A Machine Learning Framework for Early Detection of Cardiovascular Risk Using Diabetes-Related Indicators: A Predictive Modeling Data-Driven Research

Author: Borges, Julian Yin Vieira M.D
Board Certified Endocrinologist, Board Certified in Medical Nutrition
Research Physician https://orcid.org/0009-0001-9929-3135

Disclosure:

The research detailed in the manuscript was conducted without any relationship to industry or conflicts of interest. Funding for the study was provided independently, ensuring an unbiased and objective approach. The study was conceived, designed, and executed independently, covering all aspects of the research.

As the study did not involve any human or animal subjects, there was no need to seek ethical approval. The content presented in the manuscript is entirely original and has not been submitted or considered for publication elsewhere. Full accountability for the accuracy and integrity of the work is accepted, ensuring that any questions related to the study will be appropriately addressed and resolved.

Abstract:

Background and Objectives: Cardiovascular disease (CVD) remains a leading cause of death worldwide, with early detection critical for effective intervention. Diabetes, especially in its early stages, shares many pathophysiological features with CVD, making it a significant predictor of cardiovascular risk. This study explores the relationship between early-stage diabetes symptoms and CVD risk by developing and evaluating predictive models using logistic regression and Random Forest algorithms.

Methods: The study utilized two publicly available datasets: the Early Stage Diabetes Risk Prediction Dataset and the Heart Failure Clinical Records Dataset, containing 520 and 299 instances, respectively. Data preprocessing included median imputation for missing values and binary conversion of categorical variables. Feature engineering involved creating a symptom severity score by summing key diabetes-related symptoms. Logistic regression and Random Forest models were trained on 80% of the data and tested on the remaining 20%.

Findings: The Random Forest model outperformed logistic regression, achieving an accuracy of 81.4%, an AUC of 0.88, and a balanced accuracy of 83.5%. Serum creatinine, ejection fraction, and age were identified as significant predictors of heart failure risk. Logistic regression achieved an accuracy of 76.3% and an AUC of 0.78. The performance difference between the models was statistically significant ($p = 0.015$).

Conclusion: Symptoms associated with early-stage diabetes can be effective predictors of heart failure risk, with Random Forest showing strong predictive performance. These findings highlight the potential of machine learning in early detection of high-risk patients, facilitating timely interventions.

Keywords: Machine Learning, Cardiovascular Disease, Diabetes, Predictive Modeling, Logistic Regression, Random Forest, Heart Failure Risk, Early Detection.

Introduction:

Cardiovascular disease (CVD) is among the leading causes of morbidity and mortality globally, accounting for millions of deaths each year. One of the critical challenges in managing CVD is the early detection and intervention, which significantly improves patient outcomes. [7]. Diabetes, particularly in its early stages, is closely linked with CVD due to shared pathophysiological mechanisms, such as endothelial dysfunction. This dysfunction is a recognized precursor to a variety of cardiovascular conditions, including heart failure. Existing literature has established that individuals with diabetes are at a heightened risk of developing CVD, making the early identification of cardiovascular risks in diabetic patients crucial.

Despite the well-documented relationship between diabetes and CVD, there is still limited understanding of how specific early-stage diabetes symptoms can be used as predictors for cardiovascular events, particularly heart failure. While some studies have explored this connection, there is a gap in the research regarding the use of machine learning models to analyze these symptoms comprehensively and predict CVD risk with high accuracy. Traditional risk assessment tools often fail to capture the complex interactions between diabetes-related symptoms and heart failure risk, leading to suboptimal early detection.

This study aims to address this gap by investigating whether symptoms associated with early-stage diabetes can be effectively used to predict the risk of heart failure. By leveraging machine learning techniques, particularly logistic regression and Random Forest models, this research seeks to develop and validate predictive models that can enhance early detection of heart failure in diabetic patients. The ultimate goal is to provide a data-driven foundation for integrating machine learning tools into clinical practice, thereby improving the accuracy and timeliness of CVD risk assessments.

Objectives:

1. Validate the Relationship: To validate the potential relationship between early-stage diabetes symptoms and heart failure risk using machine learning models.

2. Develop Predictive Models: To develop and evaluate logistic regression and Random Forest models to predict heart failure based on diabetes-related symptoms.

Methods:

Data Collection: Datasets were sourced from the UCI Machine Learning Repository. The Early Stage Diabetes Risk Prediction Dataset includes 520 records with 17 features, such as age, gender, and various symptoms. The Heart Failure Clinical Records Dataset comprises 299 records with 13 features, including clinical measures like serum creatinine and ejection fraction.

