Interpretable Scorecard Model for Early Diabetes Classification

Devin Setiawan, Sien Gong, Anto Felix, Maisoon Rahman

Abstract

Background: Interpretable machine learning is crucial in clinical settings, where high-stakes decisions require transparency and trust. In applications like early diabetes diagnosis, models that provide clear, actionable insights are essential for improving patient outcomes. This project explores the use of the FasterRisk algorithm to generate interpretable risk scorecards for early diabetes classification. Scorecard models, which offer simple, additive predictions, are well-suited for clinical environments due to their human-readable format. Drawing inspiration from key works like FasterRisk and RiskSLIM, we apply these methods to the Early Diabetes dataset, validating the model's feature selection and comparing its performance with other machine learning approaches.

Methods: This study uses the Early Diabetes dataset from Kaggle, consisting of 520 observations with 17 characteristics, to explore machine learning models for early-stage diabetes classification. The methodology follows the concept of simpler models, testing various algorithms like Logistic Regression, CART, Random Forest, Gradient Boosted Trees, and SVM using 5-fold cross-validation. Two feature importance methods—correlation-based and SHAP values—validate the relevance of selected features for the scorecard model. The risk scorecard is generated using the FasterRisk algorithm, which efficiently creates sparse, interpretable risk scores through a beam-search and optimization process, with cross-validation to determine the optimal sparsity parameter for model performance.

Findings: The evaluation of the scorecard model using the FasterRisk algorithm demonstrated a strong performance, achieving a test accuracy of 91.3% and an AUC of 0.973. This performance is competitive with other models, such as Random Forest (97.5% accuracy), though the latter sacrifices interpretability for higher accuracy. The scorecard, on the other hand, strikes a balance between predictive power and clinical usability by offering a transparent, interpretable risk prediction tool suitable for healthcare settings. The scorecard's design incorporates the top features identified by SHAP values, such as polyuria, polydipsia, and gender, ensuring its consistency with prior feature importance analysis. These factors make the scorecard a valuable clinical decision-support tool, offering good accuracy while maintaining clarity and ease of use for healthcare professionals.

Interpretation: Our scorecard provides interpretable ways of diagnosing early diabetes that offers easy-to-use but also accurate way of assessment. The scorecard format allows health professionals to easily remember the features and use the model without needing the complex decision making process of a complex model.

1. Background

Interpretable machine learning has become increasingly vital in the clinical domain, where decisions often involve high stakes and life-altering consequences. Unlike black-box models, interpretable models enable clinicians to understand and trust the predictions made by algorithms, fostering greater confidence in their integration into medical decision-making. In high-stakes scenarios such as early diagnosis and treatment planning, explainable predictions are crucial for accountability, regulatory compliance, and ensuring that decisions are aligned with clinical knowledge and ethical standards. Early classification of diabetes is especially critical, as timely detection can significantly improve patient outcomes by enabling earlier intervention, reducing the risk of severe complications, and improving quality of life. This underscores the need for machine learning tools that not only deliver accurate predictions but also provide transparent and actionable insights. In this project, we will discuss how the FasterRisk algorithm can be used to efficiently produce interpretable risk scorecards for early diabetes classification (Liu, et al., 2022).

A scorecard model is a type of interpretable machine learning model that provides predictions in a simple, additive form, often expressed as a set of weighted rules or scores. These sparse additive models are particularly suited for clinical applications because of their human-readable format, allowing healthcare professionals to easily understand, apply, and even memorize key decision rules. By converting complex relationships into simple, intuitive scores, scorecard models facilitate easier diagnoses, quicker decision-making, and improved communication among clinical teams and patients.

This project draws inspiration from two significant works. The first, FasterRisk: Fast and Accurate Interpretable Risk Scores by Liu et al., introduces a state-of-the-art algorithm for generating interpretable risk scores that can be used for constructing scorecards. FasterRisk advances the efficiency and accuracy of creating these models, making it a powerful tool for real-world applications. The second, Learning Optimized Risk Scores by Ustun and Rudin, presents RiskSLIM, a predecessor to FasterRisk (Ustun and Rudin, 2019). RiskSLIM was successfully deployed in collaboration with Massachusetts General Hospital to develop a customized risk score for predicting ICU seizures, demonstrating the practical utility and transformative potential of such interpretable models in critical medical settings. These foundational works have inspired the development of this interpretable scorecard model for early diabetes classification.

