Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 09, September-2020

Building Explainable and Interpretable model for Diabetes Risk Prediction

Anshuman Guha
Johns Hopkins University
Department of Computer Science
Baltimore, Maryland, USA

***Abstract*—Diabetes is becoming a prolific disease and a global pandemic. The capability to detect diabetes early can improve the scope of disease management. For disease prognosis, artificial intelligence techniques are used widely. In this paper, probability and prediction for diabetes are discussed along with its model explanability and interpretability. A UCI repository's new diabetes dataset, which is collected from Bangladesh patients, is used. Various tools were used to interpret the model outcome, like Local Interpretable Model-Agnostic Explanations (LIME), Shapley Additive Explanations (SHAP), tree-based model feature importance, partial dependency plots. A plethora of tree-based and other popular classification algorithms were used to build the diabetes prediction model. The findings indicated that using kernel-PCA before fitting random forest (RF) gives slightly better results than vanilla-RF. Our findings indicate a strong positive correlation with the 3Ps of diabetes, i.e., polyuria, polydipsia, and polyphagia, along with genital thrush, visual blurring, and delayed healing. Being a female in this dataset (derived from Bangladesh patients) have a high chance of diabetes. This represents the model's accuracy and reliability that eventually will be crucial in the trust of medical practitioners on the model outcome. The counter-intuitive outcome of individual data attributes was examined with model agnostic interpretability of machine learning techniques, which can be explained with domain knowledge of the data. The outcome also suggests that the model built on one domain, like population demographics, cannot be directly applied to other demographics. Any patterns observed in the data relevant to the model outcome should be well studied and explained, along with model accuracy metrics.**

Keywords— Diabetes, Interpretability, Explanability, Local Interpretable Model-Agnostic Explanations, Shapley Additive Explanations, Partial Dependence Plots, T-SNE, Random Forest, Decision Tree, Machine Learning

I. INTRODUCTION

Diabetes is emerging as a concerning global health problem that is nearing epidemic proportions globally. According to an estimate of the International Diabetes Federation, the comparative prevalence of diabetes during 2007 was 8.0 % and likely to increase to 7.3% by 2025. The number of people with diabetes is 246 million and likely to increase to 380 million by 2025 [1]. The three Ps of diabetes is Polyuria, Polydipsia, and Polyphagia. It is common for several symptoms to appear together, i.e., thirst (polydipsia), and an increased need to urinate (polyuria) will often come as a pair. The earlier diagnosis of diabetes would help in better management of the disease. Artificial Intelligence has helped make an early diagnosis of diabetes [2][3][4]. A robust interpretable model representing a good correlation with cardinal features of diabetes can provide more confidence to medical practitioners and patients in the model prognostics. A better understanding of how an algorithm works can help better align the model outcome with the critical questions required for validation. Amongst a wide variety of machine learning algorithms that were used to predict diabetes, some include the traditional machine learning method [5], linear models, tree-based models and support vector machine. Polat et. al. [6] used principal component analysis with neuro inference for diabetes detection. Yue et al. [7] used quantum particle swarm optimization algorithm and weighted least squares support vector machine Razavian et al. [8] implement linear models. Georga et al. [9] used multivariate prediction approach and Ozcift and Gulten [10] used futuristic model of rotation forest for diabetes prediction. Alama Muhammad et al. (2019) implemented neural network, tree-based methods and clustering techniques are used to build prediction models for diabetes.

In this study, along with various tree-based and other popular classification algorithms, a wide range of techniques are used to make the model more explainable and interpretable. The focus is to explain the underlying data rather than purely focusing on model prediction accuracy. The paradox of correlation and causation is explored. The random forest model on this dataset already achieves 98% AUROC, so more emphasis is given on investigating the underlying data. The practical relevance of features with the problem statement, i.e., diabetes prediction, is emphasized, and results w.r.t to contributing features are explained with wide variety of techniques.

II. MATERIALS & METHODS

A. Dataset

Sylhet Diabetes Hospital patients in Sylhet, Bangladesh, diabetes dataset is used in this paper (UCI Repository [11]). The dataset has patient's critical attributes which are helpful for diabetes prediction. The Diabetes (PID) dataset having: 17 = 16+ 1 features, 328 male patients, 192 female patients and 268 positive instances (61.5%)). The detailed description of all attributes is given in Table 1.

