TAMU Datathon 2023

Understanding Patient Survival Factors - TD Hospital

(Submission Mail Id: pragatinaikare311@tamu.edu)

Introduction

Automated decision-making systems play a significant role in our lives, and understanding how these models work is crucial. Modern AI and machine learning algorithms often pose challenges in explaining their outcomes, leading to the "black box" problem. This report addresses the need to gain insights into the survival factors of patients and improve the transparency of the decision-making process.

∞ Team Infinite Matrix ∞

1. Pragati Naikare
2. Shivesh Kodali
3. Jack Wooley
4. Vinay Chandra

Description

TD Hospital has sought assistance in rectifying issues with their patient data, accumulated over several years, without clear documentation of flaws. The objective is to understand the information contained in the dataset and identify the factors contributing to patient survival. This process may involve making modifications to the dataset and utilizing machine learning techniques to predict outcomes.

Data Preprocessing:

We made the following modifications to the dataset:

Data Cleaning: We addressed missing values, outliers, and inconsistencies within the dataset to ensure data integrity. Certainly, here's a concise and clear way to explain the decision to drop specific columns from the dataset due to a high number of missing values in your report:

A. **Removing columns with a high proportion of missing values:**

During the data cleaning process, we encountered columns with a substantial number of missing values, which could compromise the quality and reliability of the dataset. After careful consideration, we made the decision to drop the following columns:

Cost, bloodchem1, glucose, psych2, bloodchem3, bloodchem4, sleep, bloodchem5, pdeath, psych3, psych4, disability, urine, bloodchem6, dose

The rationale behind this decision is to maintain data integrity and ensure the dataset's suitability for further analysis. By removing columns with a high proportion of missing values, we reduce noise and enhance the overall quality of the dataset. This, in turn, will contribute to more robust and accurate analyses and predictions.

B. Handling Missing Values:

In our effort to prepare the dataset for analysis, we addressed the challenge of missing values. A thorough evaluation revealed that certain columns contained missing data. To maintain data completeness and enable meaningful analyses, we employed imputation techniques to fill in these gaps.

Imputation Methods

- Mean Imputation: For numerical features, we adopted mean imputation. This approach involved replacing missing values with the mean of the respective feature. This method is particularly useful for preserving the overall statistical characteristics of the data.
- Mode Imputation: In the case of categorical features, we chose mode imputation. Mode imputation involves replacing missing values with the mode (most frequent value) of the corresponding feature. This method is suitable for maintaining the distribution of categorical data.

C. Handling Duplicate Records

Upon thorough analysis, we identified a total of 26 duplicate records in the dataset. These duplicates were characterized by identical values across all features and were instances where the same data appeared more than once.

D. Outlier Reduction: Handling Extreme Values

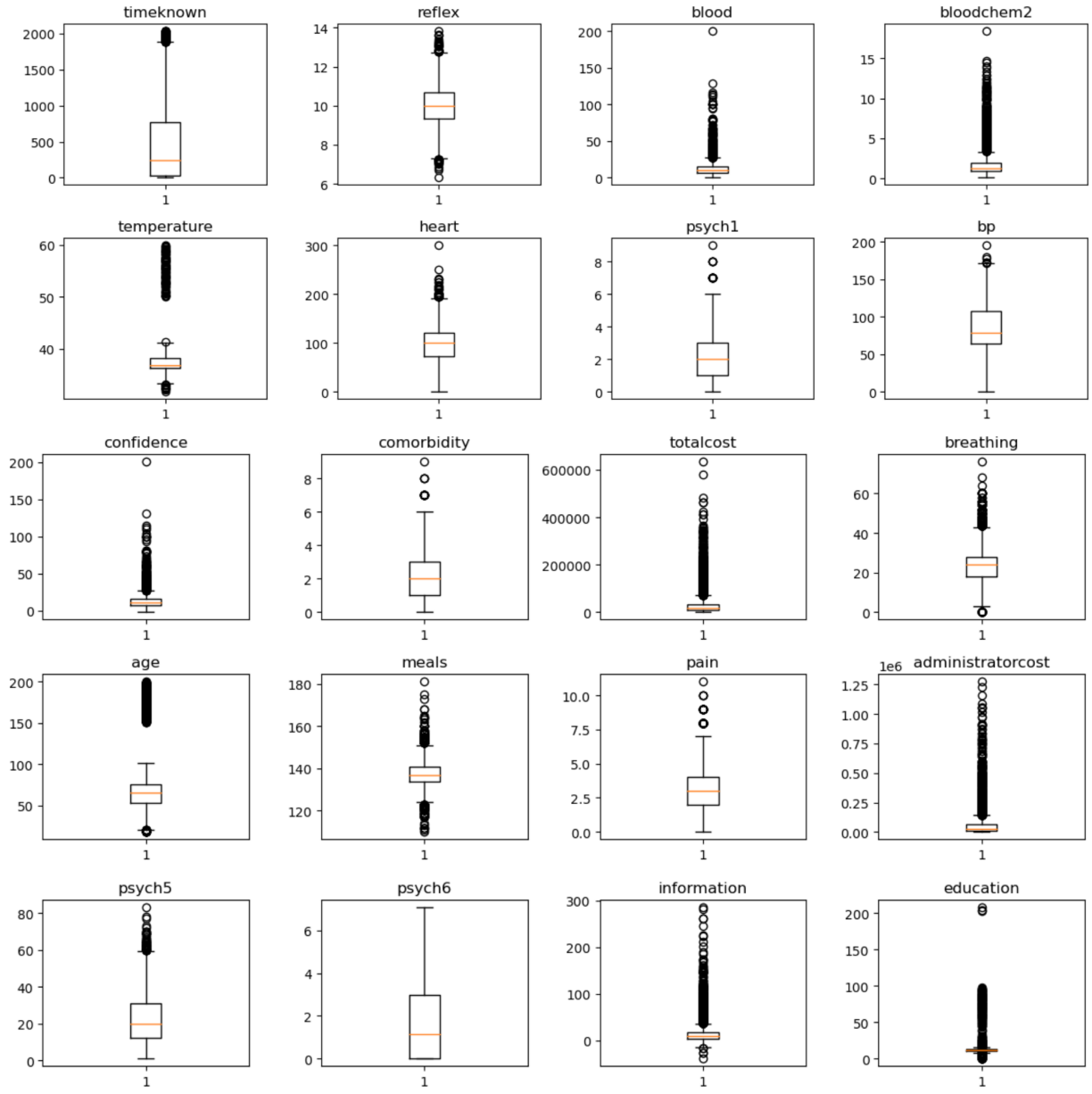
As part of our data preprocessing efforts, we conducted an examination of potential outliers within the dataset.

