# Applied Data Mining

## Project Report

Empowering Healthcare in Developing Nations:

Machine Learning for Triage Enhancement

## Description of the Problem We Aim to Solve

In our research project, we tackle the critical challenge of improving triage systems within emergency departments (EDs) in resource-limited settings. Triage, the process of determining the priority of patients' treatments based on the severity of their condition, is a cornerstone of effective emergency healthcare. It ensures that limited medical resources are allocated efficiently and that patients receive care in a timely manner based on their medical needs.

The existing triage methods, such as the Korean Triage and Acuity Scale (KTAS), heavily rely on manual assessments by medical personnel, which can vary significantly in accuracy due to human factors and the subjective nature of quick decision-making under pressure. This variability can be particularly problematic in developing nations where healthcare systems often struggle with overcrowding, underfunding, and staff shortages.

Our project seeks to address these challenges by integrating machine learning (ML) techniques into the triage process. By doing so, we aim to develop a more robust, data-driven approach that can assist healthcare providers in making more accurate and faster triage decisions. The ultimate goal is to improve patient outcomes by ensuring that critical cases receive immediate attention while efficiently managing the flow of less urgent cases.

## Our Motivation

Our motivation stems from a profound commitment to enhance healthcare delivery in developing nations, where medical resources are scarce and the burden on healthcare systems is intense. By harnessing the power of machine learning, we believe we can significantly improve the triage process, thus enabling hospitals to better manage patient intake and prioritize care based on objective, data-driven insights.

This project is driven by the potential of machine learning to transform healthcare practices by providing predictive insights that are not easily achievable through traditional methods. Machine learning models can analyze vast amounts of data from past patient interactions and outcomes to identify patterns and predictors of urgency that may not be immediately obvious to human assessors.

In improving triage accuracy, we not only aim to enhance patient outcomes but also to reduce the strain on hospital resources, thereby creating a more sustainable healthcare environment. Better triage methods can lead to decreased waiting times, more appropriate use of medical personnel, and ultimately, a higher standard of care. Through this research, we aspire to contribute to a global effort to empower healthcare systems in developing nations, making them more resilient and responsive to the needs of their communities.

## The Dataset

The data used in this project was obtained from an anonymized dataset released by the authors of the research paper titled ‘Triage accuracy and causes of mistriage using the Korean Triage and Acuity Scale’ which aimed to understand triage effectiveness in emergency departments. The dataset encompasses a wide range of diagnostic variables, patient complaints, and clinical outcomes, systematically selected from 1,267 adult patient records admitted to two emergency departments over a year.

The data includes 24 diverse variables, such as basic patient information (age, gender), vital signs (systolic and diastolic blood pressure, heart rate, respiratory rate, body temperature), chief complaints, mental state, pain description, and more detailed clinical metrics like oxygen saturation and arrival mode. Retrospective features within the dataset also include length of stay, patient outcome, and error grouping, along with triage scores assigned by both nurses and expert assessments (KTAS_nurse and KTAS_expert). Here is a detailed breakdown of the dataset's contents and how each type of data supports our machine learning model development:

### Basic Demographics and Triage Context

- **Group**: Categorical variable indicating patient grouping based on undisclosed criteria, possibly related to triage categorization.
- **Sex**: Binary indicator (1 for male, 2 for female), providing fundamental demographic information useful for pattern recognition in disease prevalence and response to treatment.
- **Age**: Continuous variable providing the age of the patient, crucial for adjusting triage priority as age can significantly influence medical urgency and outcomes.
- **Patients number per hour**: This variable indicates the volume of patients per hour, which helps in understanding and predicting resource needs and patient flow in the ED.

### Arrival and Condition Specifics

- **Arrival mode**: Categorical variable indicating how patients arrived (e.g., walk-in, ambulance), which can be an indicator of the severity of their conditions.
- **Injury**: Binary indicator (1 for no injury, 2 for injury), essential for distinguishing between traumatic and non-traumatic cases.
- **Chief Complaint Translated**: Textual descriptions of the primary complaint in English, critical for initial machine learning text analysis and classification.

### Clinical Measurements

- **Mental State**: Indicator of the patient's mental status, important for assessing the urgency and nature of care required.
- **Pain and NRS Pain**: Both a binary indicator of pain presence and a numerical rating scale (NRS) for pain, providing insights into the severity of discomfort and potential medical conditions.

- **Vital Signs**: Systolic and diastolic blood pressure (**SBP, DBP**), heart rate (**HR**), respiratory rate (**RR**), body temperature (**BT**), and oxygen saturation levels. These are fundamental for assessing patient condition and are often used to calculate triage levels.

**Outcome and Diagnostic Data**

- **Diagnosis in ED**: The diagnosis made in the emergency department, crucial for training our models to correlate symptoms and measurements with medical conditions.
- **KTAS Expert**: The triage score assigned by an expert, using the Korean Triage and Acuity Scale, serves as the ground truth for training our models to accurately predict triage levels.

## The Dataset supports our product idea.

The dataset is instrumental for our machine learning project as it provides a rich basis for training, validating, and testing our models. The diversity and depth of the data allow us to develop and refine algorithms that can predict triage levels with higher accuracy by recognizing patterns and correlations between the variables and the outcomes.

Specifically, the inclusion of both nurse-assigned and expert-assigned KTAS scores offers a unique opportunity to measure and improve the accuracy of machine learning predictions against established human expertise. By comparing these scores, we can train our models to align closely with the expert-level triage decisions, ensuring that the machine learning system learns to replicate and potentially enhance the decision-making process used by seasoned practitioners.

Moreover, the detailed clinical and demographic data supports our product idea by enabling the identification of key predictors of triage urgency. This allows us to focus on the most impactful features, reducing the dimensionality of the problem and improving the speed and efficiency of our algorithms. By harnessing this data, our product aims to deliver a triage tool that is not only accurate but also adaptable to the varying conditions and constraints of healthcare systems in developing nations.

This dataset, therefore, not only supports but is central to the development of our machine learning-based triage enhancement tool, providing a grounded and robust foundation for our efforts to innovate and improve emergency healthcare delivery.

## Exploratory Data Analysis

In our dataset, we identified certain completeness and missing data issues across various columns. The dataset comprises 1267 entries, with complete data in critical fields such as Group, Sex, Age, Arrival mode, Injury, Mental, Pain, Chief Complaint Translated, Diagnosis in ED, and KTAS_expert ratings. However, notable gaps exist in several medical measurement fields, which are crucial for precise triage decision-making. Specifically, the 'NRS_pain' and 'Saturation' columns exhibit significant missing data, with 552 and 694 missing entries respectively. Other vital signs such as Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Heart Rate (HR), Respiration Rate (RR), and Body Temperature (BT) also show smaller but still significant numbers of missing values. Addressing these gaps through appropriate data imputation methods is essential for maintaining the integrity of our machine learning model and ensuring that it is built on a robust and comprehensive foundation of patient data.

### Descriptive Statistics Overview

- **Count**: All columns were assessed for completeness, with the 'KTAS_expert' column and several others showing a full count of 1267 entries, indicating no missing data in these critical fields.

- **Mean and Standard Deviation**: The mean age of patients in the dataset is approximately 54.42 years, with a standard deviation of 19.73, reflecting a moderate spread of ages. This variability is crucial for understanding age-related trends in triage needs.

- **Minimum and Maximum Values**: Age data ranges from a minimum of 16 years to a maximum of 96 years, showcasing the broad age spectrum of patients treated in emergency settings.

- **Percentiles**: The 25th, 50th (median), and 75th percentiles for age are 37, 57, and 71 years respectively, providing a deeper insight into the age distribution and helping identify the core demographic most frequently encountered in EDs.

**Skewness Analysis**

- **Symmetric Distributions**: Columns like 'Group', 'Sex', and 'Arrival mode' displayed skewness close to zero, indicating nearly symmetric distributions and suggesting balanced categorical representations.

- **Age Distribution**: The age column exhibited negative skewness, implying a distribution with a longer tail on the left, suggesting a smaller number of younger patients compared to the elderly.

- **Patient Flow**: The 'Patients number per hour' column showed positive skewness, indicating occasional spikes in patient inflow, which could impact triage operations.

- **Injury and Mental State**: Both 'Injury' and 'Mental' columns demonstrated positive skewness, indicating more frequent occurrences of injury and higher mental state assessments, critical factors in urgent triage classification.

- **Pain Assessments**: Interestingly, the general 'Pain' column had negative skewness, suggesting that lower pain assessments were more common, whereas the 'NRS_pain' column showed positive skewness, pointing to higher pain ratings when specifically measured.

- **Vital Signs**: The skewness for systolic and diastolic blood pressure, heart rate, respiration rate, and body temperature was positive, indicating a tendency towards higher measurements in these vital signs.

- **Oxygen Saturation**: A significant negative skew in the 'Saturation' column highlighted a potential area for further investigation, as lower saturation levels are critical in emergency medicine.

- **Triage Scores**: The 'KTAS_expert' ratings showed negative skewness, indicating a distribution favoring lower expert ratings, which could suggest a conservative approach in triage scoring by experts.

### Kurtosis Analysis

In our dataset analysis, we also examined kurtosis, which measures the "tailedness" of the distribution relative to a normal distribution. The kurtosis metric helps us understand the extremity of data points (outliers) and the peak sharpness, which are critical for identifying trends and anomalies in medical data.

Our analysis revealed that variables like Group, Sex, Age, Patients number per hour, Pain, and KTAS_expert exhibited platykurtic distributions, indicating fewer extreme values than a normal distribution. Conversely, variables such as Arrival mode, Injury, Mental, NRS_pain, Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Heart Rate (HR), Respiration Rate (RR), Body Temperature (BT), and Oxygen Saturation showed leptokurtic distributions, suggesting a higher likelihood of outliers and more pronounced peaks, which are vital for identifying cases requiring immediate attention.

### Correlation Analysis

**Significant correlations found include:**

Positive Correlation:

- Age and Patients number per hour showed a weak positive correlation (0.211540), suggesting that as age increases, the patient number per hour slightly increases.
- Injury and Mental displayed a very weak positive correlation (0.000868), indicating almost no linear relationship.

