

Course Project Report

Predicting the readmission risk of diabetic patients

Submitted By

Anush Revankar (221AI009)

Akhilesh Negi (221AI008)

as part of the requirements of the course

Data Science (IT258) [Dec 2023 - Apr 2024]

in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology in Artificial Intelligence

under the guidance of

Dr. Sowmya Kamath S, Dept of IT, NITK Surathkal

undergone at



**DEPARTMENT OF INFORMATION TECHNOLOGY
NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA, SURATHKAL**



DEC 2023 - APR 2024

DEPARTMENT OF INFORMATION TECHNOLOGY
National Institute of Technology Karnataka, Surathkal

C E R T I F I C A T E

This is to certify that the Course project Work Report entitled **“Predicting the readmission risk of diabetic patients”** is submitted by the group mentioned below -

Details of Project Group

Name of the Student	Register No.	Signature
Anush Revankar	221AI009	
Akhilesh Negi	221AI008	



this report is a record of the work carried out by them as part of the course **Data Science (IT258)** during the semester **Dec 2023 - Apr 2024**. It is accepted as the Course Project Report submission in the partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Artificial Intelligence**.

(Name and Signature of Course Instructor)
Dr. Sowmya Kamath S
Associate Professor, Dept. of IT, NITK

DECLARATION

We hereby declare that the project report entitled “**Predicting the readmission risk of diabetic patients**” submitted by us for the course **Data Science (IT258)** during the semester **Dec 2023 - Apr 2024**, as part of the partial course requirements for the award of the degree of Bachelor of Technology in Artificial Intelligence at NITK Surathkal is our original work. We declare that the project has not formed the basis for the award of any degree, associateship, fellowship or any other similar titles elsewhere.

Details of Project Group

Name of the Student	Register No.	Signature
Anush Revankar	221AI009	
Akhilesh Negi	221AI008	

Place: NITK, Surathkal

Date: **15th April 2024**

Predicting Readmission Risk in Diabetic Patients

Akhilesh Negi

Department of Information Technology
National Institute of Technology
Surathkal, Karnataka

Anush Revankar

Department of Information Technology
National Institute of Technology
Surathkal, Karnataka

Abstract—Predicting hospital readmission risk in diabetic patients is crucial for improving healthcare outcomes and reducing healthcare costs. In this study, we propose novel enhancements to traditional feature selection and classification techniques to improve the accuracy of readmission risk prediction models. Instead of relying solely on Chi-square analysis to identify factors affecting readmission risk, we apply multiple feature selection algorithms including ReliefF, Correlation, Information Gain, and SelectKBest. By combining these techniques, we rank features based on their importance scores and evaluate their performance using Random Forest classifiers with varying numbers of selected features. Additionally, we address the issue of imbalanced datasets by applying the SMOTE for data balancing. Furthermore, we perform hyperparameter tuning using grid search cross-validation to optimize the performance of our models. Our experimental results demonstrate that our proposed enhancements significantly improve the accuracy of readmission risk prediction models. We compare the performance of various classifiers including Random Forest, Decision Tree, Logistic Regression and XGBoost, and identify the most effective models for predicting readmission risk in diabetic patients.

Keywords: readmission risk prediction, diabetic patients, feature selection, ReliefF Algorithm, Correlation, Information Gain, SelectKBest, SMOTE, hyperparameter tuning, Random Forest, Support Vector Machine, Decision Tree, Logistic Regression, XGBoost
Github Link to Code: [Github](#)

I. INTRODUCTION

Hospital readmission is a critical concern in managing diabetic patients, as it not only poses significant financial burdens on healthcare systems but also indicates potential lapses in patient care and management. Predicting readmission risk accurately is paramount for healthcare providers to implement timely interventions and reduce avoidable readmissions. Traditional approaches to feature selection and classification often rely on simplistic methods like Chi-square analysis, which may fail to capture the nuanced relationships between features and readmission risk.

In this study, we propose novel enhancements to address these limitations and improve the accuracy of readmission risk prediction models for diabetic patients. Our approach involves leveraging advanced feature selection techniques, including ReliefF, Correlation, Information Gain, and SelectKBest algorithms, to identify the most relevant factors influencing readmission risk. By combining these methods.