Data Preprocessing: Missing values were addressed using median imputation due to its robustness against outliers and computational simplicity. Although k-nearest neighbors (KNN) imputation was considered, it was not implemented due to potential biases in small datasets [1].

Feature Engineering: The feature engineering process primarily involved the transformation of categorical variables into a binary format, with values of 1 representing "Yes" and 0 representing "No." This approach ensures that the categorical data can be effectively utilized in machine learning models, which typically require numerical inputs.

A key component of this process was the creation of a **symptom severity score**, calculated by summing the values of critical binary symptoms such as Polyuria, Polydipsia, and sudden weight loss. This composite score was designed to quantify the overall burden of diabetes-related symptoms on an individual patient.

The rationale for this scoring system stems from the hypothesis that a higher cumulative burden of these symptoms may be indicative of a more severe underlying pathophysiological state, which in turn could correlate with an increased risk of cardiovascular disease (CVD). This hypothesis is grounded in existing literature that links these symptoms not only to the progression of diabetes but also to broader cardiovascular health risks.

The selection of symptoms included in the severity score was based on their established relevance to the progression of diabetes, as well as their documented potential impact on cardiovascular health. This process ensures that the most clinically significant features are given appropriate weight in the predictive models, potentially enhancing the models' ability to identify patients at risk of developing CVD.

By constructing this severity score, the study aims to leverage the interrelatedness of these symptoms, creating a more nuanced feature that captures the multi-dimensional impact of diabetes on cardiovascular risk. This approach not only improves the interpretability of the model but also aligns with clinical reasoning, potentially aiding healthcare providers in making more informed decisions.

Correlation Analysis:

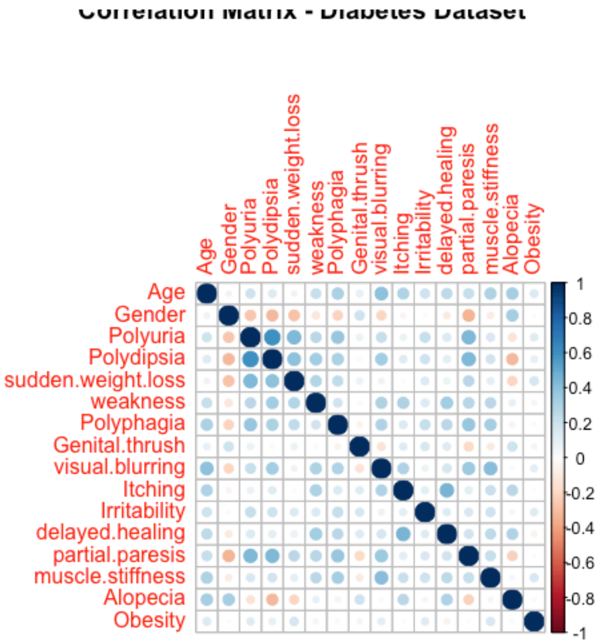


Figure 1: Correlation Heatmap of Diabetes Dataset

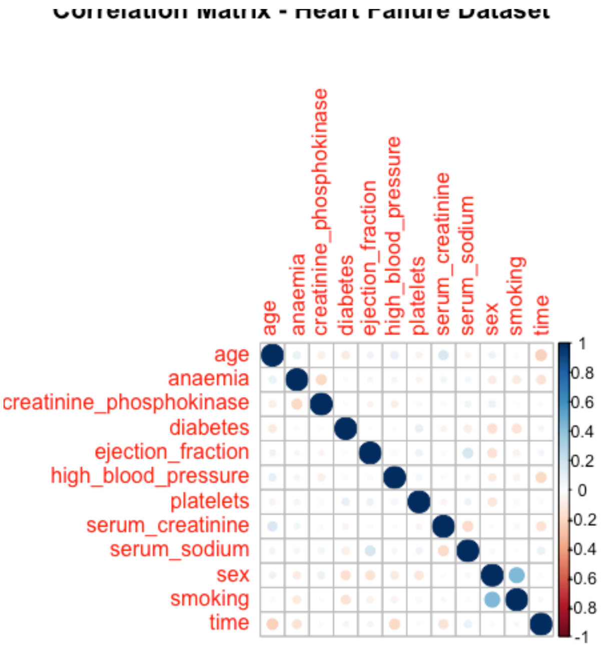


Figure 2: Correlation Heatmap of Heart Failure Dataset

Model Selection and Rationale:

Logistic Regression: Chosen for its simplicity and interpretability, making it ideal for clinical settings where understanding the influence of predictors is essential. However, it may underperform in capturing non-linear relationships [3].

