BAREKENG: Jurnal Ilmu Matematika dan Terapan
June 2022 Volume 16 Issue 2 Page 615–624
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol16iss2pp615-624

PREDICTING DIABETES MELLITUS USING CATBOOST CLASSIFIER AND SHAPLEY ADDITIVE EXPLANATION (SHAP) APPROACH

Novia Permatasari [1*], Shafiyah Asy Syahidah[2], Aldo Leofiro Irfiansyah[3], M. Ghozy Al-Haqqoni[4]

[1,2,4]BPS - Statistics Indonesia
Dr. Sutomo St., No. 6-8, Pasar Baru, Sawah Besar, Jakarta, 10710, Indonesia
[3]BPS Kabupaten Kepulauan Sula - Statistics Sula Islands District
Yos Sudarso St., No. KM10, Pohea, North Sanana, North Maluku, 97796, Indonesia

Corresponding author's e-mail: [1] novia.permatasari@bps.go.id*

Abstract. *Diabetes mellitus as a metabolic disease characterized by hyperglycemia can be dangerous if it cannot be handled properly. Early detection of existing symptoms can reduce the impact of delays in treatment. This study aims to carry out early-detection patients with diabetes mellitus using a machine learning approach through data from MIT's GOSSIS (Global Open Source Severity of Illness Score). By using Shapley Additive Explanation (SHAP) which enables prioritization of feature that determine compound classification, this study shows that the CatBoost classifier has 14 features that significantly can be used for classification with feature 'd1_glucose_max' or the highest glucose concentration of the patient in their serum or plasma during the first 24 hours of their unit stay has the highest impact to classify diabetes mellitus patients, then followed by age and glucose APACHE. The selected features are then classified and get the validation AUC score of 86.86%.*

Keywords: *CatBoost, Classification, Diabetes Mellitus, Machine Learning, SHAP Value.*

Article info:

Submitted: 28th February 2022 *Accepted: 4th May 2022*

How to cite this article:
N. Permatasari, Shafiyah A.S., A. L. Irfiansyah and M. G. Al-Haqqoni, "PREDICTING DIABETES MELLITUS USING CATBOOST CLASSIFIER AND SHAPLEY ADDITIVE EXPLANATION (SHAP) APPROACH", *BAREKENG: J. Il. Mat. & Ter.*, vol. 16, iss. 2, pp. 615-624, June, 2022.

1. INTRODUCTION

Diabetes is a disease that occurs due to a lack of insulin in the blood, which is a hormone to regulate blood sugar. This is due to the pancreas being unable to produce enough insulin (type 1 diabetes) or the body not being able to use insulin effectively (type 2 diabetes). Diabetes is one of the causes of death in the world, which causes 1.5 million deaths in 2019 [1].Unfortunately, the early symptoms of this are very difficult to detect, even by experienced doctors [2].

Recent research in bioinformatics shows that early detection of diabetes mellitus using machine learning (ML) is better and more efficient than manual detection [2]. Several ML methods that have been implemented in diabetes mellitus detection include: SVM and random forest [2], LightGBM, Glmnet, and XGBoost [3], and decision tree, random forest and neural network [4]. According to Zou et al. [4], random forest predicts the Luzhou hospital physical examination diabetes data with 0.8084 accuracy. Results from Srinivasa et al. [5] using PIMA Indian diabetes dataset found an area under Receiver Operating Characteristic (ROC) curve of neural network with 10-fold using percentage split prediction achieved 0.8452. Similarly, by using Canadian Primary Care Sentinel Surveillance Network (CPCSSN), Lai et al. [6] show that Gradient Boosting Machine (GBM) perform 0.847 of area under ROC curve.

Another study conducted by Kopitar et al. [3] who performed five models comparing regression models and ensemble methods using machine learning from a Slovenian primary healthcare institution. Their experiments perform 0.852, 0.847, and 0.844 area under ROC curve using random forest, LightGBM, and XGBoost method. Using a regression approach, they show AUC values of Glmnet and lm are 0.859 and 0.854. Besides that, a logistic regression model at the National Institute of Diabetes, John Hopkins University which was built by Joshi and Dhakal [7] show 0.7826 accuracy and cross-validation error rate of 0.2286. Similar pattern found from Rajendra and Latifi [8] who compared regression approach with ensemble models using Vanderbilt dataset (a study of rural African Americans in Virginia). Their result perform an accuracy of 0.8889 and 0.9341 using logistic regression and ensemble model. Recently, Kumar et.al [9] also compare the performance of CatBoost, K-Nearest neighbor, Multi-layer perceptron, Logistic regression, Gaussian Naive Bayes, and Stochastic. It results that Catboost classifier gets better performance than other machine learning method. From the literature we found, it can be concluded that ensemble model had outstanding performance.

