Research Article

XAI Implementation on Preliminary Data Analysis Phase: Explainable Output Application with Prediction of Diabetes Mellitus at Early Stage

Mohanad M.Alsaleh[a], Kyung-Mo Yeon[b], SohailAkhtar[a], Qazi Mohammad Sajid Jamal[a*]

[a] Department of Health Informatics, College of Public Health and Health Informatics, Qassim University, Al Bukayriyah, Saudi Arabia;

[b] Department of Big-data AI, Namseoul University, South Korea

***Corresponding author**: Qazi Mohammad Sajid Jamal, Associate Professor, Department of Health Informatics, College of Public Health and Health Informatics, Qassim University, P.O. Box: 52741, Al Bukayriyah, Saudi Arabia; E-mails: m.quazi@qu.edu.sa; ORCID:0000-0001-5525-708X

Abstract: Background: This study aims to create a machine learning model that produces explainable, interpretable, and trustable predictions for diabetes using an XAI approach. Objective: In the study, we have utilized an earlier approach to implementing explainable Machine Learning. Methods: In order to apply XAI technique, we follow a brief version of CRISP-DM. (i) Data Understanding, (ii) Data Preparation, (iii) Model Planning and Building (iv) SHAP Implementation for Interpretability. Results: Global interpretability shows us that two major contributors are symptoms of Polydipsia and Polyuria. An algorithm doesn't "know" prior information, which is highly specific domain knowledge. Local interpretability-based single-instance explanation showed decent multivariate reasoning capability. If the reasoning was based on a simple univariate approach, positive polyuria alone should result in a high probability of positive model output, considering the positive SHAP value of polyuria. Conclusion: The model output results 99.7% confidence to be classified as negative makes much sense since polyuria is also a common symptom of many different situations, such as diabetes insipidus, Kidney disease, Liver failure, Medications that include diuretics, Chronic diarrhea, Cushing's syndrome, Psychogenic polydipsia, Hypercalcemia, Pregnancy.

Keywords: Diabetes Mellitus; XAI Implementation; Machine Learning

1. Introduction

Diabetes mellitus (DM), which results from a complex interaction of genetic and environmental factors, has become a very serious metabolic disorder.DM is principally characterized by hyperglycemia, polyuria, and polyphagia (Alam et al., 2021). Resulting from defects in insulin production, insulin action, or both, Diabetes mellitus showcases persistent hyperglycemia (Nicchio G Ingra et al. 2021). Uncontrolled high blood sugar levels can lead to diabetic complications, some of which are life-threatening (Alam S et al., 2021). The pervasiveness of diabetics throughout the world is rising to epidemic proportions. DM is one of the most common chronic diseases that affect the metabolism in the body (CDC, 2020).

Diabetic patients experience high blood sugar levels due to a lack of insulin production in the pancreas or inadequate amounts of insulin.The main types of diabetes include type 1 diabetes, type 2 diabetes, and gestational diabetes (Singh et al.2016).In general, symptoms of diabetes include weight loss, obesity, frequent urination, blurred vision, vomiting, nausea, acetone breath, and extreme fatigue. However, type 2 diabetes mellitus is the most common among diabetic patients, and it is also referred to as adult-onset diabetes (Singh et al., 2016). Type 2 diabetic patients are usually resistant to insulin as It affects a significant percentage of the global population (Chaudhary & Tyagi, 2018).Because of its high prevalence and complications, which can affect multiple parts of the body, diabetes has the potential to harm or even destroy the healthcare system. Numerous risk factors can increase the possibility of developing diabetes in humans.

Nevertheless, researchers have identified a couple of predictors that should be considered while assessing people who may develop diabetes during their lifetime. Family history of diabetes, current smoking status, and an increased body-mass index are some of those predictors (Lyssenko et al., 2008). These predictors can facilitate the earlydetection of the disease process and prevent the disease from progressing. Due to the advancement in disease detection technologies such as artificial intelligence and machine learning, researchers now have the leverage to predict conditions at an early stage. These technologies facilitate disease prediction by converting the genetic and clinical data into valuable knowledge (Li et al., 2020). Currently, machine learning is a promising technology for enhancing the life quality of diabetic patients. A technique known as predictive analysis integrates various machine learning algorithms, statistical techniques, and data mining methods and utilizes past and current data to extract knowledge and forecast future events (Mir &Dhage, 2018).