Furthermore, we recognize the challenge posed by imbalanced datasets commonly encountered in healthcare data,

where non-readmissions may significantly outnumber instances of readmission. To mitigate this issue, we employ the Synthetic Minority Oversampling Technique (SMOTE) to balance the dataset and ensure the robustness of our models.

Moreover, we optimize the performance of our models through hyperparameter tuning using grid search cross-validation, thereby fine-tuning the parameters of our classifiers to achieve optimal predictive accuracy.

Through our proposed enhancements, we seek to provide healthcare providers with more accurate and reliable tools for predicting readmission risk in diabetic patients, ultimately leading to improved patient outcomes and more efficient resource allocation within healthcare systems.

II. DATASET

A. Dataset Description

This study utilized a dataset spanning a decade (1999-2008) containing information from 130 hospitals and integrated delivery networks across the United States. The dataset primarily centers on hospital records of individuals diagnosed with diabetes. The dataset can be accessed from the UCI Machine Learning Repository at the following link: UCI Diabetes Dataset.

B. Dataset Characteristics:

- Multivariate
- Feature Type: Categorical, Integer
- Instances: 101,766
- Features: 47

The dataset includes attributes such as patient demographics (age, gender, race), admission details (admission type, time in hospital), medical information (physician specialty, lab tests, HbA1c results), medication history, and healthcare utilization metrics (outpatient, inpatient, and emergency visits). The dataset also contains missing values.

III. EXPLORATORY DATA ANALYSIS

1. Bar Plot: The bar plot provides insights into the distribution of numerical features within the dataset (Fig. 1). This aids in determining suitable scaling techniques for data preprocessing. Notably, columns such as "time_in_hospital," "num_lab_procedures," and "num_medications" exhibit skewed normal distributions, indicating potential areas for normalization or transformation to enhance model performance.

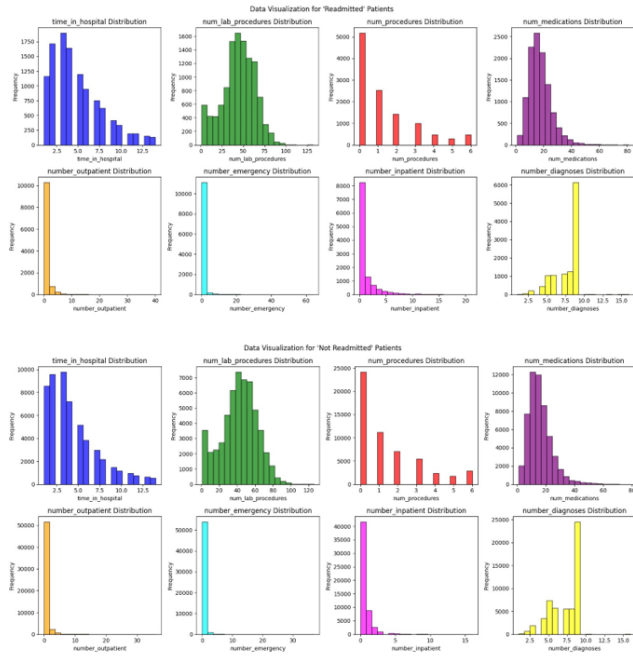


Fig. 1: Bar Plot: Shows the trend followed by the numerical columns.

2. Histogram: The histogram (Fig. 2) illustrates the distribution of classes within the target variable "readmitted" across its categories ("No", "<30", ">30"). Understanding this distribution is crucial as it directly informs our predictive task, allowing us to gauge the balance or skewness among the classes.

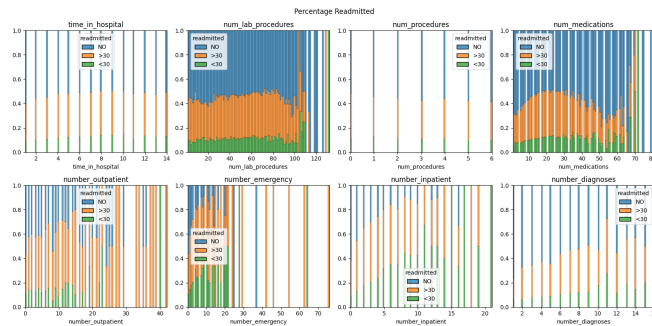


Fig. 2: Histogram: Shows the proportion of each category in the Target Column ("readmitted") against the values of respective numerical columns.