In this work, we use the Early Diabetes dataset from Kaggle to train and evaluate our models (Islam, et al., 2020). We first experimented with a variety of machine learning models to explore whether a simpler model exists for this dataset, following the methodology proposed in On the Existence of Simpler Machine Learning Models (Semenova, et al., 2022). We then conduct two

feature importance analyses—correlation-based feature importance and SHAP-based feature importance—to validate that our scorecard selects the most relevant features rather than arbitrary or insignificant ones (Lundberg and Lee, 2017). Finally, we compare the performance of our scorecard with other machine learning models in terms of accuracy and AUC, demonstrating its competitiveness while retaining interpretability.

2. Methods

2.1. Early Diabetes Dataset

The dataset we used is cited from the work of Islam et al. (2020): Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques, published in Computer Vision and Machine Intelligence in Medical Image Analysis. This dataset is publicly available on kaggle and can be downloaded and accessed freely on the internet from the following link: https://www.kaggle.com/datasets/andrewmvd/early-diabetes-classification. The dataset used in this study comprises 520 observations with 17 characteristics. These data points were collected through direct questionnaires and diagnostic results from patients at the Sylhet Diabetes Hospital in Sylhet, Bangladesh. The dataset provides a valuable resource for analyzing the factors contributing to early-stage diabetes and building predictive models.

2.2. On the Existence of Simpler Models

The concept of testing for the viability of simpler models is inspired by Rashomon Set Theory, which posits that if multiple models with differing complexities perform similarly and generalize well to validation data, then it is likely that a simpler model exists. This methodology provides a systematic way to explore the trade-off between model complexity and performance, encouraging the use of interpretable models when feasible.

To investigate this for the Early Diabetes dataset, we implemented a 5-fold cross-validation procedure across the following machine learning models, each with varying levels of complexity:

- Logistic Regression
- Classification and Regression Tree (CART)
- Random Forest
- Gradient Boosted Tree
- Support Vector Machine (SVM)

For each model, we recorded both the training and validation accuracy across the folds. This comprehensive approach enabled us to assess the generalization performance of models ranging from simple linear classifiers to more complex ensemble and kernel-based methods. The training and test accuracies were then plotted to evaluate the consistency and generalization ability of the models. This visualization helps identify whether a simpler, interpretable model can achieve

comparable performance to more complex alternatives, providing a foundation for further refinement and scorecard development.

2.3. Feature Importance Analysis

The purpose of the feature importance analysis is to validate whether the features selected by the scorecard model align with the most important features identified through independent feature importance techniques. This ensures that the generated scorecard uses relevant and meaningful features, enhancing its credibility and reliability in clinical decision-making. The analysis was conducted using two different methods for determining feature importance:

1. **Correlation-Based Feature Importance:** The first approach involves calculating the correlation of each feature with the target variable. This method provides a linear assessment of feature importance, as it directly measures the strength and direction of a linear relationship between individual features and the outcome. While straightforward and computationally efficient, this method may fail to capture more complex, non-linear interactions between features and the target variable, limiting its comprehensiveness.

2. **SHAP Values for Feature Importance:** The second approach uses SHAP (SHapley Additive exPlanations) values, a post hoc model-agnostic explanation method. SHAP values provide a detailed explanation of a model's predictions by quantifying each feature's contribution to the output for individual predictions. In this study, we trained a logistic regression model and used SHAP to explain its decisions. This method is more robust than correlation-based analysis, as it can capture both linear and non-linear interactions between features and the target variable, offering a more accurate representation of feature importance.

By combining these two approaches, we aim to validate that the scorecard model prioritizes features that are consistently deemed important by both simple linear and more nuanced, non-linear analyses. This step is crucial to ensure the scorecard's reliability and interpretability in a clinical context.

