HOSPITAL READMISSION

INTRODUCTION:

Readmission analysis is a well-established area of research in healthcare, with numerous studies conducted to understand and reduce readmission rates. Here are some typical focuses of readmission analysis work:

Risk Factor Identification: Studies often identify and analyze risk factors for readmission, which can include clinical factors like the severity of the disease, comorbidities, and complications during the initial hospital stay, as well as socio-demographic factors such as age, sex, socioeconomic status, and access to healthcare.

Predictive Modeling: Many projects use statistical and machine learning models to predict the likelihood of patient readmission. Commonly used techniques include logistic regression, decision trees, random forests, and more recently, deep learning models. These models help in understanding which variables most significantly predict readmissions and can be targeted to mitigate risks.

Intervention Strategies: Research is also directed towards developing and evaluating intervention strategies to reduce readmissions. These might include improved discharge planning, patient education programs, enhanced follow-up care, and integrated care pathways.

Policy and Management: Studies often assess the impact of policy changes and management practices on readmission rates. This includes analyzing the effects of new healthcare policies or changes in hospital practices on the quality of care and readmission rates.

Economic Impact Analysis: Some research focuses on the cost implications of readmissions, providing insights into how readmissions affect healthcare spending and how reducing readmissions can lead to cost savings.

Quality Improvement Studies: These studies typically involve interventions at the healthcare facility level and measure changes in readmission rates to assess the effectiveness of specific quality improvement initiatives.

METHODS:

In this section we will provide a brief information about approaches taken, exploratory data analysis, the preprocessing techniques, model preparations, machine learning models, and evaluation metrics we are intending to use for our project.

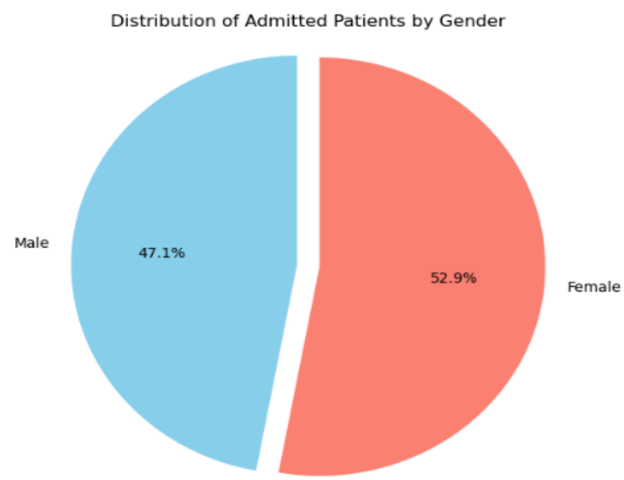
Importing Libraries, Packages, and Dataset:

In our project, we utilized variety of Python libraries to efficiently manage the data. We relied on Pandas for its strong data handling capabilities and Numpy for comprehensive operations. For data visualizations, we used Matplotlib, Plotly and Seaborn. Our machine learning processes were supported by Scikit-learn and enhanced with XGBoost for more complex modeling. To address class imbalances, we integrated SMOTE from imblean library. Additionally, Ipywidgets was used for creating interactive charts within our environment.

Exploratory Analysis of the Data:

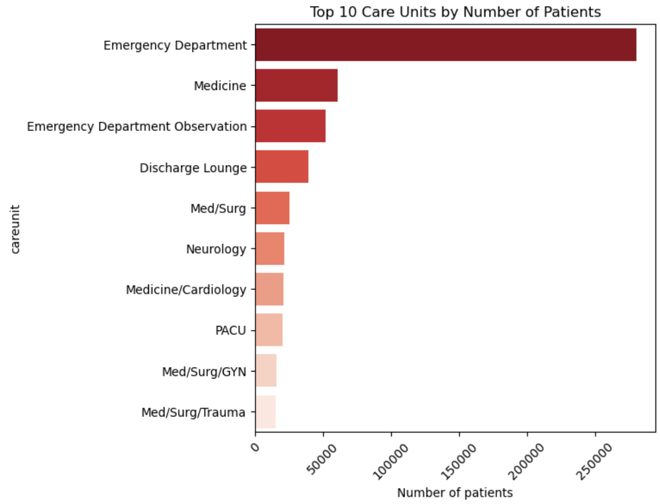
We initiated our exploratory data analysis by loading "admissions, patients and diagnoses dataset", examining their structure and content with functions such as head(), describe(), and info(). This exploration enabled us to detect critical trends and patterns, such as demographic distributions and healthcare utilization. We also graphically represented patient health trends, acuity levels, and other relevant metrics to derive deeper insights into the data.

- Furthermore, as part of our EDA, we delved into distribution of admitted patients by gender, frequency of hospitalizations by admission location, ranking of care unites by number of patients, and monthly trends in patient admissions. Each visual representation revealed valuable insights :
    1. **Distribution of Admitted Patients by Gender:** We observed slight predominance of female patients, representing 52.9% of admissions compared to 47.1% for males. Pie chart emphasized importance of considering gender difference in our analysis and potential implications for gender specific healthcare services.

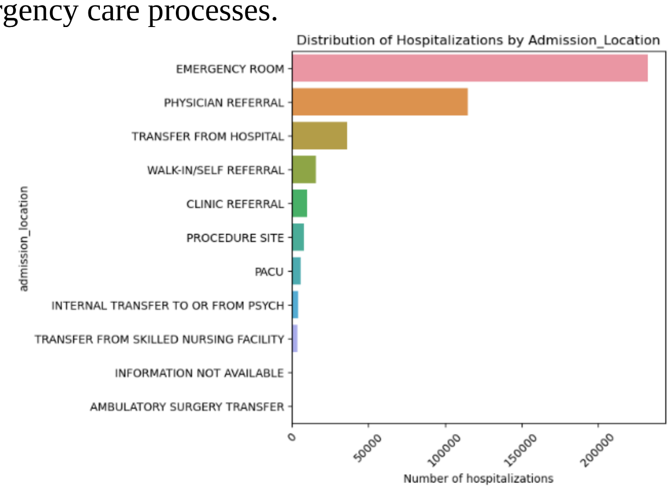


    2. **Top 10 care units by Number of Patients:** Our analysis highlighted that Emergency Department, followed by Medicine unit, saw highest volume of

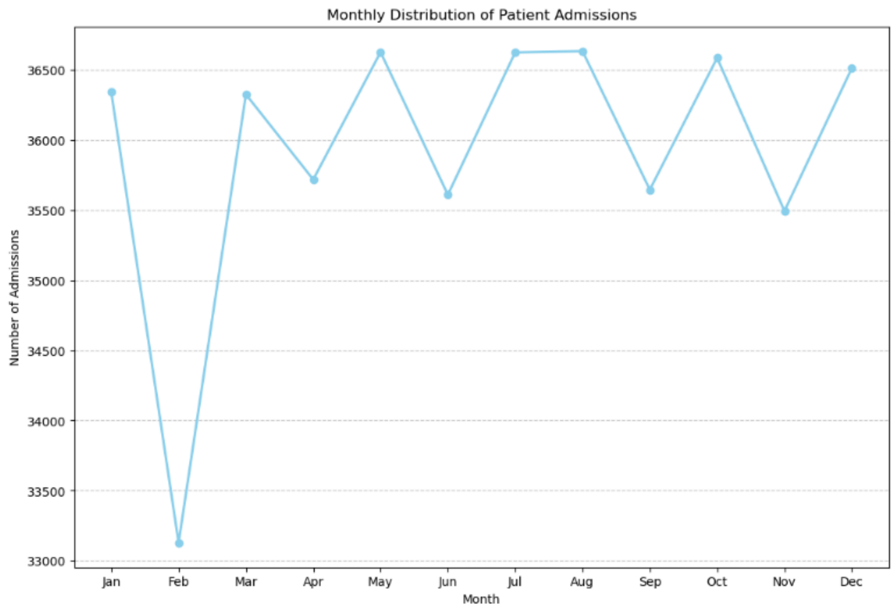
patients, pointing to critical need for resources in these areas. This insight informs resource allocation and operational strategies for managing patient flow and improving care delivery.



