Predicting the Relationship Between Hemodialysis Frequency and Patient's Demographics, Comorbidities, CKD Severity, Lab Parameters, and Other Parameters

Alekhya Jilla, Likhitha Kantipudi, Sai Shakti Rao Lendale, and FNU Sahrash Fatima

Indiana University-Purdue University, Indianapolis, IN 46202, USA

Abstract. This project focuses on predicting the relationship between hemodialysis frequency and a range of patient factors, including demographics, comorbidities, CKD severity, lab parameters, and other relevant variables. By developing a classification model, we aim to provide valuable insights for clinical practitioners, facilitating improved clinical decision-making and treatment optimization to enhance patient care outcomes. The findings of this study have the potential to significantly impact clinical practice by advancing our understanding of the factors influencing hemodialysis frequency and guiding the development of personalized treatment plans.

Keywords: Chronic Kidney Disease · Hemodialysis · Comorbidities · Lab Parameters · Classification Models.

1 Project Scope

1.1 Introduction

In the United States, around 37 million suffer from Chronic Kidney Disease(CKD), which accounts for more than 14 percent of all the adults in the country. The risk is observed to be higher for people with preexisting diabetes or blood pressure, which are considered to be the two primary causes of kidney disease. About one-third of individuals with diabetes and one-fifth of those with high blood pressure develop kidney disease. Other factors contributing to kidney disease include heart disease and a family history of kidney failure (Kidney Disease Statistics for the United States, 2024). In the U.S., nearly 808,000 people live with end-stage renal disease (ESRD), also known as ESKD, with 69 percent undergoing dialysis and 31 percent having received a kidney transplant (Kidney Disease Statistics for the United States, 2024).

End-stage renal disease (ESRD) represents a final stage in kidney function decline, necessitating either hemodialysis or a kidney transplant to prolong the patient's life. Chronic kidney disease (CKD) is an ongoing condition that advances to ESRD when the glomerular filtration rate (GFR) falls below 15mL/min. The

transition from CKD to ESRD is influenced by various factors such as diabetes, glomerular diseases, vascular diseases, urinary tract obstruction, and tubulointerstitial disease (Hashmi, 2023).

Hemodialysis is the primary treatment offered to ESRD patients who are not candidates for renal transplantation. In developed countries, hemodialysis is typically performed three times a week (Chauhan and Mendonca, 2015).

Hemodialysis is a medical intervention utilized to remove waste products and excess fluids from patients when kidney function is no longer viable. This procedure involves the use of a dialyzer, which facilitates the circulation of the patient's blood outside the body while maintaining electrolyte balance (National Institute of Diabetes and Digestive and Kidney Diseases, 2019). Individuals with end-stage renal disease (ESRD) typically necessitate three hemodialysis sessions weekly, each lasting approximately three to five hours, to thoroughly eliminate toxins and filter the blood (Cleveland Clinic, n.d.).

1.2 Aim

We aim to analyze the relationship between hemodialysis frequency and patient data such as demographics, comorbidities, severity, lab parameters, mode, and process of hemodialysis. This project seeks to develop a classification model that accurately forecasts hemodialysis frequency based on these variables. The results of this analysis can greatly assist clinical practitioners in making informed decisions and optimizing treatment plans, ultimately leading to improved patient care.

1.3 Hypothesis

Based on our research aim, we have formulated a hypothesis regarding the relationship between various patient factors and hemodialysis frequency.

Null Hypothesis H0: There is no significant association between hemodialysis frequency with one or more of the patient factors such as demographics, comorbidities, CKD severity, lab parameters, and other parameters.

Alternate Hypothesis: H1: There is a significant association between hemodialysis frequency and one or more of the patient parameters such as demographics, comorbidities, CKD severity, lab parameters, and other parameters.

1.4 Purpose

The purpose of this study is to examine the relationship between hemodialysis frequency and various patient factors, such as demographics, comorbidities, CKD severity, lab parameters, and other relevant variables. Through the development of a predictive model, our goal is to provide actionable insights for clinical practitioners, enabling them to make informed decisions and optimize treatments to

enhance patient care outcomes. Ultimately, this research aims to advance our understanding of the factors influencing hemodialysis frequency and guide the development of personalized patient management strategies to improve clinical practice in treating End Stage Renal Disease (ESRD).

2 Methodology

The steps followed during the project are as follows:

- Data Collection, Storage, and Extraction: The relevant dataset is obtained from Kaggle and stored as an SQL file to ensure data integrity and data consistency. Python is the tool used for extracting the required data from SQL.
- Data Preprocessing: Data consistency was maintained by checking for duplicates and standardizing the data formats.
- Exploratory Data Analysis (EDA): The Interquartile Range (IQR) method was utilized to handle outliers, and it was observed that no outliers were present in our dataset. Summary statistics, such as measures of central tendency and measures of dispersion, provided insights into the distribution of numerical variables in the dataset.
- Statistical analysis: In our statistical analysis, we utilized several tests to assess different aspects of the data. Firstly, the Shapiro-Wilk test evaluated the normality of the distribution of numerical variables, determining whether the data adhered to a normal distribution. Since the data did not follow a normal distribution, non-parametric tests like the Mann-Whitney U test were considered for comparing two independent groups and handling ordinal data. Additionally, the chi-square test for independence was employed to examine associations between categorical variables.
- Data Visualizations: We generated a variety of visualizations including pie charts, bar charts, histograms, line charts, and heat maps to effectively represent the data clearly and concisely.
- Machine learning models: In our machine learning analysis, classification models were employed due to the ordinal nature of the target variable. Specifically, logistic regression, random forest, and gradient boost models were utilized. Various performance metrics such as accuracy, precision, recall, F1 score, ROC curve, and AUC were computed to evaluate the effectiveness of these models.

2.1 Team Members Responsibilities

Name	Background	Responsibilities
Alekhya Jilla	Bachelors of Dental Surgery	Machine Learning, Project Report
Likhitha Kantipudi	Pharm D	Statistical Analysis, Project Report
Sai Shakti Rao Lendale	Pharm D	Data Collection, Storage, and Extraction, Data Preprocessing.
FNU Sahrash Fatima	Pharm D	EDA,Data Visualization,

2.2 Project Challenges

- Navigating the initial stages of the project proved challenging for us as healthcare professionals without programming experience. Yet, with determination, we learned MySQL and Python, overcoming obstacles under the guidance of our professor, Saptarshi, and our teaching assistant. Together, we successfully completed the project within the allotted time.
- Low Accuracy of Machine Learning Models: The Logistic Regression, Random Forest, and Gradient Boost models demonstrated an average accuracy of approximately 33 percent. This indicated that they correctly predicted the outcome only about one-third of the time. Despite attempts to fine-tune hyperparameters, their accuracy did not show significant improvement, suggesting that the models may not have maximized their potential with the available data and features.
- Statistical analyses, such as the chi-square test and Spearman correlation, revealed that most patient parameters did not exhibit a statistically significant association with hemodialysis frequency per week. This unexpected finding led to the conclusion that the null hypothesis could not be dismissed, which proposed no significant relationship between hemodialysis frequency and the measured patient parameters.

