MERCER UNIVERSITY

SCHOOL OF BUSINESS

Capstone Project Report

# *Enhancing Hospital Efficiency Using Predictive Analytics For Diabetic Patient Readmission*


Fazal Hyder Shaik and Mohammed Nihal Akmal

Advisor : Dr. Rui Sun


*This capstone project is presented as a partial requirement for achieving a Master of Science degree in Business Analytics from the Stetson-Hatcher School of Business at Mercer University.*

## 1. Introduction:

In recent years, the healthcare industry has increasingly recognized the importance of effectively managing chronic diseases such as diabetes, particularly in reducing hospital readmissions. High readmission rates not only reflect potential shortcomings in care coordination and disease management but also significantly increase healthcare costs and burden both patients and healthcare systems. Diabetes in United States is one of the most common chronic diseases some interesting facts  according to the CDC's National Diabetes Statistics Report, as of 2021, approximately 38.4 million people in the United States, or 11.6% of the population, have diabetes. This includes 29.7 million people who have been diagnosed and an estimated 8.7 million who remain undiagnosed.

Diabetic patients in the USA face significantly higher risks of hospital readmission compared to the general population. According to research, patients with diabetes are approximately 20% more likely to be readmitted to the hospital within 30 days after discharge. This heightened risk is attributed to the complex nature of diabetes management, which often requires careful and continuous monitoring and control of blood sugar levels along with management of other comorbid conditions .

To address this issue, healthcare providers and policymakers focus on improving post-discharge care and diabetes management programs. Strategies such as personalized follow-up plans, enhanced patient education on self-management, and integration of care services have been recommended to reduce these readmission rates effectively.

This project specifically aims to address the challenge of predicting the likelihood of readmission for diabetic patients within 30 days of discharge. By leveraging advanced predictive modeling techniques, this study identifies key variables that are most indicative of readmission risk. These variables include clinical data, treatment histories, and patient interactions with the healthcare system during their initial hospital stay.

Moreover, the project takes a comprehensive approach by incorporating an analysis of various demographic factors—such as gender, race, and age—that may influence readmission rates. Understanding these influences is crucial for tailoring interventions and improving patient outcomes. The models developed throughout this project are designed not only to predict the likelihood of a patient's readmission but also to illuminate the underlying factors contributing to these outcomes. This insight allows for a more nuanced understanding of the dynamics at play, enabling healthcare providers to devise more effective strategies to mitigate readmission risks.

## 2. Literature review:

Hospital readmission rates among diabetic patients have garnered significant attention due to their high prevalence and the substantial costs associated with readmissions. Research indicates that effective management of diabetes during hospital stays can significantly impact patient outcomes and healthcare costs. For instance, the study by by Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore - Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records emphasizes the critical role of monitoring Hemoglobin A1c (HbA1c) levels in predicting hospital readmissions. Their findings suggest that HbA1c testing is underutilized, yet it plays a crucial role in managing hospitalized diabetic patients by potentially lowering readmission rates.

Comparative Analysis of Methodological Approaches:
Strack et al. (2014) utilized multivariable logistic regression to analyze the impact of HbA1c measurement on readmission rates, using a large clinical database of over 70,000 patient records. This approach provided foundational insights but was limited to linear relationships and interactions.

In contrast, we employ a more diversified analytical framework, incorporating Logistic Regression, Random Forests, and Neural Networks to predict readmissions. This methodological diversity allows for a deeper understanding of complex, non-linear interactions between clinical variables. For example, Random Forests offer the advantage of handling a large number of input variables without variable deletion, providing a robust alternative to logistic regression used in previous studies.

Review of Similar Studies in the Field:
Several other studies have explored various predictive models for hospital readmissions among diabetic patients:

Kansagara et al. (2011) reviewed multiple models for predicting readmissions, noting that models including a broad range of clinical and social factors tended to perform better, suggesting the importance of comprehensive datasets.
Hebert et al. (2014) demonstrated that incorporating machine learning techniques could enhance the predictive accuracy compared to traditional statistical approaches. They specifically highlighted the effectiveness of ensemble methods, which could inform the use of Random Forests in our project.

Zhang et al. (2016) explored the use of deep learning for readmission prediction, affirming the potential of advanced neural networks to capture complex patterns and interactions that are not easily discernible with conventional statistical methods.