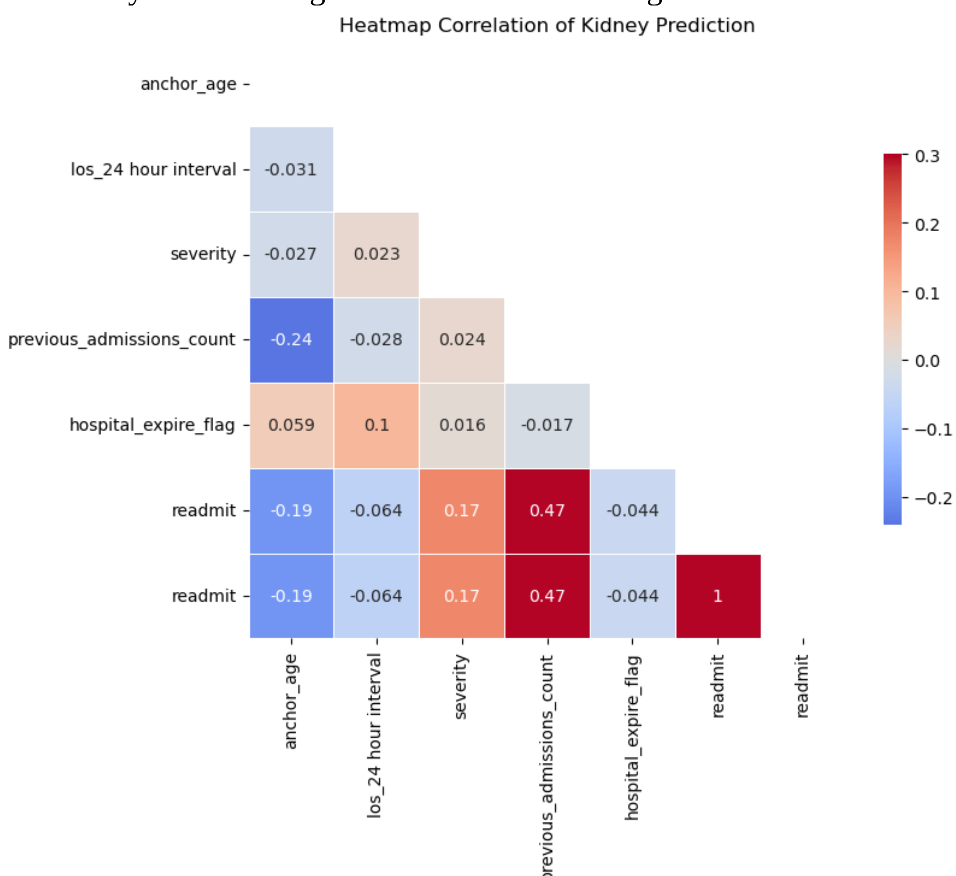
3. **Distribution of Hospitalization by Admission location:** Bar chart illustrated that most hospitalizations originated from Emergency room, followed by Physician referral and Transfer from other Hospitals. This underscores emergency room's role as key factor for hospital admissions and importance of streamlining emergency care processes.



4. **Monthly Distribution of Patient Admissions:** Line graph depicted monthly admission trends indicated fairly consistent number of admissions throughout the year, with noticeable peaks and troughs. For example, there was significant dip in admissions in February, which might correlate with seasonal factors or healthcare policy changes that merit further investigation.



5. **Heatmap correlation of kidney predictions:** We can notice that previous_admissions_count has moderate negative correlation with anchor_age. Readmit shows positive correlation with severity, indicating higher the severity of illness might be associated with a higher chance of readmission.



These visualizations were integral to our understanding of healthcare landscape within which our predictive models operated. They provided backdrop for our subsequent analysis and model training, enabling us to tailor predictive algorithms to nuances uncovered in initial phase of exploration.

Data Preparation and Preprocessing:

In our project, we initiated on multi-step analytical phase to better understand and predict patient readmissions.

1. **Data Selection and Merging:**
   We selected and merged essential tables such as admissions, patients and diagnoses from MIMIC-IV database. Python query was executed to filter and merge data from mentioned tables based on subject_id and icd_code which served as unique identifier for each record. This comprehensive process was our first step in creating unified dataset that would serve as foundation for our subsequent analysis.
2. **Readmission Indicators Analysis:**
   Next, we computed critical readmission indicators such as **length of stay** for each patient based on their admit and discharge time for consecutive visits; occurrence of **readmission within 30 days** by analyzing consecutive patient visits, determining time difference between admission time of the current visit and discharge time of the previous visit for each patient; and **count of past visits**. These indicators were important in constructing more detailed picture of patient readmission patterns, serving as key predictor in our models.

3. **Repeat Visit Analysis:**
   We then proceeded to query our dataset for patients who had multiple visits for the same diagnoses. By identifying these repeat visits, we aimed to uncover patterns and correlations that might influence likelihood of patient being readmited. Additionally, we added new feature in dataframe, **previous_admission_count** which represents number of past admissions for each patient up to their current hospital stay. Also, created feature called **vist_order** to indicate order of patients visit to hospital without using actual dates since our MIMIC-IV data was deindentified.

4. **Targeting key readmission Diagnoses:**
   Our analysis helped us determine that kidney Failure was among the top five diagnoses leading to patient readmissions. Rather than concentrating solely on kidney failure diagnosis, we broadened our focus to encompass any diagnoses related to kidney issues. This focus allowed us to tailor our analysis to kidney patients specifically, ensuring that our predictions would be as relevant and accurate as possible for this group.

5. **Severity Scoring using NLP:**
   Finally, we utilized rule based system from NLP to calculate severity scores for each kidney related diagnosis based on predefined set of medical criteria, further merging traditional medical knowledge with ML model to refine predictions.
   a. We created comprehensive dictionary (‘severity_keywords’) mapping specific keywords related to medical conditions and their stages - for example, “Stage iv (severe), chronic kidney disease, acute, kidney failure”; to severity scores ranging

from 1 to 5 based on medical foundation. This dictionary serves as base for evaluating the severity of various kidney diagnoses.

b. We implemented python function that dynamically assessed the severity score of diagnosis by searching highest-priority keyword contained in description. Function further iterates through each keyword in the dictionary defined, applying default score of 1 (least severe) if no specific keyword found, ensuring that each diagnosis receives severity evaluation.

c. By intergrating severity scores into dataset as new column, we habe augmented our data with additional dimension that captures clinical significance of each diagnoses. This have enabled our predictive model to consider severity of patient conditions, potentially improving accuracy of readmission predictions.