3 Data Collection, Storage and Preprocessing

3.1 Data Collection

The dataset was collected from Kaggle website. This dataset is named 'Hemodialysis Realtime Hospital Dataset' and is licensed by CC0: Public Domain. The data was downloaded in the csv format. The dataset included patient demographic details, comorbidities, severity of the disease and various lab parameters that associate with the dialysis frequency.

3.2 Data Storage

After completing the Excel review, we proceeded to import the data into the group's shared database using phpMyAdmin.Initially, a table is created in the phpmyadmin with the appropriate data structure to import the csv file into it.

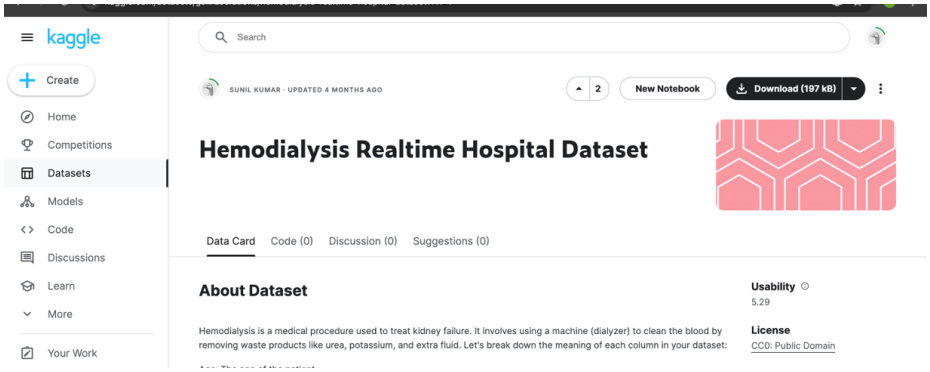


Fig. 1. Illustration of Kaggle page displaying the dataset.

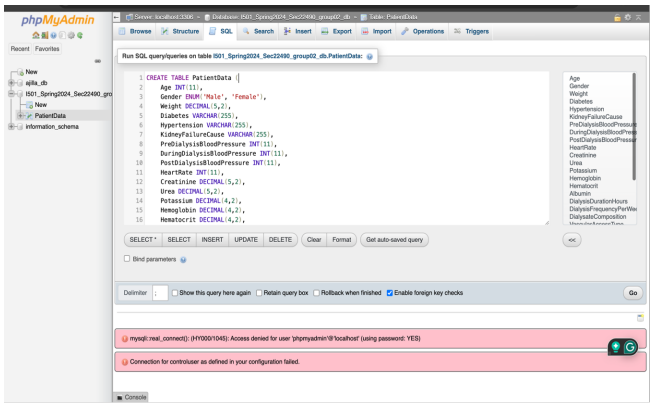


Fig. 2. Created table for data import

Description of Columns Our dataset has 27 columns, 19 numerical columns and 8 categorical columns. The columns are listed as follows:

Numerical Variables:

- Age of the patient
- Weight of the patient
- Pre-Dialysis Blood Pressure which represents the blood pressure of the patient before dialysis procedure.
- During-Dialysis Blood Pressure which represents the blood pressure of the patient during dialysis procedure.
- Post-Dialysis Blood Pressure which represents the blood pressure of the patient after dialysis procedure.
- Heart rate of the patient
- Creatinine measured in patient's blood
- Urea which is eliminated during dialysis
- Potassium which is regulated in kidneys.
- Hemoglobin measured in patient's RBCs.
- Hematocrit measured in patient's RBCs.
- Albumin measured from patient's plasma.
- Dialysis Duration is the duration of dialysis in hours.
- Dialysis Frequency per week is the frequency of dialysis undergone by the patient every week.
- Kt/V assesses the adequacy of dialysis.
- Urea Reduction Ratio (URR) is the amount of urine produced in a day.
- Urine Output is the amount of urine produced in ml/day.
- Dry Weight is th epatientt's weight without the excess fluid that is measured in kilograms.
- Fluid removal rate is the amount of fluid removed per hour during dialysis.

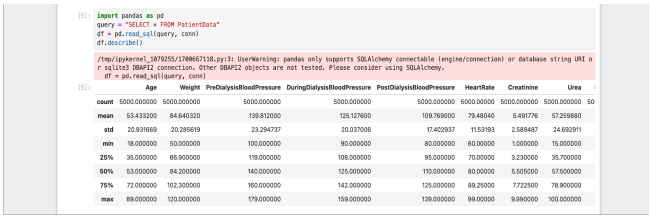
Categorical Variables:

- Gender of the patient
- Diabetes of the patient (True/False)
- Hypertension of the patient (True/False)
- Disease Severity of the patient (Mild/Moderate/Severe)
- Kidney Failure cause of the patient (Hypertension/Diabetes/Other)
- Dialysate Composition used in dialysis (Customized)
- Vascular Access Type created for the dialysis (e.g., Fistula, Graft, Catheter)
- Dialyzer Type dialyzer(Low-flux, High-flux).

3.3 Data Connection

To ensure data integrity, we established a secure connection between the SQL file and the Python notebook. The connection was facilitated using MySQLdb as the connecting tool.

Fig. 3. Data connection with Python



	Age	Weight	PreDialysisBloodPressure	DuringDialysisBloodPressure	PostDialysisBloodPressure	HeartRate	Creatinine	Urea	
count	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.00000	5000.000000	5000.000000	50
mean	53.433200	84.640320	139.812000	125.127600	109.769000	79.48040	5.491776	57.259880	
std	20.931669	20.285619	23.294737	20.037006	17.402937	11.53193	2.589487	24.692911	
min	18.000000	50.000000	100.000000	90.000000	80.000000	60.00000	1.000000	15.000000	
25%	35.000000	66.900000	119.000000	108.000000	95.000000	70.00000	3.230000	35.700000	
50%	53.000000	84.200000	140.000000	125.000000	110.000000	80.00000	5.505000	57.500000	
75%	72.000000	102.300000	160.000000	142.000000	125.000000	89.25000	7.722500	78.900000	
max	89.000000	120.000000	179.000000	159.000000	139.000000	99.00000	9.990000	100.000000	

Fig. 4. df.describe

3.4 Data Description

df.describe df.describe generates descriptive statistics summarizing key characteristics of the data in the DataFrame, which is useful for understanding its distribution and central tendencies (Fig 4).

df.dtypes df.dtypes revealed the data types of each column in the DataFrame, which aids in understanding the nature of the variables within the dataset.

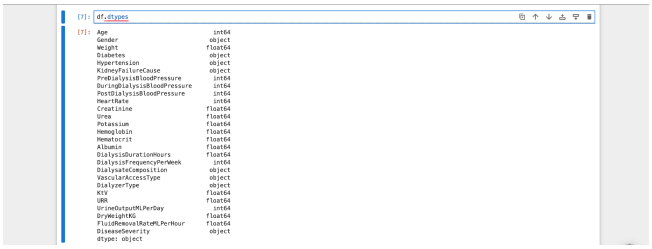


Fig. 5. df.dtypes

3.5 Data Preprocessing

Null values The command df.isnull().sum displays the count of missing values in each column of the DataFrame, aiding in data quality assessment and preprocessing. It revealed that there were no null values present in the dataset.

