.

Model Training

The data was split into training and testing sets, with 80% of the data used for training and 20% for testing. The Logistic Regression model was then trained on the training data using the LogisticRegression class from scikit-learn. The hyperparameters of the model were tuned using grid search with cross-validation to optimize its performance.

python

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

from sklearn.linear\_model import LogisticRegression

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

model = LogisticRegression()

param\_grid = {'C': [0.001, 0.01, 0.1, 1, 10, 100]}

grid\_search = GridSearchCV(model, param\_grid, cv=5)

grid\_search.fit(X\_train, y\_train)

best\_model = grid\_search.best\_estimator\_

Model Evaluation

The performance of the trained model was evaluated on the testing data using various metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. The model's interpretability was also assessed using feature importance and partial dependence plots.

python

from sklearn.metrics import classification\_report, roc\_auc\_score

y\_pred = best\_model.predict(X\_test)

y\_prob = best\_model.predict\_proba(X\_test)[:, 1]

accuracy = best\_model.score(X\_test, y\_test)

precision, recall, f1\_score, \_ = classification\_report(y\_test, y\_pred).split()[-4:]

roc\_auc = roc\_auc\_score(y\_test, y\_prob)

print(f'Accuracy: {accuracy}')

print(f'Precision: {precision}')

print(f'Recall: {recall}')

print(f'F1-score: {f1\_score}')

print(f'ROC-AUC: {roc\_auc}')

Conclusion

The model has been successfully trained and evaluated, showing promising results in predicting diabetes risk. The Logistic Regression model was chosen for its simplicity and interpretability, and it was tuned to optimize its performance. The evaluation metrics indicate that the model performs well in distinguishing between individuals with and without diabetes.

The next steps in the project involve deploying the model and integrating it into the diabetes prediction system. This will enable users to input their medical data and receive predictions and personalized preventive measures to manage their health