2.4. Scorecard Generation

The risk scorecard was generated using the FasterRisk algorithm, which is known for efficiently creating high-quality risk scores from data. We utilized the publicly available implementation of FasterRisk from the GitHub repository referenced in the paper. The algorithm operates by first using a beam-search process to produce a collection of nearly optimal sparse continuous solutions, each with a unique set of selected features. These solutions are then transformed into risk scores through a “star ray” search, where a range of multipliers is explored. The coefficients are sequentially rounded to maintain low logistic loss, resulting in a variety of interpretable and

sparse risk scores. This method is highly efficient, completing the process within minutes and delivering multiple high-quality models for user consideration.

To customize the FasterRisk implementation, we incorporated our own cross-validation procedure to optimize the sparsity parameter *k*. Specifically, we conducted 5-fold cross-validation on *k* and used the Area Under the Receiver Operating Characteristic Curve (AUC) as the evaluation metric. This ensured that the chosen sparsity level resulted in the best-performing model, balancing simplicity and predictive performance.

At a high level, the procedure for generating the scorecard was as follows:

1. **Import and preprocess data**: Continuous variables were binarized to simplify the interpretation and align with the risk score structure.
2. **Cross-validation for sparsity selection**: We conducted 5-fold cross-validation to identify the optimal sparsity *k* based on the AUC metric.
3. **Scorecard generation and evaluation**: Using the entire dataset and the optimal sparsity *k*, we generated the final scorecard model. This model was then evaluated on a separate 20% held-out test set to assess its performance.

This process ensures that the generated scorecard is not only interpretable but also optimized for accuracy and generalizability.

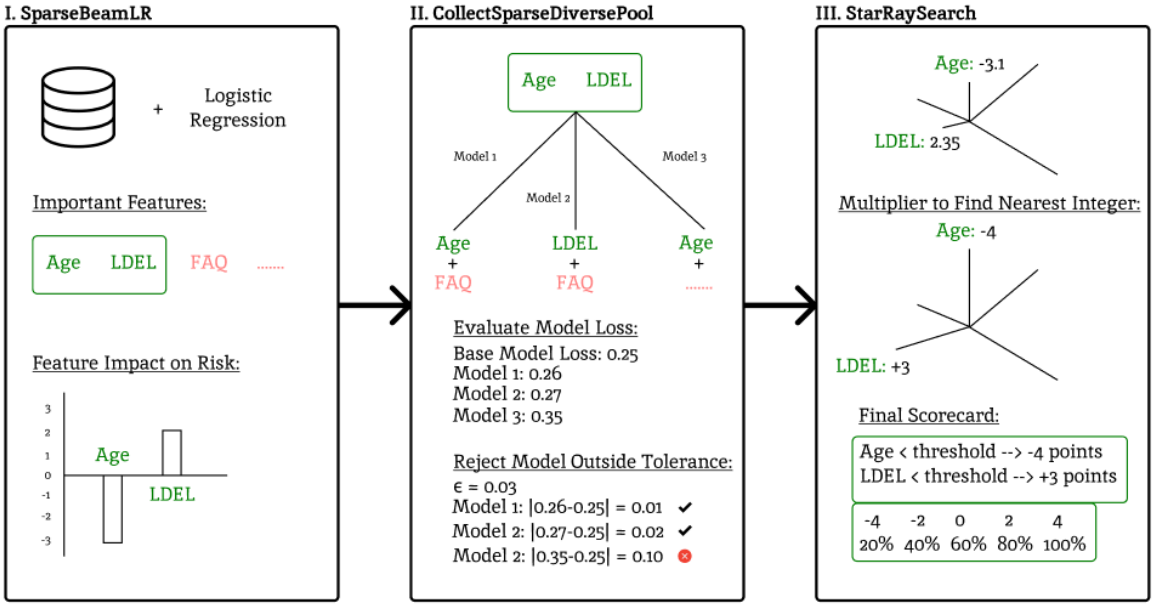


Figure 1: **FasterRisk Algorithm.** This figure presents a visual representation of the FasterRisk algorithm. In SparseBeamLR, we find the best sparse model based on a logistic loss metric. Then, we manipulate the features to find a pool of "good enough" models that doesn't differ too

much in loss. Finally, we find the best multiplier to convert the coefficients to integer coefficients.