Fig. 6. Null values

Duplicate values Identifying and removing duplicates is essential for maintaining data quality, as it eliminates redundant records, streamlines data processing, and enhances the reliability of insights derived from the dataset. No duplicates were found in this case.

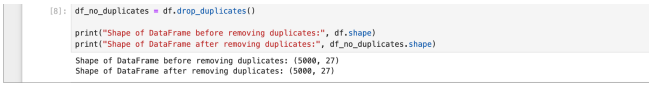


Fig. 7. Duplicate values

Outliers Outliers were identified through the Interquartile Range (IQR) method, which entails detecting data points falling below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR. Here, the IQR represents the spread of the middle 50 percent of the data. Upon applying the IQR method, it was noted that there are no outliers present in any of the numerical columns.

4 Exploratory Data Analysis(EDA)

4.1 Box and Whisker Plots

Box and Whisker plots faciliate in visualizing the distribution of numerical variables, by depicting the median, quartiles, and outliers.

4.2 Bar Charts

Bar charts visually depict the frequency or count of categorical variables, enabling straightforward comparison of distribution among different categories and identification of patterns or trends in the data.

```
Number of outliers in column 'Age': 0
Number of outliers in column 'Weight': 0
Number of outliers in column 'PreDialysisBloodPressure': 0
Number of outliers in column 'DuringDialysisBloodPressure': 0
Number of outliers in column 'PostDialysisBloodPressure': 0
Number of outliers in column 'HeartRate': 0
Number of outliers in column 'Creatinine': 0
Number of outliers in column 'Urea': 0
Number of outliers in column 'Potassium': 0
Number of outliers in column 'Hemoglobin': 0
Number of outliers in column 'Hematocrit': 0
Number of outliers in column 'Albumin': 0
Number of outliers in column 'DialysisDurationHours': 0
Number of outliers in column 'DialysisFrequencyPerWeek': 0
Number of outliers in column 'KtV': 0
Number of outliers in column 'URR': 0
Number of outliers in column 'UrineOutputMLPerDay': 0
Number of outliers in column 'DryWeightKG': 0
Number of outliers in column 'FluidRemovalRateMLPerHour': 0
```

Fig. 8. Outliers in numerical columns

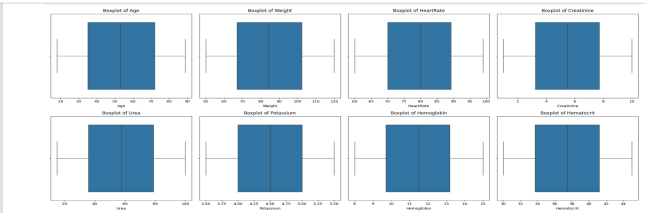


Fig. 9. Box and Whisker plots

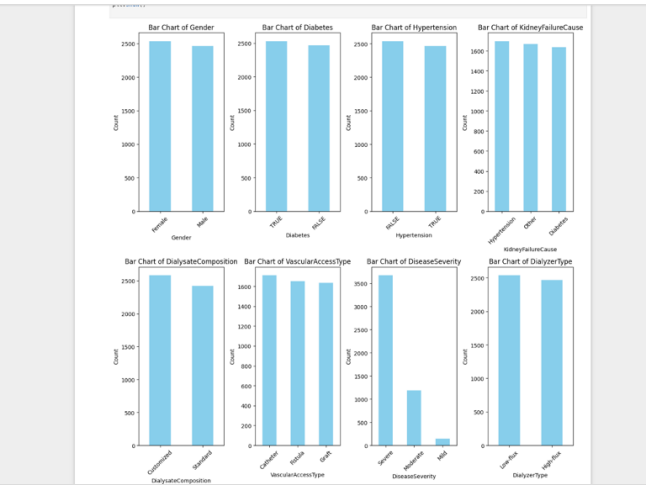


Fig. 10. Bar Charts

4.3 Pie charts

Pie charts offer a concise overview of the proportional distribution of categorical variables, aiding in understanding the relative sizes of different categories within the dataset.

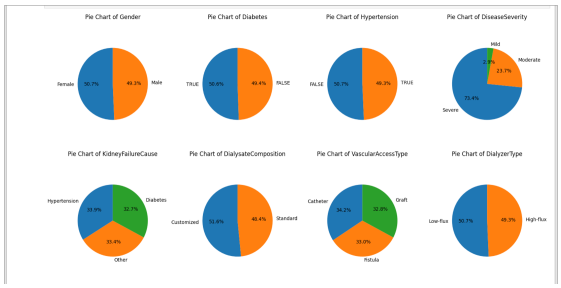


Fig. 11. Pie Charts

5 Statistical Analysis

5.1 Normality Testing

Shapiro-Wilk normality test was conducted to assess the normal distribution of continuous variables.The test results signified that the distribution of these is not normally distributed.

Fig. 12. Results of Shapiro-Wilk Test

5.2 Non-Parametric Test

Since our data did not demonstrate a normal distribution, we utilized non-parametric tests like the Mann-Whitney U test to identify any significant differences between two groups.

```
from scipy.stats import mannwhitneyu

x = df[['Age', 'Weight', 'PreDialysisBloodPressure',
        'DuringDialysisBloodPressure', 'PostDialysisBloodPressure', 'HeartRate',
        'Creatinine', 'Urea', 'Potassium', 'Hemoglobin', 'Hematocrit',
       'Albumin', 'DialysisDurationHours',
       'KtV', 'URR', 'UrineOutputMLPerDay', 'DryWeightKG',
        'FluidRemovalRateMLPerHour']]
y = df['DialysisFrequencyPerWeek']

for feature in x.columns:
    u_statistic, p_value = mannwhitneyu(x[feature], y)
    print(f"Mann-Whitney U test for feature '{feature}':")
    print(f"P-value: {p_value}")
    if p_value < 0.05:
        print("Reject the null hypothesis: There is a significant difference between groups.")
    else:
        print("Fail to reject the null hypothesis: There is no significant difference between groups.")
```