Age: The column "age" exhibited values that ranged from 150 to 200, which appeared to be extreme outliers. To ensure the dataset's consistency and the reliability of our analysis, we made the decision to reduce these outliers by limiting the age range to 50-100. This modification is based on domain knowledge and the understanding that such extreme age values are likely erroneous.

Temperature: Similarly, the "temperature" column displayed values ranging from 60 to 70, which were considered outliers due to their extreme nature. To enhance data quality and maintain consistency, we reduced these outliers by constraining the temperature range to 30-40. This adjustment aligns with typical human body temperature ranges.

Heart Rate: In the "heart rate" column, a few values exceeded 200, indicating extreme outliers. To maintain data integrity, we further reduced these outliers to the upper bound value.

Other Columns: It's important to note that, due to the lack of detailed information about certain columns in the dataset, we refrained from considering them as outliers. In cases where data characteristics and clinical relevance were not well-defined, we exercised caution and did not perform outlier reduction.

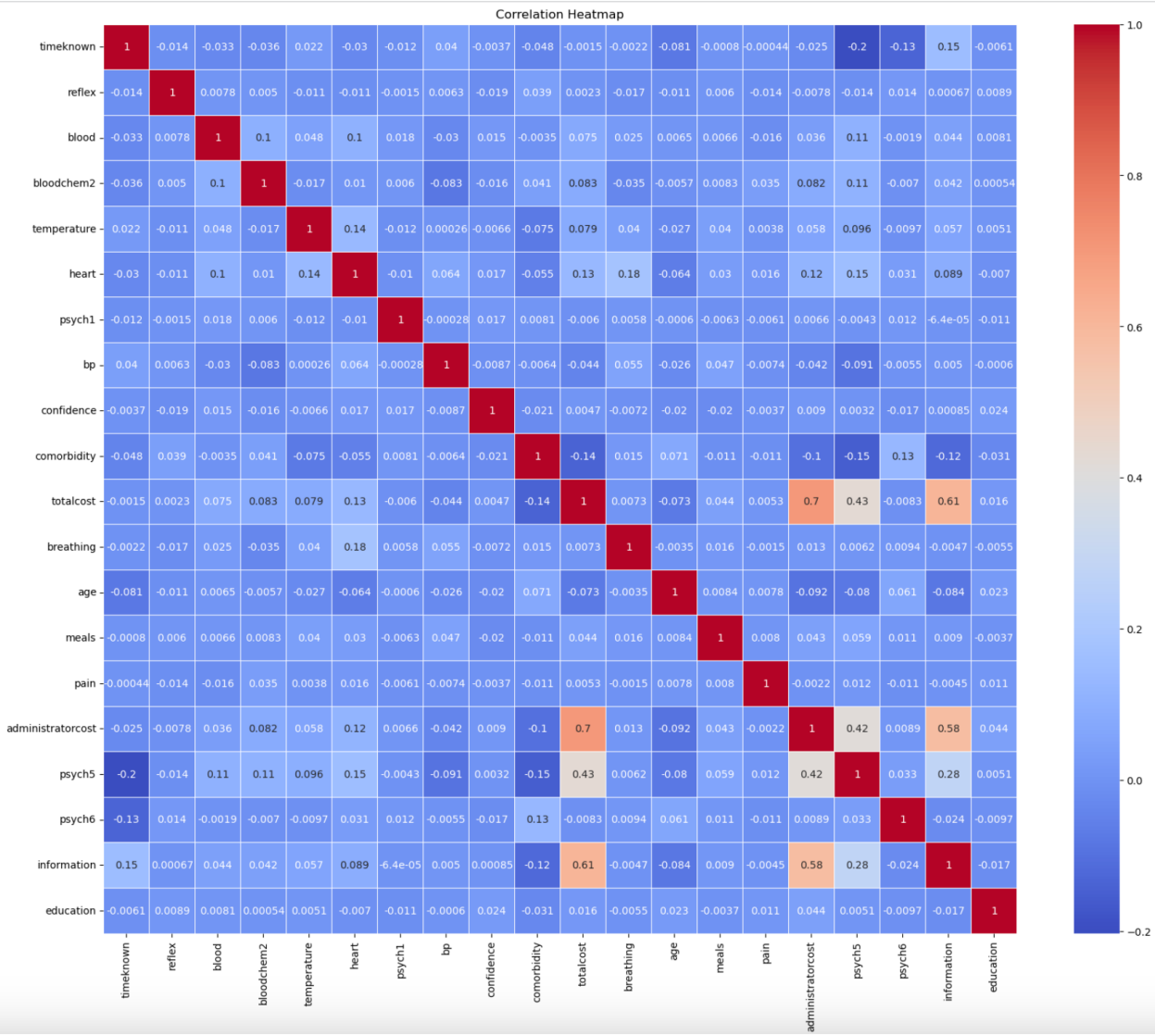


E. Data Scaling:

In cases where numerical features had varying scales, we applied scaling to normalize them.

Correlation Analysis: Heatmap

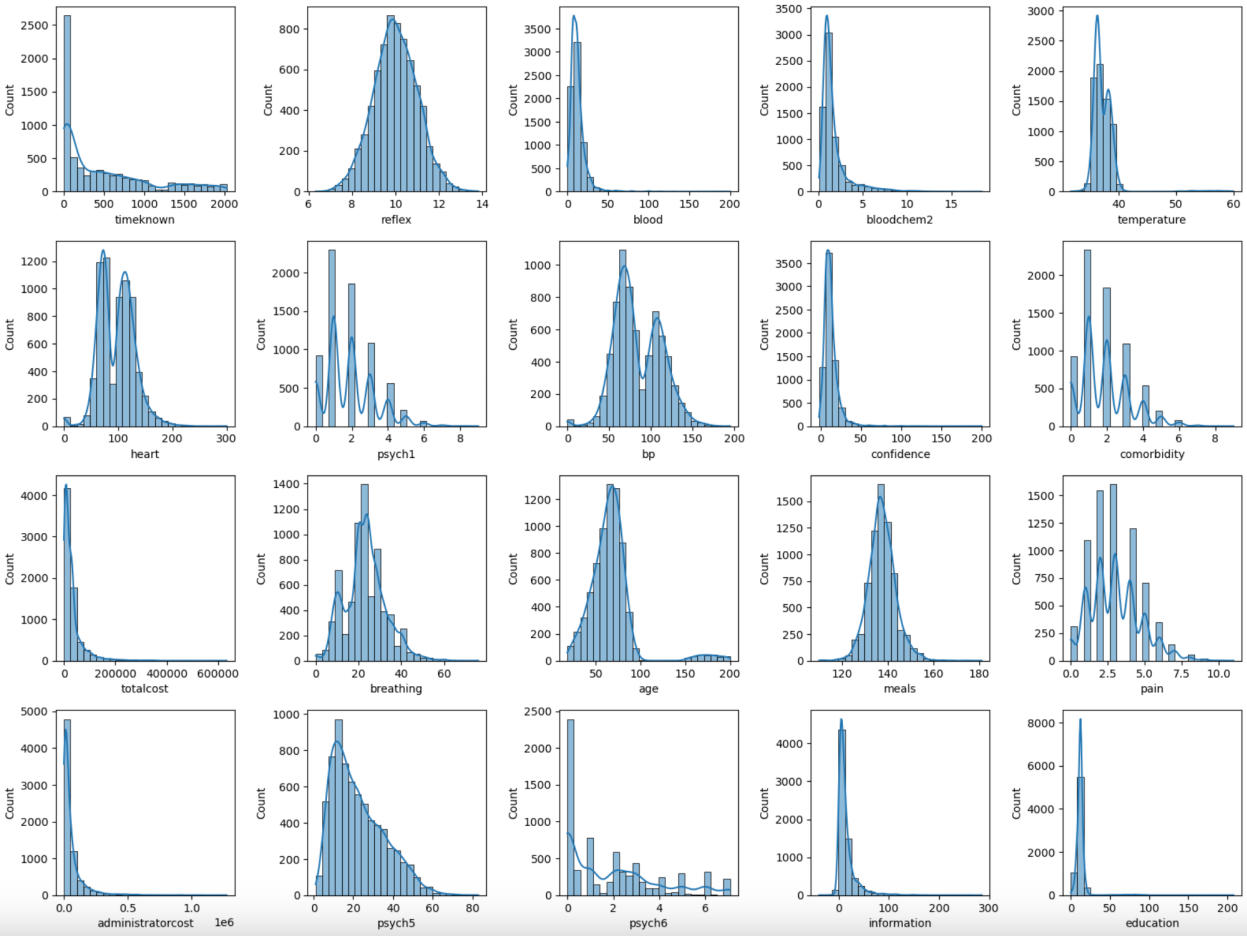
To gain a deeper understanding of the relationships between various numeric columns in the dataset, we conducted a correlation analysis. The primary objective was to identify patterns of association and assess the strength of correlations and accordingly make the decisions on columns.