Feature Engineering:

After prepping and preprocessing our data, we focused on feature engineering to identify which variables could best predict readmissions. We identified key variables including mix of demographics such as gender and age, clinical data such as length of stay, and administrative information such as admission location and insurance status. Additionally, we engineered new features to capture historical patterns of hospital usage as well as patient journey such as number of previous admissions and visit order to reflect the sequence of visits for each patient. Our objective was to include as much relevant information as possible to create informative picture of patient profiles.

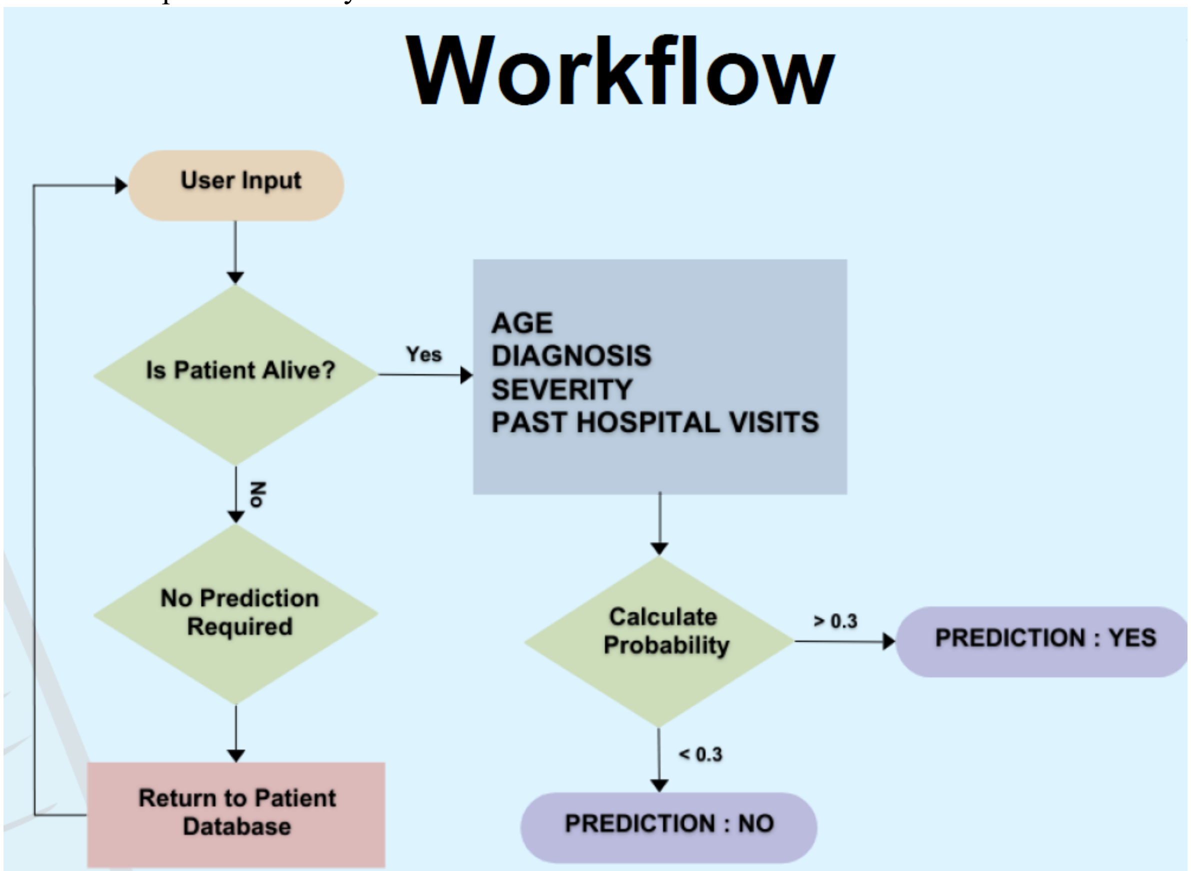
For our model targets, we chose **readmit_under_30_days** for predicting whether patient would be readmitted within next 30 days, a significant measure for assessing quality of care and patient outcomes. We labeled these potential outcomes as binary variables, forming strong base of our predictive model.

Model Preparation:

Now that our features selected, we proceeded to prepare them for analysis. Categorical features like gender, icd_code, admission location and insurance were transformed using OneHotEncoder, which converted them into binary matrix necessary for ML algorithm. Numerical features such as age, length of stay, severity, previous admission count, hospital expire flag and readmits were standardized, which means we scaled them to have mean of zero and standard deviation of one, to ensure that all features contributed equally to model's predictions.

- Interactive Predictive Widget :
  - We designed interactive widget within Python Jupyter to predict readmission based on subject_id input.
  - Created prediction function "**predict_readmission**" to estimate risk of individual patients readmissions within 30days. This function incorporates predictive model and applies custom logic to adjust predictions based on severity of patient's condition and previous admissions. This function first examines patients latest hospital record for **hospital_expire_flag** to determine if patient is alive; only then does it continue with further prediction analysis. It also factors in number of patient

visits, age, diagnosis, and severity before providing probability assessment of whether patient is likely to be readmitted.



Data Pipeline Construction:

To streamline our model building process, we constructed data pipelines for each classifiers. Each pipeline began with **ColumnTransformer**, which combined encoding and standardization processes into single step. This helped to maintain efficiency and consistency across our models.

For **RandomForestClassifier** and **XGBClassifier**, we followed preprocessing with **SimpleImputer** to handle any potential missing values, ensuring that our datasets were complete. Subsequently, we included classifiers themselves in the pipeline, setting their respective hyperparameters such as random_state for reproducibility.

For **LogisticRegression** model, we deployed similar pipleline, adjusting imputer strategy to **mean** to accommodate specific requirements of this algorithm. We also set **max_iter** to higher number to allow model to coverage properly, considering complexity of our dataset.

Constructed pipeline not only streamlined preprocessing and model training process, but also safeguard against data leakage, ensuring that our models learned to generalize from training data without being inadvertently influenced by test set. This is crucial as it allowed us to confidently assess each model's predictive power on unseen data.

Throughout this process, we split our data into training and testing sets, ensuring that we could both train our models effectively as well as evaluate their performance objectively. With each step carefully executed, our pipelines stood as strong tool to identify necessary patterns within our data and offer predictions on patient readmissions.

Machine Learning Models:

Upon constructing our data pipeline, we looked into variety of algorithms to handle complex patterns within our large dataset. We explored following key models, each offering unique strength to address our analytical challenges:

- **Logistic Regression**
- **Random Forest Classifier**
- **XGBoost**

Logistic Regression served as our initial baseline model, valued for its simplicity and interpretability. It allowed us to establish foundational understanding of relationships between our features and likelihood of patient readmissions. Next, we looked into RandomForestClassifier, which is less prone to overfitting and adept at capturing nonlinear relationships making it strong choice of our complex dataset. Lastly, we utilized XGBoost algorithm, which is known for its performance and speed.