3. Pi Chart: The pie chart (Fig. 3) visually represents the proportions of categories within categorical columns. By examining these proportions, we gain insights into the relative frequencies of different categorical values, aiding in feature importance assessment and understanding the dataset's categorical structure.

4. Correlation Matrix: The correlation matrix (Fig. 4) provides a quantitative measure, specifically Pearson's correlation coefficient, for assessing the linear relationship between pairs of numerical features. This analysis enables identifica-

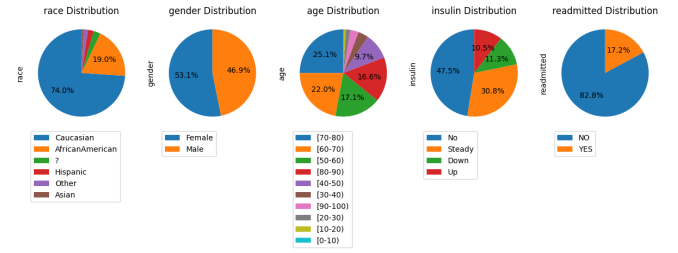


Fig. 3: Pi Chart: Shows the proportion of each category in respective categorical column.

tion of potentially correlated features, which informs feature selection or dimensionality reduction strategies to prevent multicollinearity issues in predictive modeling.

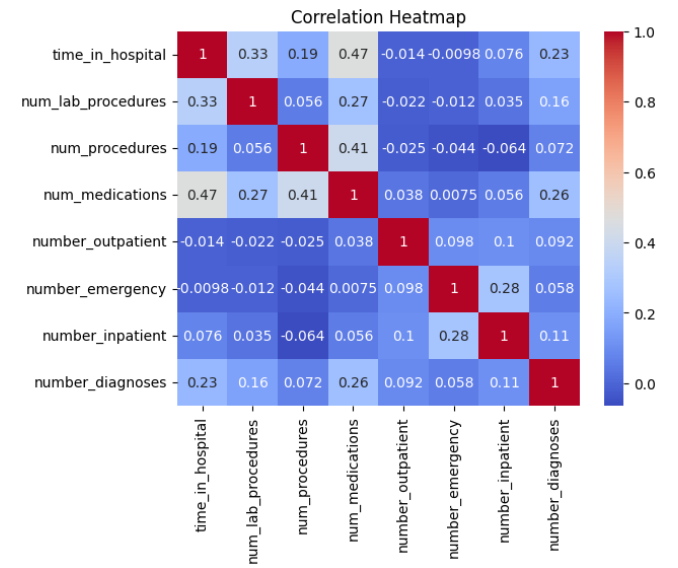


Fig. 4: Correlation Matrix

5. Pair Plot: The pair plot (Fig. 5) offers a visual depiction of the relationships between pairs of numerical features through scatter plots. This visualization facilitates the identification of potential patterns, trends, or outliers within the data. Understanding these relationships aids in feature engineering decisions and can uncover nonlinear associations that may not be apparent through correlation analysis alone.

6. Box Plot: The box plot (Fig. 6) has been utilized to identify and potentially remove outliers within the numerical columns:

- 1) 'time_in_hospital',
- 2) 'num_lab_procedures',
- 3) 'num_procedures',
- 4) 'num_medications', and
- 5) 'number_diagnoses'.

Outliers can significantly impact model performance and predictive accuracy, hence addressing them is crucial for robust analysis and modeling.



Fig. 5: *Pair Plot*: Shows the scatter plot of every pair of numerical columns.