1.1Explainable Artificial Intelligence (XAI) Approach:

Nowadays, machine learning and predictive models are widely used in healthcare to achieve long-term strategic objectives. However, most predictive models produce results that non-specialist people do not easily comprehend. Thus, experts have been implementing techniques and algorithms to bridge the gap between ML experts and non-experts. One of the techniques is known as Explainable Artificial Intelligence (XAI), which is a field of Artificial Intelligence (AI) that advances algorithms and predictive models, thus generating high-quality interpretable, trustable, accountable, and understandable AI systems (Das, A., & Rad, P. (2020)).

ML algorithms usually generate vague and unclear explanations of decisions made by the ML model. The expression used for such a situation is known as a "black box", wherein decisions made by AI systems are not explicitly explained, and thus the end-used of the system will be unable to understand why this particular decision is made. As a result, the XAI is often implemented to overcome the end-users understandability and interpretability issues (Calegari, R., Ciatto, G., Dellaluce, J., &Omicini, A. (2019)). However, an enhancement in the understanding of an AI system can increase the utilization of AI systems and mitigate Al ethical issues. Conventional ML/DL approaches typically focus on improving evaluation metrics, thus providing model output performance. But it lacks interpretability and results in trust issues. This paper applies a conventional GBDT-based XGBoost algorithm for fine performance and XAI technique for interpretability. Ergo, this is an attempt to satisfy model performance and explainability.

To the best of the researchers' knowledge, no studies evaluating and interpreting predictive models using XAI for diabetes have been conducted. Thus, there is a lack of explainable ML models for diabetes prediction. This study aims to cover the gap by implementing XAI in a ML model.

This study aims to create a machine learning model that produces explainable, interpretable, and trustable predictions for diabetes using an XAI approach with the standardization learning performance. We have adopted previously developed conceptual propositions summarized previous efforts to define explainability in Machine Learning by establishing a novel definition of explainable Machine Learning (Barredo Arrieta, A. et al., 2020).

2. Materials and Methods

To apply XAI technique, we follow a brief version of CRISP-DM. (i) Data Understanding, (ii) Data Preparation, (iii) Model Planning and Building (iv) SHAP Implementation for Interpretability

2.1 Data collection:

The dataset “Early Stage Diabetes Risk Prediction Dataset” used in this project was obtained from the UCI Machine Learning Repository (UCI, 2020).

The dataset used contains signs and symptoms of recent diabetics or would be a diabetic patients. The researchers used direct questionnaires distributed among the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh. To ensure patients' privacy and data completeness and consistency, the collected data was approved by a doctor. In Jun 2020, the data was donated to be publicly available at the UCI Machine Learning Repository as a downloadable CSV file.

Nonetheless, the dataset contains 17 attributes "columns" and 520 instances "records." All the answers "data" of the questionnaires were either "Yes" or "No" except for the last column, "Class," in which the answers were either "Positive" or "Negative," where positive means having diabetes and negative means not having diabetes. Also, the dataset has been classified as a multivariate dataset due to having multiple data variables, including certain types of medical conditions related to diabetes.

However, Table 1 includes all the dataset variables and the available attributes adopted from Alpan, K., and Ilgi, G.S. (2020) work.