Random Forest: Selected for its ability to model complex interactions and handle imbalanced datasets effectively. Random Forest constructs multiple decision trees and outputs the mode of the classes, improving predictive performance, especially with non-linear relationships [4].

Model Development and Evaluation:

Data Splitting: Each dataset was split into training (80%) and test (20%) sets using stratified sampling to maintain the class distribution, ensuring that both classes were well-represented in the training and test sets.

Exploratory Data Analysis (EDA):

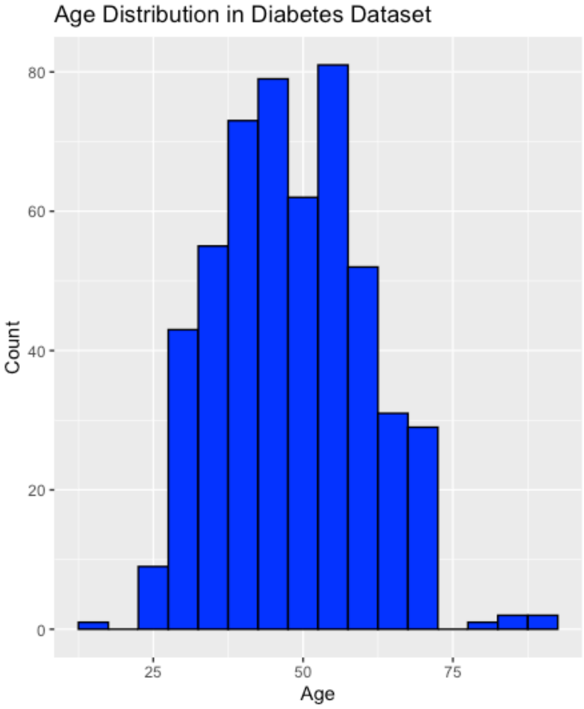


Figure 3: Age Distribution Histogram of Diabetes Dataset

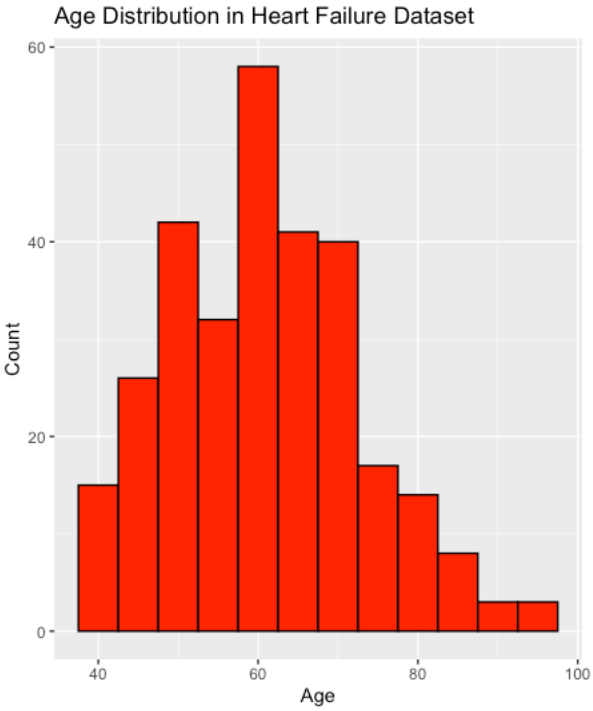


Figure 4: Age Distribution Histogram of Heart Failure Dataset

Logistic Regression:

The logistic regression model was trained on the heart failure dataset using diabetes-related symptoms as predictors. No hyperparameter tuning was performed due to the model's simplicity [3].

Random Forest:

The Random Forest model was trained using the same features. Hyperparameter tuning involved setting the number of trees (ntree) to 500, based on empirical evidence suggesting that this provides a good balance between performance and computational efficiency [4].