3. Findings and Interpretation

3.1. Performances Across Different Models

The results of the cross-validation can be seen in Figure 2, which highlight that simpler machine learning models can achieve performance levels comparable to those of more complex models. For example, Logistic Regression, a simple linear model, achieved a test accuracy of 92.69%. In comparison, more complex models like Random Forest and Gradient Boosting Classifier achieved slightly higher test accuracies of 97.50% and 96.92%, respectively. Despite the marginal differences in performance, the results indicate that simpler models are not drastically inferior in their predictive capabilities. This finding suggests the presence of a large Rashomon Set within the dataset, as multiple models of varying complexity perform similarly well. The Rashomon Set theory posits that when a dataset has many models with comparable performance, it is likely that simpler, interpretable models exist that can effectively capture the underlying patterns.

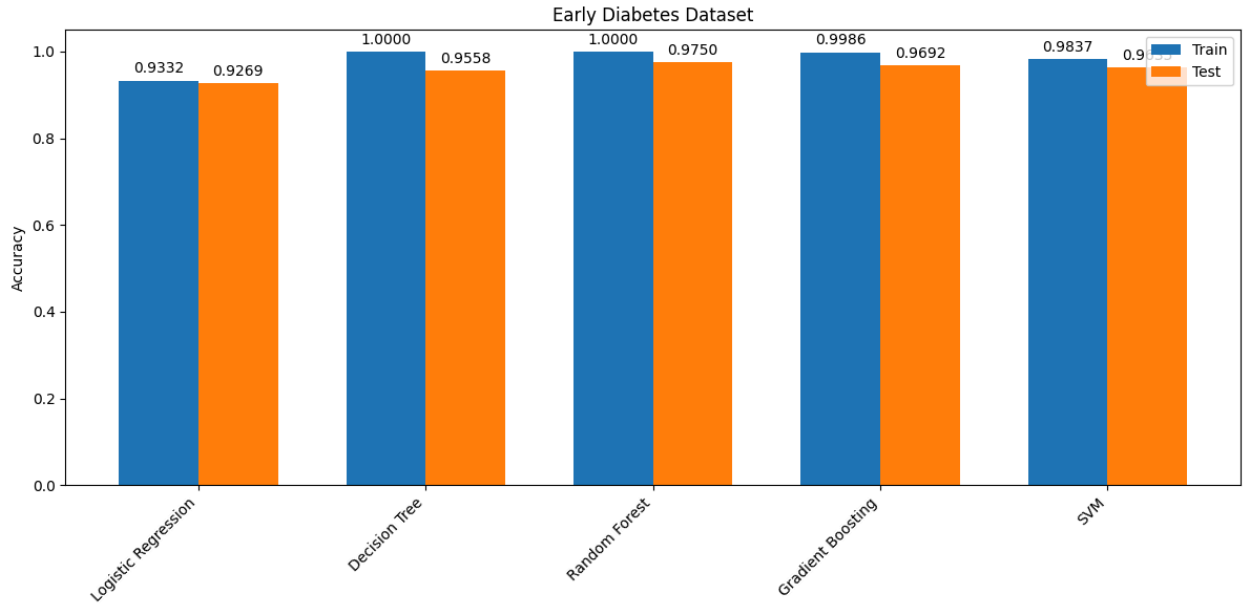


Figure 2: **Early Diabetes Dataset Model Performances.** This figure presents a comparative overview of the accuracy score across different models of differing complexities. The train and test results were obtained by averaging the accuracies of 5-fold cross validation on each fold.

3.2. Feature Importance Analysis Results

The feature importance analysis conducted through two methods, Correlation Coefficient Analysis and SHAP Value Analysis, offers valuable insights into the most influential features in predicting early diabetes. In the correlation analysis, the absolute ranking of features based on

their correlation with the target variable places *polyuria* and *polydipsia* at the top, indicating their strong linear relationship with diabetes. Other relatively important features include *gender*, *sudden_weight_loss*, and *partial_paresis*. In contrast, features such as *itching*, *delayed_healing*, and *obesity* exhibit the lowest correlation values, suggesting weaker linear relationships with the target variable.

