Interpretability and Explainability in AI

(DSAI 305) Final Project

ExplainAI

Sama Mohamed - 202201867
Bosy Ayman - 202202076
Zeyad Sherif - 202201220

Research Gap and Project Aim

Explainable AI has significantly improved in healthcare filed but there is still a gap in the combination of interpretability and performance of machine learning models. Our research aims to fill this gap by using 9 different models, applying them on one dataset to be able to see the difference of performance on using each model, and evaluating the models by using interpretability and explainability frameworks.

Goal

Combine model accuracy with interpretability to improve the adoption of ML in healthcare.

Impact

Provide insights for the used models aiming to balance performance and transparency.

Dataset: Pima Indian Diabetes Dataset:

It consists of 9 columns, which are pregnancies, glucose, blood pressure, skin thickness, Insulin, BMI, diabetes pedigree function, age, and Outcome

Data Preprocessing and Exploration

1

Data Loading & Cleaning

Dataset from Kaggle. Handle nan values, and zeros in key columns replaced with median values to handle missing data realistically.

2

Outlier Detection

Using boxplots to reveal skewness and outliers aiming to data validation.

3

Exploratory Analysis

Implementing Univariate and multivariate analysis on the data as correlation matrix and pair plots indicates the strong feature relationships which is important for diabetes prediction.

4

Train-Test Split

Data split 80/20 for training and testing to ensure model generalization while testing on the 20% of unseen data.

Feature Selection Techniques

Fisher’s Score & Correlation Coefficient

They work on Selecting features that are strongly related to diabetes outcome using ANOVA and Pearson correlation.

Variance Threshold & Multicollinearity

They work on removing low variance features and detecting possible multicollinearity to avoid redundant features.

Chi-Square & Backward Elimination

They work on assessing the categorical feature independence and iteratively removes insignificant features for the models.

Machine Learning Models, and its interpretability

Logistic Regression

It is used for diabetes classification with 64% accuracy on the Pima Indians dataset.
It was interpreted using the 6 assumptions of linear models. The assumptions show that the data has an appropriate outcome type, the VIF value was more than 10, so there was multicollinearity. Also, it shows that the data is sufficiently large.

Neural Networks

It Captures complex nonlinear relations, achieving 68% accuracy.

It was interpreted and explained using the global and local agnostic models. The results indicate that Glucose, and BMI are the most affecting 2 features on the diabetes predictions. Also, it shows that feature like skin thickness have less effect on the predictions.

XGBoost

It was used as it is efficient for high-dimensional data, and it achieved 71% accuracy.
It was interpreted and explained using the global and local agnostic models. The results also indicate the importance of Glucose and BMI features compared to Skin thickness and Blood pressure.

Machine Learning Models, and its interpretability

K-Nearest Neighbors (KNN)

It Classifies based on the nearest to neighbors using the input features like glucose and BMI. It achieved an accuracy of 79%.
It was interpreted and explained by using global and local agnostic models. The results show that high glucose values significantly increase the likelihood of being classified as diabetic. Also, It shows that the blood pressure has the lowest feature importance.

Support Vector Machine (SVM)

It Uses kernels like linear and RBF to classify the predictions. It achieved AUC of 0.86, and accuracy of 78%.

It was interpreted and explained by using both of the assumptions of linear models, and global and local agnostic models. The results show the consistency of the importance of Glucose feature. Also, It shows that the number of pregnancies have an effect on the diabetes prediction.

Naive Bayes

It uses the probabilistic classifier achieving an AUC of 0.86, and accuracy of 75%.

It was interpreted and explained by using global and local agnostic models. The results show the consistency of the importance of Glucose feature, and the high effect of Insulin. Also, It shows that the age has the least effect on predicting the diabetes outcome.

Machine Learning Models, and its interpretability

Decision Tree

Its mechanism is splitting the data based on feature thresholds, which achieving 76% accuracy on diabetes prediction.

It was interpreted and explained by using global and local agnostic models. All results validates that Glucose, BMI, and Age are the most important and affecting features on the predictions, and Blood Pressure, Skin Thickness are the least affecting ones.

Random Forest

Its mechanism is aggregating multiple trees to improve the accuracy, which scores 81% accuracy on diabetes prediction.

It was interpreted and explained by using global and local agnostic models. The results show that the most affecting features on predictions are Glucose, BMI, and Age, and the least is skin thickness. Also the model has high surrogate accuracy of 1.00 which means that the complex model can be approximated linearly.

LightGBM

Its mechanism is optimizing for speed and efficiency, which achieves 82% accuracy on diabetes prediction.

It was interpreted and explained by using global and local agnostic models. The results show that Glucose, Age, and BMI showed stable and monotonic effects across their ranges. While Skin Thickness had some Flat areas on their ranges in plots like ALE, and it was the least important feature affecting predictions.

Conclusion

The Model with the highest accuracy was LightGBM, as it achieved 82% with a precision of 0.82, recall of 0.81, and f1score of 0.81.

The Most significant features allover the models were Glucose and BMI.

Thanks !

ExplainAI

Sama Mohamed - 202201867
Bosy Ayman - 202202076
Zeyad Sherif - 202201220