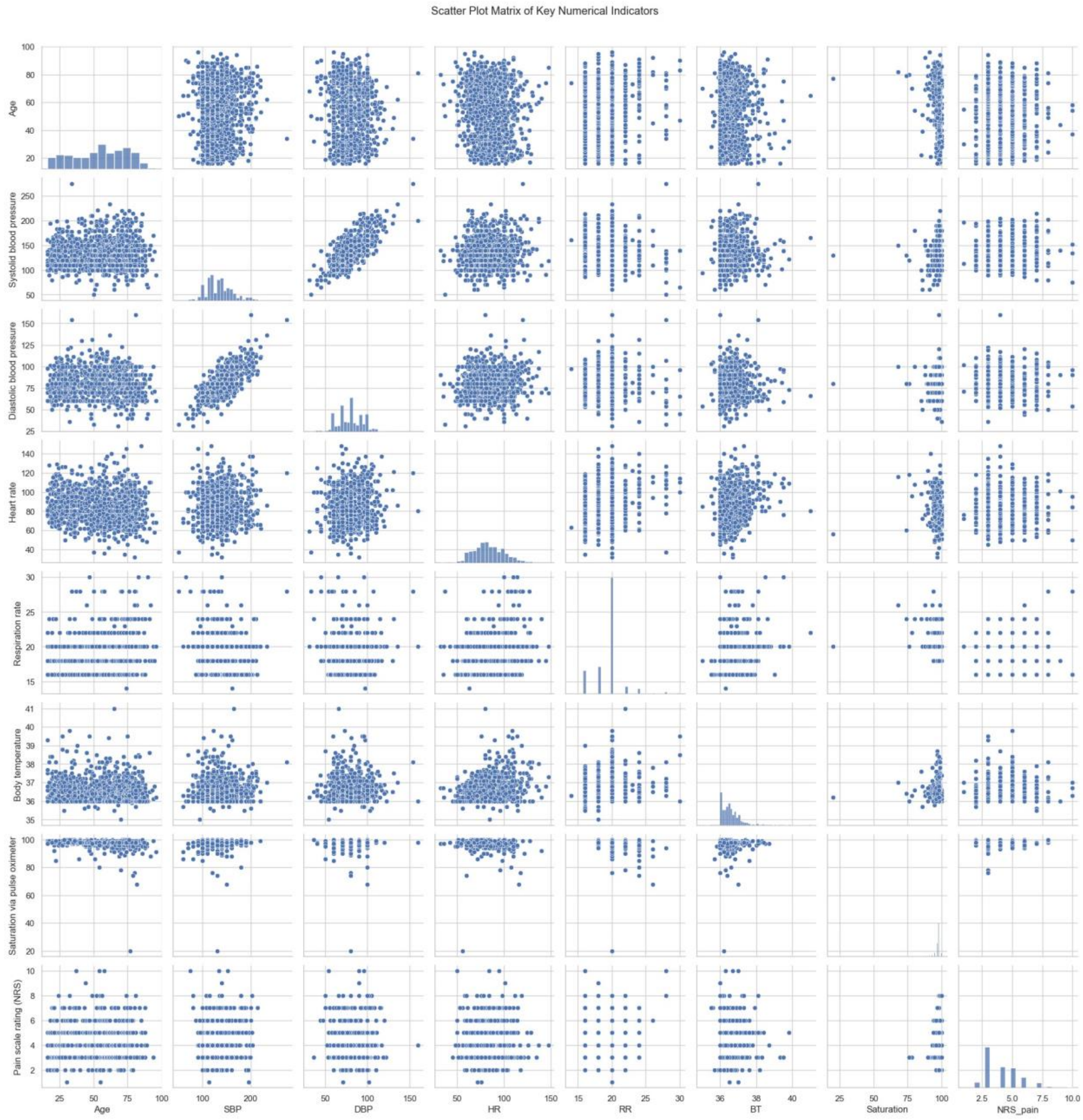
Negative Correlation:

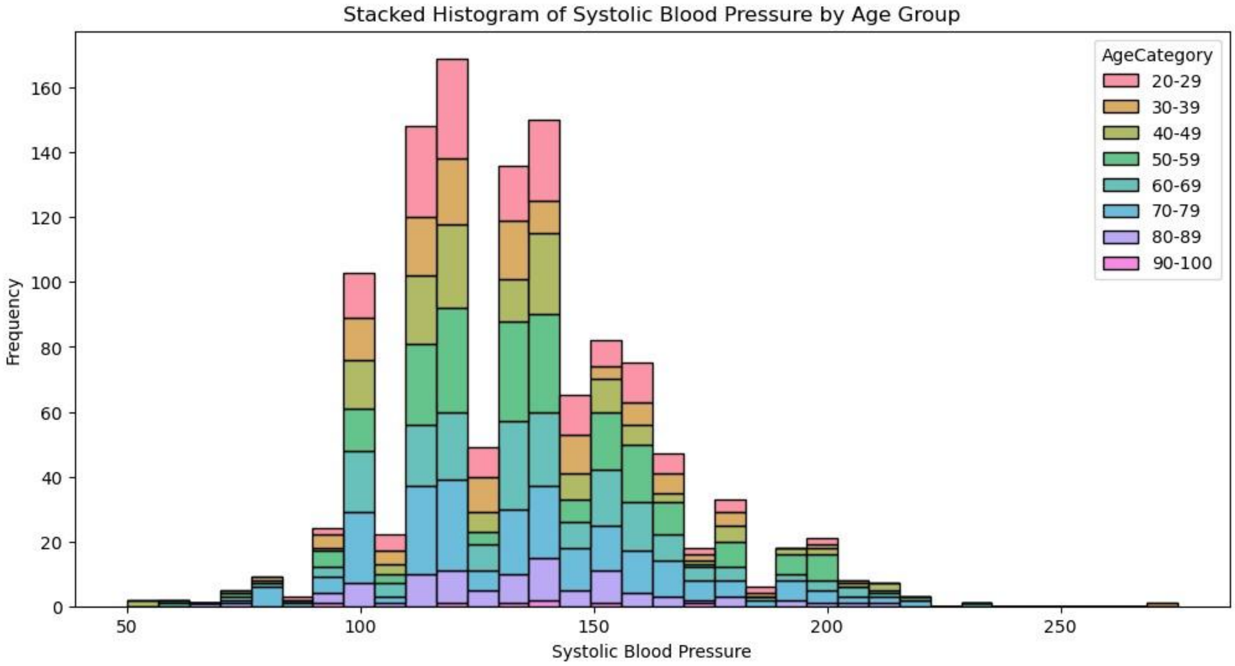
- Pain and NRS_pain had a moderate negative correlation (-0.571965), unusual given both are measures of pain intensity but may reflect different reporting methods or scales.
- Pain and DBP exhibited a very weak negative correlation (0.014824), suggesting virtually no relationship.

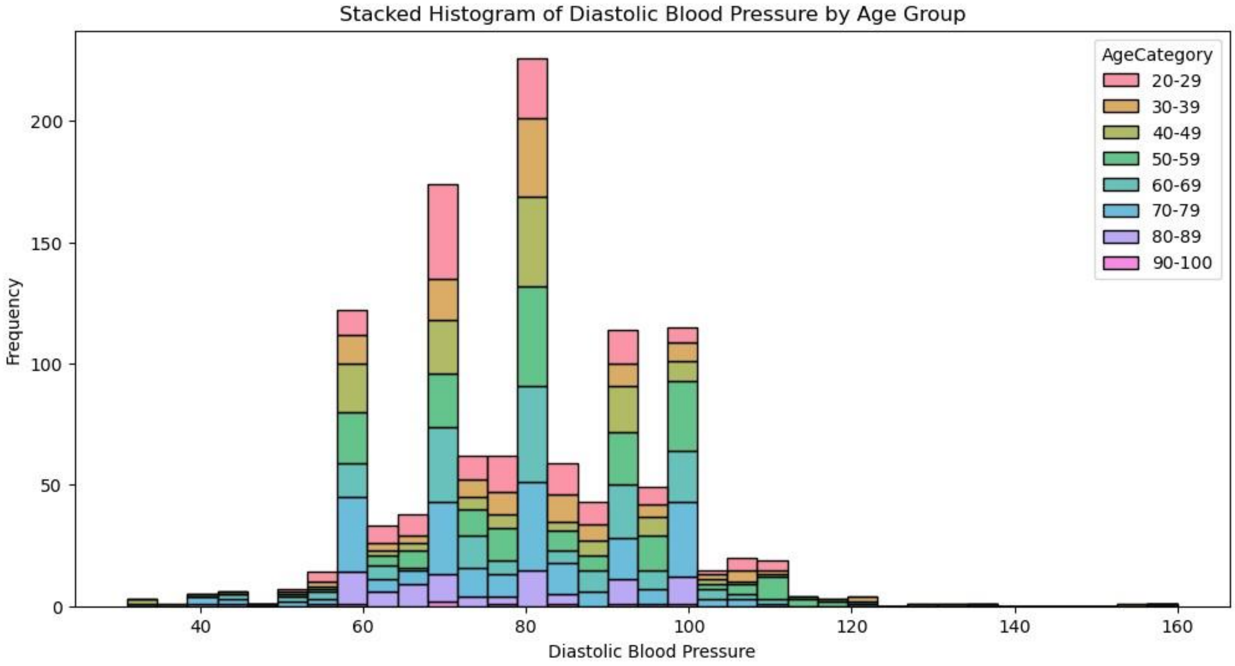
Strength of Correlation:

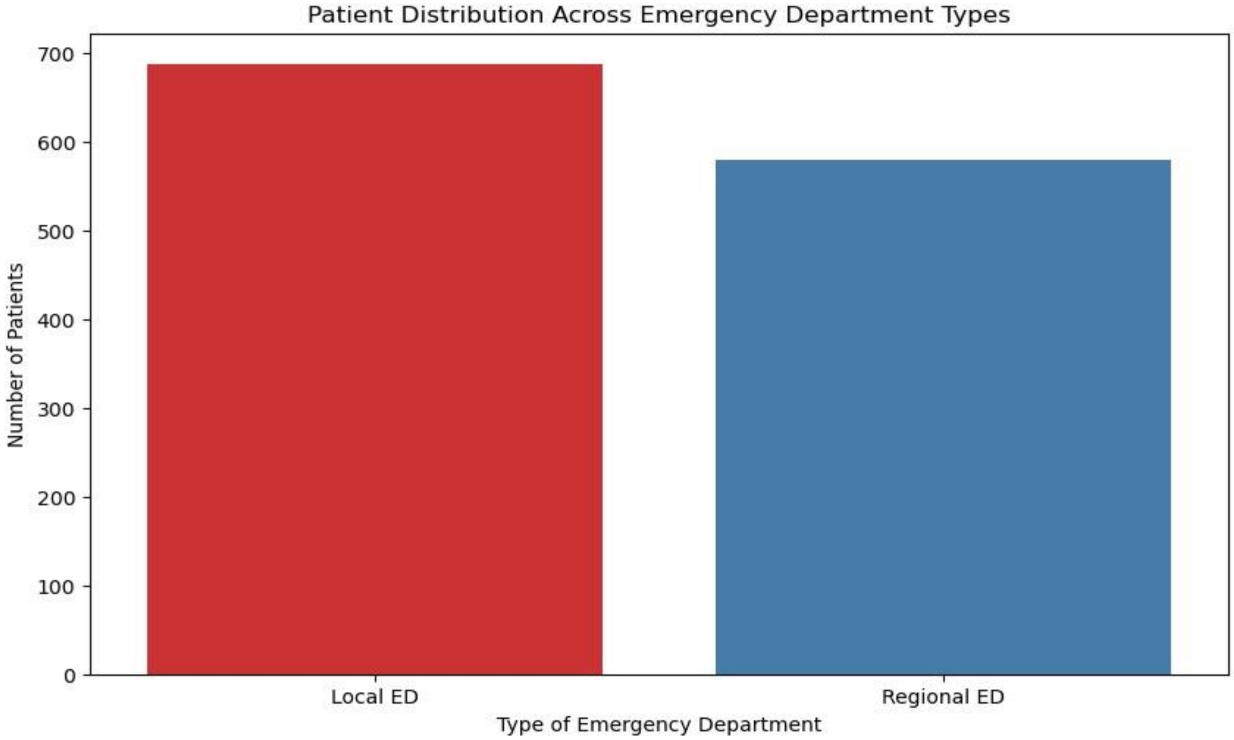
- KTAS_expert and Mental showed a moderate negative correlation (-0.349787), important for understanding how mental assessment impacts expert triage decisions.
- Saturation and BT had a very weak positive correlation (0.004443), indicating negligible linear association.

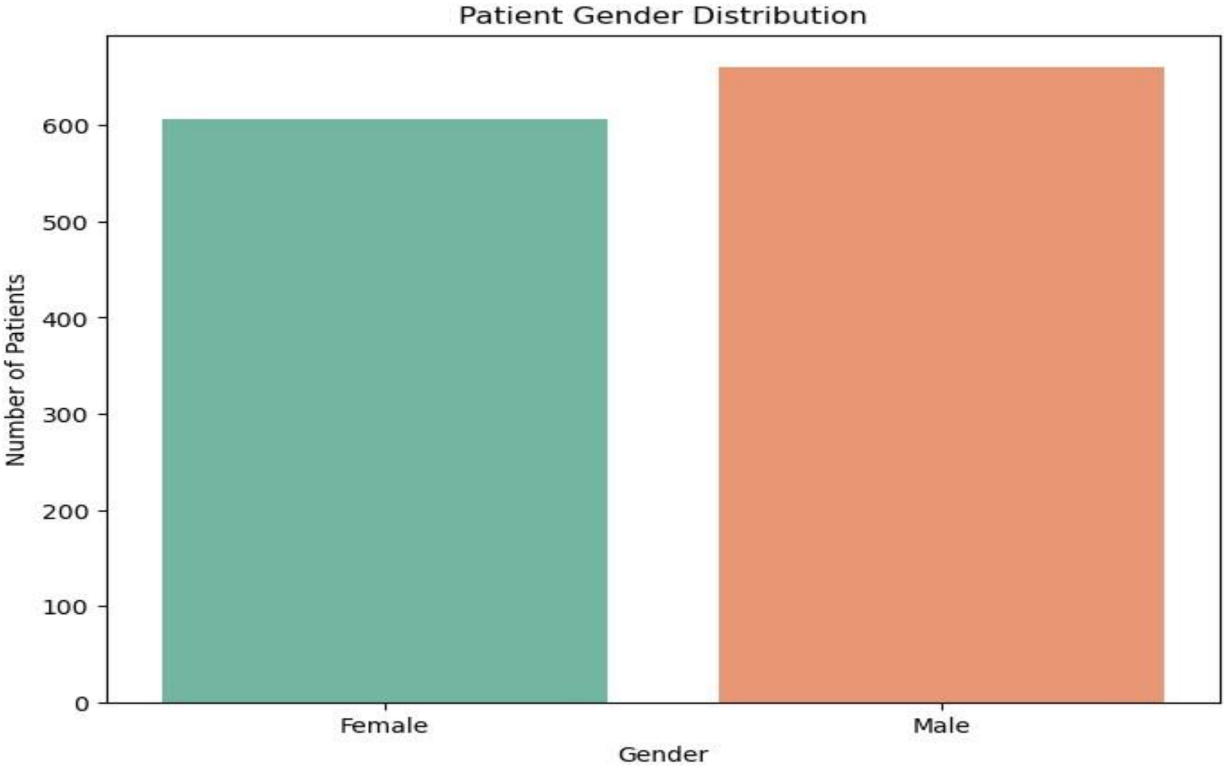
## Dataset Visualizations

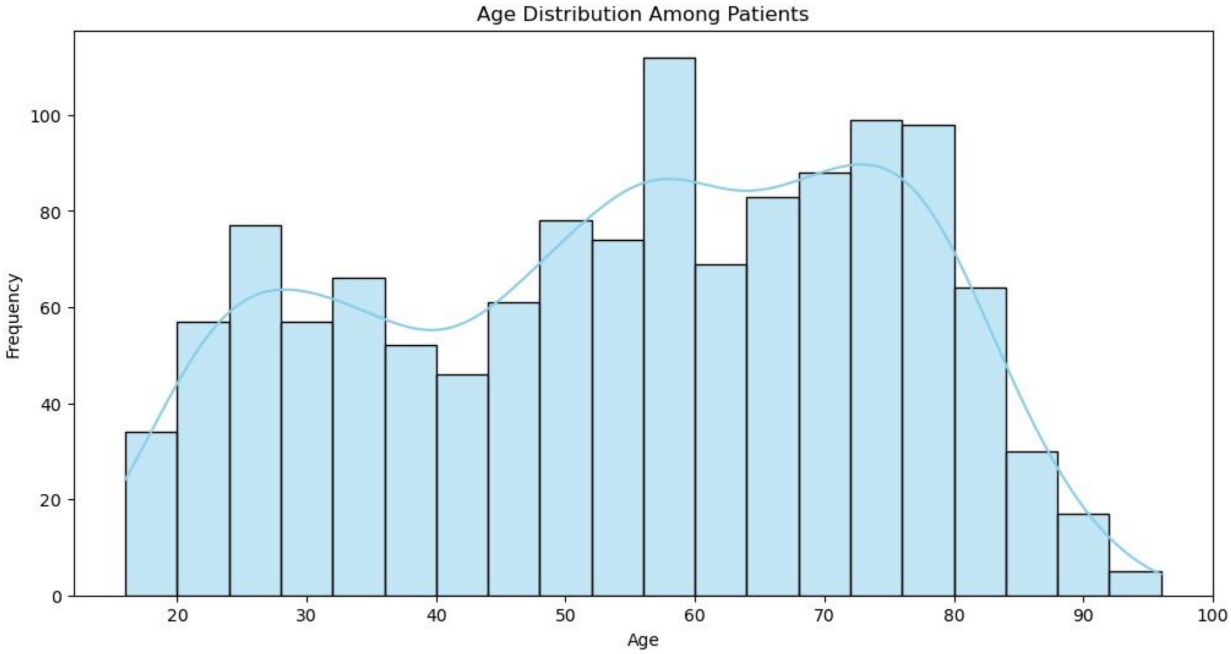


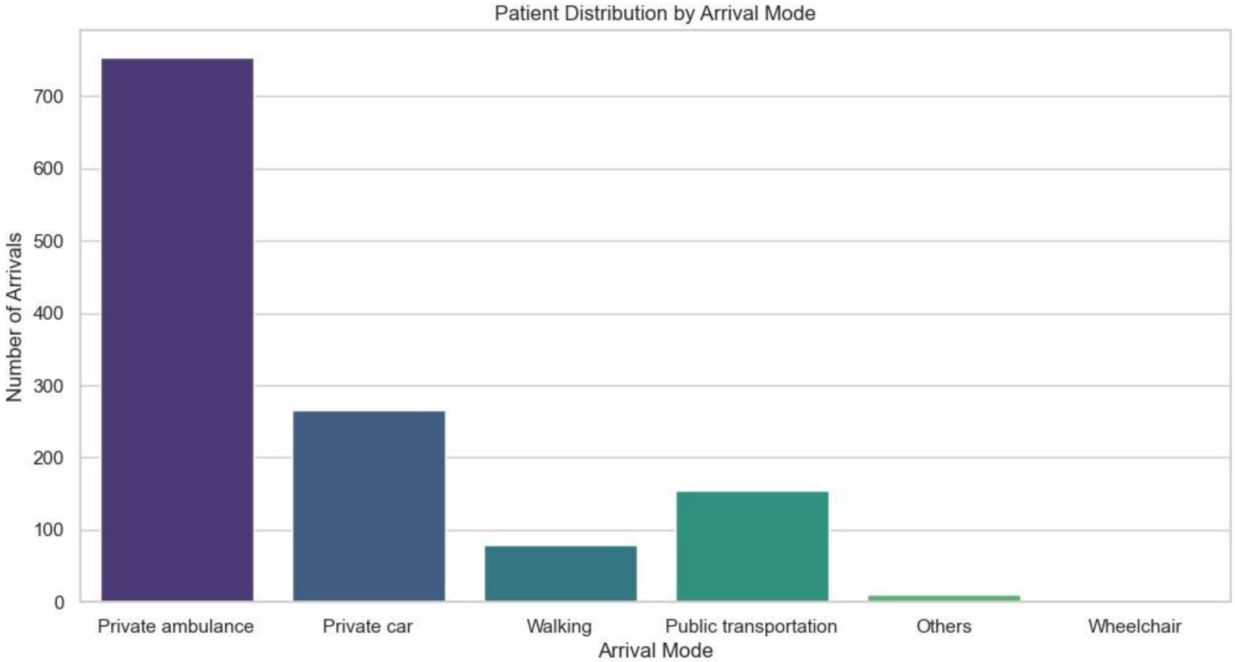

Stacked Histogram of Systolic Blood Pressure by Age Group
Frequency
0
20
40
60
80
100
120
140
160
50
100
150
200
250
Systolic Blood Pressure
AgeCategory
20-29
30-39
40-49
50-59
60-69
70-79
80-89
90-100