Fig. 13. Mann-Whitney U Test

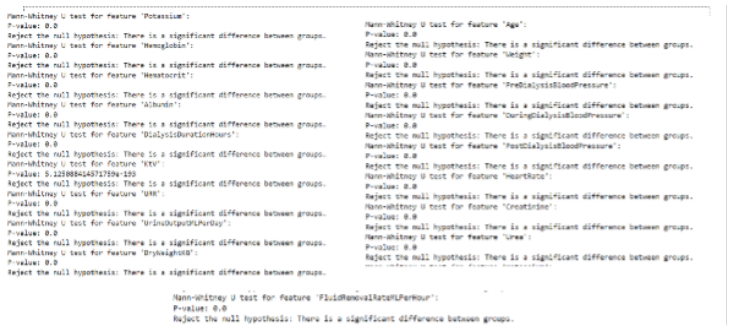


Fig. 14. Results of Mann-Whitney U Test

5.3 Chi-Square Test

The chi-square test for independence was employed to examine associations between categorical variables.This test evaluates whether there is a significant association between each categorical variable and the target variable.

```
Gender: Chi-square statistic = 0.2151891216177439, p-value = 0.8979916036234705
Diabetes: Chi-square statistic = 0.009722141166217401, p-value = 0.9951507252992192
Hypertension: Chi-square statistic = 7.476376817977098, p-value = 0.023797174853415457
DiseaseSeverity: Chi-square statistic = 8.018336968920936, p-value = 0.09090879259088606
KidneyFailureCause: Chi-square statistic = 3.8673799211694977, p-value = 0.4242514497416163
DialysateComposition: Chi-square statistic = 1.1158168728589382, p-value = 0.5724050342059385
VascularAccessType: Chi-square statistic = 3.111130647642323, p-value = 0.539403239325601
DialyzerType: Chi-square statistic = 0.755719202168039, p-value = 0.6853267130392591
```

Fig. 15. Results of Chi-Square Test

5.4 Spearman Correlation Analysis

We performed Spearman correlation analysis, a non-parametric method, because of the non-linear and non-continuous characteristics of our dataset. The null hypothesis is accepted for variables with a p-value greater than the selected significance level (0.05). This suggests that for these variables, there is no significant association with the frequency of dialysis per week. However, for the variable "KtV," the null hypothesis is rejected as its Spearman correlation p-value is below 0.05. Thus, there exists a significant association between "KtV" and the frequency of dialysis per week. This analysis is visualized through heatmap which presents the correlations among all variables considered in our hypothesis analysis.

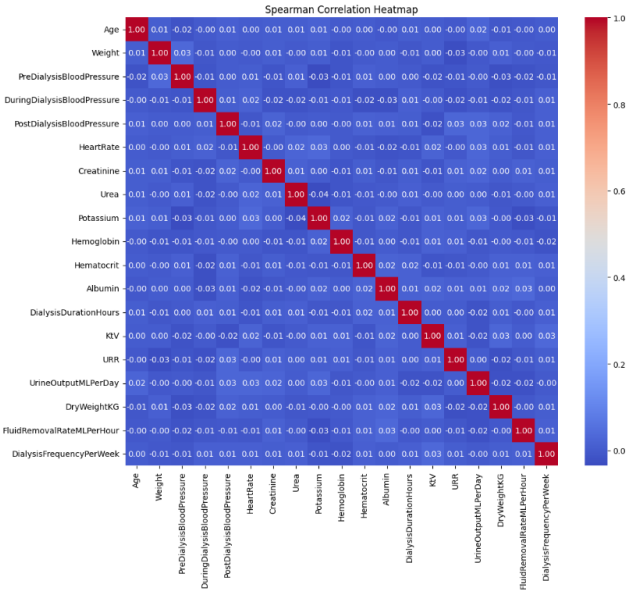


Fig. 16. Heatmap

6 Label Encoding

Label encoding is done to convert categorical into numerical columns so that they can be fitted by machine learning models that use numerical data. It is one of the crucial steps to perform before machine learning.

Fig. 17. Label Encoding

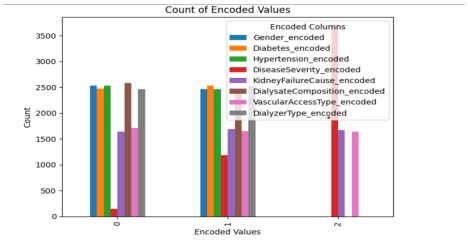


Fig. 18. Count of Encoded Variables

7 Data Visualization

7.1 Distribution of Numerical Variables

These bar charts aid in understanding the distribution of continuous variables. These images depict that the distribution is not following a Gaussian pattern, hence the requirement of non-parametric tests for further analysis.

7.2 Age

A bar chart representing the frequency of dialysis is displayed for two age groups: under 45 and over 45. It displays the total number of patients receiving dialysis at different intervals during the week. Most people in both groups receive dialysis three times a week.

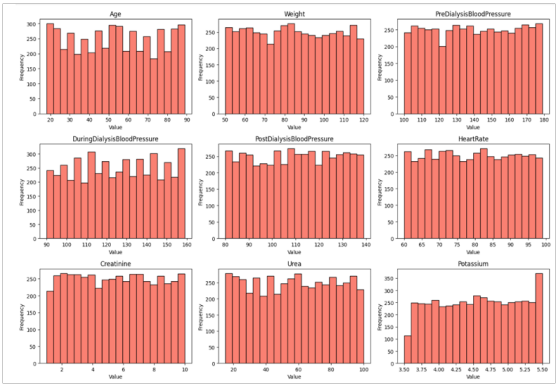


Fig. 19. Distribution of Numerical Variables

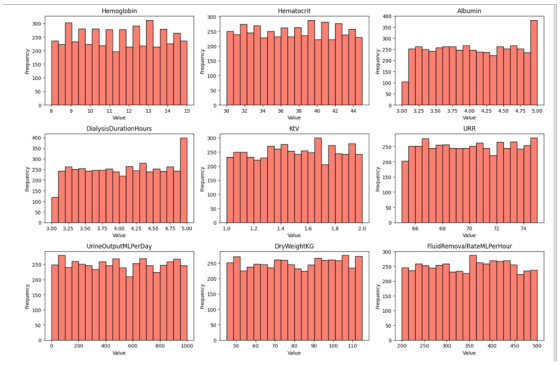


Fig. 20. Distribution of Numerical Variables

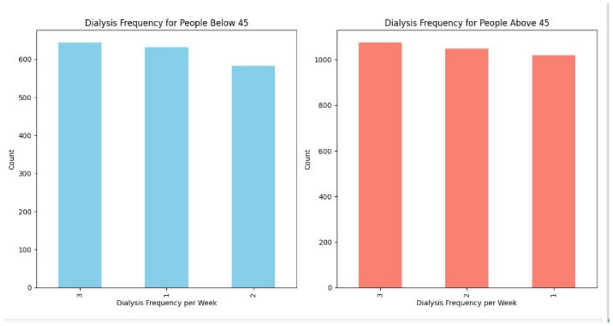


Fig. 21. Dialysis frequency per week vs Age

7.3 Kidney Failure Cause

Line charts depict the occurrences of kidney failure causes: diabetes alone, hypertension with other causes, and diabetes alone. Bar charts detail dialysis frequency for each cause, highlighting hypertension as the most common cause followed by diabetes alone and a combination of causes. All groups exhibit a predominant frequency of three dialysis sessions per week.