Table. 1. variables and attributes of the dataset:

No.	Variable	Attribute
1	Age	20-65 years
2	Sex	1= male, 2= female
3	Polyuria	1= Yes, 2= No
4	Polydipsia	1= Yes, 2= No
5	Sudden weight loss	1= Yes, 2= No
6	Weakness	1= Yes, 2= No

7	Polyphagia	1= Yes, 2= No
8	Genital thrush	1= Yes, 2= No
9	Visual blurring	1= Yes, 2= No
10	Itching	1= Yes, 2= No
11	Irritability	1= Yes, 2= No
12	Delayed healing	1= Yes, 2= No
13	Partial paresis	1= Yes, 2= No
14	Muscle stiffness	1= Yes, 2= No
15	Alopecia	1= Yes, 2= No
16	Obesity	1= Yes, 2= No
17	Class	1= Positive, 2= Negative

2.2 Data understanding:

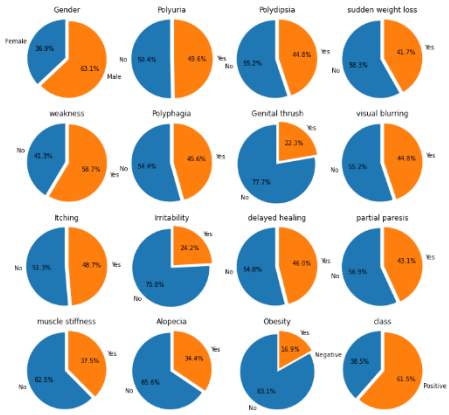

Figure.1. percentages of values in the dataset

Figure 1 illustrates 16 categorical columns, including one label(class), and there are two things we should look into:

1. No extremely imbalanced categorical variables are spotted; therefore, we will not deselect features based on frequency.
2. Label(class) attribute could result in a class-imbalance issue, so the over/under-sampling method could be applied to maximize model performance.

To deal with the class-imbalance issue, we apply the synthetic oversampling method in the following step.

2.3 Data preparation pre-processing

Dataset integrity must be considered in order to follow proper data modeling steps. We applied a two-step approach:

1. Missing data exploration and employ data cleaning if necessary

2. Handle class-imbalance issue: Apply SMOTE technique

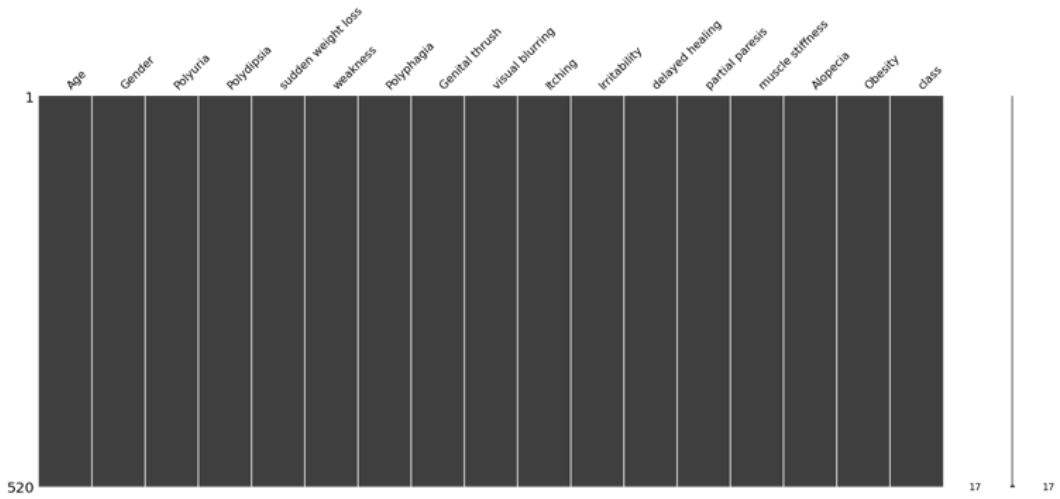

Figure.1. analysis of missing data

Figure 2 is the missing data matrix plot visualization obtained from the python implementation called missing no package. A solid black bar indicates zero missing values, and the plot should show a white horizontal line if any missing data is spotted. Since there are no missing values, we don't apply further missing data handling technique methods.

