# An Explainable Approach to Diabetes Mellitus Prediction Using ICU Records

Hena Ghonia
*Associate Data Scientist*
*Ecolab*
Pune, India
henaghonia2015@gmail.com

Shanaya Mehta
*Department of Data Science*
*Harbour Space University*
Barcelona, Spain
msdec.ict@gmail.com

Vyoma Patel
*Department of Computer Science*
*University of Windsor*
Toronto, Canada
vyomapatel01@gmail.com

***Abstract*—Patients admitted to the intensive care unit (ICU) are under distress and unable to provide information on pre-existing chronic health conditions such as heart disease, allergies or diabetes. ICUs often lack verified medical records for admitted patients. Transferring medical records from one medical provider to another may also take a long time. The study of patients with diabetes mellitus is vital, especially in the Coronavirus pandemic where the hospitals are overloaded and getting a rapid understanding of overall patient health is crucial. Based on the analysis of ICU data, this article studies and analyzes the various factors that determine diabetes in an admitted patient. Finally, this article develops a predictive machine learning model and explores a method to interpret the results to improve our understanding of medical data.**

***Index Terms*—machine learning, prediction, interpretability**

## I. Introduction

The momentous advances in the fields of health services and big data have led to a significant production of data, such as detailed patient health information, generated from Electronic Health Records (EHRs). The data stored in an EHR includes the demographics information, medication, laboratory results and clinical observations [10]. By utilizing this data, it is feasible to determine chronic illnesses, such as cancer and diabetes mellitus, in patients and in certain cases, prevent the illness before it manifests.

Diabetes mellitus is a common chronic condition affecting a large population across the world. It is a metabolic disease in humans with high morbidity. It is characterized by hyperglycemia which results from defects in insulin secretion, or insulin action, or both.

For individual patient care in non-ICU (Intensive Care Unit) settings, physicians are likely well-equipped to identify diabetes and monitor the patient's health. However, in an ICU where there are a number of patients needing urgent care and a limited staff available to provide care to these patients, it becomes more challenging to keep a track of unknown chronic ailments that might interfere with the treatment being provided to the patient. Failure to provide the right medication to a patient in an ICU may prove to be fatal. Knowledge of this disease can lead to better patient outcomes.

Data mining and predictive modelling techniques have made it possible to study and explore various risk factors for diabetes [7] [5] [4]. We propose a system that can predict if a patient has diabetes by using various machine learning approaches and attempt to generalize the system for unseen data.

Much of the research in diabetes mellitus prediction has been done in the recent years. Hui et al. [3] used neural networks to classify electronic medical records for diabetes mellitus and achieved a Receiver Operator Characteristics (ROC) score of 0.82. Their paper identified age, LDL-C, HDL-C, Triglyceride, total cholesterol, Diabolic Blood Pressure (DBP), Systolic blood pressure (SBP) and Body Mass Index (BMI) as the salient variables for prediction. Ioannis Kavakiotis et al [5] studied and compared the results of six classifiers, Logistic regression, Support Vector Machine (SVM), Naive Bayes, Gradient Boosting and Random Forest Classifier on the Pima Indian Diabetes Database. The Random Forest classifier achieved the highest accuracy of 98.48% from all 6 classifiers. The techniques of Logistic Regression and Gradient Boosting Machine (GBM) were used in the study by [7], and these models were compared to other machine learning techniques such as Decision Tree and Random Forest. The AROC for the proposed GBM model is 84.7%, with a sensitivity of 71.6%. Finally, in the article by Maria Athanasiou et al. [1], a novel approach to study the results of machine learning predictions for Type-2 diabetes mellitus is proposed. The article discusses predictions made using XGBoost Classifier and explains the results of the classifier using TreeSHAP (SHapely Additive exPlanations), a modified version of the SHAP explainer. The developed risk prediction model was evaluated by applying stratified 10-fold cross-validation with the accuracy of 71.13±11.69% and 71.00±23.85% sensitivity.

## II. Dataset

The data used for the research is made available through the Women in Data Science (WiDS) Datathon 2021 to all the participants. The training data consists of 130,157 encounters or training samples and 181 feature columns. These feature columns can be grouped as follows:

- Identifiers, such as hospital id and encounter id
- Patient Demographics
- Patient Vitals

- Lab Results, including blood gas
- APACHE Comorbidities
- APACHE Covariates
- Target column

Of these 180 features, not all are relevant or related and thus feature selection is done. Feature selection is done on the basis of statistical analysis, as well as with reference to medical literature. In this study, the identifiers variables, namely encounter ID and hospital ID can be safely dropped since identifiers do not attribute to the decision making process. The variable readmission_status is constant in both train and test datasets, so it is dropped.