Although ensemble methods get better result, there will be more challenges in terms of interpreting the results that should support the health care professional's decisions [3]. One recommended method to better understand ML results is the Shapley Additive Explanation (SHAP) approach, which is able to generate interpretation of the ML model and its predictions, and get feature importance values for individual predictions [10]. Hathaway [11] use SHAP visualization to get better interpretation of diabetes mellitus model.

In this paper, we present a machine learning model to predict the probability of a patient having diabetes mellitus. Our study highlights the using of ensemble model to get the best classification model. We also interpret the model to get better insight about how demographic information and laboratory results affect diabetes mellitus patients as an early-detection of diabetes mellitus in the future with SHAP approach.

2. RESEARCH METHODS

This research has several stages of methodology, including: collecting data, preprocessing and validation, classification using CatBoost, hyperparameter tuning, and model interpretation using feature importance/contribution.

2.1 Data Collection

The data used in this research were obtained from Massachusetts Institute of Technology's Global Open Source Severity of Illness Score (MIT's GOSSIS) which provides a new in-hospital mortality prediction algorithm created by the GOSSIS consortium with emphasis on chronic conditions of diabetes. There are 130,157 records used as training data and 10,234 records used as testing data from another dataset. It consists of 181 features of demographic information and laboratory results of diabetes mellitus. The features in the dataset can be grouped into 8 groups shown in Table 1.

Table 1. Features group from the dataset

No	Group	Description
1	Identidier	Unique identifier from hospital
2	Demographic	Demographic variables include the location or type of unit admission
3	APACHE covariate	Disease classification system, ICU scoring system and its component related to diabetes mellitus
4	Vitals	Vital sign
5	Labs	Laboratory result
6	Labs blood gas	Laboratory blood gas result
7	APACHE comorbidity	Status of comorbidity
8	Target Variable	Status of diabetes mellitus

The case definition for diabetes mellitus is already labeled in the training dataset then we predict it using a testing dataset. After we get the data, Figure 1 shows the flow of this paper work and a detailed explanation shows below.

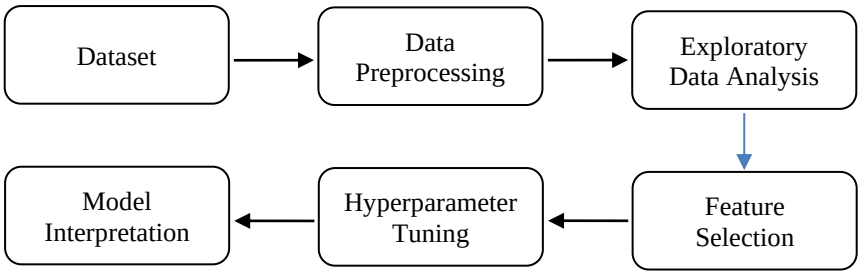

Figure 1. Framework research

2.2 Data Pre-processing

Data preprocessing is the first step of data mining, including: data cleaning, feature engineering, and feature selection. In data cleaning steps, we drop 73 features with its missing values greater than 50%, and impute the others missing values (median values for numerical features and mode values for categorical features). Then, we encoding categorical variables using Target Encoding [12] and using the rank of CatBoost [13] feature importance for variable selection.

Instead of another popular encode algorithm, such as One Hot encode, target encodes can deal with high cardinality categorical variables. Target encoding replaces the categorical value with the posterior probability of target value given the categorical variable's value, added with the prior probability of overall target value. It can be written in Equation (1).

$$S_i = P(X = X_i).\lambda(N_i) + P(Y) \,.(1 - \lambda(N_i)) \tag{1}$$

where S_i is the value to replace the categorical variable X_i, $P(Y|X = X_i)$ is a posterior probability of target value given $X = X_i$, $P(Y)$ is a prior probability of overall target value and $\lambda(N_i)$ is a monotonically non-decreasing function of sample size. In this case, Y is defining the patient's status of diabetes mellitus. The X notation defined factors that probably determine the patient's status of diabetes mellitus.