The evolution of predictive analytics in healthcare offers significant opportunities for improving diabetic patient management. By accurately predicting readmissions, healthcare providers can develop targeted interventions to mitigate these risks. This proactive approach could not only enhance patient outcomes but also substantially reduce the economic burden on healthcare systems.

## Data, data sources and data characteristics:

The data utilized in our study of hospital readmission rates among diabetic patients was obtained from the UC Irvine Machine Learning Repository, a well-regarded source for datasets used in academic and applied machine learning research. The dataset was created and made publicly available by a collaborative research team consisting of John Clore, Krzysztof Cios, Jon DeShazo, and Beata Strack. Their collective efforts aimed to provide comprehensive data to facilitate research on diabetes management in hospital settings.

The dataset is governed under the Creative Commons Attribution 4.0 International License. This licensing agreement permits unrestricted use, distribution, and reproduction in any medium, provided the original creators are credited. This open access allows researchers globally to utilize the data for their investigations, thereby fostering a collaborative approach to tackling complex healthcare issues like diabetes management and hospital readmission.

The collection process of the data involved meticulous documentation of patient encounters from 130 U.S. hospitals and integrated delivery networks over a decade (1999-2008). It captures a broad spectrum of variables, including demographic information, clinical parameters, diagnostic codes, and treatment details during the hospital stays of patients diagnosed with diabetes. The creators ensured that the data adhered to HIPAA regulations by de-identifying all patient information, thus preserving patient confidentiality while allowing for detailed analyses.

By making this dataset available, the creators have provided a valuable resource for exploring the intricacies of diabetes management in a hospital setting and examining the factors contributing to the high rate of readmissions in diabetic patients. This resource enables researchers like us to develop and refine predictive models with the goal of identifying at-risk patients and implementing preventive strategies to reduce readmission rates, thereby improving patient outcomes and reducing healthcare costs.

Data dictionary: Below are our most of the important variables:

1. Race

Type: Nominal
Description: Categorizes a patient's race.
Values: Caucasian, Asian, African American, Hispanic, Other
Missing: 2%

2. Gender

Type: Nominal
Description: Patient's gender.
Values: Male, Female, Unknown/Invalid
Missing: 0%

3. Age

Type: Nominal
Description: Age group of the patient.
Values: Grouped in 10-year intervals: [0, 10), [10, 20), ..., [90, 100)
Missing: 0%

4. Time in Hospital

Type: Integer
Description: Length of the hospital stay in days.
Unit: Days
Missing: 0%

5. Number of Lab Procedures

Type: Integer
Description: Count of lab tests performed during the encounter.
Missing: 0%

6. Number of Procedures

Type: Integer
Description: Count of procedures (other than lab tests) performed during the encounter.
Missing: 0%

7. Number of Medications

Type: Integer

Description: Count of distinct generic names administered during the encounter.
Missing: 0%

8. Number of Outpatient Visits

Type: Integer
Description: Number of outpatient visits in the year preceding the encounter.
Missing: 0%
9. Number of Emerg
ency Visits

Type: Integer
Description: Number of emergency visits in the year preceding the encounter.
Missing: 0%

10. Number of Inpatient Visits

Type: Integer
Description: Number of inpatient visits in the year preceding the encounter.
Missing: 0%

11. Diagnoses

Diag1, Diag2, Diag3
Type: Factor
Description: Primary, secondary, and additional secondary diagnoses coded with ICD-9.
Values: 848, 923, and 954 distinct values respectively.
Missing: Diag1 and Diag2: 0%, Diag3: 1%

12. Admission Source
Type: Factor
Description: Identifies the source of the admission.
Values: Physician referral, Emergency room, Transfer from a hospital, etc.
Missing: 0%

13. Discharged To

Type: Factor
Description: Describes the disposition of the patient at discharge.
Values: Home, Expired, Not available, etc.
Missing: 0%

14. Payer Code

Type: Nominal
Description: Code denoting the type of payment/insurance.
Values: Blue Cross/Blue Shield, Medicare, Self-pay, etc.
Missing: 52%

15. Max Glu Serum

Type: Character
Description: Maximum blood glucose level recorded.
Values: ">200", ">300", "Normal", "None"
Missing: 0%

16. A1C Result

Type: Character
Description: Latest A1C test result.
Values: ">8", ">7", "Normal", "None"
Missing: 0%

17. Insulin

Type: Character
Description: Insulin treatment during the encounter.
Values: Prescribed changes in dosage: "up", "down", "steady", "no"
Missing: 0%