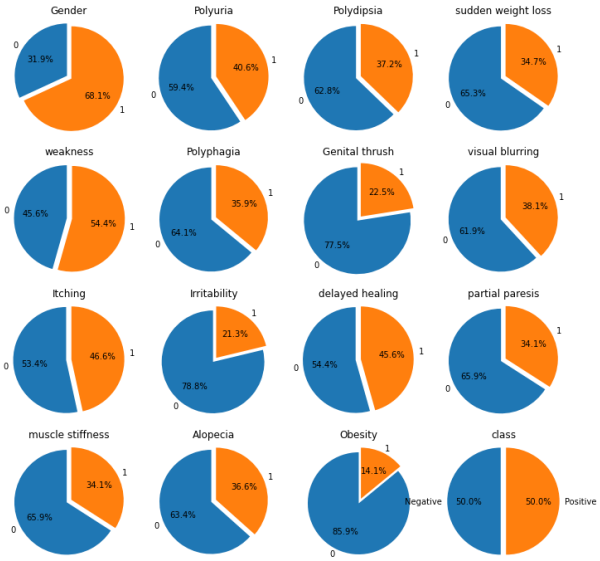

Figure.2. application of SMOTE

Figure 3 is the SMOTE-applied train dataset's categorical column value counts, and there are two things to consider:

1. Class attribute(Label)’s 50:50 balance is achieved by SMOTE synthetic oversampling technique, which was not present in the original dataset.
2. SMOTE is applied only to the training dataset.

No 1 is to handle class-imbalance issues. The reason for no 2 is that since there could be an over-fitting issue from overlapping instances(rows), the intersection between the train/test dataset should be zero.

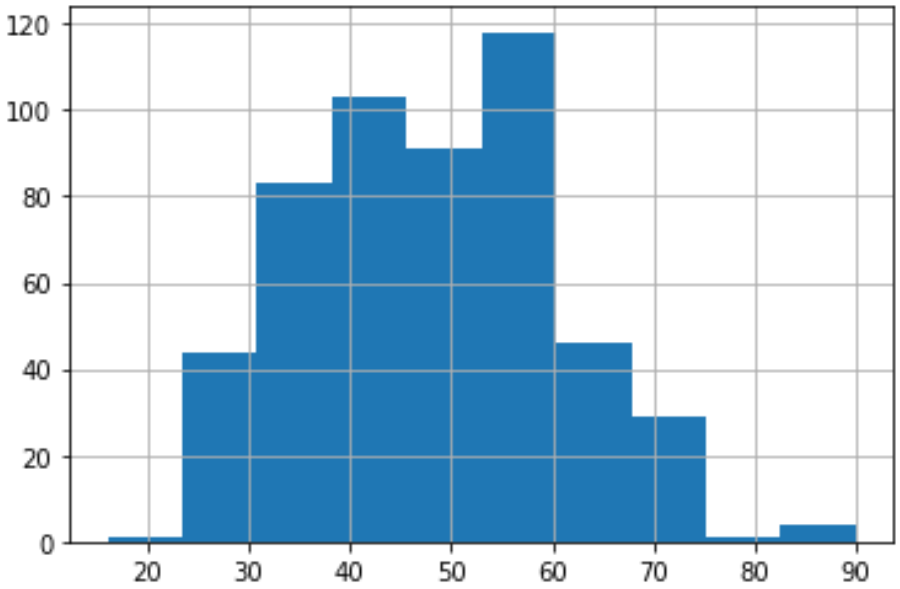

Figure. 3. distribution of data based on the age attribute

Figure 4 is one and only numerical attribute in this dataset, the age column. From this histogram info, the age column shows a roughly gaussian distribution. Therefore, we employ no further data manipulation techniques such as categorization, transformation, or winsorization.

2.4 Model planning and building

The "class" attribute indicates that the goal here is to find the relationship between feature and categorical label(binary) value. Therefore, model candidates should be supervised classification type. The type of dataset is "tabular", so conventional ml models should outperform dl models.

Typical ML models that are used for classification problems with outstanding performance are GBDT type ones such as XGBoost, CATBoost, and lightgbm. Since this research is not about comparing various models' performance, we select one model(XGBoost) without comparing each other. The reason for using XGBoost is that CATBoost and lightgbm IS AN UPGRADED VERSIONS OF XGBoost, making XGBoost a baseline model.We also implemented a hyperparameter-tuning technique. Typical methods that are used here are random search and grid search, but we used the Bayesian optimizer technique, which ensures the best hyper-parameter combination within a limited time.