The SHAP analysis of the logistic regression model also emphasizes *polydipsia* and *polyuria* as the most significant features, aligning with the results from the correlation analysis. However, SHAP places a higher importance on *gender*, *itching*, and *irritability* compared to correlation analysis, while giving less weight to *alopecia*, *delayed_healing*, and *obesity*. The disparities in feature importance rankings between the two methods stem from the fundamental differences in how they assess the relationship between features and the target variable. Correlation analysis measures the direct linear relationship between each feature and the target variable, treating each feature independently without considering any potential interactions. In contrast, SHAP values provide a post hoc explanation of a trained model's predictions, accounting for interactions, non-linear effects, and the combined contribution of features to the final decision. This nuanced approach allows SHAP to capture complexities that simple correlation analysis cannot. Thus, while correlation analysis offers a more basic, linear perspective, SHAP provides a deeper, model-specific understanding of feature importance.

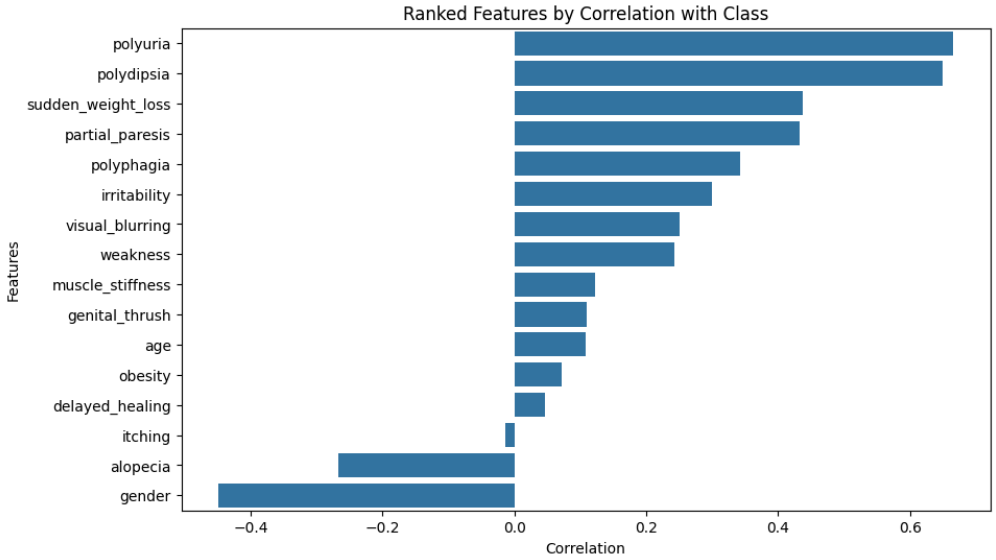


Figure 3: **Correlation-based Feature Importance.** This figure presents a comparative overview of the correlation score across different features. The higher the correlation regardless of the sign of the correlation indicates a stronger feature importance. A positive correlation indicates that when the feature value increases, the output also does the same. A negative correlation indicates that when the feature value increases, the output does the opposite.

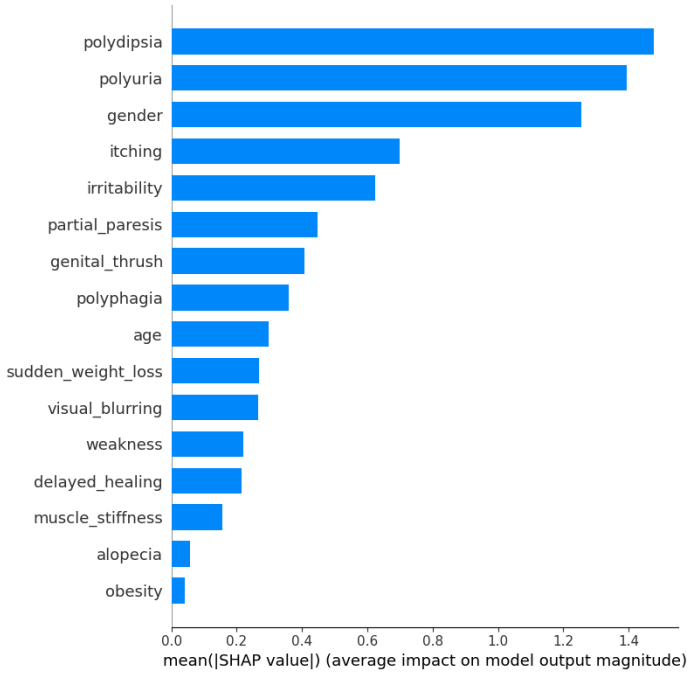


Figure 4: **SHAP-based Feature Importance.** This figure presents a comparative overview of the absolute mean SHAP value across different features. The higher the value indicates a stronger feature importance.