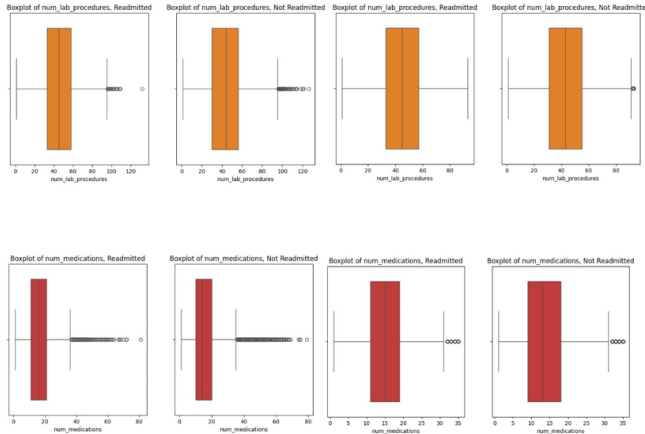


Fig. 6: *Box Plot*: The figure shows the box plot of two columns before outlier removal (on left) and after outlier remover (on right)

IV. METHODOLOGY

A. Data Preprocessing

1) *Removing Unnecessary Rows*: This step involves eliminating rows from the dataset that are deemed irrelevant for the predictive task at hand. In this specific case, rows with a value of ">30" in the "readmitted" column were removed, indicating patients who were readmitted after one month. By excluding such instances, the dataset is refined to focus solely on patients with readmission occurrences within a specified timeframe, enhancing the clarity of predictive modeling objectives.

2) *Handling Missing Values*: Missing data can significantly impact the efficacy of predictive modeling. To address

this issue, several strategies were employed:

- For columns with relatively low percentages of missing values ('race', 'diag_1', 'diag_2', 'diag_3'), missing values were treated as a separate category to retain potentially informative data.
- For columns with higher percentages of missing values ('weight', 'payer_code', 'medical_specialty'), missing values were imputed using the mode of the respective columns. This imputation strategy ensures minimal disruption to the dataset's overall distribution while filling in missing entries with plausible values.

3) *Drop Unnecessary Columns*: Certain columns within the dataset were deemed extraneous for the predictive task and were subsequently removed:

- 'encounter_id' and 'patient_nbr': These columns serve as unique identifiers for encounters and patients respectively, providing no discernible predictive value for readmission risk assessment.
- 'weight' and 'payer_code': These columns contained a substantial proportion of missing values (98.2

4) *Label Encoding*: To convert categorical columns into a numerical format, label encoding was implemented. This technique assigns a distinct numerical label to each unique category present in a categorical variable, facilitating the interpretation and processing of categorical data by machine learning algorithms.

5) *Feature Ranking*: Feature ranking techniques were employed to identify and prioritize features with significant predictive power. This has been discussed in detail in the next subsection.

6) *Normalization*: Min-max scaling, a type of normalization, was applied to standardize numerical features within a common range (typically 0 to 1). This scaling technique ensures that all features contribute proportionally to the model's learning process, preventing features with larger magnitudes from dominating the predictive model.

B. Feature Selection

In this section, we discuss the feature selection process employed in our project, which involves the application of various techniques followed by the evaluation of a Random Forest classifier with different subsets of selected features.

- *ReliefF Algorithm*: Features were ranked using ReliefF scores, with higher scores indicating greater relevance. Top-ranked features identified by ReliefF included 'glyburide-metformin', 'glimepiride', 'A1Cresult', 'admission_type_id', and 'number_inpatient'.
- *Correlation Based Feature Selection*: Correlation analysis was performed to measure the linear relationship between each feature and the target variable. Top features identified through correlation analysis included 'time_in_hospital', 'number_inpatient', 'glimepiride', 'diag_1', and 'nateglinide'.
- *Information Gain Based Feature Selection*: Features were ranked based on their information gain scores, with higher scores indicating greater predictive

power. Top features identified using information gain included 'insulin', 'race', 'miglitol', 'tolazamide', and 'troglitazone'.

- **SelectKBest:** Features were ranked based on their chi-square scores, representing the strength of association with the target variable. Top features identified using SelectKBest included 'discharge_disposition_id', 'number_inpatient', 'glimepiride', 'diag_1', and 'nateglinide'.
- **Combination of Rank Matrices and Model Evaluation:** After obtaining individual feature rankings from different methods, the rankings were combined into a rank matrix. This matrix provided a comprehensive view of feature importance across multiple selection techniques. Subsequently, we evaluated the performance of a Random Forest classifier with different subsets of selected features, ranging from the top-ranked features to the complete feature set. By analyzing the model's performance metrics across different feature subsets, we determined the optimal subset of features that maximized model performance while minimizing computational complexity.