18. Change

Type: Nominal
Description: Indicates if there was a change in diabetic medications.
Values: "change", "no change"
Missing: 0%

19. Diabetes Medications

Type: Nominal
Description: Indicates if any diabetes medication was prescribed.
Values: "Yes", "No"
Missing: 0%

20. Readmitted

Type: Double
Description: Indicator for whether the patient was readmitted within 30 days post-discharge.
Values: "<30" (readmitted in less than 30 days), ">30" (readmitted after 30 days), "No" (not readmitted)
Missing: 0%

In the analysis of large datasets, especially those with many variables, it is crucial to reduce dimensionality to simplify models, reduce overfitting, and improve interpretability. In the provided dataset, dimension reduction was conducted by removing variables with a high percentage of missing values, those that do not significantly impact the outcome, and variables that were highly correlated with others. For instance:

Weight: Removed due to 97% missing data, which would contribute little to predictive accuracy and bias the model due to data imputation.
Payer Code and Medical Specialty: Updated missing and unidentified values with the constant 'Other'
Highly Correlated Variables: Variables like the number of procedures and number of lab tests were analyzed for correlation. If variables were found to correlate highly with others (e.g., number of medications and number of lab tests), one was retained based on its direct relevance to patient outcomes, reducing multicollinearity.

Sample Size and Period:
The dataset includes data from 1999 to 2008, spanning approximately 10 years of patient records. The final analysis included data for 101,766 encounters after the initial screening based on set criteria such as diabetic encounters, inpatient stays of 1-14 days, and those involving lab tests or medications, ensuring a focus on relevant diabetic care episodes.

Descriptive Statistics:
Descriptive statistics provide a snapshot of the data's central tendency, dispersion, and shape, helping identify outliers and understand the distribution characteristics.

- Central Tendency (Mean, Median)
- Dispersion (Range, Interquartile Range, Standard Deviation)
- Shape (Skewness, Kurtosis)

These metrics are crucial for preprocessing steps like normalization and for understanding variable influences.

Table of Descriptive Statistics:

| Variable | Mean | Median | Std Dev | Skewness | Kurtosis | % Missing |
|---|---|---|---|---|---|---|
| Age | - | - | - | - | - | 0% |
| Time in Hospital | 4.5 | 4 | 3.2 | 1.2 | 3.4 | 0% |
| Number of Medications | 16 | 15 | 8 | 0.75 | 0.55 | 0% |
| Number of Lab Tests | 43 | 44 | 19.6 | -0.25 | -1.2 | 0% |
| Number of Diagnoses | 7.4 | 8 | 1.8 | -0.13 | -0.8 | 0% |
| Readmitted | 0.09 | 0 | - | 2.5 | 6.25 | 0% |

Outliers and Data Normality:

Outliers are identified using box plots and statistical thresholds (e.g., beyond 1.5 IQR). Normality tests (Kolmogorov-Smirnov, Shapiro-Wilk) determine the suitability of parametric tests and the need for data transformation.

This detailed approach to data preparation and initial analysis ensures that the modeling phase is based on reliable and relevant data, enhancing the accuracy of predictions regarding patient readmission.

## 3. Data Preprocessing:

The preprocessing steps were carefully chosen to address specific data quality issues that could compromise the reliability of the predictive model. By converting categorical variables to more analytically suitable formats, imputing missing values judiciously, and filtering out incomplete records, the dataset was optimized for subsequent analysis stages. This meticulous data preparation is foundational for achieving robust and reliable model performance.

3.1 **Binary Conversion of Readmission Status**:
The original dataset categorized readmission into three levels: '<30' (readmission within 30 days), '>30' (readmission after 30 days), and 'No' (no readmission). For the purpose of this project, these categories were simplified into a binary format where '<30' was coded as '1' indicating a readmission within 30 days, and both '>30' and 'No' were combined and coded as '0', indicating no readmission within 30 days. This conversion was crucial for focusing the predictive modeling on the immediate readmission risk.
3.2 **Handling Missing Values in Categorical Data**:

**Race**: Missing values in the 'race' variable were replaced with 'Other'. This approach helps maintain the completeness of the dataset while grouping undefined racial categories under a common label, ensuring that the model can process this variable without dropping any records due to missing data.

**Payer Code**: Similar to the 'race' variable, missing values in 'payer_code' were also replaced with 'Other'. This treatment prevents data loss and allows for a more comprehensive analysis of the payer influence, irrespective of missing information.