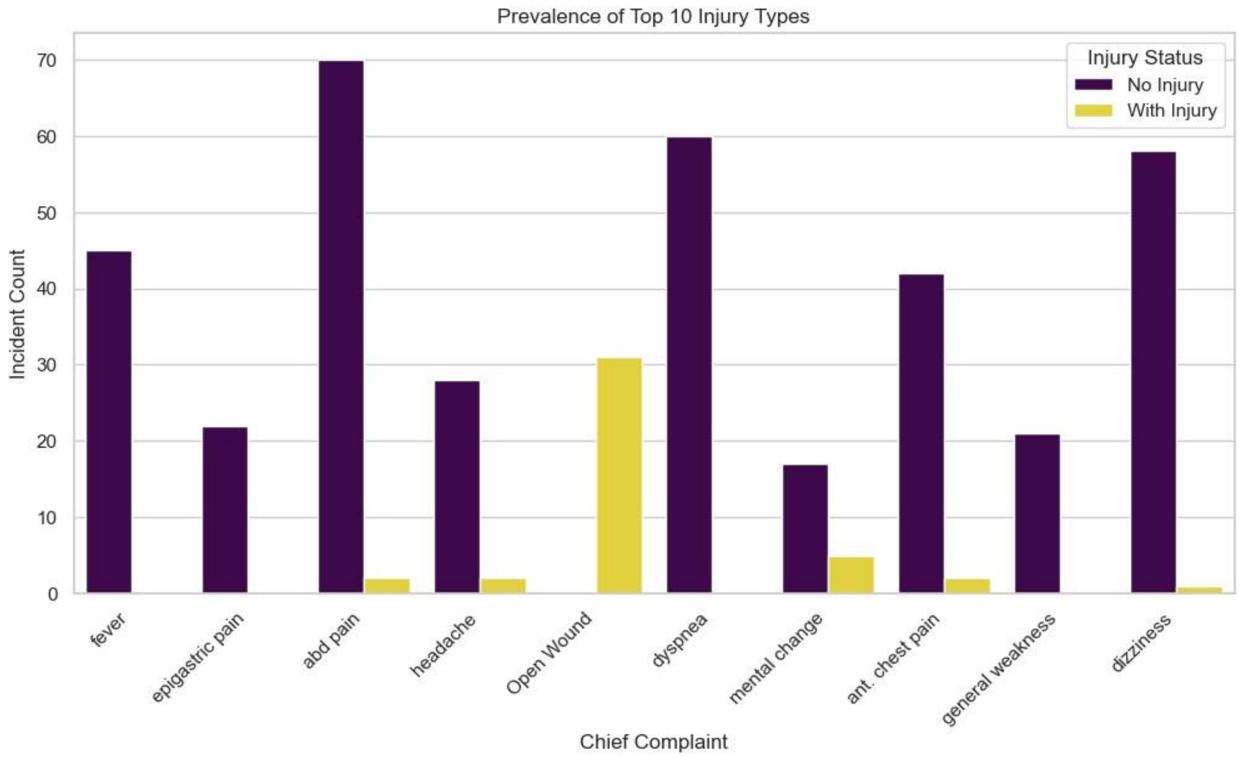

Stacked Histogram of Diastolic Blood Pressure by Age Group
Frequency
0
50
100
150
200
40
60
80
100
120
140
160
Diastolic Blood Pressure
AgeCategory
20-29
30-39
40-49
50-59
60-69
70-79
80-89
90-100

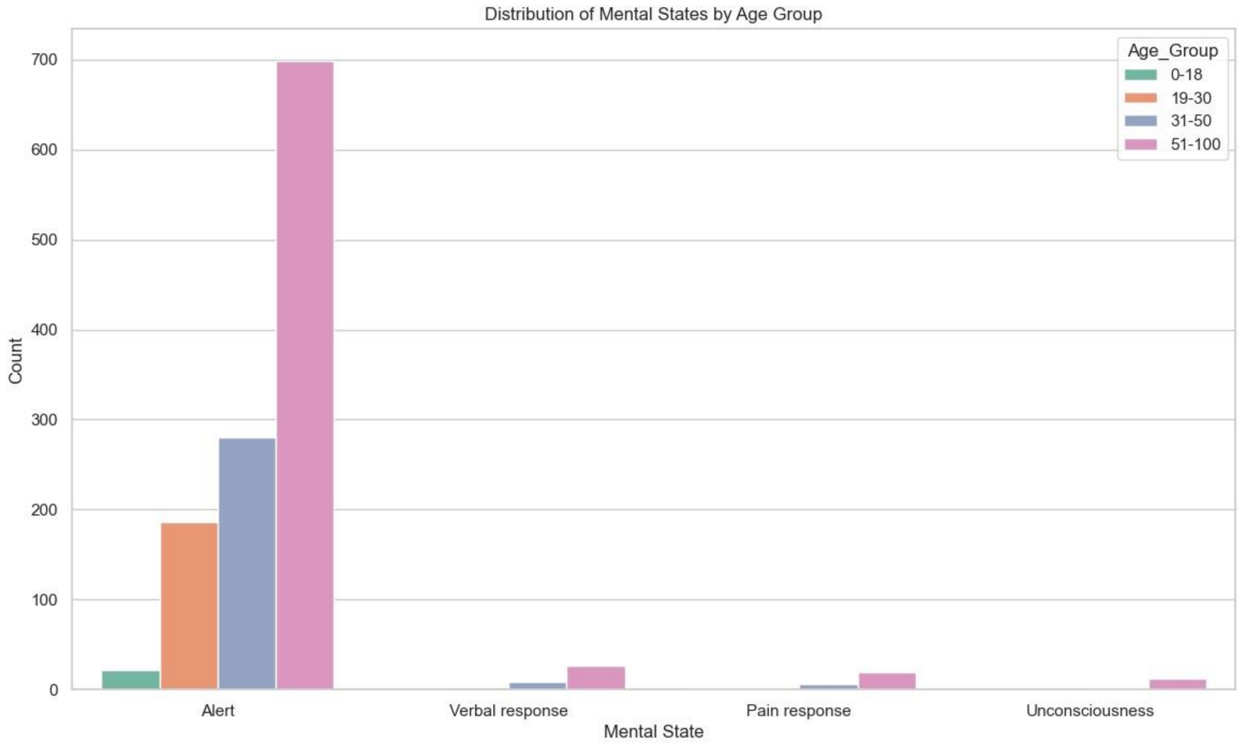

Patient Distribution Across Emergency Department Types
700
600
500
400
300
200
100
0
Number of Patients
Local ED
Regional ED
Type of Emergency Department

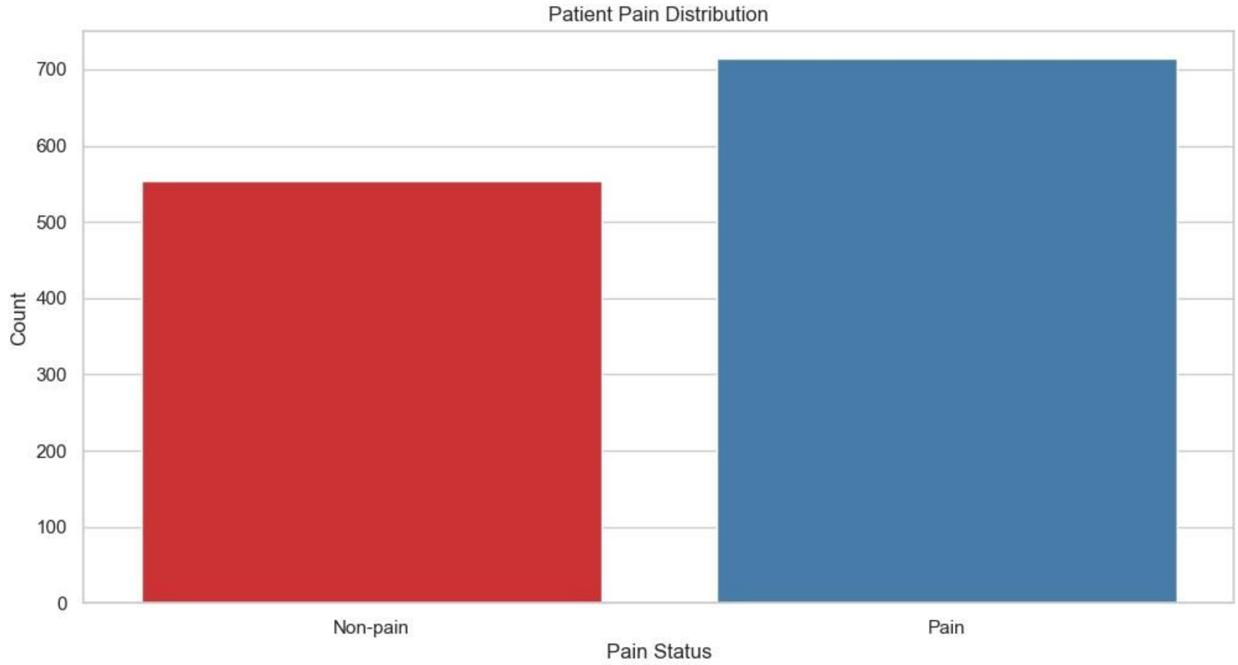

Patient Gender Distribution
600
500
400
300
200
100
0
Number of Patients
Female
Male
Gender