B. Maintaining Model Explainability and Interpretability

Interpretability describes the cause and effect that can be observed within a system. Explainability is the extent to which a machine or deep learning system's internal mechanics can be explained in human terms. Interpretability provides the ability to understand the mechanics without necessarily knowing why. Explainability is able quite literally to explain what is happening. Regulated domains like healthcare and finance are

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 09, September-2020

looking to deploy artificial intelligence and deep learning systems, where questions of model interpretability and compliance are particularly important.

Attribute Name	Data Dictionary
Sex	1: Male, 2: Female
Polyuria	1: Yes, 2: No
Polydipsia	1: Yes, 2: No
Sudden Weight Loss	1: Yes, 2: No
Weakness	1: Yes, 2: No
Polyphagia	1: Yes, 2: No
Genital Thrush	1: Yes, 2: No
Visual Blurring	1: Yes, 2: No
Itching	1: Yes, 2: No
Irritability	1: Yes, 2: No
Delayed Healing	1: Yes, 2: No
Partial Paresis	1: Yes, 2: No
Muscle Stiffness	1: Yes, 2: No
Alopecia	1: Yes, 2: No
Obesity	1: Yes, 2: No
Class	1: Positive, 2: Negative

Table 1. Dataset description and characteristics

III. EXPLANIBILTY METHODS

A. *Shapely Plots*

SHAP (Shapley Additive Explanations) related game theory with explanations and represents locally accurate additive feature attribution method. [12]

For example, in Fig. 1, this patient has a very high risk of diabetes with a probability of 0.81. The contributing factors are Polyuria-yes, age-41 and the top factors which are against having a diabetic patient are Polydipsia-no, Gender-Male, and Alopecia-No.

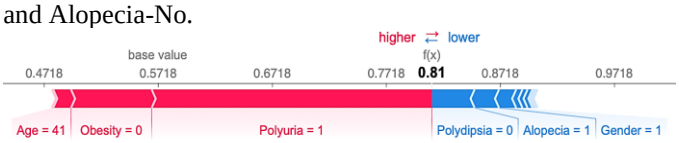

Fig. 1. SHAP plot - patient with high risk score

The above SHAP plot shown in Fig. 2. explains the following phenomenon:

1. Polyuria and Polydipsia - Yes is the most significant contributor to diabetes, and No value strongly reduces the probability of diabetes.
2. Irritability, Genital thrush, sudden weight loss are moderate contributors to diabetes.
3. Gender - Diabetes afflicts women more than it does men in Bangladesh, according to the British medical journal "The Lancet." Double rate of diabetes in urban areas than the villages of Bangladesh is mentioned in this study [13]

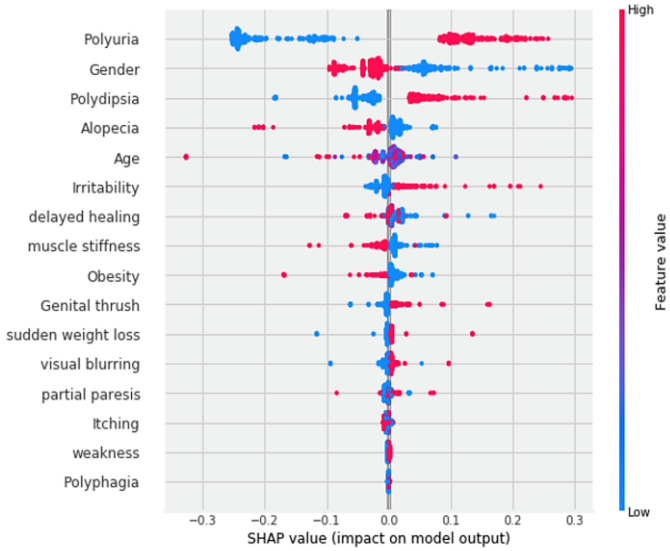

Fig. 2. SHAP plot – explaining feature importance w.r.t. to diabetes risk scores

B. *GLM Interpretation*

A generalized linear model is fitted to explore coefficients for their positive and negative impact on diabetes probability. A positive coefficient value indicates the presence of an attribute or high value will result in a high probability of diabetes and vice versa for negative coefficient. Further, examining the p-values i.e., less than 0.05, will suggest attributes are important to diabetes prediction. Table 2. exhibits a GLM model output with all variables. The attributes which have statistically significant for diabetes prediction are gender, polyuria, polydipsia, irritability, itching, genital thrush, visual blurring, and delayed healing.