Evaluation Metrics:

Evaluation metrics played key role in our machine learning implementaion, serving as guiding based of model optimization which offered different perspective on each models predictive abilities.

Comprehensive classification report gave us overview of **precision**, **recall** and **F1 scores**, allowing us to balance trade off between model sensitivity and specificity. Precision measures accuracy of positive predictions, recall gauges models ability to find all positive samples and F1 scores showcases both, providing single score to express overall performance. Accuracy reflected proportion of total correct predictions. It gave us quick snapshot of effectiveness across all classes.

Confusion matrix showed models performance in more granular detail, revealing instances of true positives, true negatives, false positives, and false negatives. This was crucial for understanding types of errors our models were making.

Lastly, Receiver operating characteristics (**ROC**) curve, coupled with area under curve (**AUC**) score, allowed us to assess model’s discriminating capacity as to how well it could distinguish between classes at various threshold settings. This helped us further refine our models by ensuring it could not only predict accurately but also with higher degree of confidence.

RESULTS:

In this section we will provide a brief information about approaches taken, exploratory data analysis, the preprocessing techniques, model preparations, machine learning models, and evaluation metrics we are intending to use for our project.

1. Readmits Key Indicators : LOS, time_diff, readmission_under_30_days

	subject_id	admittime	dischtime	los_24 hour interval
0	10000032	2180-05-06 22:23:00	2180-05-07 17:15:00	18.87
1	10000032	2180-06-26 18:27:00	2180-06-27 18:49:00	24.37
3	10000032	2180-07-23 12:35:00	2180-07-25 17:55:00	53.33
2	10000032	2180-08-05 23:44:00	2180-08-07 17:50:00	42.10

	subject_id	hadm_id	admittime	dischtime	time_diff_admit	time_diff_disch	time_diff	readmission_under_30_days
1	10000032	22841357	2180-06-26 18:27:00	2180-06-27 18:49:00	50.0	51.0	50 days 01:12:00	0
3	10000032	29079034	2180-07-23 12:35:00	2180-07-25 17:55:00	26.0	27.0	25 days 17:46:00	1
2	10000032	25742920	2180-08-05 23:44:00	2180-08-07 17:50:00	13.0	12.0	11 days 05:49:00	1

2. Patient hospital admission records and readmission tracking

	subject_id	hadm_id	los_24 hour interval	admission_type	hospital_expire_flag	time_diff_admit	time_diff_disch	gender	anchor_age	readmission_under_30_days	readmission_under_60_days
0	10000032	22595853	18.87	URGENT	0	0.0	0.0	F	52	0	0
1	10000032	22841357	24.37	EW EMER.	0	50.0	51.0	F	52	0	1
2	10000032	29079034	53.33	EW EMER.	0	26.0	27.0	F	52	1	1
3	10000032	25742920	42.10	EW EMER.	0	13.0	12.0	F	52	1	1
4	10000068	25022803	7.17	EU OBSERVATION	0	0.0	0.0	F	19	0	0

3. Identifying recurrent readmissions by subject_id and icd_code :

subject_id	icd_code	readmit
10000032	2761	3
10000032	2767	2
10000032	29680	2
10000032	3051	3
10000032	496	4
...	...	...
19999828	Y838	2
19999828	Z87891	2
19999840	2724	2
19999840	4019	2
19999840	43811	2

4. Filtering patient data for for kidney-related diagnoses and readmission analysis :

subject_id	hadm_id	admittime	dischtime	los_24 hour interval	admission_location	insurance	hospital_expire_flag	time_diff_admit	time_diff_disch	time_diff	readmit_under_30_days	readmit_under_60_days	gender	anchor_age	icd_code	icd_version	long_title	readmit
10000032	22595853	2180-05-06 22:23:00	2180-05-07 17:15:00	18.87	TRANSFER FROM HOSPITAL	Other	0	0.0	0.0	NaT	0	0	F	52	78959	9.0	Other ascites	4
10000032	22841357	2180-06-26 18:27:00	2180-06-27 18:49:00	24.37	EMERGENCY ROOM	Medicaid	0	50.0	51.0	50 days 20:04:00	0	1	F	52	78959	9.0	Other ascites	4
10000032	29079034	2180-07-23 12:35:00	2180-07-25 17:55:00	53.33	EMERGENCY ROOM	Medicaid	0	26.0	27.0	26 days 18:08:00	1	1	F	52	78959	9.0	Other ascites	4
10000032	25742920	2180-08-05 23:44:00	2180-08-07 17:50:00	42.10	EMERGENCY ROOM	Medicaid	0	13.0	12.0	13 days 11:09:00	1	1	F	52	78959	9.0	Other ascites	4
10000032	22595853	2180-05-06 22:23:00	2180-05-07 17:15:00	18.87	TRANSFER FROM HOSPITAL	Other	0	0.0	0.0	NaT	0	0	F	52	5715	9.0	Cirrhosis of liver without mention of alcohol	4

5. Patient admission trends : Analyzing previous admissions and visit orders :

subject_id	hadm_id	los_24 hour interval	time_diff	readmit_under_30_days	readmit_under_60_days	icd_code	readmit	long_title	previous_admissions_count	visit_order
10000980	29654838	47.82	NaT	0	0	40390	4	Hypertensive chronic kidney disease, unspecifi...	0	1
10000980	26913865	139.37	540 days 13:57:00	0	0	40390	4	Hypertensive chronic kidney disease, unspecifi...	1	2
10000980	26913865	139.37	540 days 13:57:00	0	0	5854	3	Chronic kidney disease, Stage IV (severe)	4	3
10000980	24947999	43.02	497 days 13:19:00	0	0	40390	4	Hypertensive chronic kidney disease, unspecifi...	2	4
10000980	24947999	43.02	497 days 13:19:00	0	0	5854	3	Chronic kidney disease, Stage IV (severe)	5	5

6. Assessment of Diagnosis severity in Patient :

subject_id	hadm_id	los_24 hour interval	anchor_age	icd_code	readmit	long_title	severity
10000980	29654838	47.82	73	40390	4	Hypertensive chronic kidney disease, unspecifi...	2
10000980	26913865	139.37	73	40390	4	Hypertensive chronic kidney disease, unspecifi...	2
10000980	26913865	139.37	73	5854	3	Chronic kidney disease, Stage IV (severe)	4
10000980	24947999	43.02	73	40390	4	Hypertensive chronic kidney disease, unspecifi...	2
10000980	24947999	43.02	73	5854	3	Chronic kidney disease, Stage IV (severe)	4
10000980	24947999	43.02	73	28521	2	Anemia in chronic kidney disease	2
10000980	25242409	189.55	73	40390	4	Hypertensive chronic kidney disease, unspecifi...	2
10000980	25242409	189.55	73	5854	3	Chronic kidney disease, Stage IV (severe)	4
10000980	25242409	189.55	73	28521	2	Anemia in chronic kidney disease	2
10000980	25911675	25.68	73	N184	3	Chronic kidney disease, stage 4 (severe)	4