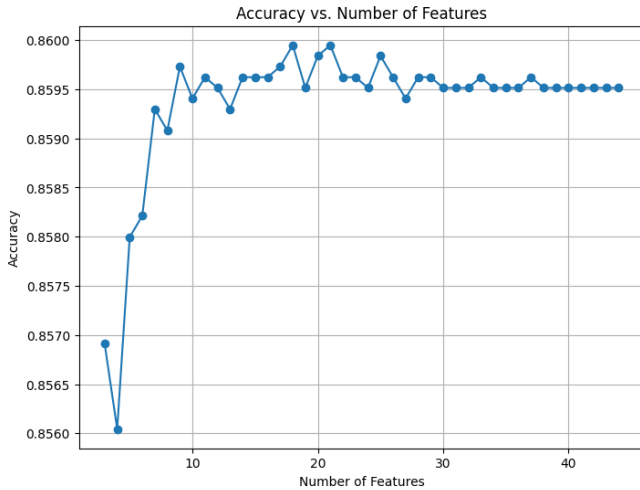


Fig. 7: Graph of Accuracy vs. Number of Features

Features Selected:

- admission_type_id
- number_inpatient
- num_lab_procedures
- number_diagnoses
- medical_specialty
- number_emergency
- num_procedures
- race
- discharge_disposition_id
- number_outpatient
- max_glu_serum
- A1Cresult
- glimepiride
- glipizide-metformin
- admission_source_id

- num_medications
- time_in_hospital
- repaglinide

C. Data Balancing using SMOTE

After preprocessing the dataset, we proceeded to split it into training and testing sets to evaluate the performance of our models.

Imbalanced datasets, where one class significantly outweighs the other, can lead to biased models that perform poorly on minority classes. To address this issue, we applied the Synthetic Minority Over-sampling Technique (SMOTE).

Before SMOTE:

- **Readmitted:** 5167 instances
- **Not Readmitted:** 31705 instances

After applying SMOTE, the dataset was rebalanced:

- **Readmitted:** 31705 instances
- **Not Readmitted:** 31705 instances

SMOTE generates synthetic samples of the minority class to balance the distribution of classes in the dataset. This ensures that our models are trained on a more representative dataset, leading to better generalization and performance on both classes.

D. Modeling

We employed GridSearchCV to search for the optimal hyperparameters for each model expect for KNN model. For KNN model We performed iterative hyperparameter varying the number of neighbors (n_neighbors) from 1 to 15

1) *Logistic Regression (LR)*: The best parameters obtained were:

- 'C': 1.0
- 'fit_intercept': True
- 'penalty': 'l2'

The accuracy achieved by the Logistic Regression model was 0.65.

2) *KNN (K-Nearest Neighbors)*: The best parameter obtained was:

- 'n_neighbors': 2

The accuracy achieved by the KNN model was 0.75.

3) *Decision Trees (DT)*: The best parameters obtained were:

- 'ccp_alpha': 0.001
- 'criterion': 'entropy'
- 'max_depth': None
- 'max_features': 'auto'

The accuracy achieved by the Decision Trees model was 0.74.

4) *Random Forest*: The best parameters obtained were:

- 'criterion': 'gini'
- 'max_depth': None
- 'n_estimators': 200

The accuracy achieved by the Random Forest model was 0.85.

TABLE I: Explanation of the selected features.