2.5 Proposed system architecture

The proposed system architecture is displayed in Figure 5. The ML model “XGBoostclassifier" will be fed with the dataset consisting of diabetes symptoms and signs of the patients for training purposes. Then the explainable ai model will make decisions or recommendations based on the input data into the ml model. After that, the end-user will receive the decision along with its explanations in the interface.

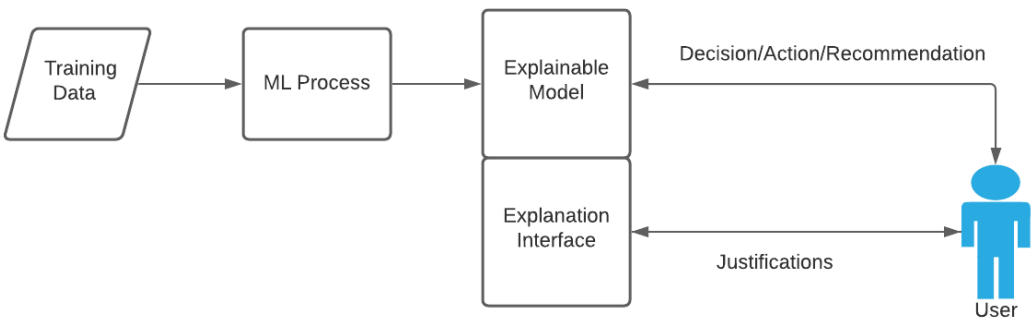

Figure. 4. proposed ML model architecture

5. Results

The evaluation metric "AUC - ROC Curve" used on the built model showed a high-performance score of 0.99, which means that the model is 99% accurate in distinguishing between patients with diabetes and patients without diabetes. Figure 6 is about the AUC_ROC score of the fitted model on the test dataset, which shows outstanding performance.

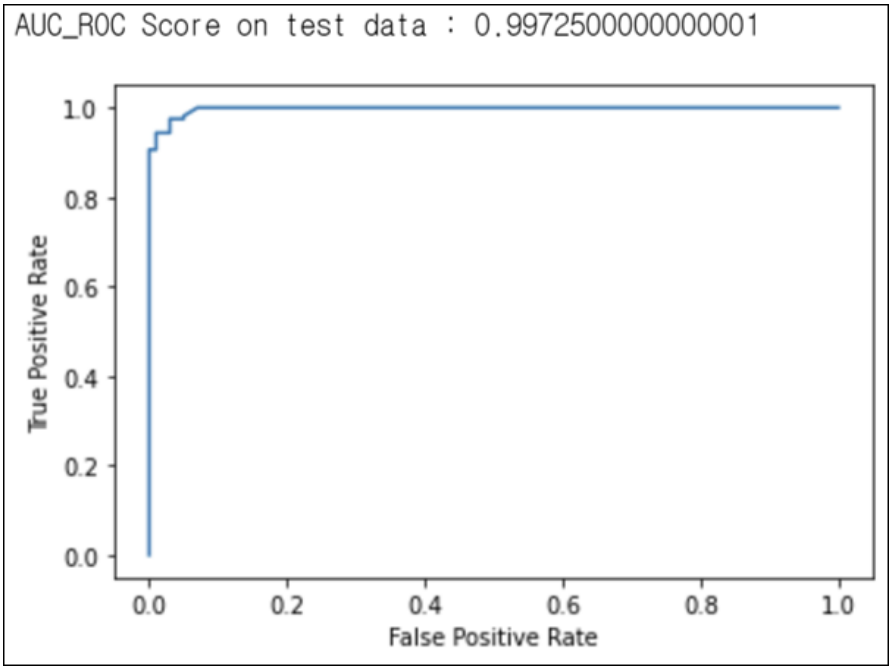

Figure.5. AUC_ROC of the model

Table.1. Confusion matrix of a classification model

Confusion Matrix		Model Output	
		Positive	Negative
Dataset Label	Positive	155 (60%)	5 (2%)
	Negative	3 (1%)	97 (37%)