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.4313	0.036	12.098	0.000	0.361	0.501
Gender	-0.2630	0.031	-8.512	0.000	-0.323	-0.202
Alopecia	-0.0007	0.033	-0.022	0.982	-0.066	0.065
Obesity	-0.0545	0.036	-1.518	0.129	-0.125	0.016
Polyuria	0.3180	0.035	9.018	0.000	0.249	0.387
Polydipsia	0.2797	0.037	7.648	0.000	0.208	0.351
Polyphagia	0.0416	0.031	1.360	0.174	-0.018	0.101
Irritability	0.1546	0.032	4.789	0.000	0.091	0.218
Itching	-0.1164	0.031	-3.798	0.000	-0.176	-0.056
sudden_weight_loss	0.0460	0.031	1.461	0.144	-0.016	0.108
Genital_thrush	0.1775	0.035	5.138	0.000	0.110	0.245
visual_blurring	0.0546	0.032	1.710	0.087	-0.008	0.117
delayed_healing	-0.0856	0.032	-2.680	0.007	-0.148	-0.023
partial_paresis	0.0627	0.033	1.902	0.057	-0.002	0.127
muscle_stiffness	-0.0240	0.031	-0.772	0.440	-0.085	0.037
weakness	0.0249	0.031	0.806	0.420	-0.036	0.085

Table 2. GLM Results with all features

The GLM coefficients in Table 3. suggests that female patients and the absence of itching increase the probability of

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 09, September-2020

diabetes. The presence of these positive attributes with positive coefficients i.e., polyuria, polydipsia, irritability, genital thrush, visual blurring, and delayed healing, increases the chances of diabetes.

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.4666	0.033	14.142	0.000	0.402	0.531
Gender	-0.2833	0.029	-9.677	0.000	-0.341	-0.226
Polyuria	0.3530	0.033	10.703	0.000	0.288	0.418
Polydipsia	0.3077	0.034	9.069	0.000	0.241	0.374
Irritability	0.1596	0.031	5.068	0.000	0.098	0.221
Itching	-0.1192	0.030	-3.971	0.000	-0.178	-0.060
Genital_thrush	0.1624	0.033	4.883	0.000	0.097	0.228
visual_blurring	0.0589	0.029	2.016	0.044	0.002	0.116
delayed_healing	-0.0629	0.029	-2.136	0.033	-0.121	-0.005

Table 3. GLM model predictions with significant features of p-value<0.05

C. LIME Plots

LIME (Local Interpretable Model-agnostic Explanations) is used for explaining model outputs and locally approximating the selected model with an interpretable one. The interpretable models are trained on small samples of the original observation. Thus, they only provide an excellent local approximation.

The LIME plot presents two contrasting individuals. The first one with no Polyuria and Polydipsia has a much lower probability of diabetes. The second person has both of these conditions have a 72% chance of diabetes. Further, the counter-intuitive fact that having alopecia (hair loss) lowers diabetes can be explained with the fact that hair loss is prominent in aging men in Bangladesh, but they are less prominent in having diabetes.

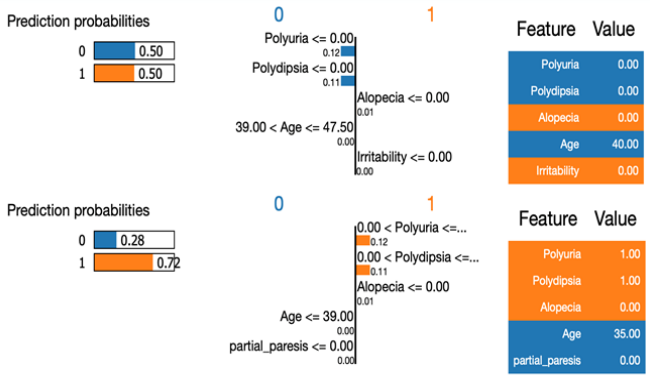

Fig. 4. LIME plot – Explaining feature contribution to patients with high and low risk probability

D. Equations Random Forest and Tree-Based Feature Importance

Random forests are practical algorithms for feature ranking, and they use mean decrease impurity is exposed in most random forest libraries. However, the concern arises with correlated features. Robust features can end up with low scores, and the method can be biased towards variables with many categories.