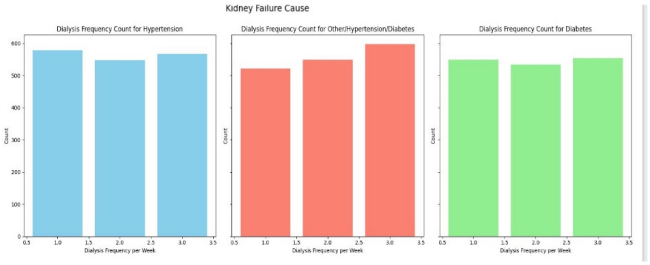


Fig. 22. Dialysis frequency per week vs Kidney Failure Cause

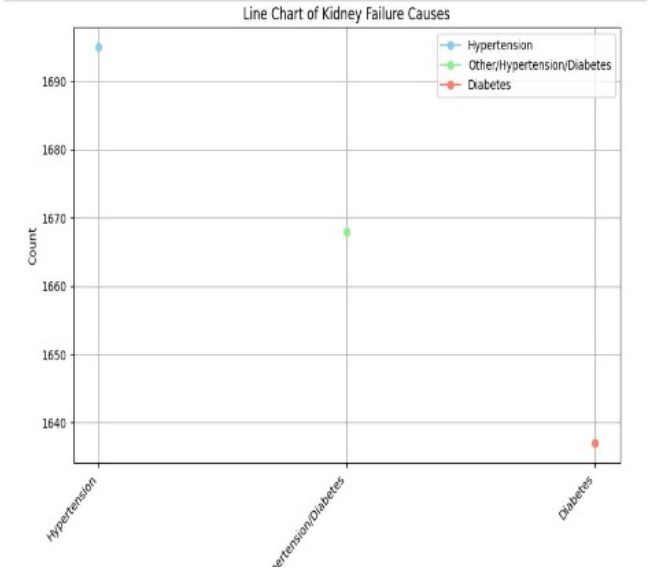


Fig. 23. Dialysis frequency per week vs Kidney Failure Cause

7.4 Creatinine

In the line charts, the left plot indicates higher creatinine levels correspond to a sharp rise in weekly dialysis frequency, peaking at around 2.0 sessions. The right plot indicates Dialysis frequency decreases after reaching a peak of approximately two sessions per week for creatinine levels below 1.2.

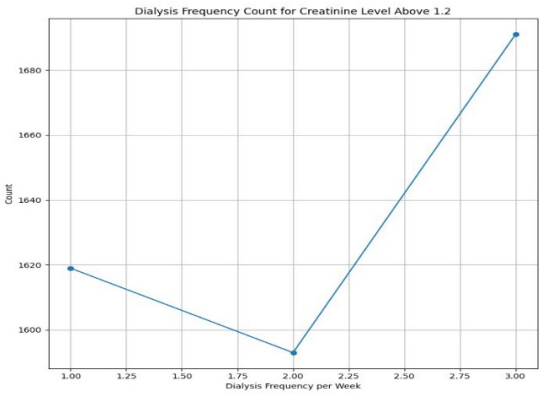


Fig. 24. Line chart of Creatinine(left)

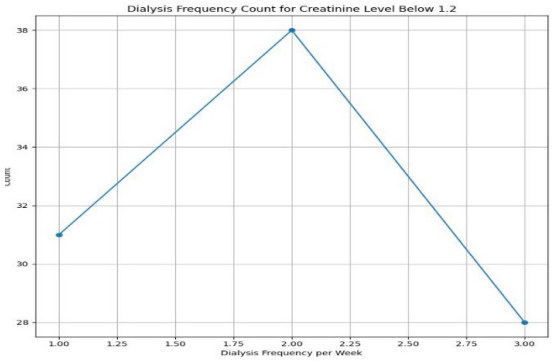


Fig. 25. Line Chart of Creatinine(right)

7.5 Potassium

In the dialysis frequency chart, higher potassium levels correlate with a sharp increase in frequency, peaking around 2.0 sessions, while lower levels show a decline after reaching a peak of approximately 2.0 sessions per week.

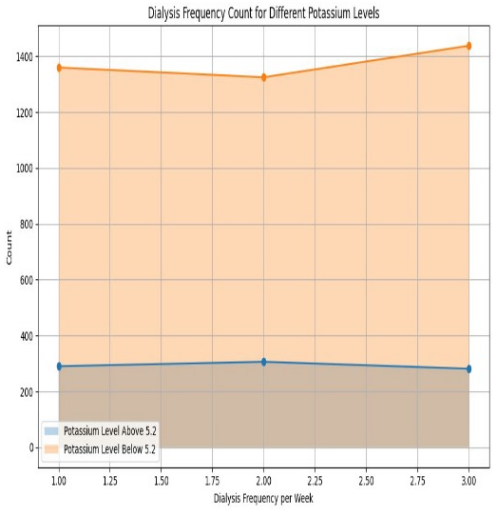


Fig. 26. Dialysis frequency per week vs potassium chart

7.6 Hemoglobin

Stacked bar charts displaying the number of patients by hemoglobin level (above or below 10) and dialysis frequency (once, twice, or three times per week). For patients with hemoglobin levels exceeding 10, dialysis frequency 1 (blue bar) had the greatest count, followed by frequency 2 (orange bar) and frequency 3 (green bar). The count shows lower total numbers for those whose hemoglobin levels are below 10. Still, the pattern is comparable.

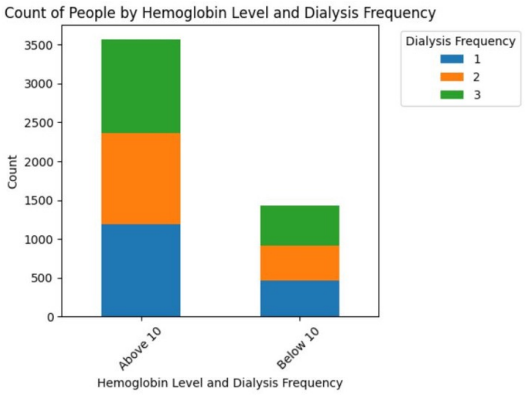


Fig. 27. Dialysis frequency vs Hemoglobin stacked bar chart

7.7 Disease Severity

In the line chart, severe cases show a gradual increase in dialysis frequency, peaking at around 3.0 sessions per week, while mild cases have the lowest count and least variation across frequencies, and moderate cases remain relatively constant.

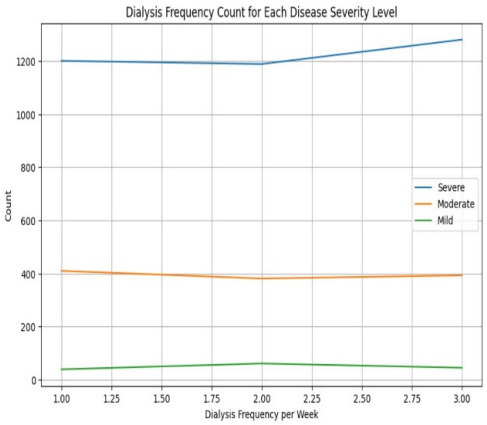


Fig. 28. Dialysis frequency vs Disease Severity

7.8 Gender,Diabetes, Hypertension,Dialysate Composition,and Vascular Access Type

The distribution of gender appears to be relatively balanced across all dialysis frequencies, the same for diabetics.This chart shows that a higher proportion of patients who undergo dialysis 3 times per week have no hypertension. The use of customized dialysate composition appears to be more common than standard solutions across all dialysis frequencies. However, the proportion of patients on customized dialysate increases slightly with higher dialysis frequency. Catheters seem to be the most commonly used vascular access type, followed by fistula and grafts. The distribution remains relatively higher in catheters across different dialysis frequencies. Low-flux dialyzers are more widely used than high-flux dialyzers, and their usage appears to increase with higher dialysis frequency.