2.3 CatBoost Algorithm and Validation

In this research we use a CatBoost algorithm which is based on gradient boosted decision trees. CatBoost uses ordered target statistics and ordered boosting that make it good for categorical values of heterogeneous data and has a strong performance relative to other gradient boosting decision tree implementations [13]. It uses symmetric trees for good prediction speed. Each successive tree in the CatBoost algorithm is built with reduced loss compared to the previous trees. It reduces the need for extensive hyper-paramater tuning, uses categorical features directly and scalably, and allows specifying custom functions [14].

A recent study found one of the important issues from their interdisciplinary research is its sensitivity to hyper-parameters and the importance of hyper-parameter tuning [13]. Researcher can set settings for the

maximum number of iterations for CatBoost to use, the maximum depth of constituent Decision Trees, and the maximum number of combinations of categorical features to boost model performance. The values that researcher uses for these hyper-parameters may explain discrepancies in performances of CatBoost.

So, after encode the variables, we use the rank of CatBoost using feature importance for variable selection. The equation of CatBoost feature importance shown in Equation (2) [14].

$$feature\_importance_F = \sum_{tree,leafs_F} (v_1 - avr)^2 . c_1 + (v_2 - avr)^2 . c_2 \quad (2)$$

Where:

- $avr = \frac{v_1 . c_1 + v_2 . c_2}{c_1 + c_2}$
- c_1, c_2 represent the total weight of objects in the left and right leaves respectively. This weight is equal to the number of objects in each leaf if weights are not specified for the dataset
- v_1, v_2 represent the formula value in the left and right leaves respectively.

If the model uses a combination of some of the input features instead of using them individually, an average feature importance for these features is calculated. If the model uses a feature both individually and in a combination with other features, the total importance value of this feature is defined using the formula in Equation (3).

$$feature_{total_{importance_j}} = feature_{importance} + \sum_{i=1}^{N} average\_feature\_importance_i \quad (3)$$

Where:

- $feature\_importance_j$ is the individual feature importance of the j-th feature.
- $average\_feature\_importance_i$ is the average feature importance of the j-th feature in the i-th combinational feature.

To examine the model, we use stratified K-Fold with out of fold validation. Stratified K-Fold validation divides the dataset into non-overlapping folds and preserve the probability of each class in each fold. We use fold as a test dataset to evaluate the model which was built by the other dataset. Overall performance of the model is the average of all folds. Also, the validation we used solves the imbalance data problem in machine learning.

For interpreting the importance of each feature with a better understanding, we use Shapley Additive exPlanations (SHAP) values [11]. The SHAP value ranges between a condition being true (>0.0) and it being false (<0.0). The more sample with a specific value influences the composition of model, the farther the point will get away from zero SHAP value. If the sample doesn't give a meaningful value to the diabetes mellitus, its SHAP value will be near or at zero. Besides that, the darker of the color, which defines by red (has a positive impact on the diabetes mellitus) and blue (has a negative impact on the diabetes mellitus), the stronger its effect to the diabetes mellitus.

2.5 Evaluation Metrics

We use evaluation metrics to evaluate model performance. Hajuan-Tilaki [15] recommend AUC (Area Under Curve) Score to be used as a single number evaluation for disease classification from healthy subjects. A good classification model is indicated by an AUC value close to 1 with a true positive rate close to 1 and a false positive rate close to 0.

2.4 Hyperparameter Tuning

We also use Optuna hyperparameter optimization to improve model performance [16]. The objective of the Optuna optimization is to minimize the Area under the Receiver Operating Characteristic (ROC) curve and the hyperparameter values is defined in Figure 2.

```
params = {
    'max_depth': trial.suggest_int('max_depth', 3, 10),
    'learning_rate': trial.suggest_float('learning_rate', 0.005, 0.1),
    'n_estimators': trial.suggest_int('n_estimators', 50, 3000),
    'max_bin': trial.suggest_int('max_bin', 200, 400),
    'min_data_in_leaf': trial.suggest_int('min_data_in_leaf', 1, 300),
    'l2_leaf_reg': trial.suggest_float('l2_leaf_reg', 0.0001, 1.0, log = True),
    'subsample': trial.suggest_float('subsample', 0.1, 0.8),
    'random_seed': 42,
    'task_type': 'GPU',
    'loss_function': 'Logloss',
    'eval_metric': 'F1',
    'bootstrap_type': 'Poisson' }
```