Fig. 4. depicts the order of features important, contributing to the prediction of diabetes in patients. It is essential to predict which patients have a high propensity to have diabetes, but presenting feature importance would help in early detection and maybe even improve the treatment.

Weight	Feature
0.1246 ± 0.0246	Polydipsia
0.1123 ± 0.0396	Gender
0.1077 ± 0.0389	Polyuria
0.0169 ± 0.0151	muscle stiffness
0.0154 ± 0.0097	Polyphagia
0.0138 ± 0.0062	partial paresis
0.0123 ± 0.0185	weakness
0.0123 ± 0.0075	sudden weight loss
0.0108 ± 0.0075	delayed healing
0.0077 ± 0.0138	Itching
0.0062 ± 0.0062	Irritability
0.0046 ± 0.0075	Genital thrush
0.0031 ± 0.0075	Alopecia
0 ± 0.0000	Obesity
-0.0000 ± 0.0138	Age
-0.0031 ± 0.0075	visual blurring

Fig. 4. RF Feature Importance plot

Below, Fig. 5. presents the diabetes classification decision tree as a flowchart-like structure where the node represents an attribute, the branch explains a decision rule, and the leaf node presents the outcome. The top node in a decision tree is polydipsia. If a Polydipsia and Polyuria is true, then 166 out of 170 patients have diabetes. If Polydipsia and Polyuria are both negative, then 138 out of 170 samples will not have diabetes. This flowchart representation helps us in decision making. It is a visualization like a flowchart diagram that easily mimics human-level thinking.

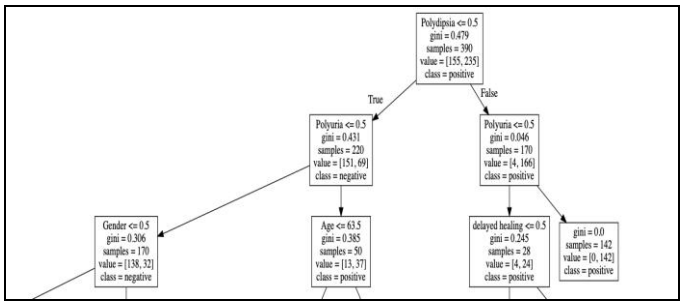

Fig. 5. Decision Tree Feature Importance plot

E. Partial Dependence Plots

The partial dependence plot presents the marginal effect one feature at a time on the predicted outcome of a model (J. H. Friedman 200125). The following figures show some example PDPs. Fig. 6A and 6B exhibit that both Polyuria and Polydipsia contribute to a higher risk of diabetes and being a male in Bangladesh dramatically reduce diabetes chances.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 09, September-2020