7.9 Class Distribution of Target Variable

This bar graph depicts that the three class variables of the target variable (dialysis frequency per week) are evenly distributed, so it will not present any class imbalance that might impact machine learning performance.

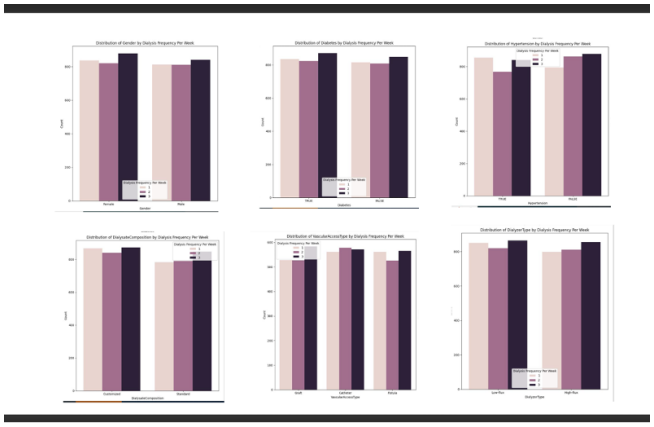


Fig. 29. Dialysis frequency vs Categorical Variables

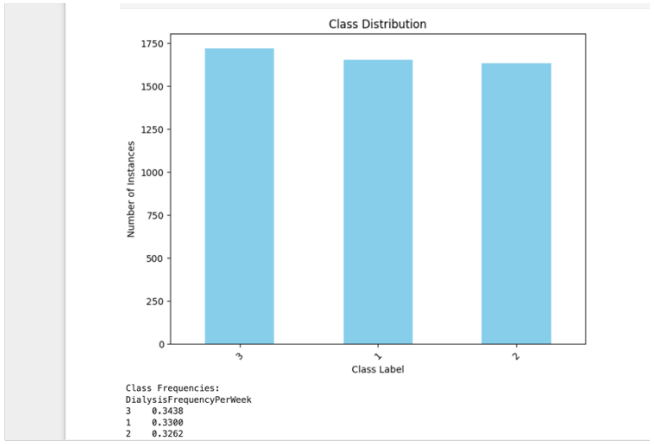


Fig. 30. Class Distribution of Target Variable

8 Machine Learning Models

In machine learning, partitioning data into training and testing sets is pivotal for model development and evaluation. The robust capabilities of the Scikit-learn (sklearn) library employs the train test split function to partition the dataset into training and testing sets. A fixed random state of 42,ensures reproducibility and facilitates comparisons across different methodologies and implementations.

Careful consideration is given to the selection of features that will inform the classification model. These features, based on their relevance and predictive power, can uncover meaningful patterns and relationships within the data. The features selected for our classification model based on their predictive power and statistical significance to the target variable (Dialysis Frequency per Week) are as follows:

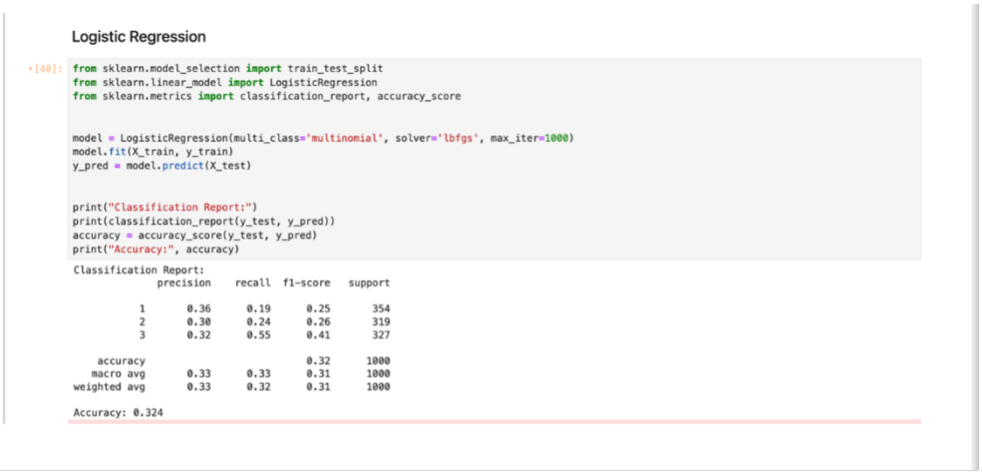
- VascularAccessType(encoded)
- DiseaseSeverity(encoded)
- KidneyFailureCause (encoded)
- DialysateComposition (encoded)
- FluidRemovalRateMLPerHour
- URR
- DryWeightKG
- Heart rate
- PostDialysisBloodPressure
- DuringDialysisBloodPressure
- KtV
- DialysisDurationHours
- Hematocrit
- Urea
- Creatinine

Fig. 31. Classification Model Features and Train-Test Split

8.1 Logistic regression

In our project, we employed logistic regression as one of the classification modeling technique to predict the target variable. Logistic regression is a statistical technique widely used in machine learning for classification tasks, was utilized to predict outcomes based on provided features.

However, the logistic regression model yielded an accuracy score of 0.324 which might be because of lack of significant statistical relationship between the features and the target variable. This indicates potential limitations in the logistic regression model's ability to capture underlying patterns within the data effectively.


Logistic Regression

```
[40]: from sklearn.model_selection import train_test_split
      from sklearn.linear_model import LogisticRegression
      from sklearn.metrics import classification_report, accuracy_score

      model = LogisticRegression(multi_class='multinomial', solver='lbfgs', max_iter=1000)
      model.fit(X_train, y_train)
      y_pred = model.predict(X_test)

      print("Classification Report:")
      print(classification_report(y_test, y_pred))
      accuracy = accuracy_score(y_test, y_pred)
      print("Accuracy:", accuracy)
```

```
Classification Report:
              precision    recall  f1-score   support

           1       0.36      0.19      0.25       354
           2       0.30      0.24      0.26       319
           3       0.32      0.55      0.41       327

    accuracy                           0.32      1000
   macro avg       0.33      0.33      0.31      1000
weighted avg       0.33      0.32      0.31      1000

Accuracy: 0.324
```

Fig. 32. Logistic regression and Classification Report

Cross validation Despite implementing cross-validation techniques to optimize the logistic regression model, the accuracy only increased marginally to 33.2 percent. This can suggest that the model's performance remains limited by the underlying complexity and distribution of the data.

Confusion Matrix The confusion matrix offers valuable insights into the performance of classification model, revealing patterns of correct and incorrect predictions across different classes.The confusion matrix highlights areas of concern such as high false positive or false negative rates, indicating potential challenges in accurately classifying certain instances.