Figure 2. Hyperparameter tuning using optuna hyperparameter

2.5 Feature Importance / Contribution

Rodríguez-Pérez and Bajorath [10] explained that the Shapley Additive Explanations (SHAP) approach is a methodology that enables the identification and prioritization of features that determine compound classification and activity prediction using any machine learning model. In the context of activity predictions, Shapley values can also be rationalized as a fair or reasonable allocation of feature importance given a particular model output. It calculates features that contribute to the model's prediction with different magnitudes and signs. Features with positive sign contribute to the prediction of activity, whereas features with negative sign contribute to the prediction of inactivity.

The importance of a feature is defined by the Shapley value using the formula in Equation 4 [10].

$$\phi_i = \frac{1}{|N|!} \sum_{S \subseteq N \setminus \{i\}} |S|! \, (|N| - |S| - 1)! \times \left[f(S \cup \{i\}) - f(S) \right] \tag{4}$$

Here $f(S)$ corresponds to the output of the ML model to be explained using a set S of features, and N is the complete of all features. The final contribution or Shapley value of feature $i(\phi_i)$ is determined as the average of its contributions across all possible permutations of a feature set. In this research, we accommodate the SHAP value to determine contribution of each feature.

3. RESULTS AND DISCUSSION

There are 130,157 records in the dataset, consisting of 28,151 of diabetes patients and 102,006 of non-diabetes patients. Demographically, most of the DM patients are over the age of 35 years, which is 86% of all DM patients. However, there is no significant difference in the proportion of DM patients based on gender and ethnicity. DM patients have an average BMI of 31.8 which is slightly higher than non-diabetics of 28.4. The biggest difference between DM patients and non-patients is the glucose and blood urea nitrogen concentration in their serum or plasma in first hour and first day of their unit stay. The glucose and blood urea nitrogen concentration of DM patients are higher than non-patients.

Using CatBoost feature importance selection and filter the features with the importance more than one, we get 14 features of selection variables, shown in Table 2.

Table 2. Result of selection variables using feature importance.

Feature	Feature Importance
d1_glucose_max	20.85047
icu_id	6.82597
Age	5.58073
glucose_apache	4.92322
Bmi	4.50577
hospital_id	4.22718
d1_glucose_min	3.77577
Weight	2.23078
apache_3j_diagnosis	2.13222
apache_2_diagnosis	1.87202
pre_icu_los_days	1.26991
d1_hemaglobin_max	1.25523
creatinine_apache	1.19689
d1_wbc_max	1.15638

Table 2 shows that feature 'd1_glucose_max' has the highest feature importance in this best model with the value of 20.85, it means that the highest glucose concentration of the patient in their serum or plasma during the first 24 hours of their unit stay has the biggest effect for classifying whether the patient has been diagnosed with diabetes mellitus or not. Then followed by 'icu_id', a unique identifier for the unit to which the patient was admitted, has 6.83 for its value of importance. The third is 'age' with the importance value is 5.58. Also, the result from Table 2 shows 14 important features obtained that represents several groups of features, including demographic, APACHE covariate, identifier, and labs features.

Using optuna hyperparameter, the optimal set of parameters we obtained for this CatBoost model is as follows: depth of the tree (max_depth) is 4; the learning rate (learning_rate) is 0.09883; number of tree (n.estimators) is 2950; number of splits for numerical features (max_bin) is 400; minimum number of training samples in a leaf (min_data_in_leaf) is 25; coefficient at the L2 regularization (l2_leaf_reg) is 0.76056; and the sample rate for bagging (subsample) is 0.27346.

We obtained a best AUC Score of 86.86% which is indicating a good performance at distinguishing between the positive and negative classes, in this study predicting the positive and negative of diabetes mellitus. The AUC Score visualized with the ROC curve show in Figure 3.