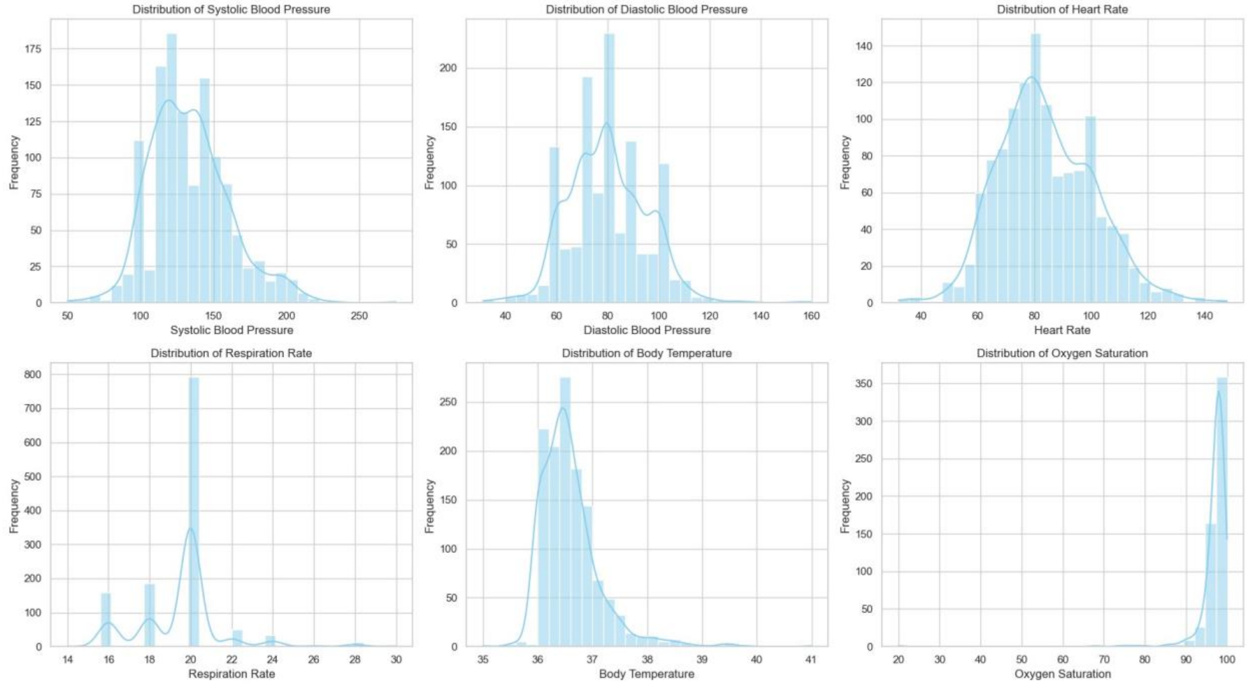

Age Distribution Among Patients
Frequency
0
20
40
60
80
100
20
30
40
50
60
70
80
90
100
Age

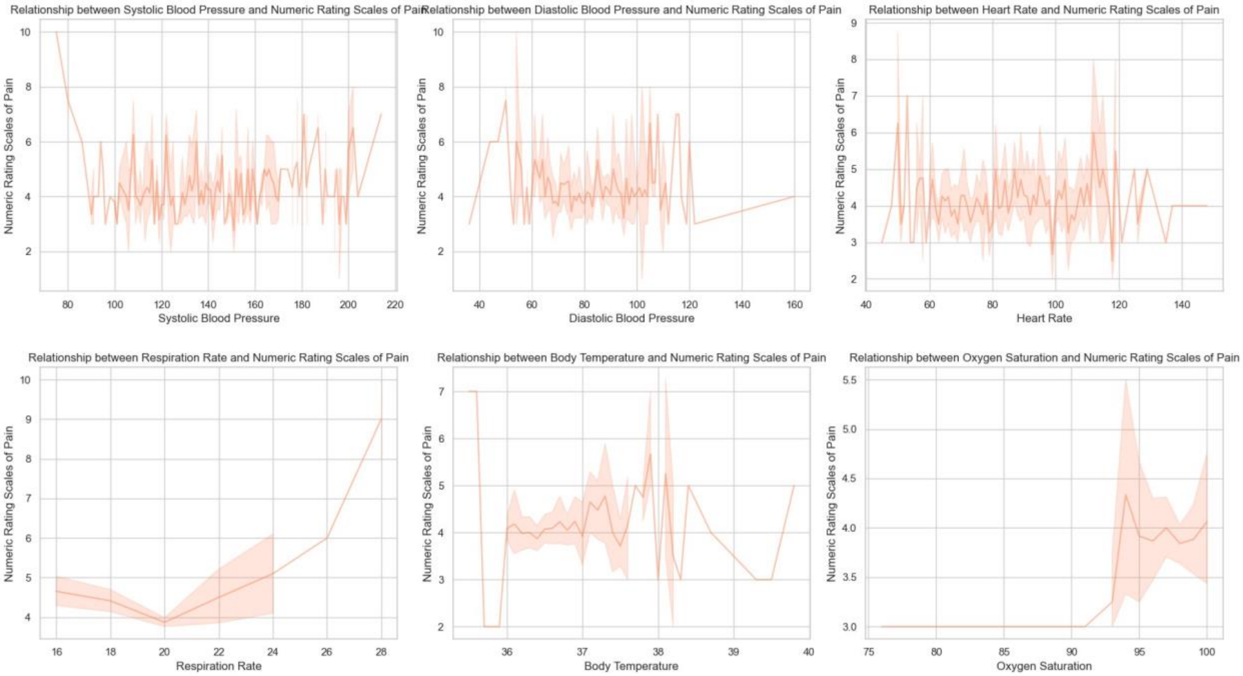

Patient Distribution by Arrival Mode
Number of Arrivals
0
100
200
300
400
500
600
700
Private ambulance
Private car
Walking
Public transportation
Others
Wheelchair
Arrival Mode

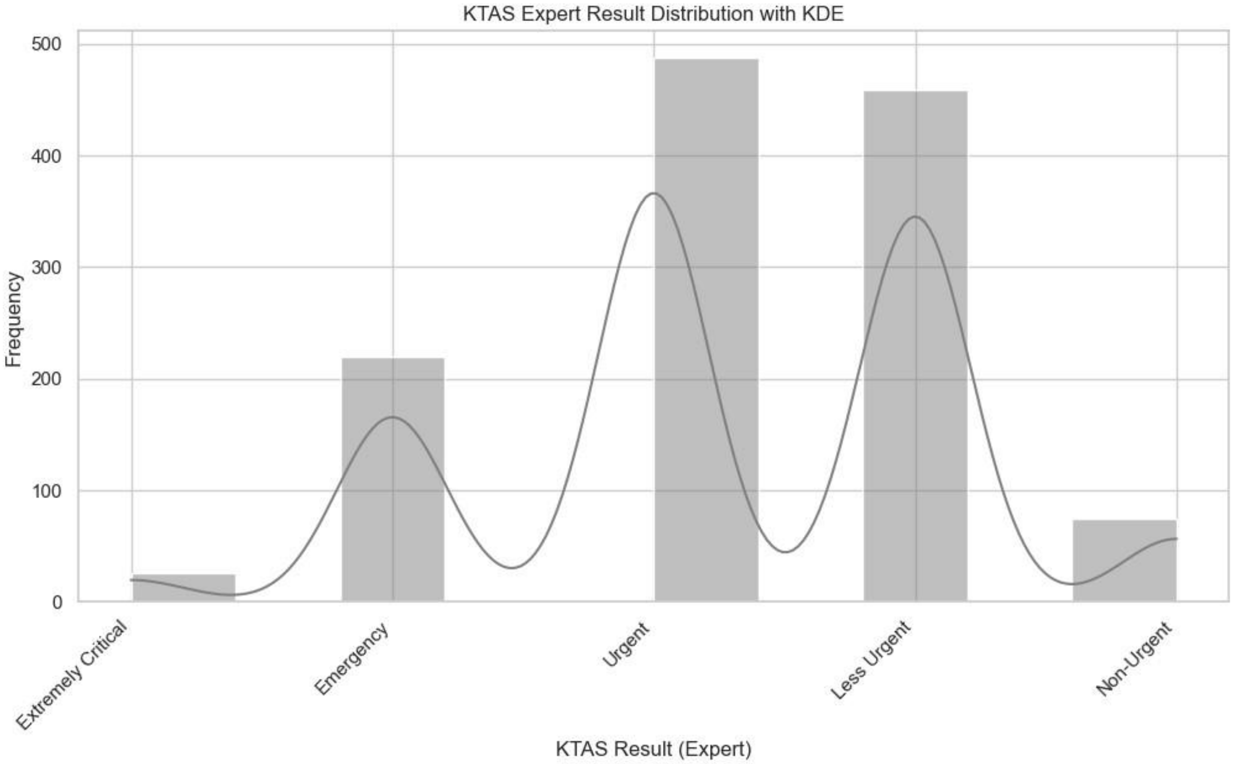

Prevalence of Top 10 Injury Types
Injury Status
No Injury
With Injury
Incident Count
0
10
20
30
40
50
60
70
fever
epigastric pain
abd pain
headache
Open Wound
dyspnea
mental change
ant. chest pain
general weakness
dizziness
Chief Complaint


Distribution of Mental States by Age Group
Age_Group
0-18
19-30
31-50
51-100
Count
0
100
200
300
400
500
600
700
Alert
Verbal response
Pain response
Unconsciousness
Mental State


Patient Pain Distribution
Count
0
100
200
300
400
500
600
700
Non-pain
Pain
Pain Status


Distribution of Systolic Blood Pressure
Frequency
0
25
50
75
100
125
150
175
50
100
150
200
250
Systolic Blood Pressure
Distribution of Diastolic Blood Pressure
Frequency
0
50
100
150
200
40
60
80
100
120
140
160
Diastolic Blood Pressure
Distribution of Heart Rate
Frequency
0
20
40
60
80
100
120
140
40
60
80
100
120
140
Heart Rate
Distribution of Respiration Rate
Frequency
0
100
200
300
400
500
600
700
800
14
16
18
20
22
24
26
28
30
Respiration Rate
Distribution of Body Temperature
Frequency
0
50
100
150
200
250
35
36
37
38
39
40
41
Body Temperature
Distribution of Oxygen Saturation
Frequency
0
50
100
150
200
250
300
350
20
30
40
50
60
70
80
90
100
Oxygen Saturation


Relationship between Systolic Blood Pressure and Numeric Rating Scales of Pain
Relationship between Diastolic Blood Pressure and Numeric Rating Scales of Pain
Relationship between Heart Rate and Numeric Rating Scales of Pain
Relationship between Respiration Rate and Numeric Rating Scales of Pain
Relationship between Body Temperature and Numeric Rating Scales of Pain
Relationship between Oxygen Saturation and Numeric Rating Scales of Pain
Numeric Rating Scales of Pain
Systolic Blood Pressure
80 100 120 140 160 180 200 220
Diastolic Blood Pressure
40 60 80 100 120 140 160
Heart Rate
40 60 80 100 120 140
Respiration Rate
16 18 20 22 24 26 28
Body Temperature
36 37 38 39 40
Oxygen Saturation
75 80 85 90 95 100
2 4 6 8 10
2 3 4 5 6 7 8 9
4 5 6 7 8 9 10
2 3 4 5 6 7
3.0 3.5 4.0 4.5 5.0 5.5


KTAS Expert Result Distribution with KDE
Frequency
0
100
200
300
400
500
Extremely Critical
Emergency
Urgent
Less Urgent
Non-Urgent
KTAS Result (Expert)

## Data Preprocessing

In the project, comprehensive data preprocessing was critical to ensuring the accuracy and effectiveness of our machine learning models.preprocessing steps implemented to prepare our dataset are:

**Translation and Initial Data Cleaning:**

- We initiated the preprocessing by translating entries from Korean to English using the Google Translate function in Google Sheets, ensuring all textual data was uniformly understandable.
- Our dataset contained relatively few missing values across most variables, allowing for straightforward handling of these instances. We completely excluded the first column as it did not contain relevant information.