ROC Curve of Logistic Regression Predicted probabilities for each class are generated for the test data using the predictproba method. The ROC curves for

```
from sklearn.model_selection import cross_val_score, KFold
from sklearn.linear_model import LogisticRegression
from sklearn.pipeline import make_pipeline
from sklearn.preprocessing import StandardScaler

logReg = LogisticRegression()

n_folds = 10

kf = KFold(n_splits=n_folds, shuffle=True, random_state=42)
pipeline = make_pipeline(StandardScaler(), logReg)
cv_scores = cross_val_score(pipeline, x, y, cv=kf)
print("Cross-validation scores:", cv_scores)
print("Mean accuracy:", np.mean(cv_scores))
```

```
Cross-validation scores: [0.332 0.324 0.332 0.338 0.342 0.328 0.312 0.286 0.378 0.352]
Mean accuracy: 0.3324000000000003
```

Fig. 33. Cross-Validation of Logistic regression

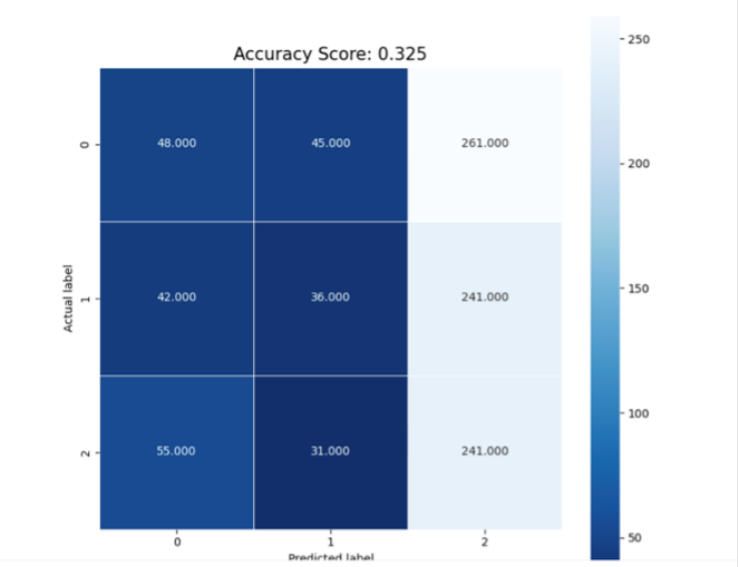


Fig. 34. Confusion Matrix of Logistic regression

each class are plotted on a single graph using Matplotlib. Given that the AUC values for all three classes hover around 0.5, akin to random chance, it indicates a subpar performance of the model.

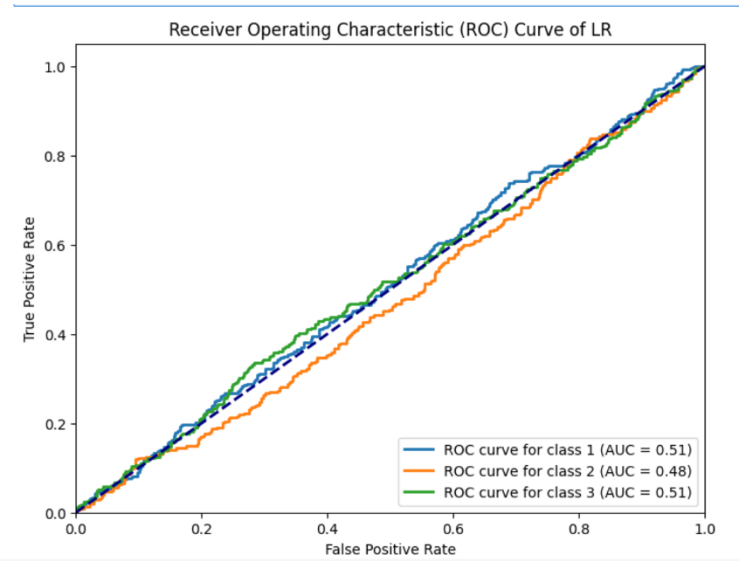


Fig. 35. ROC Curve of Logistic Regression model

8.2 Random Forest Classifier

Random Forest Classifier is a classification model that uses powerful ensemble learning method through constructing a multitude of decision trees during training and outputs the mode of the classes as the prediction. In our analysis, the Random Forest Classifier yielded an overall accuracy of 0.337. This suggests that the model did not achieve strong accuracy accuracy in classifying instances across the dataset.

Similarly, the classification report for the Random Forest Classifier model was generated. The report provided the precision, recall, and F1 scores for each class. For Class 1, the precision was 0.34, the recall was 0.27, and F1 score was 0.30. For Class 2, the precision was 0.34, the recall was 0.31, and F1 score was 0.32. For Class 3, the precision was 0.34, the recall was 0.43, and F1 score was 0.38. The overall accuracy, macro average, and weighted average were also reported.

Cross validation A 10-fold cross-validation was performed on the dataset using the KFold method from the sklearn module. The cross-validation scores for the

model were obtained, and the mean cross-validation accuracy was calculated to be 0.33625.This can suggest that the model's performance remains limited by the underlying complexity and distribution of the data.

ROC Curve of RFC The ROC curves for each class are plotted on a single graph using Matplotlib. Additionally, a diagonal dashed line representing a random classifier is included for reference. The resulting plot visualizes the model's discriminative power across different thresholds. The ROC curves for class 1 have an Area Under the Curve (AUC) of 0.51, and class 2 has an AUC of 0.50, which indicates a fair classification performance, while class 3 has a slightly lower AUC of 0.48, suggesting a less accurate classification result.

```
[48]: from sklearn import metrics
      print(metrics.classification_report(expected, predicted))
      print(metrics.classification_report(rfc_expected, rfc_predicted))
```

	precision	recall	f1-score	support
1	0.42	0.26	0.32	354
2	0.33	0.25	0.28	319
3	0.34	0.56	0.42	327
accuracy			0.36	1000
macro avg	0.36	0.36	0.34	1000
weighted avg	0.36	0.35	0.34	1000

	precision	recall	f1-score	support
1	0.36	0.32	0.34	354
2	0.32	0.34	0.33	319
3	0.33	0.35	0.34	327
accuracy			0.34	1000
macro avg	0.34	0.34	0.34	1000
weighted avg	0.34	0.34	0.34	1000

Fig. 36. Classification Report of RFC

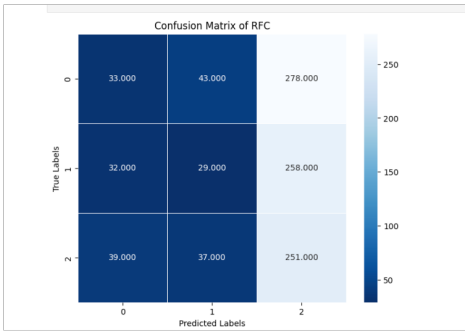


Fig. 37. Confusion Matrix of RFC

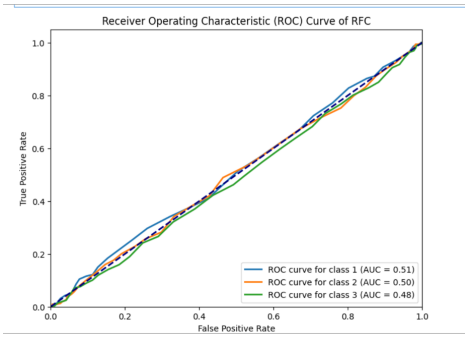


Fig. 38. ROC Curve of RFC

8.3 Gradient Boost Classifcation

The Boosting Classifier is an algorithm that amalgamates numerous weak learning models (decision trees) to formulate a robust predictive model. Initially, the model imports requisite modules such as GradientBoostingClassifier from sklearn module. It then proceeds to instantiate a GradientBoostingClassifier and trains it on the training data. Following this, predictions are generated for the test data utilizing the trained model.