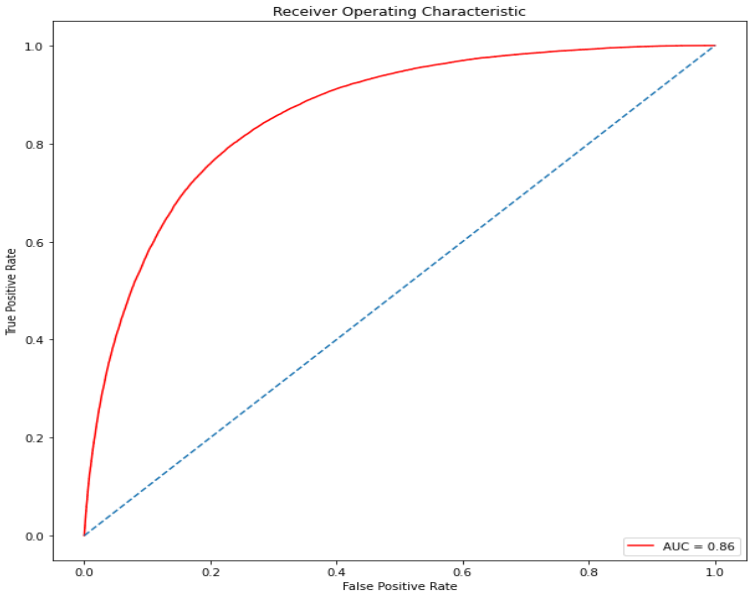

Figure 3. Area under receiver operating curve from catboost model

For better understanding, we use SHAP value visualization to explain the details, shown in Figure 4.

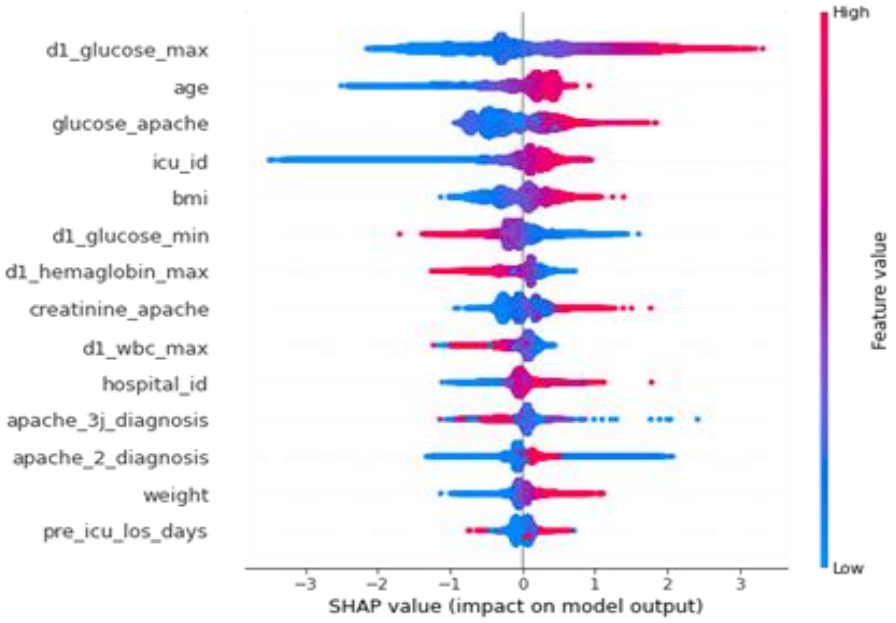

Figure 4. SHAP value of each feature for interpreting the probability of a patient diagnosed as diabetes mellitus patient.

Figure 4 shows the feature importance of all selected features across the data. SHAP values are important to define how big the feature affects the classifiers. The red color indicates the positive correlation of the feature to the model and blue otherwise. The farther line graphic from zero also indicates the feature to have a strong influence in classifying the diabetes mellitus patients. As we can see from Figure 4, it shows that d1_glucose_max has a red line far from the zero with red dark. Also, there are several features that has a dark blue with the value quite far from zero. For better understanding, we simplify the graph shown in Figure 5.