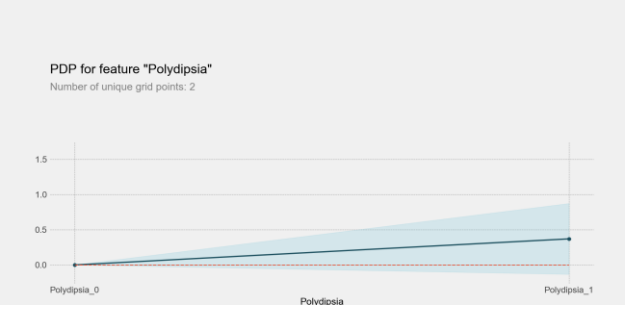

Fig. 6A. Partial Dependence Plot for Polydipsia

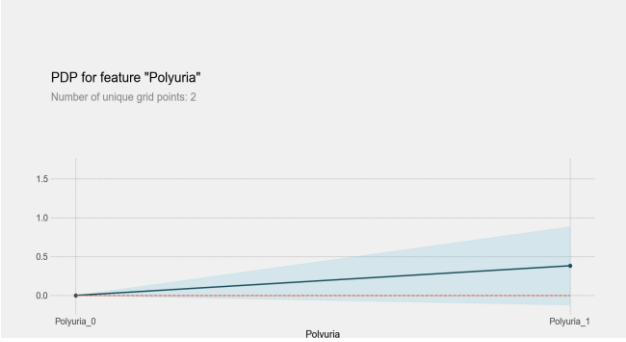

Fig. 6B. Partial Dependence Plot for Polyuria

F. *PCA and T-SNE Plots*

Both Fig. 7 and Fig. 8 represent PCA and t-SNE (t-Distributed Stochastic Neighbor Embedding) representation respectively and show good separability in the data, which suggests that any classification algorithm will have tremendous accuracy in predicting diabetes in this group of patients. The decision boundary is non-linear. Therefore, tree-based models will perform well. This often serves a good starting point to visualize the data in two-dimensional space and choose an appropriate classification algorithm.

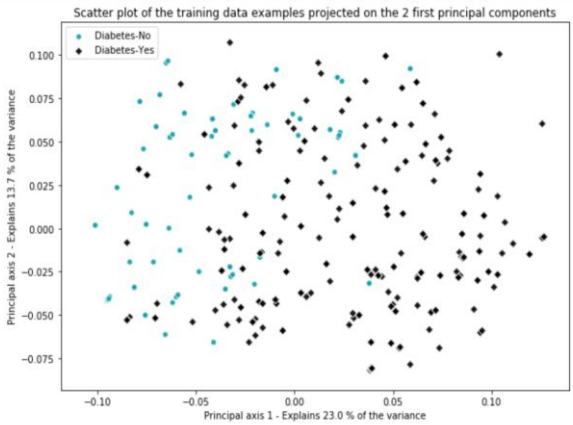

Fig. 7. PCA plots showing good separability between classes

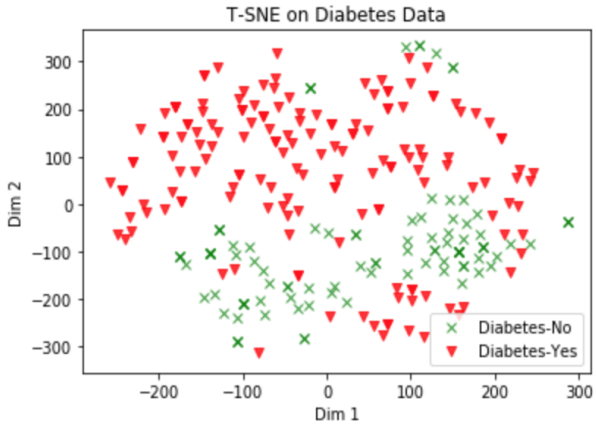

Fig. 8. T-SNE plots showing good separability between classes

IV. MODELING

A. *Covariance, Correlation, and Collinearity*

Both terms measure the relationship and the dependency between variables. Covariance explains the direction of the relationship between variables, and correlation measures both the strength and direction of the relationship between any two attributes. Correlation is a function of the covariance. The difference between correlation and covariance is the fact that correlation values are standardized, whereas covariance values are not.

As described in Fig.10, both Polydipsia and Polyuria have a high correlation of 0.6. Therefore, one of these features will be dropped during the modeling stages.