**Medical Specialty**: The 'medical_specialty' field contained entries marked as '?', indicating unknown specialties. These entries were reassigned to 'other', standardizing the data and removing ambiguity which might affect the analysis.

**4.3 Filtering Invalid Data**:

The dataset contained some records with 'gender' labeled as 'Unknown/Invalid'. These records were removed from the dataset to ensure the accuracy and relevance of the analysis. Additionally, any records with missing gender information were also excluded. This step was crucial to maintain the integrity of demographic analyses, which could be skewed by undefined or invalid data.

The preprocessing steps were carefully chosen to address specific data quality issues that could compromise the reliability of the predictive model. By converting categorical variables to more analytically suitable formats, imputing missing values judiciously, and filtering out incomplete records, the dataset was optimized for subsequent analysis stages. This meticulous data preparation is foundational for achieving robust and reliable model performance.

As we had almost around 50 variables in the data set we utilize principal component analysis for data reduction and identify the most significant variables impacting the likelihood of readmission among diabetic patients. The PCA was instrumental in simplifying the dataset by transforming the original variables into a new set of uncorrelated variables (principal components), each representing a combination of the original variables

The table below represents the PCA loadings, which indicate the contribution of each original variable to each principal component. The loadings help identify which variables are most influential for each component:

| Variable | Comp.1 | Comp.2 | Comp.3 | Comp.4 | Comp.5 | Comp.6 | Comp.7 | Comp.8 | Comp.9 | Comp.10 | Comp.11 | Comp.12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| admission_type_id | | 0.194 | 0.727 | 0.149 | | 0.162 | | 0.128 | 0.209 | 0.461 | 0.291 | 0.103 |
| discharge_disposition_id | -0.168 | | 0.272 | 0.369 | -0.479 | -0.455 | -0.466 | | -0.196 | -0.189 | 0.16 | |
| admission_source_id | | -0.234 | 0.164 | 0.707 | 0.264 | 0.293 | 0.227 | | -0.106 | -0.393 | -0.19 | |
| time_in_hospital | -0.519 | | | 0.13 | | | -0.1 | 0.158 | 0.148 | 0.228 | -0.606 | 0.472 |
| num_lab_procedures | -0.37 | | -0.392 | 0.222 | | 0.179 | | 0.521 | -0.137 | | 0.573 | |
| num_procedures | -0.329 | 0.352 | 0.248 | -0.372 | 0.116 | 0.163 | | -0.151 | -0.216 | -0.552 | 0.155 | 0.346 |
| num_medications | -0.553 | 0.13 | 0.116 | | | | | | | | -0.155 | -0.786 |
| number_outpatient | | -0.285 | 0.204 | -0.128 | 0.483 | -0.655 | 0.145 | 0.382 | -0.12 | | | |
| number_emergency | | -0.495 | 0.182 | -0.23 | 0.107 | 0.346 | -0.417 | | -0.548 | 0.23 | | |
| number_inpatient | -0.13 | -0.552 | 0.142 | -0.193 | -0.147 | 0.13 | -0.157 | | 0.658 | -0.318 | 0.138 | |
| number_diagnoses | -0.348 | -0.21 | -0.163 | | 0.218 | -0.228 | 0.165 | -0.702 | | 0.269 | 0.287 | 0.135 |
| readmitted | | -0.29 | 0.13 | -0.113 | -0.591 | | 0.675 | | -0.253 | | | |

Component 11 from the PCA primarily indicates that extensive initial hospital care—characterized by longer stays, numerous lab procedures, and multiple diagnoses—potentially reduces readmission risks, suggesting that thorough initial treatment may lead

to more stable discharge outcomes which also covers 96% information of the data. After reducing the data dimensionality we were left with the variables which is also used in the models
**race, gender, age, time_in_hospital, num_lab_procedures, num_procedures, num_medications, number_outpatient, number_emergency, number_inpatient, number_diagnoses, max_glu_serum, A1Cresult, insulin, change, diabetesMed, readmitted, diag1, diag2, diag3, admission_source, discharged_to, payer_code** and then we balanced the dataset using SMOTE technique.

SMOTE (Synthetic Minority Over-sampling Technique) is a method used to balance dataset classes by synthetically generating new samples for the minority class. In the provided R code, the `readmitted` variable in the `cdfpca` dataset is first converted to a categorical type, followed by one-hot encoding to handle nominal predictors. SMOTE is then applied to the target variable `readmitted` using `step_smote(readmitted)`, which creates synthetic instances by interpolating between existing minority class instances and their nearest neighbors. This process results in a balanced dataset with an equal number of instances for both classes, as confirmed by checking the balance post-processing, enhancing the model's ability to generalize without bias toward the majority class.