**Handling Sparse Data:**

- Certain 'Arrival mode' categories (specifically modes 5 and 7) were sparsely populated and thus removed to maintain data consistency and relevance.

**Missing Data Imputation:**

- Textual data in 'Diagnosis in ED' with missing entries were filled with empty strings to prepare for vectorization.
- Numerical missing values were imputed differently based on their nature: 'NRS_pain' was filled with zeros, considering the clinical relevance of having no reported pain as a baseline; other vital signs like 'SBP', 'DBP', 'HR', 'RR', 'BT', and 'Saturation' had their missing values replaced with the mean of their respective columns to maintain statistical integrity.

**Normalization and Encoding:**

- We applied a StandardScaler to normalize age and patient numbers per hour, along with other numerical data that required zero or mean imputation. This normalization helps in mitigating the influence of outlier values and scaling the data to a uniform range, which is essential for many machine learning algorithms.

- Categorical data such as 'Group', 'Sex', 'Arrival mode', 'Injury', and 'Pain' were encoded using OneHotEncoder. This step converts categorical variables into a form that could be provided to ML algorithms to do a better job in prediction.

**Text Data Vectorization and Dimension Reduction:**

- The chief complaints and diagnoses were transformed using a CountVectorizer, which converts text data into a matrix of token counts. This step is pivotal for analyzing text data by turning the qualitative data into quantitative forms that can be evaluated by our models.
- We then applied Truncated Singular Value Decomposition (TruncatedSVD) to these vectorized texts to reduce their dimensionality, thereby enhancing computational efficiency and focusing on the most informative aspects of the text data.

**Integration into a Unified Preprocessing Pipeline:**

- All preprocessing steps were integrated into a single ColumnTransformer pipeline. This pipeline structured the various transformations for different types of data (e.g., imputation for missing values, scaling for numerical data, encoding for categorical data, and vectorization for text data), ensuring that each type of data was optimally prepared for subsequent analysis.

## Methodology

In our project, we adopted a methodological approach centered around regression analysis to predict KTAS scores, which are integral for efficient triage. The KTAS scores range from 1 to 5, with each integer representing a level of urgency. Initially, we considered classification models due to the discrete nature of these scores. However, the preliminary results were not satisfactory as they failed to capture the ordinal relationship between the scores effectively.

Given that KTAS scores approximate the severity of a patient's condition, we opted for regression models to better capture the continuous relationship and nuances between different severity levels. We then mapped the continuous predictions back to the closest discrete KTAS scores to align with the practical application needs of triage systems.

The dataset was split into training and testing sets in a 7:3 ratio. We applied 20 machine learning algorithms to the training set, each chosen for its potential to handle different aspects of the regression task:

1. **Linear Regression**: A basic approach that models the relationship between a dependent variable and one or more independent variables using a linear equation. This model is straightforward but often effective in predicting outcomes where a linear relationship is assumed.

2. **Ridge Regression**: An extension of linear regression that includes regularization. It introduces a penalty term to the loss function, which helps in reducing model complexity and preventing overfitting, particularly useful when dealing with multicollinearity.

3. **Lasso Regression**: Similar to ridge regression, lasso also adds a regularization term to the loss function, but it uses the L1 norm, which can lead to zero coefficients for less important variables. This results in feature selection within the model, making it excellent for models with high dimensionality.

4. **Stochastic Gradient Descent (SGD) Classifier**: While primarily used for classification, SGD can be adapted for regression. It minimizes a chosen loss function by iteratively updating the model parameters, making it suitable for large-scale and sparse data.

5. **Support Vector Regression (SVR)**: SVR uses the same principles as SVM for classification but for regression. It attempts to fit the error within a certain threshold and is robust to outliers, making it powerful for complex regression problems with a clear margin of separation.

6. **K-Nearest Neighbors Regressor (KNN)**: A non-parametric method that predicts the dependent variable value based on the 'k' closest training examples in the feature space. It is intuitive and effective, particularly when the assumption of similarity based on closeness holds true.

7. **Partial Least Squares Regression (PLS)**: PLS projects both the predictors and the responses to a new space and performs linear regression in that space. It is particularly effective when the predictors are highly collinear and when the dimensionality is high compared to the number of observations.

8. **Bayesian Ridge Regression**: It extends ridge regression by treating its parameters as random variables and following Bayesian approaches. It estimates a probability distribution rather than a point estimate, providing a measure of uncertainty in the predictions.

9. **Decision Tree Regressor**: A non-linear model that splits the data into smaller subsets while at the same time an associated decision tree is incrementally developed. The final model is a tree with decision nodes and leaf nodes, which is straightforward to interpret.

10. **Random Forest Regressor**: An ensemble method that builds multiple decision trees and merges them together to get a more accurate and stable prediction. It combines the simplicity of decision trees with flexibility, reducing the risk of overfitting.

11. **AdaBoost Regressor**: AdaBoost (Adaptive Boosting) works by fitting a sequence of weak learners (typically simple models) on repeatedly modified versions of the data. It adapts by focusing more on the instances that were previously mispredicted, aiming to improve the accuracy by concentrating on the harder cases. This method often leads to improved performance by combining the strengths of multiple weak models to form a strong regressor.

12. **Gradient Boosting Regressor**: This algorithm builds models sequentially, with each new model being trained to correct the errors made by the previous ones. It uses decision trees as the base learners and fits new models to provide a more accurate estimate of the response variable. Gradient boosting can optimize on different loss functions and provides several hyperparameter tuning options that can make the model robust to overfitting.

13. **Elastic Net Regression**: A regularization regression method that combines both L1 and L2 penalties of the lasso and ridge methods. Elastic Net is useful when there are multiple features correlated with each other. It tends to select groups of correlated variables and shrinks the coefficients of less contributive variables, thus improving the model's performance and interpretability.

14. **Quantile Regression**: Unlike ordinary least squares (OLS) regression that minimizes the mean squared error, quantile regression aims at estimating either the median or other quantiles of the response variable. This approach is particularly useful for datasets with heterogeneous variance or outliers, as it provides a more robust analysis of the relationship between variables.

15. **Bagging Regressor**: Short for Bootstrap Aggregating, it involves fitting multiple models (usually of the same type) on different subsets of the original dataset, and then averaging the predictions to improve accuracy and control over-fitting. This technique reduces variance and helps to avoid overfitting.

16. **Stacking Regressor**: Stacking involves training a new model to combine the predictions of several base models. In this method, the base models are trained based on the complete dataset, and then a meta-model is trained on the outputs of the base models as features. This type of ensemble learning can lead to better predictive performance compared to any single model due to its ability to distill the best qualities of each base model.

17. **Voting Regressor**: A type of ensemble machine learning model that combines the predictions from multiple other models. It uses majority voting (for classification problems) or averaging (for regression problems) to improve the final prediction accuracy. This method benefits from the diversity of the input models and can yield more robust predictions.

18. **Multi-layer Perceptron Regressor (MLP)**: MLP is a type of neural network that consists of at least three layers of nodes: an input layer, a hidden layer, and an output layer. MLP utilizes a backpropagation technique for training the network, suitable for complex non-linear modeling which can capture intricate patterns in the data.

19. **XGBoost Regressor**: An implementation of gradient boosted decision trees designed for speed and performance. XGBoost provides a robust solution to regression problems, handling sparse data and making efficient use of hardware resources, which is particularly useful for large datasets.

20. **CatBoost Regressor**: Developed by Yandex, CatBoost (Categorical Boosting) is an algorithm that efficiently handles categorical variables without the need for extensive pre-processing to convert them into numerical values. Its robust handling of categorical data and its advanced approach to reducing overfitting make it highly effective, especially when categorical and high-cardinality features are involved.

We employed a rigorous and systematic approach to train and evaluate various machine learning models. This process was designed to accurately predict the Korean Triage and Acuity Scale (KTAS) scores, which are integral to determining the urgency of each case in a triage scenario.

**Model Training and Validation:**

We utilized a diverse array of algorithms, and our implementation involved several key steps to ensure that each model performed optimally and adhered to the practical requirements of triage systems. The training process for each model followed these steps:

**Defining Custom Metrics:**

- We developed a triage_accuracy metric to measure how well our models' predictions align with actual KTAS scores within acceptable bounds (±1.6 for higher and ±1.2 for lower deviations).

- Our classify_predictions function mapped continuous predictions back to discrete KTAS classes, ensuring that our model outputs were directly applicable in a clinical setting.

**Model Evaluation:**

- Using GridSearchCV, we optimized model parameters for best performance based on the triage_accuracy metric, ensuring that each model was tuned to the nuances of our dataset.
- We calculated detailed precision, recall, and F1-score metrics for each model, allowing us to understand their performance across different KTAS levels.

**Pipeline Integration:**

- Each model was integrated into a Pipeline that included preprocessing steps such as data normalization, categorical encoding, and text data vectorization, ensuring that all input data was properly formatted and optimized for each algorithm.
- The use of ColumnTransformer allowed us to apply specific transformations to different types of data columns efficiently. **Model-Specific Implementations and Tuning:**
- For regression models like Linear Regression, Ridge, and Lasso, we utilized specific hyperparameters like alpha values to control model complexity and prevent overfitting.
- Advanced ensemble methods like AdaBoost, Gradient Boosting, and Random Forest were tuned for parameters such as the number of estimators and tree depth to enhance model accuracy and stability.
- Neural network-based models like MLP were configured with different hidden layer sizes to capture complex patterns in the data.
- For models dealing with categorical data robustly, such as CatBoost, we leveraged their ability to handle categorical features directly, optimizing iterations and depth for performance.