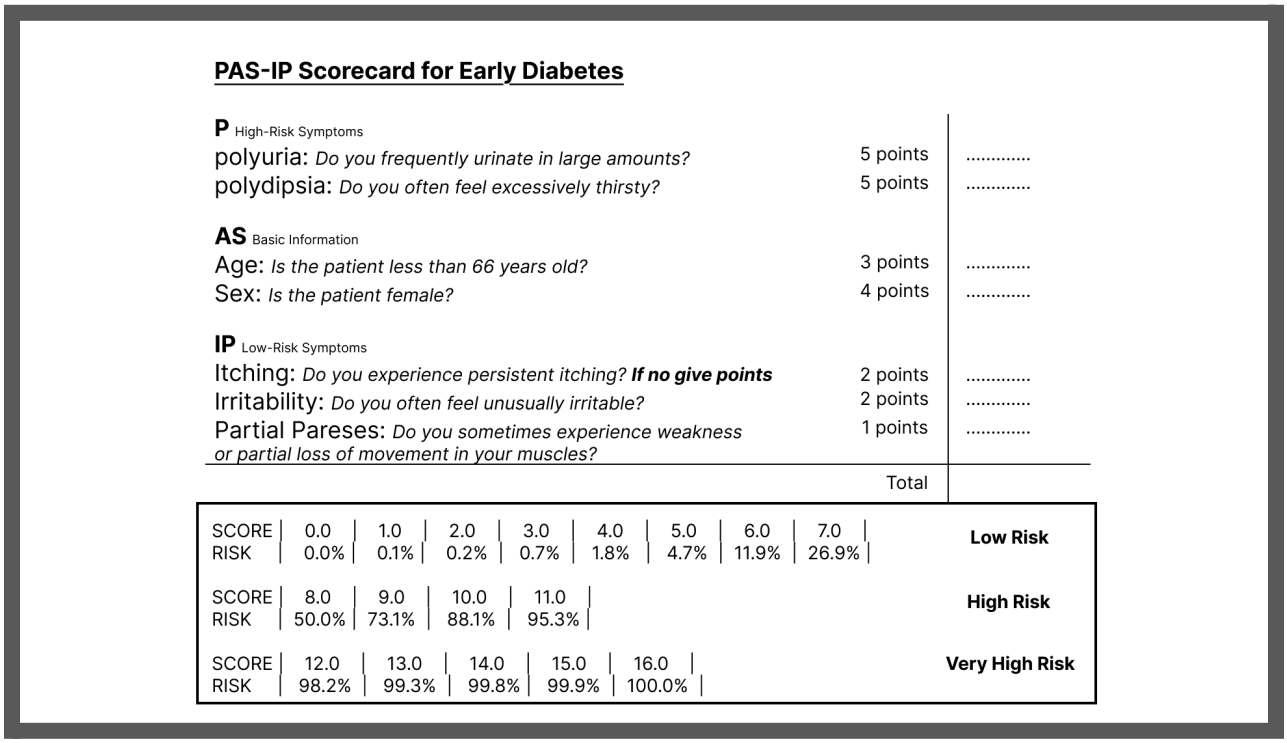
3.3. Best Scorecard Model

In our evaluation of the scorecard model, we conducted a 5-fold cross-validation with varying levels of sparsity, ranging from 1 to 7, to determine the optimal level for the model. The results indicated that a sparsity level of 7 yielded the best performance, achieving the highest test accuracy and AUC. Specifically, the model produced three different scorecards using the FasterRisk algorithm, with the highest test accuracy reaching 91.3% and an AUC of 0.973. These results demonstrate the ability of the scorecard to generate high-quality risk predictions efficiently, while maintaining interpretability. To enhance the usability of the scorecard, we transformed the risk table to eliminate negative points, making it more straightforward and easier to interpret for healthcare professionals. This modification was made to ensure that the scorecard could be effectively utilized in clinical decision-making, where clear and actionable insights are crucial.