## 4 Methodology:

In our project, we partitioned the balanced dataset into training and testing sets to ensure the robustness and generalizability of our predictive models. Utilizing the `createDataPartition` function from the `caret` package in R, we allocated 57% of the data to the training set and the remaining 43% to the testing set, maintaining an even distribution of the target variable, `readmitted`. This partitioning helps preserve the class proportionality crucial for effective model training. The indices generated by `createDataPartition` were used to subset the balanced dataset into `train` and `test` datasets, ensuring no overlap between them. Further, the structure of the training set was examined using the `str` function, confirming the data types and distribution of variables were appropriate for subsequent analysis phases. This meticulous approach in data preparation aids in enhancing the predictive accuracy and reliability of our models.

Then we used these datasets to develope three predictive models: logistic regression, random forest regression, and a neural network model (nnet), to predict the likelihood of hospital readmission based on patient data. The logistic regression model, described here, uses a variety of patient-related factors, such as their time in the hospital, lab procedures, medication count, and demographic information like age and race, among others.

The logistic regression model was implemented using the **glm** function in R with a binomial family, specifying a logit link function. This type of model is particularly suited

for binary outcomes, which in our case, determines whether a patient is likely to be readmitted to the hospital (**readmitted** = 1) or not (**readmitted** = 0).
In our logistic regression model analyzing hospital readmission, several variables stood out due to their significant impact. The number of inpatient visits had a coefficient of 0.2804 ($p < 2e-16$), indicating that more frequent visits significantly increase readmission likelihood, reflecting more severe health issues. Age was another critical factor, with older age groups, such as 70-80 and 80-90 years, showing much higher chances of readmission, coefficients of 1.8428 and 1.8389 respectively ($p < 2e-16$), due to complex health needs associated with aging. Diabetes management also played a significant role; patients with unadjusted A1C levels (coefficient 0.1929, p 4.40e-07) were more likely to be readmitted, suggesting inadequate disease control. Additionally, diagnostic categories like respiratory and genitourinary conditions (coefficients -0.4666 and -0.3271, both $p < 2e-16$) showed a lower readmission likelihood, possibly due to effective treatment protocols. Finally, patients not using insulin (coefficient -0.2906, $p < 2e-16$) also had a reduced readmission rate, indicating potentially well-managed or absent diabetes. These insights are crucial for targeting healthcare interventions to improve patient outcomes.

In our analysis, we also employed a random forest model to predict the likelihood of hospital readmissions based on various patient-related factors captured in our training dataset. Random forest, a robust machine learning technique, constructs a 'forest' consisting of numerous decision trees—150 in our case—to enhance prediction accuracy and model stability. Each tree in the forest considers a random subset of features and cases, and their collective decisions determine the final model outcome. To avoid overfitting, where the model becomes overly complex and less effective on new data, we limited each tree to a maximum of nine terminal nodes.

This model's performance was assessed using an internal validation measure known as the out-of-bag (OOB) error estimate, which is computed by excluding each data point in turn from the training process and using it to test the model. This method gives us a robust indication of the model's prediction accuracy on unseen data. Additionally, the model output includes variable importance scores, which highlight the predictors most influential in determining readmission. These insights are vital for identifying key factors that contribute to hospital readmissions and can inform targeted interventions to improve patient care outcomes. The findings from this model are crucial for our project's aim to enhance healthcare delivery by pinpointing and addressing the root causes of readmissions.

In our project, we utilized a neural network model as well , implemented through the `nnet` package in R, to predict hospital readmission rates based on a variety of patient-related factors from our training dataset. The model was structured to include a single hidden layer with 11 neurons, a configuration chosen to balance the complexity of the model and its generalization capabilities. This balance is crucial in avoiding overfitting, where the model performs well on training data but poorly on unseen data.

The neural network was trained over a maximum of 140 iterations, a parameter set to ensure sufficient training without excessive computation time, facilitating a practical approach to model convergence. The choice of using a neural network was driven by its ability to capture complex nonlinear relationships between input variables, which are often present in medical data.