Further, correlation analysis is done in order to examine the linearity of variables and features which were highly correlated with different distribution in train and test data are dropped.

TABLE I
STATISTICAL DESCRIPTION OF SELECTED FEATURES

| **Features Names** | **Feature distribution** | | |
|---|---|---|---|
| | ***Mean*** | ***median*** | ***std*** |
| d1_glucose_max | 161.12 | 145.00 | 93.88 |
| glucose_d1_value_range | 55.92 | 30.00 | 78.87 |
| age | 59.61 | 63.00 | 20.34 |
| d1_creatinine_min | 1.22 | 0.89 | 1.33 |
| d1_wbc_max | 10.86 | 10.10 | 7.62 |
| ethnicitycat_bmi | 7.53 | 5.00 | 7.64 |
| cat_bmi-arf_apache | 7.32 | 7.00 | 2.87 |

Above table shows distribution of important features where glucose_d1_value_range, ethnicitycat_bmi and cat_bmiarf_apache are calculated variables.

## III. METHODOLOGY

The next step after analyzing the data is to create a model for making predictions. The proposed system consists of defining a baseline and then improving upon the baseline model by using various ensemble methods. In particular, this article explores gradient boosting methods, LightGBM [6] and XGBoost [2]. The model used for setting the baseline in this study was the Dummy Classifier. We obtained an accuracy of 78.4% and Area Under Curve metric score as 50%.

We experimented by dropping highly correlated features, plotting importance plots and fine-tuning hyperparameters of the models to improve upon the baseline.

We experimented with various gradient boosting methods such as XGBoost, CatBoost and LightGBM. The accuracy of 84.1% was obtained with default LightGBM parameters, which was the highest from all three boosting methods. To improve on this score, we performed 5 fold cross validation using two different sets of parameters to determine which sets of parameters yield the best results for the LightGBM model.

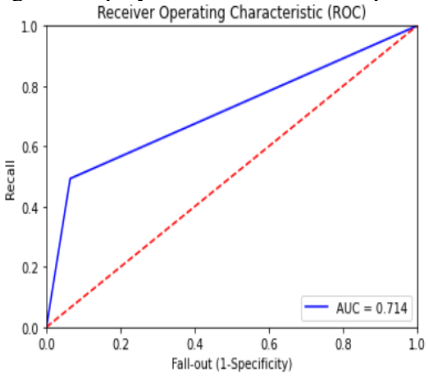
## IV. MODEL RESULTS

The training data was divided into train and test set in the ratio of training : test as 8:2. The training set had 130157 samples and the test set had 10234 samples. The result for train and test set metric is displayed in the following table.

TABLE II
MODEL METRICS ON TRAIN AND TEST SET

| **Diabetes Mellitus** | **train set** | | |
|---|---|---|---|
| | ***precision*** | ***recall*** | ***f1 score*** |
| 0 | 89% | 95% | 92% |
| 1 | 75% | 56% | 64% |
| | **test set** | | |
| 0 | 87% | 94% | 90% |
| 1 | 68% | 49% | 57% |

The performance of the proposed system is calculated using the Area Under the Curve (AUC) metric. The overall AUC score for LightGBM model 71.4%. The following figure displays AUCROC curve plot for lightGBM model

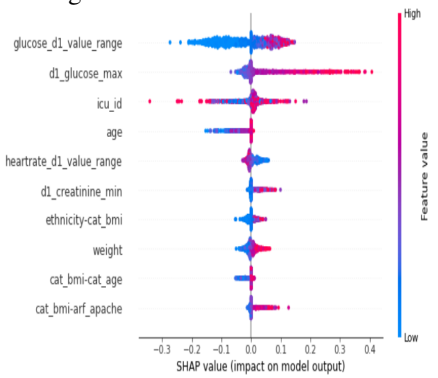



## V. MODEL EXPLAINABILITY

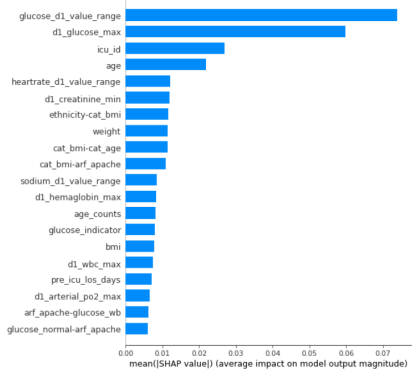
Medical data is used by health care providers such as nurses, doctors and pharmacists. To understand and explain how the model makes decisions, this article explores the concept of model explainability through various existing methodologies. Model explainability is a model-agnostic approach which means that the decisions taken by the algorithm are dependent on the data provided to the model and the decisions can be easily identified. There are several existing methods to achieve this, such as SHapely Additive exPlanations (SHAP) value plots [8], model importance plots, individual regression plots and tree visualization plots. Model explainability can be achieved with known data only since we can explain the decisions taken by the algorithm only when the end result is known to us beforehand; by making use of unseen data, it is difficult to explain the algorithmic decisions. In this article, we will be exploring SHAP plots and model importance plots in more detail.