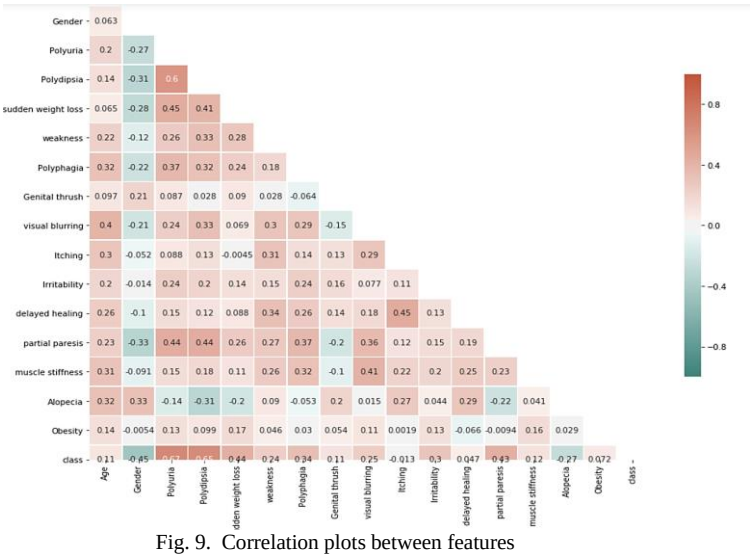

Fig. 9. Correlation plots between features

B. *Results*

A wide range of classification algorithms was used on the diabetes dataset. All algorithms were different, as the underlying mathematical principles for all these algorithms were different. The results were evaluated based on the following metrics:

a) precision score

b) recall score

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 09, September-2020

c) F1 score
d) support score
e) accuracy score
f) AUC/ROC

The cross-validation of twenty folds (k=20) is used, and the average of the metrics are reported in Table 4.

	Model	Fitting time	Scoring time	Accuracy	Precision	Recall	F1_score	AUC_ROC
5	Random Forest	0.012722	0.009724	0.948065	0.949904	0.943994	0.947882	0.984064
6	PCA + Random Forest	0.012023	0.007328	0.925516	0.931975	0.923133	0.925145	0.980925
4	Quadratic Discriminant Analysis	0.002083	0.006658	0.922850	0.932453	0.907278	0.921028	0.978788
2	Support Vector Machine	0.014543	0.006312	0.906820	0.915886	0.896131	0.904791	0.962446
0	Logistic Regression	0.002008	0.005673	0.901359	0.897620	0.902219	0.901588	0.946429
3	Linear Discriminant Analysis	0.002295	0.005710	0.876651	0.876521	0.883820	0.877453	0.946374
7	Kernel-PCA + Random Forest	0.012598	0.007004	0.887160	0.892279	0.876596	0.885752	0.942073
9	Bayes	0.001563	0.006142	0.876178	0.873006	0.868885	0.875792	0.937716
1	Decision Tree	0.001539	0.005449	0.934632	0.934683	0.933604	0.934598	0.933604
8	K-Nearest Neighbors	0.001353	0.008996	0.815471	0.825425	0.831196	0.816395	0.921266

Table 4. Model and Output Comparison

Both standard linear-PCA and Kernel-PCA have been applied for feature reduction and feature space separation between two classes, i.e., diabetes and no-diabetes. The feature space is fed into a Random Forest (RF) model and compared with vanilla-RF results. Kernel-PCA extracts non-linear features space compared to standard PCA. All performance metrics and AUROC for kernel-PCA-RF and standard PCA-RF are 94.20% and 98.09%, respectively, which surprisingly performs worse than vanilla-RF AUROC of 98.40%.

The RF algorithm does not suffer from a high number of predictors since it only takes a random subset to build each tree. Support Vector Machine (SVM) performed well with AUROC of 96.24%. Logistic regression was performed as well, with AUROC of 94.64% on this dataset. Linear Discriminant Analysis closer to Logistic regression results with AUROC of 94.63%. The Naïve Bayes (NB) classifier performed decently with AUROC of 93.77%. The K-nearest neighbor classification (KNN) performance performed worst with AUROC of 92.12%.

C. Discussion

AUROC performance with kernel-PCA and PCA before feeding the RF model exhibits that PCA does not help better data separability. One caveat is that this phenomenon is highly dependent on data. The reason being that decision trees are sensitive to the rotation of the data since the decision boundary they create is always vertical or horizontal. Therefore, if data is less separable, it will take a much more giant tree to separate these two classes, but if data aligns with its principal components, the perfect separation can be achieved with just one layer.

SVM works well in the dataset with a clear margin of separation. Logistic regression performs decently, which suggests a linear separation between the classes in this dataset. For the cases of more complex and non-linear datasets, linear-based algorithms may not be ideal for classification tasks. KNN is a simple and instance-based machine learning algorithm. KNN can be useful in the case of non-linear data.