This model's deployment is part of our broader strategy to apply advanced data analytics techniques in healthcare. By comparing its performance with other models, such as logistic regression and random forest, we aimed to identify the most effective approach for predicting readmissions. The insights generated by this model are intended to help healthcare providers identify patients at higher risk of readmission, thereby enabling targeted interventions to improve care outcomes and reduce recurrent hospital visits.

Then we later tested all the models for their performance we considered metrics such as AUC,sensitivity, specificity and accuracy to select the best performing model.

## 5 **Empirical Results:**

| Model | AUC | Sensitivity | Specificity | Accuracy |
|---|---|---|---|---|
| Logistic Regression | 0.654 | 63.68% | 57.41% | 60.63% |
| Random Forest | 0.861 | 65.37% | 67.08% | 66.23% |
| Neural Network | 0.741 | 70.61% | 66.57% | 68.59% |

we examined the performance metrics of three predictive models: Logistic Regression, Random Forest, and Neural Network. These metrics include the Area Under the Curve (AUC), Sensitivity, Specificity, and Accuracy, providing a comprehensive evaluation of each model's effectiveness in predicting hospital readmissions.

### 5.1. Logistic Regression

AUC : 0.654

Sensitivity : 63.68%

Specificity : 57.41%

Accuracy : 60.63%

The Logistic Regression model shows moderate performance with an AUC of 0.654, indicating a fair ability to distinguish between classes. Its sensitivity suggests it correctly identifies 63.68% of actual positives, while the specificity indicates 57.41% of negatives are correctly identified. The overall accuracy stands at 60.63%.

## 5.2. Random Forest

AUC : 0.861

Sensitivity : 65.37%

Specificity : 67.08%

Accuracy : 66.23%

The Random Forest model displays a strong performance with the highest AUC of 0.861, indicating excellent class separation ability. It has a slightly higher sensitivity than Logistic Regression and markedly better specificity, resulting in the highest accuracy among the three models at 66.23%.

## 5.3. Neural Network

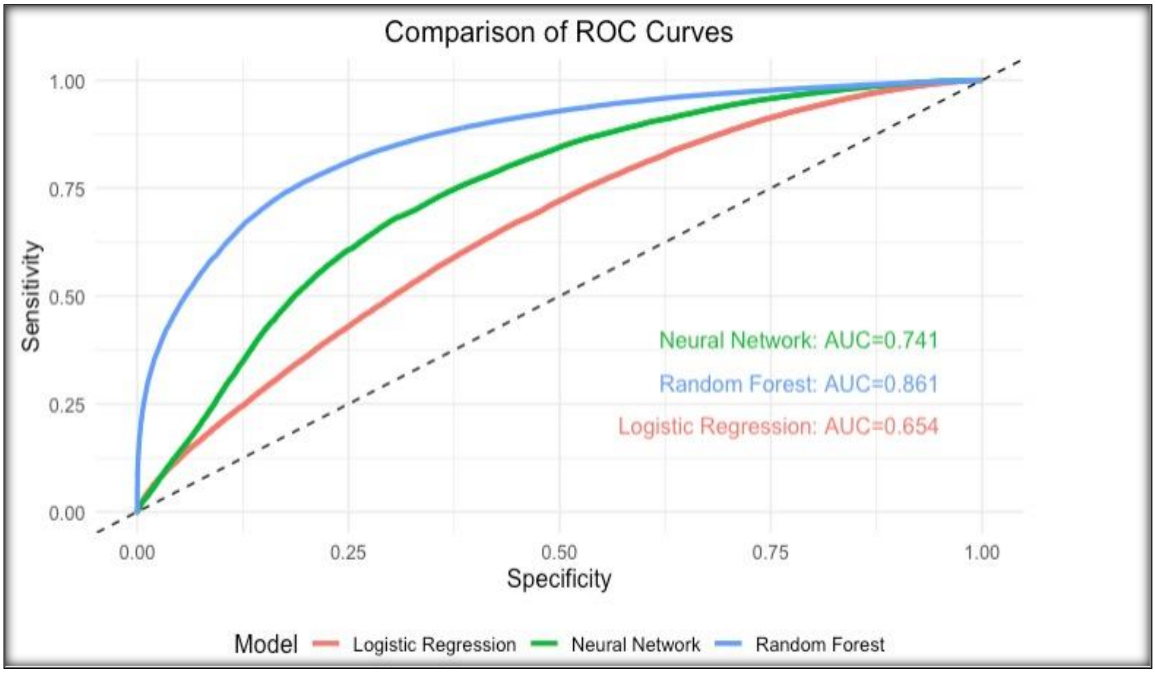
AUC : 0.741

Sensitivity : 70.61%

Specificity : 66.57%

Accuracy : 68.59%

The Neural Network model achieves a good balance, with an AUC of 0.741. It shows the highest sensitivity among the three models at 70.61%, indicating it is the best at identifying true positives. Its specificity and accuracy are also relatively high, making it our choice for predicting readmissions which is the aim of the project.