Almost all classification models' evaluation metrics are based on threshold and confusion matrix. Table 2 is a confusion matrix with a threshold level of 0.5, the default value. From here, we derived four evaluation metrics as follows:

Accuracy of 97%

Precision of 0.981

Recall of 0.968

F1 score of 0.974

Accuracy measures how accurately this model predicts, and it is perceptually simple since it resembles a simple percentage idea (number of correctly classified samples/number of all samples). But accuracy alone could mislead model performance if the dataset is imbalanced. So we adapted Precision, Recall, and F1 score for this issue. Precision and recall are similar to accuracy but implement a Bayesian approach. Precision is accurate if the model output is positive, and recall is accurate if the data label is positive. F-1 score is a harmonic mean of precision and recall.

3.1 Implementation of XAI

Typically, F1 score of 0.974 is an outstanding result considering this number is from the test dataset, which implies little over-fitting. However, we attempt to explain this model's output using XAI technique.

3.1.1 Global Interpretability

SHAP (Shapley Additive exPlanations, Lundberg, 2017) is a concept that adapts game theory to the machine learning domain. It attempts to explain the output of a fitted ML model. Using the Shapley values, this method connects optimal credit allocation with local and global explainability. Ergo, SHAP is an approach to finding Shapley value (Shapley, 1953) from a dataset.

From the original reference, the Shapley value is a number to quantitatively measure each player's contribution proportion of winning. In other words, Shapley value is "How much money the team should distribute to each player(payout) considering each player's contribution to winning.". This game-theoretic idea was first applied in the ML domain in 2017. In ML Sense, each player is a feature, and the player's contribution(payout) is the label.

A standard approach to approximate Shapley value where interested instance i have feature j is as follows:

1. Prepare fitted model f, data x, and interested instance i.

2. Select random instance z from dataset x, where z is not the same instance as i.

3. Find out output values from f(i) and f(z).

4. From instances i and z, swap values of j from i and z. We call it ij and zj.

5. Find out output values from f(ij) and f(zj).

6. From the instance I, compare f(i) and f(ij). For instance, z, compare f(z) and f(zj). This comparison gives us a difference between whether the jvalue was the same and swapped. From this approach, we understand how a value in feature j contributes to the output value.

7. By iterating steps 2 to 6, we could get the mean value that features j's contribution(affect) to the output value by each value in feature j.