7. Analysis showing Kidney Failure in top diagnosis :

icd_code character	long_title character varying (255)
41401	Coronary atherosclerosis of native coronary artery
K219	Gastro-esophageal reflux disease without esophagitis
V1582	Personal history of tobacco use
F329	Major depressive disorder, single episode, unspecified
5849	Acute kidney failure, unspecified
2449	Unspecified acquired hypothyroidism
I2510	Atherosclerotic heart disease of native coronary artery without angina pectoris
3051	Tobacco use disorder
2859	Anemia, unspecified

icd_code character	long_title character varying (255)
N179	Acute kidney failure, unspecified
5990	Urinary tract infection, site not specified
2720	Pure hypercholesterolemia
49390	Asthma, unspecified type, unspecified
V5867	Long-term (current) use of insulin
Z794	Long term (current) use of insulin
E039	Hypothyroidism, unspecified
5859	Chronic kidney disease, unspecified
Z7901	Long term (current) use of anticoagulants
E119	Type 2 diabetes mellitus without complications

DISCUSSIONS:

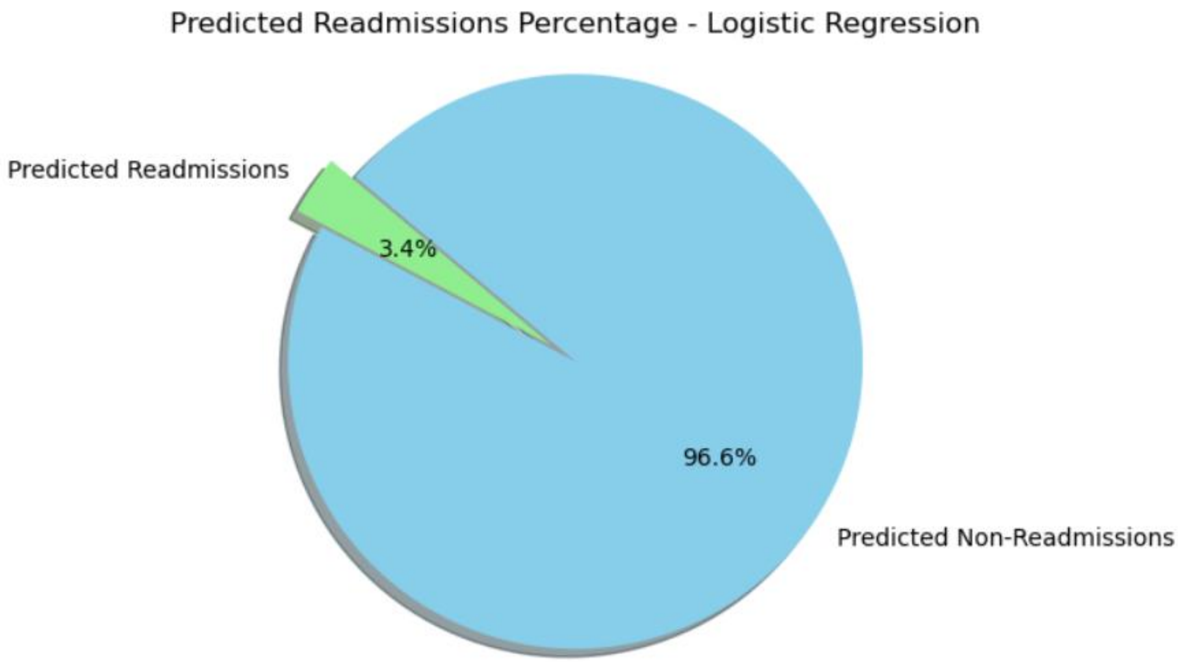
In this section we will deep dive into each of our predictive model that we employed in our project. We assessed their performance to ensure we could accurately predict kidney patient readmissions. Here's breakdown of how we went about this process:

Machine Learning Model	Accuracy Achieved
Logistic Regression	52%
Random Forest Classifier	70%
XGBoost	72%

1. **Logistic Regression Model :**
   - A logistic regression model is constructed to predict whether patients will be readmitted to the hospital within 30 days, utilizing features such as gender, age, length of stay, admission location and insurance type. These features are processed and encoded via a preprocessing pipeline that standardizes numerical features and applies one hot encoding to categorical features.
   - Despite the model being integrated into a pipeline that also handles missing values and is trained on a subset of the data, it achieves a precision score of 0.5229 and an ROC-AUC score of 0.6180. These scores reflect the model's moderate accuracy and its ability to differentiate between patients who will and will not be re-admitted. Furthermore, the pie chart indicates that the model predicted 3.4% of patients as potential admissions by categorizing the remaining 96.6% as non-readmissions.
   - However, due to its relatively lower performance metrics, we decided not to choose this logistic regression model for further development or deployment. The decision was based on the need for a more accurate and robust model to improve healthcare outcomes and decision making.

   - **Result's from Model :**

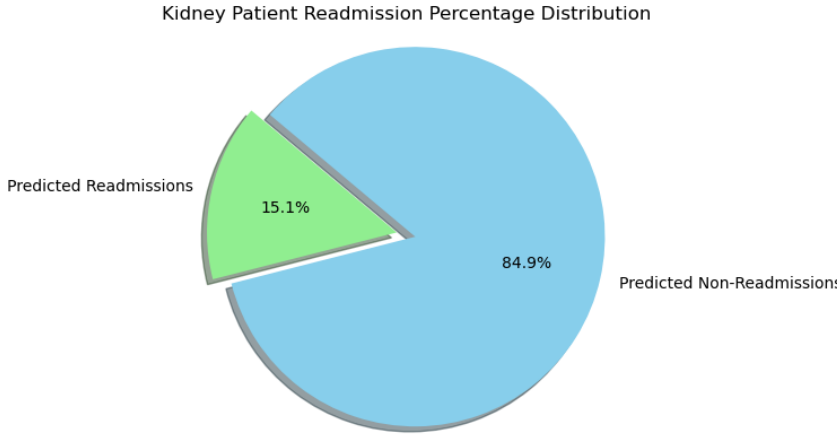
```
Precision score of Logistic Regression model: 0.5229
ROC AUC score of Logistic Regression model: 0.6180
```