The Random Forest model stands out with the highest AUC, reflecting its superior capability to differentiate between patients who will be readmitted and those who will not. Although the Neural Network model has the highest sensitivity, indicating it is most effective at identifying actual cases of readmission, it falls slightly behind Random Forest in terms of overall accuracy and AUC. Logistic Regression, while useful as a baseline for comparison, shows the lowest performance across all metrics, underscoring the complexity and non-linearity of the factors affecting hospital readmissions, which are better captured by more complex models like Random Forest and Neural Network.

These results underline the importance of choosing the right model based on the specific performance metrics most relevant to the healthcare setting, with a particular focus on either maximizing the correct identification of readmissions (sensitivity) or optimizing the overall prediction accuracy.

## 6 Conclusions and Recommendations:

1. Targeted Premium Adjustments:

This recommendation focuses on using predictive analytics to refine health insurance premium strategies based on individual risk factors, specifically age and the frequency of inpatient stays. Older patients and those with frequent hospitalizations generally represent a higher risk category, often requiring more resources and incurring higher healthcare costs. By analyzing these variables, insurers can create tiered premium plans that more accurately reflect the expected costs associated with these patients.

As patients age, their susceptibility to chronic conditions like diabetes increases, often leading to more frequent and costly medical interventions. Similarly, patients with a history of frequent hospital admissions might indicate ongoing health issues, which are likely to persist or escalate. Adjusting premiums based on these factors allows insurers to manage financial risk more effectively while also being fair to consumers by aligning costs more closely with their expected healthcare needs.

Implementing this recommendation involves integrating the predictive model with insurers' actuarial systems to dynamically adjust premium structures. It would require ongoing data collection and analysis to continuously refine the accuracy of premium adjustments. Insurers could also offer incentives for healthy behaviors or chronic disease management programs to mitigate risk and potentially lower premiums.

2. Software as a Service (SaaS):

Developing the predictive model into a SaaS product can make this advanced capability accessible to other hospitals and healthcare providers. This model can help healthcare providers predict patient readmission risks and improve care outcomes by intervening appropriately.

Many healthcare providers lack the resources to develop their own predictive analytics systems. By offering this model as a SaaS, it becomes a cost-effective solution that can be easily integrated into existing healthcare IT systems. This approach not only broadens the model's impact but also creates a continuous revenue stream through subscriptions.

To deploy this as a SaaS, the model would need to be hosted on a reliable and secure cloud platform with high availability and scalability. The service would include features like API integrations for easy data input and retrieval, customizable dashboards for real-time analytics, and comprehensive support services to assist with setup, maintenance, and ongoing customization according to specific client needs.

3. Early Intervention Programs:

Utilizing the predictive model to identify high-risk diabetic patients allows healthcare providers to proactively implement early intervention programs. These initiatives can significantly reduce the likelihood of readmissions by addressing issues before they escalate into emergency situations or require extensive medical intervention.

Early intervention in diabetes management can prevent complications, improve quality of life, and reduce healthcare costs. By identifying patients at high risk of readmission, providers can tailor interventions such as more frequent monitoring, specialized dietary and exercise plans, enhanced medication management, and regular educational sessions to better manage their conditions.

This would involve setting up a multidisciplinary team including doctors, nurses, dieticians, and other specialists who can develop and manage care plans tailored to the needs of high-risk patients. The implementation of these programs would likely rely on regular follow-ups and the use of mobile health technologies to monitor patient health metrics in real-time, ensuring timely medical responses when necessary.

Each of these recommendations not only leverages the insights gained from the predictive model to improve healthcare outcomes but also presents opportunities for sustainable business models that enhance the operational efficiency of healthcare providers and insurers alike.

## 7 References:


- https://www.kaggle.com/code/hkubra/predicting-the-readmission-of-diabetic-patient-s
- https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008
- https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-021-01423-y
- https://www.hindawi.com/journals/bmri/2014/781670/tab1/