NO.	Features	Type	Values/Explanation
1	Admition_type_id	Integer	1, 2, 3, ..., 8 Admission Type
2	Number_inpatient	Integer	The number of visits led to hospitalization during that year
3	Num_lab_procedures	Integer	Number of tests performed during the patient's hospitalization
4	Number_diagnoses	Integer	Number of diagnoses from the time of hospitalization
5	mediccal_specialty		
6	Number_emergency	Integer	Number of emergency visits before hospitalization during that year
7	Num_procedures	Integer	Number of procedures performed except the tests performed during the patient's hospitalization
8	Race	Poly	Oceania, Asia, America, Spain, and Others
9	Discharge_disposition_id	Integer	1, 2, 3, ..., 29 Clearance destination
10	Number_outpatient	Integer	Number of outpatients before hospitalization during that year
11	Max_glu_serum	Binominal	Yes, No Glucose serum test
12	A1Cresult	Real	A1C test result
13	Admission_source_id	Integer	1, 2, 3, ..., 26 Admission source
14	Num_medications	Integer	Number of different medications administered during the patient's hospitalization
15	Time_in_hospital	Integer	Number of days from the time of inception to release
16	glimepiride	Categorical	indicates whether the drug was prescribed or there was a change in the dosage.
17	glipizide-metformin	Categorical	indicates whether the drug was prescribed or there was a change in the dosage

5) *XGBoost*: The best parameters obtained were:

- 'learning_rate': 0.1
- 'max_depth': 7
- 'n_estimators': 300

The accuracy achieved by the XGBoost model was 0.84.

V. RESULTS

After selecting the most important features and determining the optimal number of features, we evaluated the performance of various classification models in predicting readmission risk. The following table presents the performance metrics for each model:

TABLE II: Performance metrics for different classification models

Model	Precision	Recall	F1-score	Accuracy
Logistic Regression	0.81	0.65	0.70	0.65
KNN	0.87	0.83	0.85	0.75
Decision Trees	0.79	0.74	0.77	0.74
Random Forest	0.79	0.85	0.81	0.85
XGBoost	0.77	0.75	0.76	0.75

VI. AN ALTERNATE APPROACH

In this approach, we applied PCA and Undersampling to the 45 features remaining after removing 'encounter_id', 'weight', 'payer_code', and 'patient_nbr' from the available 49 features.

A. Undersampling

For undersampling, we utilized the RandomUnderSampler from imbalanced-learn (imblearn) library. The table below compares the count of instances for 'Readmitted' before and after undersampling:

TABLE III: Comparison of Readmission Before and After Undersampling

Readmitted	Before Undersampling	After Undersampling
Yes	49823	10420
No	10420	10420

B. PCA

To reduce any potential bias in our dataset, we applied PCA. Firstly, we employed a random forest classifier to determine the number of principal components for which the accuracy is maximized. Subsequently, we replaced all features with principal components and evaluated the performance using various classification models.

The graph in Figure 8 depicts the accuracy of the Random Forest Classifier against the number of principal components considered. Notably, after n=9, the accuracy stabilizes.

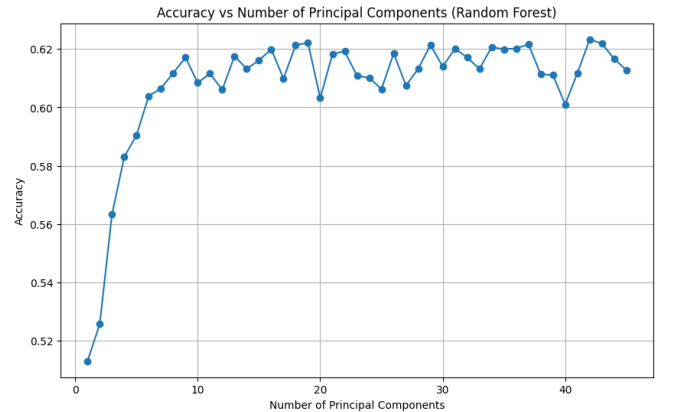


Fig. 8: Accuracy of the Random Forest Classifier vs Number of Principal Components

The table below presents the performance metrics for different classification models after PCA:

TABLE IV: Performance Metrics for Different Classification Models (With 9 Principal Components)

Model	Precision	Recall	F1-score	Accuracy
Logistic Regression	0.61	0.61	0.61	0.61
KNN	0.60	0.60	0.60	0.60
Decision Trees	0.61	0.61	0.61	0.61
Random Forest	0.63	0.63	0.63	0.63
XGBoost	0.62	0.62	0.62	0.62

Furthermore, we found the best parameters for each model:

- Logistic Regression:
'C': 1.0, 'fit_intercept': True, 'penalty': 'l2'
- Decision Trees:
'ccp_alpha': 0.001, 'criterion': 'gini', 'max_depth': 8, 'max_features': 'sqrt'
- Random Forest:
'criterion': 'entropy', 'max_depth': 9, 'n_estimators': 200
- XGBoost:
'learning_rate': 0.01, 'max_depth': 5, 'n_estimators': 200
- KNN:
'n_neighbors': 11

This shows us that PCA and undersampling did reduce the bias in the data but they reduced the accuracy of the models.