**Innovative Approaches:**

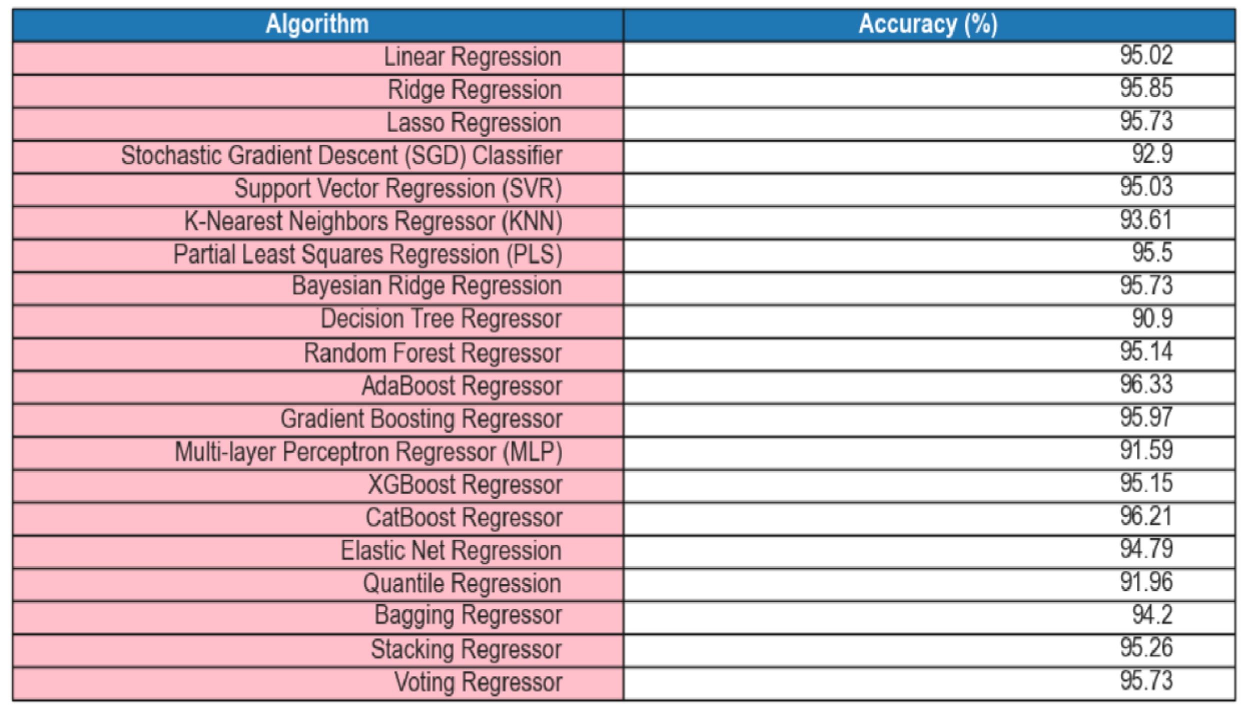
- We employed techniques like Elastic Net, which combines L1 and L2 regularization, and Quantile Regression for more robust and diverse analysis.
- Stacking and Voting Regressors were used to combine predictions from multiple models, harnessing their collective power to improve accuracy.

**Performance Analysis and Reporting:**

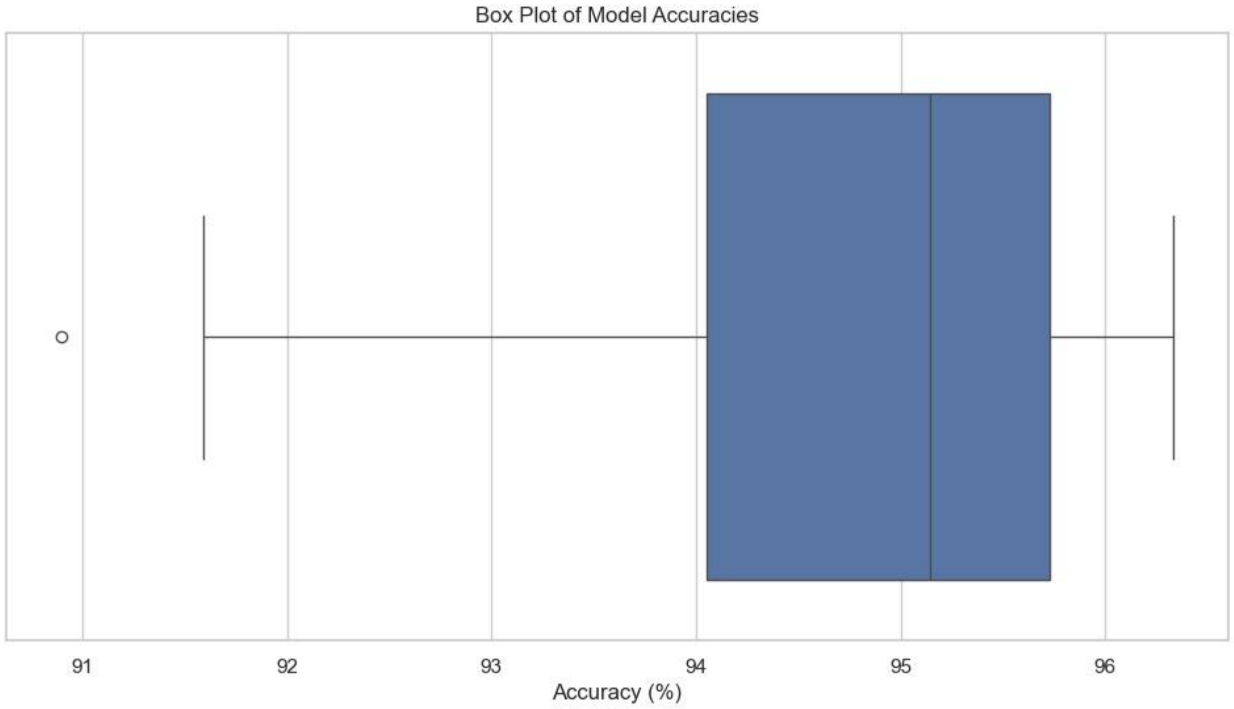
- After training, each model's performance was evaluated using our custom metrics, and results were displayed using a defined display_metrics function that provided insights into how well each model performed in terms of precision, recall, and F1 scores across different KTAS levels.
- The results of GridSearchCV were summarized and reported using display_grid_search_results, providing transparency on the best parameters and the models' effectiveness on unseen data.