The scorecard model demonstrates a satisfactory performance in terms of both accuracy and AUC, comparable to other models. While the scorecard achieves a test accuracy of 91.3% and an AUC of 0.973, the Random Forest model outperforms it with the highest test accuracy of 97.5%. However, it is important to consider that the Random Forest model, despite its superior accuracy, lacks the interpretability that the scorecard offers. In clinical contexts, interpretability is crucial, and the scorecard excels in this regard by providing a transparent and user-friendly format that can be easily understood by healthcare professionals. Thus, while the scorecard may not achieve

the highest accuracy, its combination of good predictive performance and high interpretability makes it a valuable tool for clinical decision-making.

The scorecard aligns well with the feature importance analysis conducted earlier as seen in Figure 5. The top six most important features identified by SHAP values—*polyuria*, *polydipsia*, *gender*, *sudden_weight_loss*, *partial_paresis*, and *age*—are all included in the scorecard, confirming that the model's feature selection process is consistent with the findings of the feature importance analysis. Notably, the scorecard also incorporates *age* as a significant predictor, despite it not ranking among the highest features in the SHAP analysis. This suggests that the scorecard's algorithm recognizes *age* as an important factor, even if it doesn't dominate the feature rankings in the SHAP analysis.



PAS-IP Scorecard for Early Diabetes

P High-Risk Symptoms		
polyuria: *Do you frequently urinate in large amounts?*	5 points	.....
polydipsia: *Do you often feel excessively thirsty?*	5 points	.....
AS Basic Information		
Age: *Is the patient less than 66 years old?*	3 points	.....
Sex: *Is the patient female?*	4 points	.....
IP Low-Risk Symptoms		
Itching: *Do you experience persistent itching?* ***If no give points***	2 points	.....
Irritability: *Do you often feel unusually irritable?*	2 points	.....
Partial Pareses: *Do you sometimes experience weakness or partial loss of movement in your muscles?*	1 points	.....
	Total	

SCORE	0.0	1.0	2.0	3.0	4.0	5.0	6.0	7.0	**Low Risk**
RISK	0.0%	0.1%	0.2%	0.7%	1.8%	4.7%	11.9%	26.9%	
SCORE	8.0	9.0	10.0	11.0					**High Risk**
RISK	50.0%	73.1%	88.1%	95.3%					
SCORE	12.0	13.0	14.0	15.0	16.0				**Very High Risk**
RISK	98.2%	99.3%	99.8%	99.9%	100.0%				

Figure 5: **Best AUC scorecard.** This figure presents the best scorecard in terms of test AUC score. This scorecard achieves a training accuracy of 92% and AUC of 0.978, and a testing accuracy of 91% and AUC of 0.973.

4. Conclusion

In this study, we have developed and evaluated an interpretable risk scorecard for early diabetes classification using the FasterRisk algorithm. By prioritizing transparency and simplicity, the scorecard model provides healthcare professionals with an easy-to-understand tool that can assist in early diabetes detection and decision-making. Despite a slight performance trade-off compared to more complex models, the scorecard’s interpretability makes it a promising

candidate for real-world clinical applications, offering a balance between predictive power and clinical usability.

Code Availability

The project code and all other presentation documents are available on the following link: https://github.com/DevinRS/Interpretable-Scorecard-Model-for-Early-Diabetes-Classification

References

1. Jiachang Liu, Chudi Zhong, Boxuan Li, Margo Seltzer, & Cynthia Rudin. (2022). FasterRisk: Fast and Accurate Interpretable Risk Scores.
2. Berk Ustun, & Cynthia Rudin (2019). Learning Optimized Risk Scores. Journal of Machine Learning Research, 20(150), 1–75.
3. Islam, H. (2020). Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques. In *Computer Vision and Machine Intelligence in Medical Image Analysis* (pp. 113–125). Springer Singapore.
4. Semenova, L., Rudin, C., & Parr, R. (2022). On the Existence of Simpler Machine Learning Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1827–1858). ACM.
5. Scott Lundberg, & Su-In Lee. (2017). A Unified Approach to Interpreting Model Predictions.