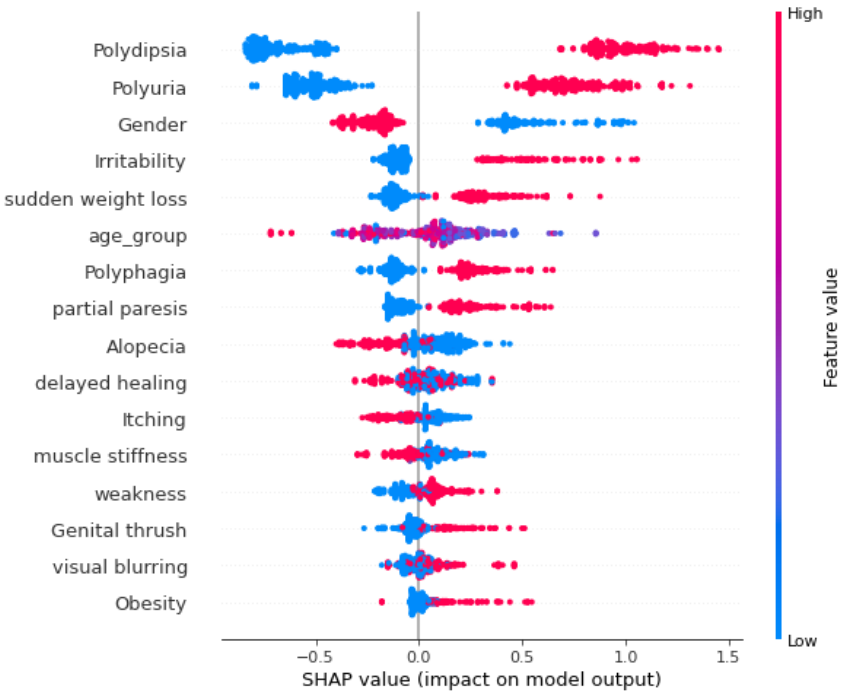

Figure .6. summary of values when Shapley is used

Figure 7 is a summary plot, and it sums up the Shapley values from all instances and feature values. The plot has two meaningful pieces of information, color, and distance from zero. Red and blue are about feature value. If a feature ranges between 0 to 1, a value close to 1 is red, and a value close to 0 is blue. All feature values are binary in this dataset, so 0 is blue,

and one is red. The distance(x-axis) is a contribution to the model output value. Since this model is binary classification, model output is the probability, and distance of positive direction means the value positively contributes to the output value (adds up the probability). The negative direction is vice-versa. Features are sorted descending-fashion considering their contribution to model output (Shapley Value).

The top 2 contributors based on Shapley value are Polydipsia and Polyuria.

3.1.2 Local Interpretability:

SHAP could also be used to explain a single instance's prediction.

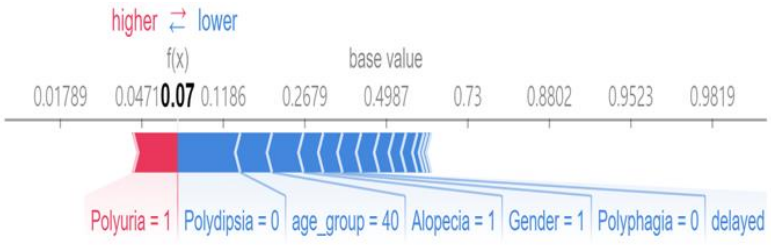

Figure. 7. individual SHAP positive value plot

Fig above is a random-chosen instance with 95% confidence that the model output is positive. Model output predicted this patient to be positive because of Shapley value's additive contribution.

Fig above is a random-chosen instance which 95% confidence of model output to be positive. Model output predicted this patient to be positive because of Shapley value's additive contribution.

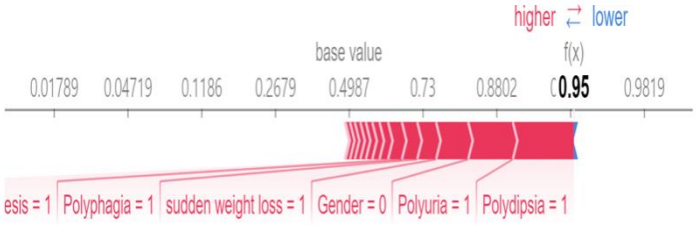

Figure. 8. Individual SHAP negative value plot

Figure 9 above is a random-chosen instance with 99.3% confidence that a model output is negative. Model output predicted this patient to be negative and one thing to notice is that even though a major symptom of polyuria is spotted, this model predicted this patient to be negative because of other features' additive Shapley values.

Global interpretability shows us that two major contributors are symptoms of Polydipsia and Polyuria. An algorithm doesn't "know" prior information, which is highly specific domain knowledge. Local interpretability-based single-instance explanation showed decent multivariate reasoning capability. If the reasoning was based on a simple univariate approach, positive polyuria alone should result in a high probability of positive model output, considering the positive SHAP value of polyuria.

4. Discussion

We obtained the top 2 contributors from the SHAP section are Polydipsia and Polyuria. From domain knowledge, polydipsia is an especially common earliest symptom of diabetes mellitus. Diabetes comes with conditions that make it harder for the patient's body to process and use blood sugar(glucose). Blood sugar levels could skyrocket when this condition occurs. Ergo, high blood sugar levels cause extreme thirst as a result.

Polyuria is also one of the major signs of diabetes mellitus. Under normal conditions, kidneys reabsorb all of the sugar when they filter blood to make urine, making zero sugar to the urine. Since diabetes is about high blood sugar levels, not all of the sugar can be reabsorbed. The resulting residual sugar component ends up in the urine, resulting in unusually large volumes of urine.