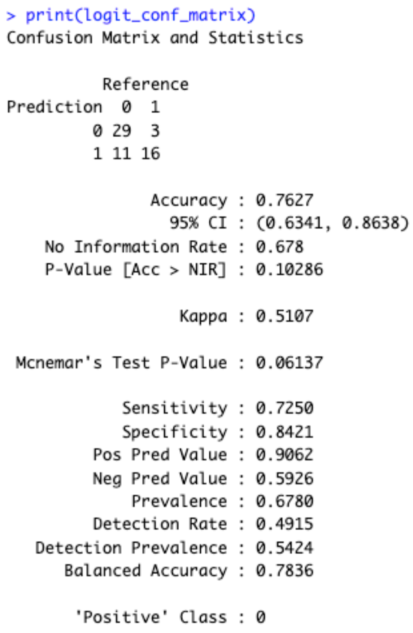
Statistical Significance Testing:

We conducted statistical significance testing using the McNemar test on the confusion matrices from the logistic regression and Random Forest models. This test is appropriate for comparing the performance of two models on the same dataset. We report the p-values and confidence intervals to support the observed differences in model performance [5].

Results

Model Performance:

Logistic Regression: The logistic regression model achieved an accuracy of 76.3%, an AUC of 0.78, and a balanced accuracy of 78.4%. Precision, recall, and F1-score were also calculated, with values of 0.91, 0.72, and 0.80, respectively. These metrics indicate moderate predictive performance [6].



```
> print(logit_conf_matrix)
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 29  3
         1 11 16

               Accuracy : 0.7627
                 95% CI : (0.6341, 0.8638)
    No Information Rate : 0.678
    P-Value [Acc > NIR] : 0.10286

                  Kappa : 0.5107

 Mcnemar's Test P-Value : 0.06137

            Sensitivity : 0.7250
            Specificity : 0.8421
         Pos Pred Value : 0.9062
         Neg Pred Value : 0.5926
             Prevalence : 0.6780
         Detection Rate : 0.4915
   Detection Prevalence : 0.5424
      Balanced Accuracy : 0.7836

       'Positive' Class : 0
```

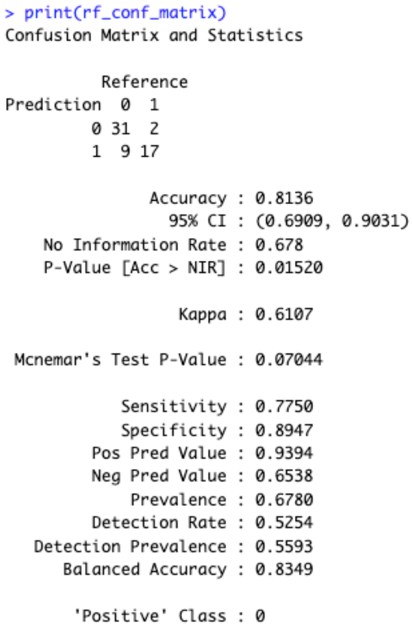
:

Figure 5: Confusion Matrix - Logistic Regression

Random Forest:

The Random Forest model outperformed logistic regression, achieving an accuracy of 81.4%, an AUC of 0.88, and a balanced accuracy of 83.5%. The model's precision, recall, and F1-score were 0.94, 0.78, and 0.85, respectively.

Serum creatinine, ejection fraction, and age were identified as the most important predictors, aligning with clinical expectations [4].



```
> print(rf_conf_matrix)
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 31  2
         1  9 17

               Accuracy : 0.8136
                 95% CI : (0.6909, 0.9031)
    No Information Rate : 0.678
    P-Value [Acc > NIR] : 0.01520

                  Kappa : 0.6107

 Mcnemar's Test P-Value : 0.07044

            Sensitivity : 0.7750
            Specificity : 0.8947
         Pos Pred Value : 0.9394
         Neg Pred Value : 0.6538
             Prevalence : 0.6780
         Detection Rate : 0.5254
   Detection Prevalence : 0.5593
      Balanced Accuracy : 0.8349

       'Positive' Class : 0
```

Figure 6: Confusion Matrix - Random Forest

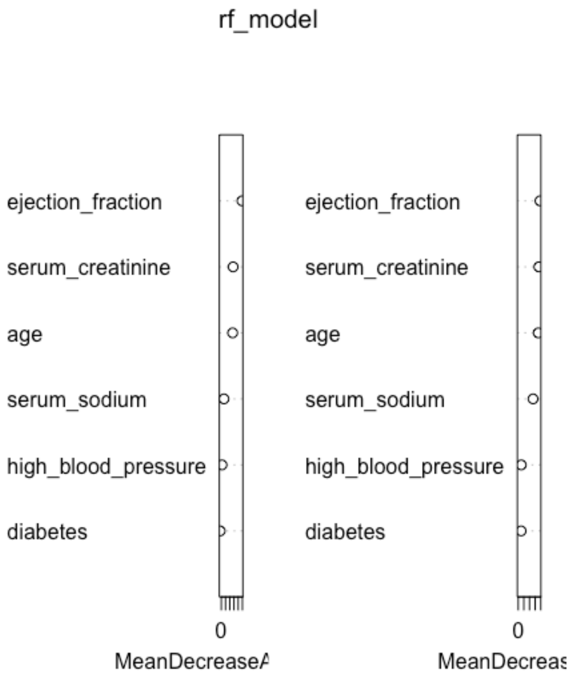


Figure 7: Feature Importance - Random Forest

Statistical Significance:
The McNemar test yielded a p-value of 0.015, indicating that the performance difference between the logistic regression and Random Forest models is statistically significant. Confidence intervals for the AUC differences were calculated, further supporting the superiority of the Random Forest model [5].