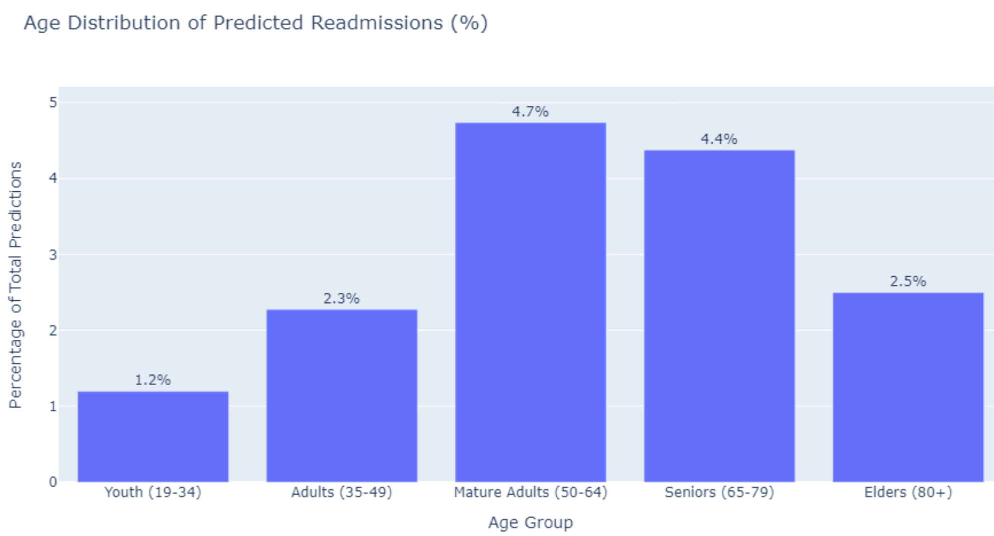
2. **RandomForestClassifier Model :**
    - With this model, we aimed to utilize its strength in handling complex and high dimensional datasets, which is necessary when dealing with diverse healthcare data. In our approach, we engineered features such as age, hospital expire flag, readmits, previous admission counts and length of stay and further augmented dataset with severity scoring based on medical terminology insights. Traget was set as readmit under 30 days. These features were important in capturing multifaceted nature of patient health records and readmission risks.
    - Upon training, our model turned our to notably efficient and effective, exhibiting accuracy of 0.70. This showed model's reliable performance in distinguishing between patients who are likely to be readmitted to the hospital and those who are not. It also demonstrated commendable precision of 0.74 when predicting non-readmissions and moderate precision of 0.47 for readmissions, indicating its strong capabiliy in identifying true negatives over true positives.
    - **Results from Model :**

```
Accuracy score: 0.70
              precision    recall  f1-score   support

           0       0.74      0.89      0.81     21592
           1       0.47      0.24      0.32      8920

    accuracy                           0.70     30512
   macro avg       0.60      0.56      0.56     30512
weighted avg       0.66      0.70      0.66     30512
```

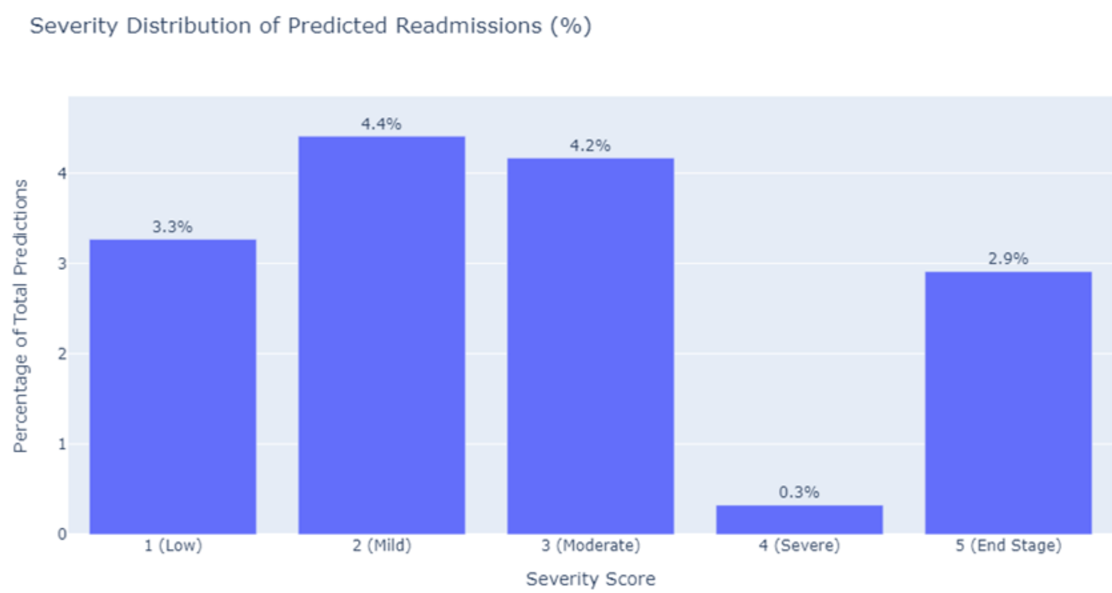
- Model predictions suggests that 15.1% of kidney patients are at risk of readmission, while 84.9% are predicted not to require readmission to hospital:



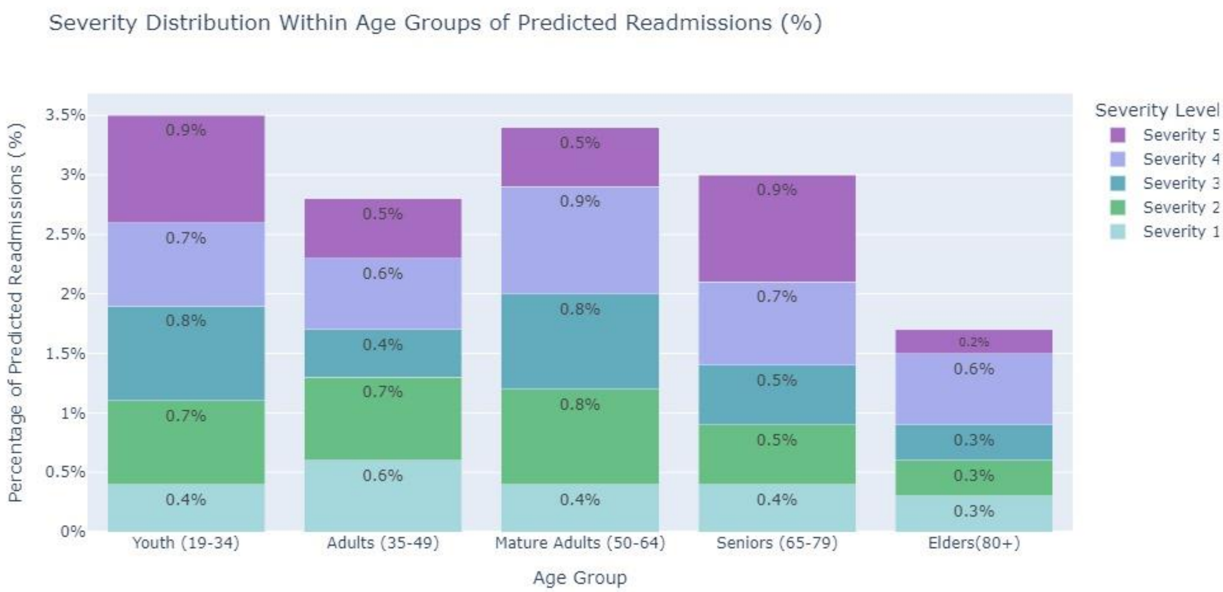
- **Age distribution of 15.1% Predicted Readmissions:** This shows that mature adults and seniors are the most likely to be predicted for readmissions with both groups at about 4.7% and 4.4% respectively. This suggests that older adults have higher risk of readmissions according to our model's predictions.