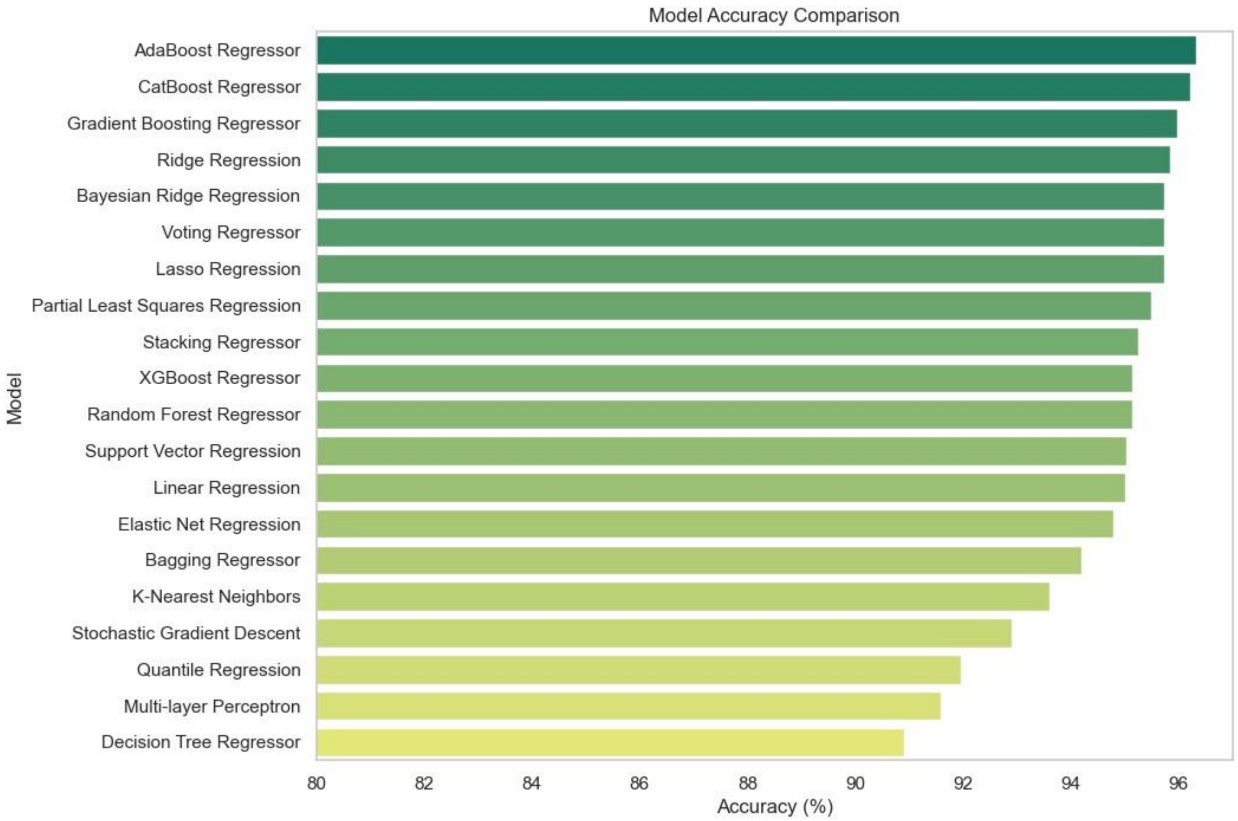
## Results And Analysis

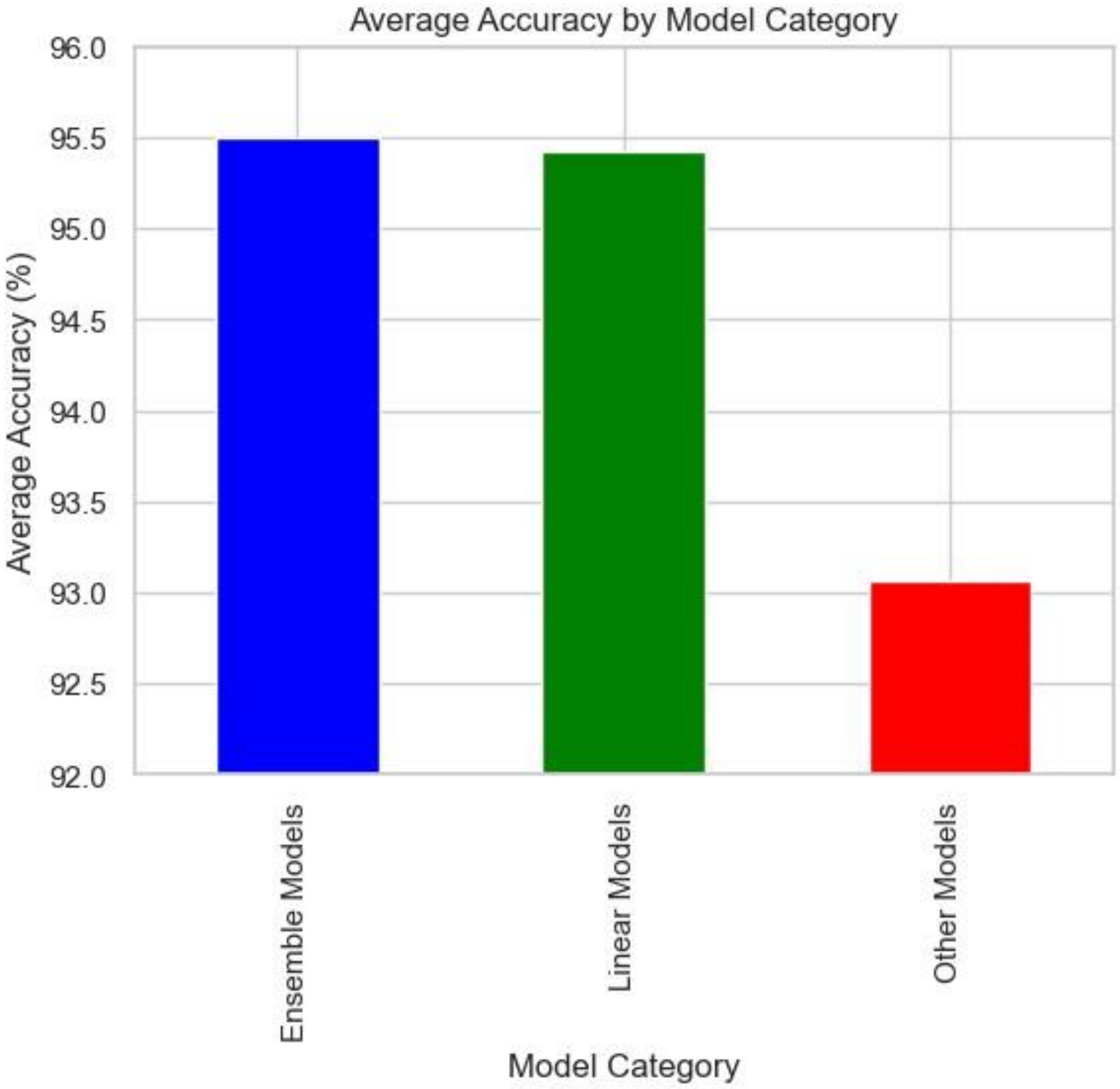
The accuracy that we got for each model is presented below.



| Algorithm | Accuracy (%) |
|---|---|
| Linear Regression | 95.02 |
| Ridge Regression | 95.85 |
| Lasso Regression | 95.73 |
| Stochastic Gradient Descent (SGD) Classifier | 92.9 |
| Support Vector Regression (SVR) | 95.03 |
| K-Nearest Neighbors Regressor (KNN) | 93.61 |
| Partial Least Squares Regression (PLS) | 95.5 |
| Bayesian Ridge Regression | 95.73 |
| Decision Tree Regressor | 90.9 |
| Random Forest Regressor | 95.14 |
| AdaBoost Regressor | 96.33 |
| Gradient Boosting Regressor | 95.97 |
| Multi-layer Perceptron Regressor (MLP) | 91.59 |
| XGBoost Regressor | 95.15 |
| CatBoost Regressor | 96.21 |
| Elastic Net Regression | 94.79 |
| Quantile Regression | 91.96 |
| Bagging Regressor | 94.2 |
| Stacking Regressor | 95.26 |
| Voting Regressor | 95.73 |




Model Accuracy Comparison
AdaBoost Regressor
CatBoost Regressor
Gradient Boosting Regressor
Ridge Regression
Bayesian Ridge Regression
Voting Regressor
Lasso Regression
Partial Least Squares Regression
Stacking Regressor
XGBoost Regressor
Random Forest Regressor
Support Vector Regression
Linear Regression
Elastic Net Regression
Bagging Regressor
K-Nearest Neighbors
Stochastic Gradient Descent
Quantile Regression
Multi-layer Perceptron
Decision Tree Regressor
Model
80
82
84
86
88
90
92
94
96
Accuracy (%)



In our evaluation of machine learning models for predicting Korean Triage and Acuity Scale (KTAS) scores, we observed varied performances across a spectrum of algorithms. Each category of performers—high, solid, moderate, and underperformers—demonstrates unique strengths and challenges.

**High Performers**

- **AdaBoost Regressor (96.33%)**: AdaBoost's top performance can be attributed to its ability to adaptively boost weak learners. This model is particularly effective in our setting due to its iterative correction of errors from the previous models, focusing increasingly on challenging cases that were initially misclassified. Its strength lies in improving decision boundaries incrementally, which is crucial in a triage context where accurate categorization can significantly impact patient outcomes.

- **CatBoost Regressor (96.21%)**: CatBoost excels in datasets with categorical and high-dimensionality features, which are prevalent in medical datasets. Its algorithm reduces overfitting without extensive hyperparameter tuning, making it highly suitable for our diverse and complex triage data. Its capability to naturally handle categorical features reduces preprocessing requirements and enhances model robustness.

- **Gradient Boosting Regressor (95.97%)**: This model's effectiveness comes from its gradient boosting framework that focuses on minimizing errors sequentially using decision trees as base learners. It is particularly adept at handling varied types of data and nonlinear relationships, which are typical in medical data sets, explaining its high accuracy in our project.

**Solid Performers**

- **Ridge and Bayesian Ridge Regression (~95.85% and 95.73%)**: These models incorporate L2 regularization, which helps manage multicollinearity among predictive features. The similarity in their performances suggests that regularization plays a crucial role in stabilizing predictions in data-rich environments, typical of medical datasets where many features may be correlated.

- **Lasso Regression (95.73%)**: Lasso’s ability to perform feature selection inherently by reducing some coefficients to zero helps in simplifying the model and focusing on the most relevant features. This is beneficial in triage settings where only a subset of features might be predictive of the urgency level.

- **Partial Least Squares Regression (PLS) (95.50%)**: PLS helps in data situations where predictors are highly correlated or the number of predictors is large compared to the number of observations. It reduces the predictors to a smaller set of uncorrelated components, which is why it performed well in our dataset.

**Moderate Performers**

- **Linear Regression (95.02%)**: As one of the simplest models, its relatively high performance underscores a potentially strong linear component in the relationship between features and triage levels. However, its simplicity may limit its ability to capture more complex patterns without overfitting.
- **K-Nearest Neighbors Regressor (KNN) (93.61%)**: KNN's moderate performance might be due to its reliance on local information, which can be a disadvantage if the sample distribution is uneven or if the feature space is too high-dimensional.

**Underperformers**

- **Decision Tree Regressor (90.90%)**: While decision trees are easy to interpret, their tendency to overfit, especially in complex datasets with many features, likely contributed to their lower accuracy.
- **Multi-layer Perceptron Regressor (MLP) (91.59%)**: MLP's performance suggests that the neural network might not have been adequately tuned or that the dataset characteristics do not favor complex non-linear models without extensive feature engineering and tuning.
- **Quantile Regression (91.96%)**: This model's lower performance could be attributed to its focus on estimating specific quantiles, which may not capture the central tendency effectively if the data distribution is skewed or has outliers.

  **Specialized Ensemble Techniques**

Bagging, Stacking, and Voting Regressors: These methods demonstrated variable performances, with Bagging slightly lower due to its straightforward approach of reducing variance by averaging multiple decision trees. Stacking and Voting showed better results by effectively combining different model strengths, thus providing a balanced approach to handling diverse data characteristics.

### Key Findings

Our comprehensive analysis and application of twenty different machine learning algorithms yielded significant insights and results:

- **High Accuracy Across Models**: Most models achieved an accuracy of over 95%, with AdaBoost and CatBoost regressors showing exceptional performance with accuracies above 96%. This high level of accuracy across diverse algorithms underscores the robustness of machine learning techniques in handling complex, real-world medical data.
- **Strength of Ensemble Methods**: Models like AdaBoost, Gradient Boosting, and Random Forest demonstrated superior performance, highlighting the strength of ensemble methods in improving prediction accuracy through model diversity and error correction mechanisms.
- **Effectiveness of Regularization**: Ridge, Lasso, and Bayesian Ridge Regression effectively managed multicollinearity and overfitting, which are common in datasets with numerous interrelated variables, typical in medical datasets.
- **Challenges in Model Tuning**: The moderate performance of models such as the Multi-layer Perceptron Regressor and Decision Tree Regressor indicated challenges in overfitting and the need for careful hyperparameter tuning and model selection in complex datasets.

### Implications for Triage Systems

The results of this project have several implications for the implementation of machine learning in triage systems, particularly in developing countries:

- **Improvement in Triage Accuracy**: The use of machine learning models can significantly enhance the accuracy of triage, potentially leading to better patient outcomes by prioritizing urgent cases more effectively.

- **Resource Optimization**: More accurate triage can lead to better resource allocation, ensuring that critical resources are reserved for the most severe cases.
- **Scalability and Adaptability**: The successful application of various machine learning models demonstrates the scalability and adaptability of these approaches across different healthcare settings.

## Conclusion

Our project demonstrated the transformative potential of machine learning in improving triage systems within emergency departments, especially in resource-limited settings typical of developing nations. By leveraging advanced algorithms to accurately predict the Korean Triage and Acuity Scale (KTAS) scores, we significantly enhanced the precision of triage decisions, thus facilitating better patient outcomes and optimizing resource allocation. The high accuracy levels achieved across a diverse array of machine learning models underscore the robustness and adaptability of these techniques in handling complex, real-world medical data, marking a promising pathway for the broader adoption and continuous refinement of machine learning technologies in global health systems.

Looking forward, the project sets the stage for further optimization of these models and their practical implementation within healthcare settings. There is potential for additional tuning of the models, particularly those that underperformed, through more intricate feature engineering or alternative algorithmic approaches. Practical integration of these models into real-world triage systems will require seamless integration with existing IT infrastructures, dedicated training for medical personnel, and rigorous ongoing evaluations to ensure that the models remain effective as healthcare conditions evolve. Moreover, expanding the diversity of datasets to encompass a broader range of medical scenarios and demographic groups will be crucial to ensuring the models' reliability and applicability across different populations and healthcare environments, paving the way for truly global health system enhancements.