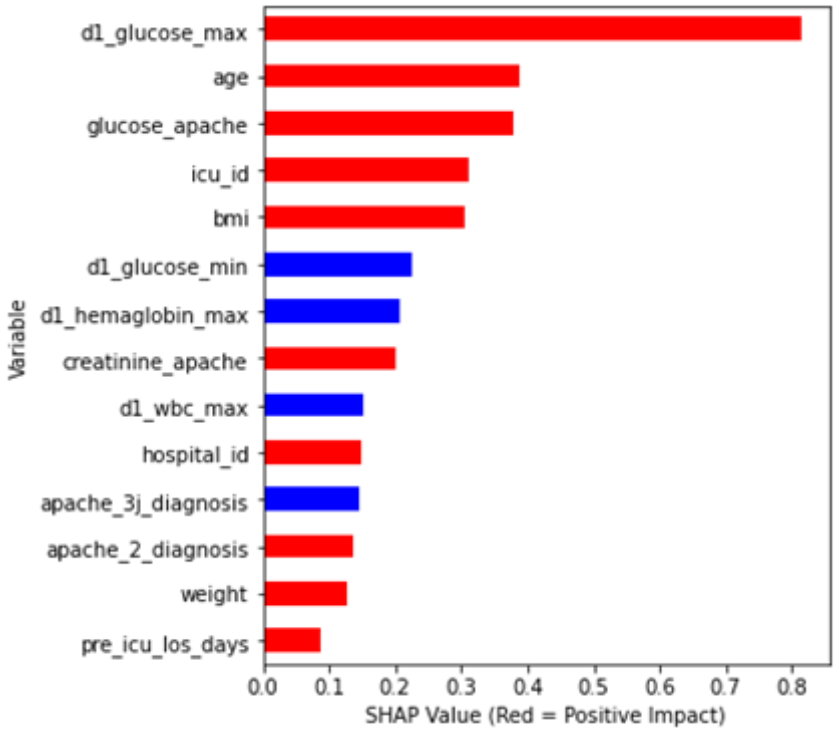

Figure 5. SHAP values and its direction of correlation that indicates probability of having diabetes mellitus for the patient.

By using another visualization of SHAP value, Figure 5 shows four features that have negative correlations to the model, which are 'd1_glucose_min' (the lowest glucose concentration of the patient in their serum or plasma during the first 24 hours of their unit stay), 'd1_hemaglobin_max' (the highest hemoglobin concentration for the patient during the first 24 hours of their unit stay), 'd1_wbc_max' (the highest white blood cell count for the patient during the first 24 hours of their unit stay), and 'apache_3j_diagnosis' (the APACHE III-J sub-diagnosis code which best describes the reason for the ICU admission). It means that by increasing the increase of selected features above by one value, it will decrease the probability of the targeted patient as a diabetes mellitus patient by one value.

The next nine features are 'd1_glucose_max' (the highest glucose concentration of the patient in their serum or plasma during the first 24 hours of their unit stay), 'age', 'glucose_apache' (the glucose concentration measured during the first 24 hours which results in the highest APACHE III score), 'bmi' (the body mass index of the person on unit admission), 'creatinine_apache' (the creatinine concentration measured during the first 24 hours which results in the highest APACHE III score), 'hospital_id', 'apache_2_diagnosis' (the APACHE II diagnosis for the ICU admission), 'weight', and 'pre_icu_los_days' (the length of stay of the patient between hospital admission and unit admission) have positive correlation. It means that every one points increase of each feature before, it will increase one value of the probability of the targeted patient as a diabetes mellitus patient. For the 'icu_id' feature, we assume that diabetes mellitus patients are being grouped in certain ICU rooms.

Figure 5 also shows that features 'd1_glucose_max' and 'd1_glucose_min' have a different color, which means one has a positive correlation and one has a negative correlation. It gives us an insight that based on the glucose concentration, whether it is too high or too low, the model will diagnose the patient to be a diabetes mellitus patient, which is true. If the patient has low glucose, we could say that the patient has hypoglycemia which happens to people with diabetes when they have a mismatch of medicine, food, and/or exercise. If the patient has high glucose concentration, we could say that the patient has hyperglycemia or high blood sugar that indicates diabetes symptoms.

Furthermore, SHAP value also explains that a feature with large absolute value shows that the feature affects the classifiers bigger than the feature with lower absolute SHAP values. Feature 'd1_glucose_max' has a significant Shapley value than the others which have the value of 20.85047, which mean that if the number of the highest glucose concentration of the patient in their serum or plasma during the first 24 hours of their unit stay is high, it means that the patient have a bigger probability to be diagnosed as a diabetes mellitus patient. The high SHAP value is then followed by the features of 'age', 'glucose_apache', 'icu_id', until 'pre_icu_los_days' that has the lowest SHAP value.