- **Severity distribution of 15.1% Predicted Readmissions:** This shows that most predicted readmissions fall within mild and moderate severity categories, represented by 4.4% and 4.2% of total predictions respectively. Notably, there is minimal proportion 0.3% in Severe category. It's noticeable that patients with severe conditions are less likely to be readmitted compared to those with milder conditions. This could be due to several reasons. Firstly, patients with severe conditions often receive more intensive and comprehensive care during their initial hospital stay. This level of care might address the immediate health issues more effectively, reducing need for a return visit. Secondly, severe conditions may lead to longer hospital stays initially, giving more time for stabilization and recovery, which could result in lower short term readmission rate. Another factor might be increased post-discharge support and resources typically allocated to patients with severe conditions, including home health care services, closed follow up appointments and better coordinated care transitions. It is also possible that for some with mose severe or end-stage conditions readmissions might not occur due to unfortunate outcome of disease propogation, where patients may move to intensive care setting or not return to hospital.



- **Severity distribution within Age groups of Predicted Readmissions:** Highest percentage of predicted readmissions occur in mature adults and seniors categories, especially for severity level 2 and 3, suggesting that middle-aged to older adults are more likely to be predicted for readmission with mild to moderate conditions.

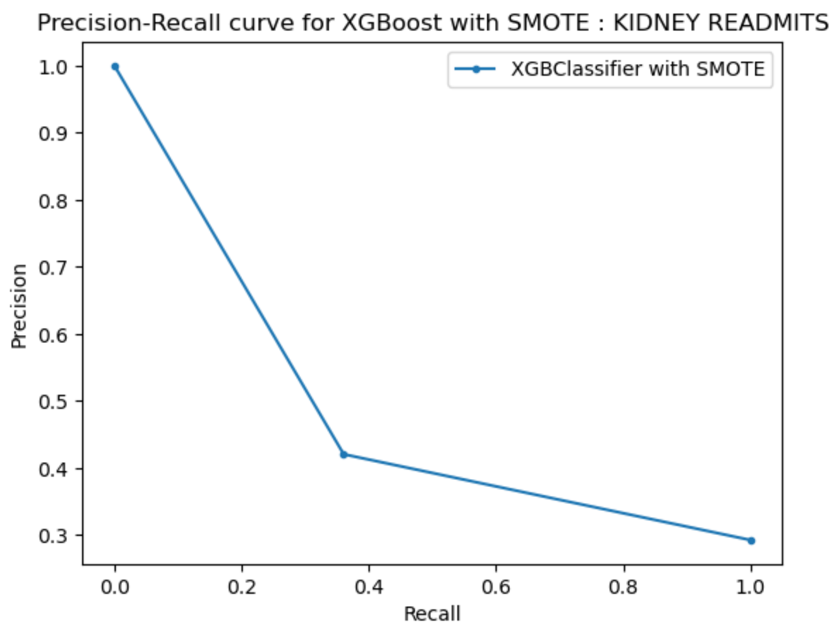


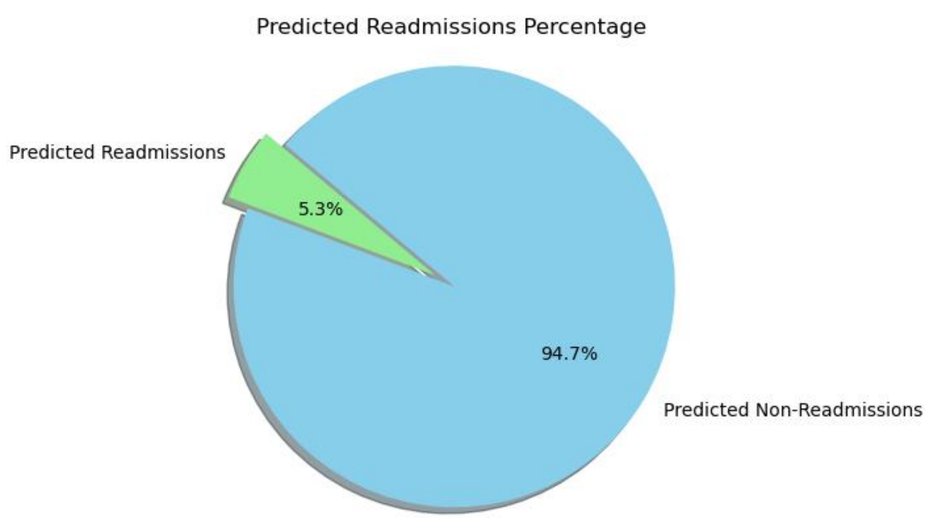
3. **XGBoost Classifier Model :**
   - Next, XGBoost classifier is employed to predict 30 days readmission rates for kidney related hospital stays. XGBoost, which stands for eXtreme Gradient Boosting is a powerful machine learning algorithm that utilizes decision trees and gradient boosting techniques to improve predictive accuracy.
   - The data features used in the model include demographic information, clinical details like severity of the condition, and historical hospitalization data. Categorical variables such as gender and insurance are covered into a format suitable for modelling using one-hot encoding, while numerical variables are standardized to have a mean of 0 under standard deviation of 1.
   - SMOTE, or synthetic minority over sampling technique, was incorporated into the modeling pipeline to address class imbalance - an issue that arises when the number of readmissions (positive class) is much lower than the number of non-readmissions (negative class), which can bias the model. SMOTE helps by creating synthetic samples of the minority class to balance the data set, thereby improving the model’s ability to learn from an otherwise underrepresented class.
   - However, despite these techniques, the XGBoost model with SMOTE did not perform satisfactorily. The pie chart indicates that only 5.3% of cases were predicted as potential re admissions by the precision recall curve that is a chart that evaluates the trade-off between the precision of a predictive model and its recall (true positive rate) showed AUC-PR score of 0.48. These evaluations, combined with the classification report suggested that while the model had some predictive power, it was not precise enough for reliable clinical applications.
   - Given the moderate accuracy score of 0.72 and the imbalanced precision and recall scores it became clear that the model's ability to accurately predict kidney readmissions within 30 days was limited. The precision of only 0.42 for the positive class in the SMOTE adjusted model indicated that it was likely to generate a high number of false positives, which is undesirable in a clinical setting. The AUC-PR score further confirmed this limitation, leading to the decision not to proceed with this model for predicting kidney readmissions. The goal in healthcare analytics is to develop a model that is both highly accurate and reliable in its predictions, to support clinical decisions and resource allocation effectively.