ClassificationReport The accuracy of the Gradient Boosting Classifier is computed using the accuracy score function and is revealed to be 0.357 based on the supplied test data. Additionally, a comprehensive classification report is generated, furnishing intricate metrics including precision, recall, F1-score, and support for each class.

Cross validation Moreover, the model undertakes cross-validation employing the crossvalscore function from sklearn model. The resulting cross-validation scores for the Gradient Boosting Classifier are presented alongside the mean cross-validation accuracy (0.338) and the standard deviation of the cross-validation accuracy (0.016).

ROC Curve of GBC The ROC curves for class 1 have an Area Under the Curve (AUC) of 0.53, and class 3 has an AUC of 0.51, which indicates a fair classification performance, while class 2 has a slightly lower AUC of 0.51, suggesting a less accurate classification result.

9 Hyperparameter Tuning

In our project, we conducted hyperparameter tuning for three models: Logistic Regression (LR), Random Forest Classifier (RFC), and Gradient Boosting Classifier (GBC). Hyperparameter tuning involves systematically searching through

Gradient Boost

```
[54]: from sklearn.ensemble import GradientBoostingClassifier
      from sklearn.metrics import accuracy_score

      gbc = GradientBoostingClassifier()

      gbc.fit(X_train, y_train)
      gbc_predicted = gbc.predict(X_test)

      gbc_accuracy = accuracy_score(y_test, gbc_predicted)
      print("Gradient Boosting Classifier Accuracy:", gbc_accuracy)

      Gradient Boosting Classifier Accuracy: 0.357

[94]: print("Classification Report:")
      print(classification_report(y_test, gbc_predicted))

      Classification Report:
                    precision    recall  f1-score   support

                 1       0.37      0.29      0.33       354
                 2       0.34      0.32      0.33       319
                 3       0.36      0.46      0.40       327

          accuracy                           0.36      1000
         macro avg       0.36      0.36      0.35      1000
      weighted avg       0.36      0.36      0.35      1000
```

Fig. 39. Classification report of GBC

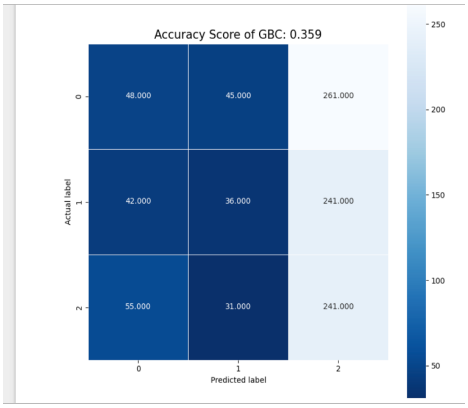


Fig. 40. Confusion Matrix of GBC

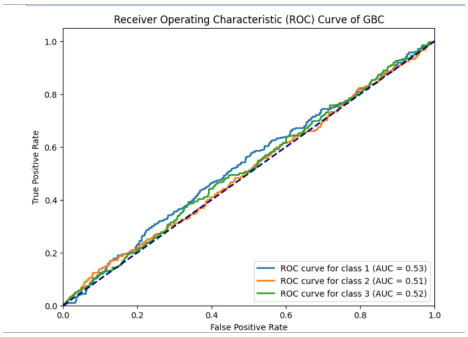


Fig. 41. ROC Curve of GBC

a predefined parameter grid, typically using techniques like grid search or randomized search, to find the optimal combination of hyperparameters.

For LR, despite our efforts in hyperparameter tuning, the accuracy decreased compared to the baseline model. This suggests that LR may not be the most appropriate algorithm for our dataset.

Conversely, for RFC and GBC, hyperparameter tuning led to an increase in accuracy compared to their respective baseline models. This signifies that the chosen hyperparameters effectively enhanced the performance of these models, resulting in improved predictive accuracy.

These findings signify the importance of thorough experimentation and optimization of hyperparameters to achieve the best possible performance from machine learning models.

Fig. 42. Hyperparametric Tuning of LR

Fig. 43. Hyperparametric Tuning of RFC

Fig. 44. Hyperparametric Tuning of GBC

9.1 Comparitive Analysis of ML Models

ML Models	Accuracy	Mean Accuracy of CV	Accuracy after HT
Logistic Regression	32.4	33.2	31.3
Random Forest	33.6	34.0	34.6
Gradient Boost	35.7	33.7	36.5

10 Conclusion

In conclusion, the machine learning models developed for the project face a significant challenge in achieving satisfactory accuracy levels for predicting hemodialysis frequency. Despite attempts, including the utilization of Logistic Regression, Random Forest, and Gradient Boost models, their accuracies remain stagnant at around 33 percent, even after hyperparameter tuning. The overall findings from various analyses suggest that the majority of patient parameters do not have a significant impact on determining hemodialysis frequency. Therefore, based on the evidence provided, we cannot reject the null hypothesis (H0) indicating no significant association between hemodialysis frequency and patient parameters, including demographics, comorbidities, CKD severity, lab parameters, and others.

11 Limitations

- Low Model Accuracy: Inconsistent data may have introduced noise or misleading patterns, hindering the models' ability to accurately learn the true relationships between variables. As a result, this inconsistency could have contributed to the models' lower accuracy during training.
- Potential Data Limitations: The unexpected absence of correlations between patient parameters and dialysis frequency could arise from deficiencies within the dataset. These may involve issues related to data accuracy, such as errors or disparities, or the omission of crucial variables influencing dialysis frequency.

12 References

- Kidney disease statistics for the United States. (2024, March 5). National Institute of Diabetes and Digestive and Kidney Diseases. https://www.niddk.nih.gov/health-information/health-statistics/kidney-disease
- Hashmi, M. F. (2023, August 28). End-Stage renal disease. StatPearls - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK499861/
- Chauhan, R., Mendonca, S. (2015). Adequacy of twice weekly hemodialysis in end stage renal disease patients at a tertiary care dialysis centre. Indian Journal of Nephrology/Indian Journal of Nephrology, 25(6), 329.https://doi.org/10.4103/0971-4065.151762

– National Institute of Diabetes and Digestive and Kidney Diseases. (2019, June 5). Hemodialysis | NIDDK. https://www.niddk.nih.gov/health-information/kidney-disease/kidney-failure/hemodialysis
– Cleveland Clinic. (n.d.). Dialysis. Cleveland Clinic. https://my.clevelandclinic.org/health/treatments/14618-dialysis