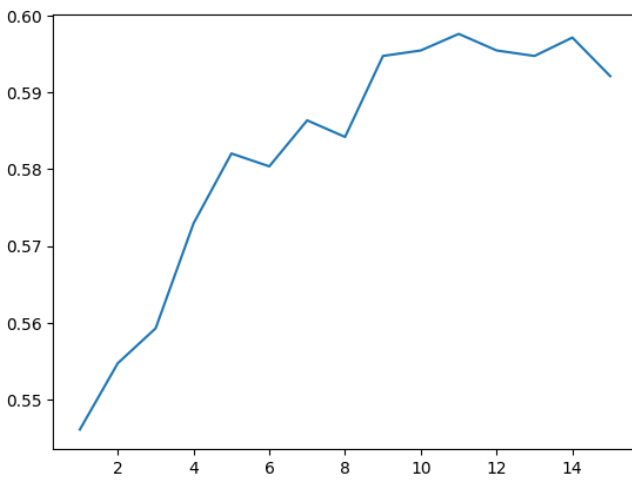


Fig. 9: Accuracy vs n_neighbours in KNN Model after applying PCA.

VII. CONCLUSION

In conclusion, our study aimed to improve the accuracy of predicting hospital readmission risk in diabetic patients by employing advanced feature selection and classification techniques. We utilized multiple feature selection algorithms, including ReliefF, Correlation, Information Gain, and SelectKBest, to identify and rank the most influential factors affecting readmission risk.

After evaluating various models, we observed that Random Forest achieved the highest accuracy, with an impressive 85% rate, closely followed by XGBoost, which demonstrated an 84% accuracy rate.

We also used an alternative approach using undersampling and PCA to remove bias from our model. However, this had led to decrease in accuracy of our models.

Moreover, our study contributes to the existing body of literature by offering a comprehensive framework for predicting readmission risk in diabetic patients, encompassing

advanced methodologies for feature selection, data balancing, and model optimization.

REFERENCES

- Hu, P., Li, S., Huang, Y.-a., and Hu, L. (2019). Predicting hospital readmission of diabetics using deep forest. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–2.
- Kumar, N. S. and Sathyanarayana, N. (2023). Prediction of diabetic patients with high risk of readmission using smart decision support framework. In *2023 Second International Conference on Electronics and Renewable Systems (ICEARS)*, pages 1664–1671.
- Ramírez, J. C. and Herrera, D. (2019). Prediction of diabetic patient readmission using machine learning. In *2019 IEEE Colombian Conference on Applications in Computational Intelligence (ColCACI)*, pages 1–4.
- Zeinalnezhad, M. and Shishehchi, S. (2024). An integrated data mining algorithms and meta-heuristic technique to predict the readmission risk of diabetic patients. *Healthcare Analytics*, 5:100292.

APPENDIX

Team_13_Anush_Akhilesh(1).pdf

ORIGINALITY REPORT

14%

SIMILARITY INDEX

8%

INTERNET SOURCES

8%

PUBLICATIONS

5%

STUDENT PAPERS

PRIMARY SOURCES

1	Masoomeh Zeinalnezhad, Saman Shishehchi. "An integrated data mining algorithms and meta-heuristic technique to predict the readmission risk of diabetic patients", Healthcare Analytics, 2023 Publication	4%
2	scholarscompass.vcu.edu Internet Source	1%
3	www.utupub.fi Internet Source	1%
4	www.mdpi.com Internet Source	1%
5	insightsociety.org Internet Source	1%
6	Submitted to University of Auckland Student Paper	1%
7	Submitted to ESC Rennes Student Paper	1%
8	Submitted to Eastern University Student Paper	