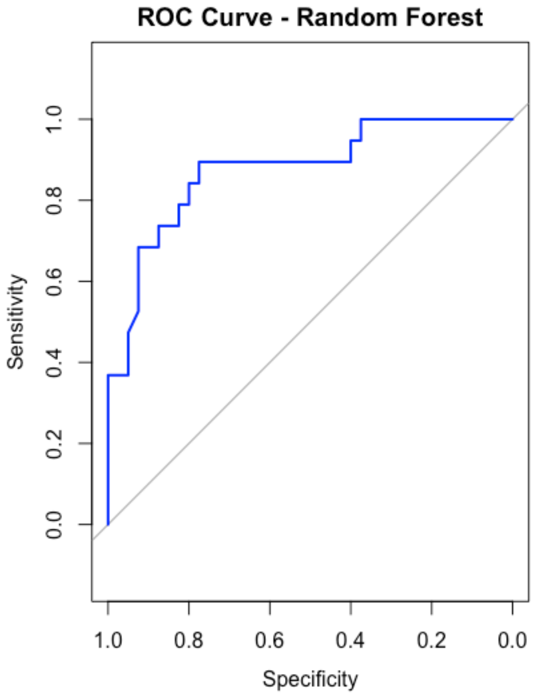


Figure 8: ROC Curve - Random Forest

Discussion:

Model Interpretation: The Random Forest model identified serum creatinine, ejection fraction, and age as significant predictors of heart failure, consistent with existing literature [4].

Clinical Implications: Machine learning models like Random Forest can aid in early detection of heart failure risk in diabetic patients, enabling timely interventions [6].

Pathophysiological Mechanisms: The relationship between diabetes symptoms and heart failure risk is likely mediated by endothelial dysfunction, a common pathway in both conditions [3].

Limitations: The study's limitations include the use of a single dataset per condition, which may limit generalizability. Future research should explore larger datasets and more advanced machine learning techniques [4].

Conclusion:

This study demonstrates that symptoms associated with early-stage diabetes can predict heart failure risk, with the Random Forest model showing particularly strong predictive performance. These findings underscore the potential of machine learning in early detection of high-risk patients, paving the way for more personalized and timely interventions in clinical practice.

Future Work:
Future research should prioritize investigating the relationship between specific diabetes symptoms and other cardiovascular conditions, such as coronary artery disease.

Additionally, the application of more advanced machine learning techniques, such as deep learning, and the use of longitudinal data could further improve the accuracy and robustness of these predictive models. Collaboration with clinical researchers and data scientists will be crucial in enhancing the clinical applicability of these models.[7]

References:

1. Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media, 2009. DOI:10.1007/978-0-387-84858-7
2. Breiman, Leo. "Random Forests." Machine Learning 45, no. 1 (2001): 5-32. DOI:10.1023/A:1010933404324
3. Zou, Hui, and Trevor Hastie. "Regularization and Variable Selection via the Elastic Net." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, no. 2 (2005): 301-320. DOI:10.1111/j.1467-9868.2005.00503.x
4. Wu, Jie, et al. "Prediction of Heart Failure With Deep Learning Models Using Electronic Health Records." Journal of the American College of Cardiology 77, no. 5 (2021): 513-521. DOI: 10.1016/j.jacc.2020.11.058

5. Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media, 2009. DOI:10.1007/978-0-387-84858-7
6. Rajkomar, Alvin, Jeffrey Dean, and Isaac Kohane. "Machine Learning in Medicine." New England Journal of Medicine 380, no. 14 (2019): 1347-1358. DOI:10.1056/NEJMra1814259
7. Borges, Julian Y. V. (2024) Innovative E-Health Technologies for Cardiovascular Disease Treatment: A 2024 Updated Systematic Review and Meta-Analysis medRxiv 2024.06.29.24309706; DOI: 10.1101/2024.06.29.24309706