Strong Positive Correlation: Cells with a color closer to 1 (e.g., dark red) indicate a strong positive correlation, suggesting that as one variable increases, the other tends to increase as well.

Strong Negative Correlation: Conversely, cells with a color closer to -1 (e.g., dark blue) represent a strong negative correlation, implying that as one variable increases, the other typically decreases.

Weak or No Correlation: Lighter cells around 0 suggest weak or no correlation between the variables, indicating that changes in one variable have little impact on the other.



Statistical Analysis: Chi-Square Test Results

To gain deeper insights into the dataset and assess potential associations between categorical variables, we conducted a chi-square test.

```
Chi-Square Test for race and dnr:
Chi-Square Value: 32.21146297027376
P-Value: 8.53581802138527e-05
There is a significant association between the columns.


Chi-Square Test for race and primary:
Chi-Square Value: 108.42247244522575
P-Value: 2.0950403211743395e-11
There is a significant association between the columns.


Chi-Square Test for race and extraprimary:
Chi-Square Value: 57.048425194173696
P-Value: 7.746686005791098e-08

Chi-Square Test for race and cancer:
Chi-Square Value: 41.6044940750675
P-Value: 1.6064493861469621e-06
There is a significant association between the columns.


Chi-Square Test for dnr and primary:
Chi-Square Value: 569.7793599665907
P-Value: 1.4253543823285652e-112
There is a significant association between the columns.


Chi-Square Test for dnr and extraprimary:
Chi-Square Value: 334.9083369125814
P-Value: 2.676255056103068e-69
There is a significant association between the columns.


Chi-Square Test for dnr and cancer:
Chi-Square Value: 82.4432369364332
P-Value: 5.287050286344088e-17
There is a significant association between the columns.


Chi-Square Test for primary and extraprimary:
Chi-Square Value: 18313.34406911296
P-Value: 0.0
There is a significant association between the columns.
```

```
Chi-Square Test for primary and cancer:
Chi-Square Value: 5991.027249977726
P-Value: 0.0
There is a significant association between the columns.


Chi-Square Test for extraprimary and cancer:
Chi-Square Value: 4700.699639678335
P-Value: 0.0
There is a significant association between the columns.
```

Model Training and Performance Comparison

We systematically trained and fine-tuned a variety of machine learning models, rigorously assessing their performance. By comparing the accuracies of these models, we identified and selected the best-performing model for our analysis.

MODEL	ACCURACY	BEST PARAMETERS
Random Forest	0.87	max_depth: None, n_estimaters: 100
Ada Boost	0.84	learning_rate: 1.0 n_estimaters: 50
Gradient Boosting	0.87	learning_rate: 0.1, n_estimators: 100
Bagging	0.87	max_samples: 0.6, n_estimators: 200
CatBoostClassifier	0.94	Iterations=100, verbose=0

Model Selection

The Gradient Boosting algorithm was selected for its capability to address complex, non-linear relationships within the data. Its ensemble approach, combining the predictions of multiple weak learners, makes it a powerful choice for predictive modeling.

Conclusion

Our analysis involved the application of various machine learning algorithms to predict patient survival based on the dataset. Among the models tested, the Gradient Boosting algorithm emerged as a standout performer, achieving an impressive accuracy of 81.02% on the testing dataset.

This report signifies our commitment to enhancing transparency in automated decision-making systems and assisting the hospital in improving patient care.