○ Results from Model :

XGBoost Model with SMOTE AUC-PR score: 0.48

	precision	recall	f1-score	support
0	0.75	0.79	0.77	21592
1	0.42	0.36	0.39	8920
accuracy			0.67	30512
macro avg	0.59	0.58	0.58	30512
weighted avg	0.65	0.67	0.66	30512



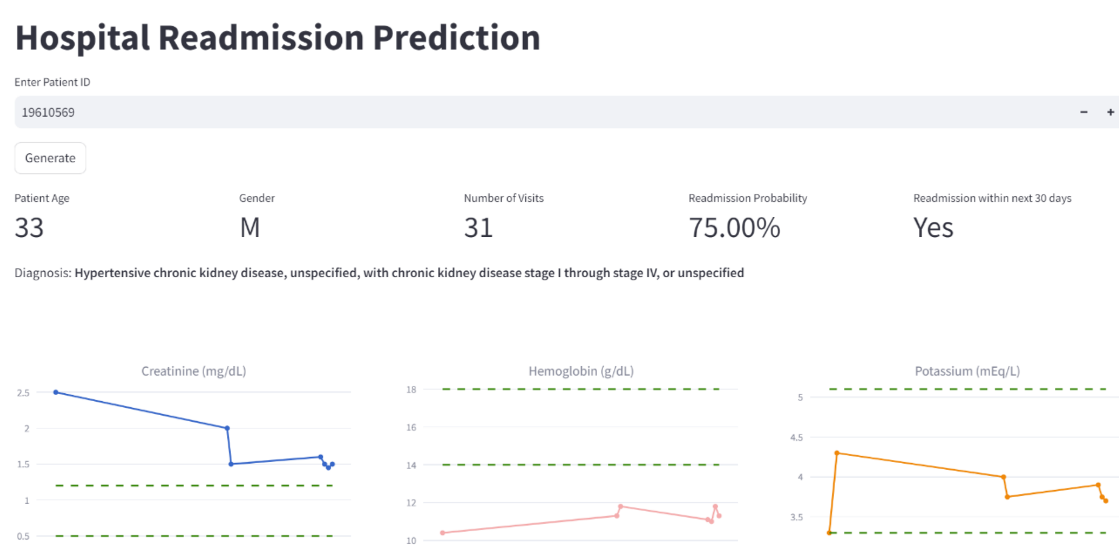


➢ STREAMLIT INTERFACE :

Streamlit interface, is a dashboard tool designed for healthcare professionals to monitor and evaluate the health status of the patients with kidney related conditions. This tool combines the clinical data presentation and predictive analytics. It features:

- Basic patient demographics such as age and gender.
- Diagnosis information.
- A summary of the patient's healthcare interaction, reflected in the number of visits, which could be indicative of the severity of their condition.
- Calculated readmission probability which is based on the patient's historical data.
- A binary prediction outcome (Yes/No) for potential readmissions within next 30 days, providing a direct actionable insight for healthcare providers.
- Vital laboratory results that are important for monitoring kidney function such as Creatinine, Hemoglobin, Potassium, Sodium, pH, and White Blood Cell (WBC) counts are presented in individual graphs. These line graphs illustrate changes over time, allowing for the visual tracking of the patient's condition and the effectiveness of treatments or interventions.

- The interface is a part of a broader application built with streamlit, a Python library that facilitates the creation of web apps for data science. The data for this dashboard is extracted from PostgreSQL database, a powerful open-source relational database system. Given the real time nature off clinical environments, such integration allows for up-to-date patient information to be retrieved and displayed dynamically.
- The use of predictive analytics in the interface, as in the readmission probability, suggests that machine learning models are running behind the scenes to provide these predictions, using the patient's historical data and current health indicators.
- To summarize, this dashboard serves as an instrument for medical professionals to efficiently evaluate patient risk, observe health patterns, and determine the best course of action in terms diagnostics, treatment alternatives, or management strategies.



Hospital Readmission Prediction

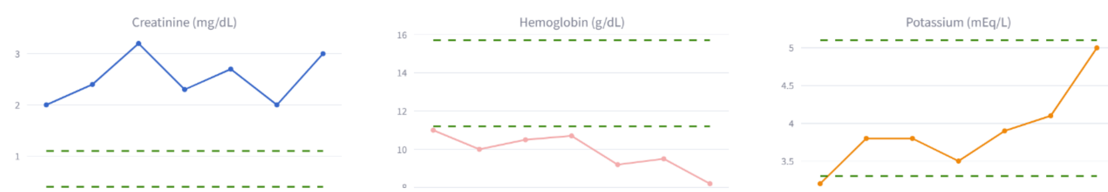
Enter Patient ID

16635089

Generate

Patient Age	Gender	Number of Visits	Readmission Probability	Readmission within next 30 days
50	F	46	28.00%	No

Diagnosis: **Kidney transplant status**



CONCLUSION:

- In conclusion, our project set out to tackle challenge of predicting kidney patient readmissions withing 30-day period, a task of significant importance in the healthcare domain for improving patient care and reducing hospital costs.
- Our exploratory data analysis provided us with valuable insights into patterns of hospital utilization and patient demographics. We performed series of visualizations that uncovered predominance of female patients in admissions, centrality of emergency departments in patient inflow, and importance of seasonality in admission rates.
- Major step of our project involved rigorous data preprocessing, where we not only merged key datasets but also engineered features to encapsulate patients historical interaction with hospital, severity of their conditions, and their demographic details. We utilized natural language processing to evaluate severity of kidney-related diagnoses, augmenting our dataset with detailed clinical insights. Further, feature engineering was important phase where we identified predictors such as age, length of stay, and labeled target outcome for readmissions. Our model preparation entailed standardization and encoding of these features, ensuring uniform contribution to predictive models performance.
- Our final choice, **RandomForestClassifier**, emerged as most suitable model, demonstrating efficient accuracy. It reliably distinguished potential readmissions with ability to predict 15.% of kidney patients as at risk of readmission, while categorizing remaining majority as non-readmissions. Moreover, age and severity distribution of predicted readmissions painted picture of higher risk among mature adults and seniors, primarily for conditions of mild to moderate severity. These findings suggest that while severe conditions are more managed during initial hospital stays, leading to fewer short-term readmissions, patients with milder conditions may benefit from increased post-discharge care and hence mitigate readmission risk.

REFERENCES:

- **Dataset (physionet):** https://physionet.org/content/mimic-iv-demo/2.2/

  https://physionet.org/content/mimic-iv-ed-demo/2.2/

- https://bmcanesthesiol.biomedcentral.com/articles/10.1186/s12871-023-02138-5
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7745694/
- https://www.kidney.org/atoz/content/gfr
- https://www.kidney.org/sites/default/files/kidney-numbers_ckd-heatmap.pdf
- https://www.mayoclinic.org/diseases-conditions/kidney-failure/symptoms-causes/syc-20369048#:~:text=Acute%20kidney%20failure%20%E2%80%94%20also%20called,people%20who%20need%20intensive%20care
- https://www.definitivehc.com/resources/healthcare-insights/average-hospital-readmission-state#:~:text=What%20is%20the%20average%20hospital,to%2022.3%25%20in%20the%20U.S.
- https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-018-2916-1
- https://www.cmaj.ca/content/194/19/E666
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5827573/
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10532623/
- https://towardsdatascience.com/predicting-hospital-readmission-for-patients-with-diabetes-using-scikit-learn-a2e359b15f0
- https://www.kaggle.com/code/iabhishekofficial/prediction-on-hospital-readmission
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9222500/
- https://ieeexplore.ieee.org/document/9368625
- https://www.medrxiv.org/content/10.1101/2021.08.10.21261858v1
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9063206/
- https://physionet.org/content/mimiciv/2.2/
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7137289/
- https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-019-4645-5
- https://www.cihi.ca/en/access-data-and-reports/health-system-performance-measurement
- https://www.cihi.ca/en/indicators/30-day-all-cause-readmission-rate-after-percutaneous-coronary-intervention-pci
- https://www.cihi.ca/en/indicators/all-patients-readmitted-to-hospital