*1) SHAP plots:* SHAP value refers to contribution of individual feature value to a prediction. Positive SHAP value indicates a prediction towards higher probability of diabetes mellitus and negative SHAP value indicates a prediction towards lower probability of diabetes mellitus. Larger the SHAP value magnitude, more important the driver is. The following figure shows that glucose_d1_value_range(Calculated as d1_glucose_max - d1_glucose_min) is most important feature and higher the feature value (color red), it has

higher chances of predicting diabetes mellitus. Similarly higher the value indicated by positive SHAP value of d1_creatinine_min, ethnicity-cat_bmi(calculated as sum of bmi mean and standard deviation over grouped by ethnicity), cat_bmi-cat_age (calculated as sum of mean and standard deviation of bmi and age), cat_bmi-arf_apache (calculated as sum of mean and standard deviation of bmi grouped over arf_apache), higher the chances(redness in SHAP plot) of predicting presence of diabetes mellitus. For feature heartrate_d1_value_range(calculated as heartrate_d1_max - heartrate_d1_min), its lower value has higher chances of prediciting diabetes mellitus.




*2) Model Importance Plot:* The model importance plot helps in identifying which feature holds more importance in the given data. Following figure shows the feature importance plot for the top 20 variables. The most important features are d1_glucose_max, icu_id and age. It is interesting to note that the ICU identifier is included in the most important features.




## VI. Conclusion

An ensemble learning approach was proposed in this paper along with a method to interpret the findings of the study. The use of LightGBM techniques achieves an acceptable performance despite the low number of unique patient situations, demonstrating the system's potential to handle the unbalanced nature of the dataset. Further, the preliminary work done to interpret the model results using the SHAP plots and importance plots demonstrates the system's ability to be useful to medical health workers.

Future work concerns the extension of different learning techniques such as artificial neural networks with more interpretability methods to facilitate the knowledge of diabetes mellitus in patients in the ICU.

## Acknowledgment

We would like to thank the authors of the notebook [9] on the Kaggle platform for sharing their original work. Their work inspired us to write this article.

## References

[1] Maria Athanasiou et al. "An explainable XGBoost–based approach towards assessing the risk of cardiovascular disease in patients with Type 2 Diabetes Mellitus". In: *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE. 2020, pp. 859–864.

[2] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 785–794. ISBN: 9781450342322. DOI: 10.1145/2939672.2939785. URL: https://doi.org/10.1145/2939672.2939785.

[3] Chen Hui et al. "Research on diabetes prediction method based on electronic medical record data analysis". In: *E3S Web of Conferences*. Vol. 185. EDP Sciences. 2020, p. 03001.

[4] T. Jayalakshmi and A. Santhakumaran. "A Novel Classification Method for Diagnosis of Diabetes Mellitus Using Artificial Neural Networks". In: *2010 International Conference on Data Storage and Data Engineering*. 2010, pp. 159–163. DOI: 10.1109/DSDE.2010.58.

[5] Ioannis Kavakiotis et al. "Machine learning and data mining methods in diabetes research". In: *Computational and structural biotechnology journal* 15 (2017), pp. 104–116.

[6] Guolin Ke et al. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 3149–3157. ISBN: 9781510860964.

[7] Hang Lai et al. "Predictive models for diabetes mellitus using machine learning techniques". In: *BMC endocrine disorders* 19.1 (2019), pp. 1–9.

[8] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[9] Siavash, Kain, and Dan Ofer. URL: https://www.kaggle.com/siavrez/2020fatures.

[10] Swati Yanamadala et al. “Electronic Health Records and Quality of Care: An Observational Study Modeling Impact on Mortality, Readmissions, and Complications”. In: *Medicine* 95.19 (2016). ISSN: 0025-7974. URL: https : / / journals . lww . com / md - journal / Fulltext / 2016/05100/Electronic_Health_Records_and_Quality_of_Care__An.10.aspx.