Diabetes prediction using ML pipeline

PALAMANICKAM S

INTRODUCTION

Diabetes is a chronic medical condition characterized by high levels of blood glucose, which can lead to serious health complications if left unmanaged. Early detection and management of diabetes are crucial to prevent its adverse effects on health. Machine learning (ML) offers a powerful set of tools for predicting the likelihood of diabetes in patients based on various medical and demographic features.

In this project, we leverage a comprehensive diabetes dataset to develop a predictive model using a machine learning pipeline. The dataset includes vital features such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level.

I have worked with hyperparameter tuning in various algorithm , Optimizing model parameters using techniques such as grid search and randomized search to enhance model accuracy.

Table: this table contain the best model and its parameter with test score of each

S.no	Supervised learning Algorithm	Best Vectorization type	Best Parameter used	Test score for optimized parameter
1.	Random Forest	MinMaxScaler()	RandomForestClassifier (n_estimators=50)	0.9146
2.	Decision Tree	StandardScaler()	DecisionTreeClassifier (max_depth=5)	0.9146
3.	Naïve bayes	StandardScaler()	GaussianNB()	0.8727
4.	SVM (SGDclassifier)	StandardScaler()	SGDClassifier (alpha=0.001, l1_ratio=0.75)	0.95885
5.	Logistic regression (sgdclassifier)	StandardScaler()	SGDClassifier (alpha=0.001, l1_ratio=0.75, loss='log_loss', max_iter=2000, penalty='l1')	0.95865
6.	KNN	MinMaxScaler()	KNeighborsClassifier(algorithm='kd_tree', n_neighbors=8, p=1)	0.9575

CONCLUSION

After extensive hyperparameter tuning and evaluation of various machine learning algorithms, the best performing models for diabetes prediction were identified. Specifically, the Support Vector Classifier (SVC) and the logistic regression model trained with Stochastic Gradient Descent (SGD) emerged as top contenders.

1. Support Vector Classifier (SVC):

- Pipeline Configuration: StandardScaler for feature scaling and SGDClassifier with alpha=0.001 and l1_ratio=0.75.

- Test Score: 0.95885

2. Logistic Regression with SGD:

- Pipeline Configuration: StandardScaler for feature scaling and SGDClassifier with alpha=0.001, l1_ratio=0.75, loss='log_loss', max_iter=2000, and penalty='l1'.

- Test Score: 0.95865

Both models demonstrated high accuracy in predicting diabetes, with the SVC slightly outperforming the logistic regression model. These results highlight the effectiveness of using a well-structured machine learning pipeline for developing predictive models in healthcare. The slight edge of the SVC model suggests its suitability for this specific task, but both models are robust options for practical implementation.

For more Insights and analysis visit my Github repository

Github Repository link: https://github.com/Palamanickam0806/Diabetes_prediction_ML_pipeline

Dataset: link

LinkedIn account: https://www.linkedin.com/in/palamanickam-s-2ab81925b/recent-activity/all/

Email: Valliaravind@gmail.com