As shown in Table 5., age explains most of the variance in principal component-1 (PC-1). Both Polydipsia and Polyuria are top differentiating features from the diabetes classification perspective but explained variability in data due to these features are less due to only two possible values, i.e., 0 and 1.

CONCLUSIONS

The preferred Machine learning techniques are incredibly useful for a diabetes diagnosis. Early prognosis of diabetes would help in patient's treatment. In this paper model, explanability and interpretability are also much emphasized. A new diabetes dataset from the UCI repository, which was collected from Bangladesh patients, is used. Shapely plots explain the extent of attribute impact in diabetes prediction in a positive or negative direction. GLM results help to understand statistically significant features. LIME plots explain contrasting risk scores for couple of diabetes patients with contributing features. Partial dependence plots and decision tree explains the directional and quantitative impact of top contributions attributes. Various agnostic tools to interpret machine learning algorithms are used to examine and explain the unusual data behavior. Both PCA and t-SNE exhibit good separation of a dataset into two classes, that can be separated using a non-linear boundary. After that, a wide variety of tree-based and other popular classification algorithms are used to build a diabetes prediction model. The latent space after kernel-PCA and PCA are also used with RF algorithms. The vanilla-RF produces the best AUROC of 98.40% after taking a mean of 20 folds. This study's scope can be applied to arthritis, cancer, and other prolonged chronic diseases where early prediction can provide a cure or reverse the condition. Other attributes, including physical activity, ancestral health condition, body weight, and energy levels, can be considered in the future to diagnose diabetes.

REFERENCES

[1] Syed Amin Tabish, “Is Diabetes Becoming the Biggest Epidemic of the Twenty-first Century?”. Int J Health Sci (Qassim). 2007 Jul; 1(2): V–VIII.
[2] Lee B. J., Kim J. Y. (2016). "Identification of type-2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning". *IEEE J. Biomed. Health Inform.* 20 39–46. 10.1109/JBHI.2015.2396520
[3] Alghamdi M. et al. "Predicting diabetes mellitus using SMOTE and ensemble machine-learning approach," PMID: 28738059 PMCID: PMC5524285 10.1371/journal.pone.0179805
[4] Kavakiotis I. et al. (2017). "Machine learning and data-mining methods in diabetes research." *Comput. Struct. Biotechnol. J.* 15 104–116. 10.1016/j.csbj.2016.12.005
[5] Kavakiotis I., Tsave O., Salifoglou A., Maglaveras N., Vlahavas I., Chouvarda I. (2017). "Machine learning and data-mining methods in diabetes research." *Comput. Struct. Biotechnol. J.* 15 104–116. 10.1016/j.csbj.2016.12.005
[6] Polat K., Kodaz H. (2005). "The medical applications of attribute weighted-artificial immune system (AWAIS), diagnosis of heart and diabetes diseases," *International Conference on Artificial-Immune Systems*, ICARIS 2005: Artificial Immune Systems pp 456-468
[7] Yue C. et al. (2008). "An intelligent diagnosis to type-2 diabetes based on QPSO algorithm and WLS SVM," book ISBN: 978-0-7695-3505-0 DOI 10.1109/IITA.Workshops.2008.36
[8] Razavian N., Blecker S., Schmidt A. M., Smith-McLallen A., Nigam S., Sontag D. (2015). "Population-level prediction of type-2 diabetes

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 09, September-2020

from claims data and analysis of risk factors". *Big Data* 3 277–287. 10.1089/big.2015.0020

[9] Georga E. I. et al. (2013). "Multivariate prediction of subcutaneous glucose concentration in type-1 diabetes patients based on support vector regression". *IEEE J. Biomed. Health Inform.* 17 71–81. 10.1109/TITB.2012.2219876

[10] Ozcift, A., Gulten A. (2011), "Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms." *Comput. Methods Programs Biomed.* 104 443–451. 0.1016/j.cmpb.2011.03.018

[11] UCI - Machine Learning Repository, "Early stage diabetes risk prediction dataset. Data Set"

[12] Diabetes Fact: Bangladesh Perspective AK Mohiuddin, "International Journal of Diabetes Research" (2019)

[13] Alama Muhammad et al. "A model for early prediction of diabetes" Informatics in Medicine. Volume 16, 2019, 100204.