4. CONCLUSION

In the end, we can conclude that our model has a good performance to classify whether a patient has diabetes mellitus or not with the validation AUC score of 86.86% using CatBoost classifier. From the model, the result shows 'd1_glucose_max' has the highest SHAP value which means the highest glucose concentration of the patient in their serum or plasma during the first 24 hours of their unit stay can be an early-detection for patient having diabetes mellitus. Besides that, the model can be wrongly detected a patient because of some reasons, such as a rare case when a patient has a non-diabetic hyperglycemia. These can be included for future research, a more in-depth experiment of classification methods and carried out extracting new features in accordance with medical principles.

REFERENCES

[1] World Health Organization, "Diabetes," [Online]. Available: https://www.who.int/health-topics/diabetes .

[2] J. Chaki, S. T. Ganesh, S. Cidham and S. A. Theertan, "Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review," *Journal of King Saud University - Computer and Information Sciences*, pp. 1-22, 2020.

[3] L. Kopitar, P. Kocbek, L. Cilar, A. Sheikh and Stiglic, "Early Detection of Type 2 Diabetes Mellitus Using Machine Learning-Based Prediction Models," *Scientific Reports*, vol. 10/11981, 2020.

[4] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju and H. Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques," *Frontiers in Genetics*, vol. 9, 2018.

[5] Y. S. j. V. K. B. Y. S. P. Srinivasa R., "Prediction of Diabetes using Machine Learning," *International Journal of Advanced Science and Technology*, vol. 29, pp. 7593-9601, 2020.

[6] H. Lai , H. Huang, K. Keshavjee, A. Guergachi and X. Gao, "Predictive Models for Diabetes Mellitus Using Machine Learning Techniques," *BMC Endocr Disord*, vol. 19, pp. 1-9, 2019.

[7] R. D. Joshi and C. K. Dhakal, "Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches," *International Journal of Enviromental Research and Public Health*, vol. 18, pp. 1-17, 2021.

[8] P. Rajendra and S. Latifi, "Prediction of diabetes using logisctic regression and ensemble techniques," *Computer Methods and Programs in Biomedicine Update*, vol. 1, pp. 1-8, 2021.

[9] P. S. Kumar, A. Kumari K, S. Mohapatra, B. Naik, J. Nayak and M. Mishra, "CatBoost Ensemble Approach for Diabetes Risk Prediction at Early Stages," in *1st Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology(ODICON)*, Bhubaneswar, India, 2021.

[10] R. Rodriguez-Perez and J. Bajorath, "Interpretation of Machine Learning Models Using Shapley Values: Application to Compound Potency and Multi-Target Activity Predictions," *Journal of Computer-Aided Molecular Design*, vol. 34, no. 10, pp. 1013-1026, 2020.

[11] Q. A. Hathway, S. M. Roth, M. V. Pinti, D. C. Sprando, A. Kunovac, A. J. Durr, C. C. Cook, G. K. Fink, T. B. Cheuvront, J. H. Grossman, G. A. Aljahli, A. D. Taylor, A. P. Giromini, J. L. Allen and Hollander John M., "Machine-Learning to Stratify Diabetic Patients Using Novel Cardiac Biomarkers and Integrative Genomics," *Cardiovasc Diabetol*, vol. 18, no. 78, 2019.

[12] W. McGinnis, "Target Encoder," 2016. [Online]. Available: https://contrib.scikit-learn.org/category_encoders/targetencoder.html.

[13] J. T. Hancock and T. M. Khosghoftaar, "CatBoost for Big Data: An Interdisciplinary Review," *Journal of Big Data*, vol. 7, no. 94, pp. 1-45, 2020.

[14] Yandex, "CatBoost," 2021. [Online]. Available: http://yandex.com/dev/catboost.

[15] K. Hajian-Tilaki, "Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation," *Caspian Journal of Internal Medicine*, vol. 4, no. 2, pp. 627-635, 2013.

[16] T. Akiba, S. Sano, T. Yanase and T. Ohta, "Optuna: A Next-generation Hyperparameter Optimization Framework," in *KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019.