But without deep academic knowledge, two major contributors "show themselves" to be common symptoms of diabetes mellitus as we apply SHAP. The model output results 99.7% confidence to be classified as negative makes a lot of sense since

polyuria is also a common symptom of many different situations, such as Diabetes insipidus, Kidney disease, Liver failure, Medications that include diuretics, Chronic diarrhea, Cushing's syndrome, Psychogenic polydipsia, Hypercalcemia, Pregnancy.

5.Conclusion

Since the dataset presents a typical class-imbalance issue, we adopted the synthetic oversampling method of SMOTE. Using the XGBoost classifier with the Bayesian optimization hyper-parameter tuning technique, we got outstanding performance with an F1 score of 0.974, a harmonic mean of precision and recall. XAI of SHAP technique is implemented to obtain global and local interpretability. Global interpretability information of the SHAP summary plot shows us two major symptoms of Polydipsia and Polyuria, information which ML model previously had no prior domain knowledge. Local interpretability also showed decent multivariate reasoning capability.

References

Alam, S., Hasan, M. K., Neaz, S., Hussain, N., Hossain, M. F., & Rahman, T. (2021). Diabetes mellitus: Insights from epidemiology, biochemistry, risk factors, diagnosis, complications and comprehensive management. *Diabetology*, *2*(2), 36–50. https://doi.org/10.3390/diabetology2020004

BarredoArrieta,A. *et al.* (2020) Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, **58**, 82–115.

Calegari, R., Ciatto, G., Dellaluce, J., &Omicini, A. (2019). Interpretable Narrative Explanation for ML Predictors with LP: A Case Study for XAI. In WOA (pp. 105-112).

Centers for Disease Control and Prevention. (2021, November 16). *What is diabetes?* Centers for Disease Control and Prevention. Retrieved November 26, 2021, from https://www.cdc.gov/diabetes/basics/diabetes.html.

Chaudhary, N., & Tyagi, N. (2018). Diabetes mellitus: An overview. *International Journal of Research and Development in Pharmacy & Life Sciences*, *7*(4), 3030–3033. https://doi.org/10.21276/ijrdpl.2278-0238.2018.7(4).3030-3033

Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (xai): A survey. arXiv preprint arXiv:2006.11371.

Li, J., Huang, J., Zheng, L., & Li, X. (2020). Application of artificial intelligence in diabetes education and management: Present status and promising prospect. *Frontiers in Public Health*, *8*. https://doi.org/10.3389/fpubh.2020.00173

lpan,K. and Ilgi,G.S. (2020) Classification of diabetes dataset with data mining techniques by using Weka Approach. *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*.

Lyssenko, V., Jonsson, A., Almgren, P., Pulizzi, N., Isomaa, B., Tuomi, T., Berglund, G., Altshuler, D., Nilsson, P., &Groop, L. (2008). Clinical risk factors, DNA variants, and the development of type 2 diabetes. *New England Journal of Medicine*, *359*(21), 2220–2232. https://doi.org/10.1056/nejmoa0801869

Mir, A., &Dhage, S. N. (2018). Diabetes disease prediction using machine learning on Big Data of Healthcare. *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*. https://doi.org/10.1109/iccubea.2018.8697439

Nicchio, I. G., Cirelli, T., Nepomuceno, R., Hidalgo, M. A., Rossa, C., Cirelli, J. A., Orrico, S. R., Barros, S. P., Theodoro, L. H., &Scarel-Caminaga, R. M. (2021). Polymorphisms in genes of lipid metabolism are associated with type 2 diabetes mellitus and periodontitis, as comorbidities, and with the subjects' periodontal, glycemic, and lipid profiles. *Journal of Diabetes Research*, *2021*, 1–21. https://doi.org/10.1155/2021/1049307

Singh, N., Kesherwani, R., Tiwari, A. K., & Patel, D. K. (2016). A review on diabetes mellitus. The Pharma Innovation, 2016; 5(7): 36-40

UCI Machine Learning Repository: Data Set. (n.d.). Retrieved November 26, 2021, from https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.