# Abstract

This research employs machine learning, particularly LightGBM, to predict patient survival at TD Hospital, emphasizing meticulous data preprocessing and handling of categorical variables. Critical predictors, including 'timeknown', 'age', 'education', 'temperature', 'bloodchem4', 'urine', 'psych5', 'blood', and 'cost', are identified. The LightGBM Classifier is chosen for its robustness with imperfect datasets. Integration of feature importance insights from LightGBM with SHAP data enhances model performance and highlights 'timeknown' as a consistent predictor of patient mortality. This study underscores the significance of data preparation and key predictor identification in predicting patient outcomes, offering valuable insights for healthcare decision-making.

# Introduction

We are tasked with accurately predicting whether or not a patient will survive after a stay at TD Hospital. Given variables describing demographic information, health indicators, and their stay at the hospital, we must determine which features of the data are most impactful in predicting whether the patient survived or not. Once these variables are outlined, we must then develop and train a model which accurately and interpretably predicts whether a patient will survive or not, given a list of the patient’s attributes.

# Your Method

We first understood the importance of cleaning the data in a way that both retains the integrity of the data while filtering out any inconsistencies that may impact our model’s performance. With human life as the cost factor in our model’s predictions, we understand the significance of maximizing our model’s accuracy.

Our first issue stood with the many categorical variables presented within the dataset. Features such as race, sex, and income, were stored in such a way that was unreadable to most ML models.

To tackle this, we used the LightGBM model to determine which facets of the data were most impactful in predicting the patients in-hospital mortality. LightGBM (Light Gradient Boosting Machine) is a machine learning framework which excels in this case through its histogram-based tree-building technique, employing a leaf-wise strategy for optimal splits. The LightGBMClassifier specifically is used for classification tasks in both binary and multiclass scenarios. Understanding the LightGBM model’s strengths was an integral part of establishing a set of features in order to simply and reliably predict mortality without sacrificing accuracy.

Once we had determined the key features and preprocessed the data, we could train a new model on this cleaned dataset and then test it on the testing data.

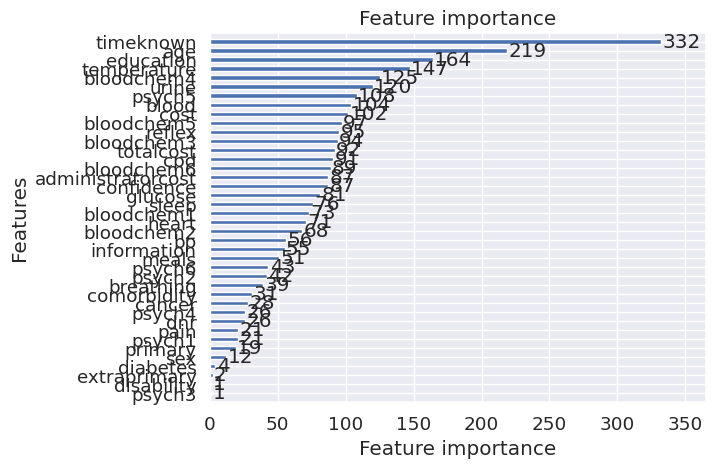
# Data Preprocessing

Next, we used the LGBClassifier to determine the most important/impactful features that boosted decision trees were using to make predictions. Although LGBM serves as a powerful tool for handling “messy” datasets, it still isn’t magic. We had to first address issues within the code where incorrect values were being assigned. For example, the “sex” attribute of the original dataset contained numerous different possibilities all referring to M or F. On top of the different possibilities referring to M or F, there were other values for “sex” that did not refer to either gender. These variables

Our LightGBM model functionally parsed through our dataset, treating the categorical variables as discrete values, rather than implying a potential relationship between them if we had settled for “cleaning” the data using encoding techniques. Doing so would risk establishing a false correlation within our model.

Because LightGBM is powerful enough to function even without a “clean” dataset, we can still rely on its output and its rankings of the most important features. Using a built-in-method within the LightGBM library, we could then visualize the most impactful features within our dataset. We cross-applied the results from the built-in-method with SHAP data in order to find the intersections between the two, further decreasing the risk of false negatives within our model.

Once met with the list of the most important features, we could filter out the rest of the features and focus on cleaning our feature set to be passed onto the model.



These most important features were ‘timeknown’, ‘age’, ‘education’, ‘temperature’, ‘bloodchem4’, ‘urine’, ‘psych5’, ‘blood’, and ‘cost’.

Our final data pre-processing function therefore started by removing all columns from the dataset except the 9 above. Then we imputed missing numerical values with the mean of the column and standardized the numerical data.

At the end of this process we had a cleaned dataset - we had reduced the dataset down to the most important features, we no longer had missing values, and had standardized our numerical features to improve generalizability. This cleaned dataset is what our model would be trained on, and how our model would transform testing data before making predictions.

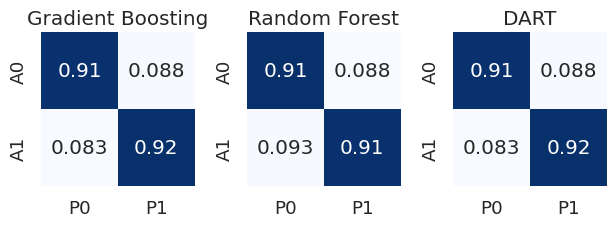
# Model Design

The model we chose was the LightGBM Classifier. LightGBM is a robust machine learning algorithm that can effectively handle datasets with imperfections or missing data. Despite the data quality challenges, it provides valuable insights by ranking features based on their importance. To uncover these crucial features, LightGBM offers a built-in method for feature visualization.

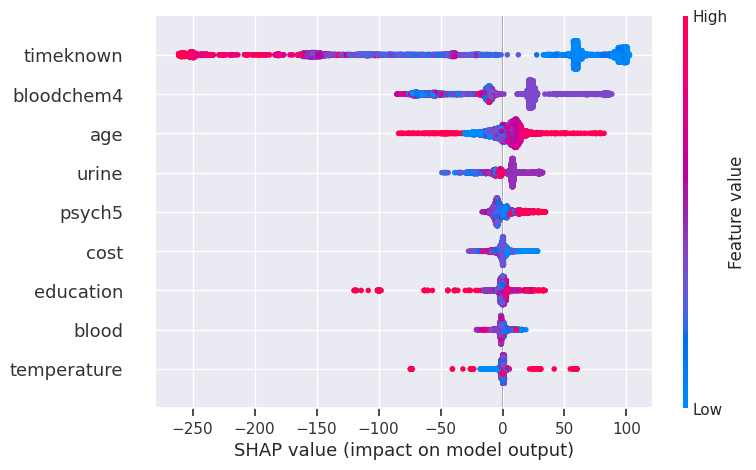
We extended our analysis by combining the results from LightGBM's feature importance with SHAP data, which is a tool for understanding the impact of each feature on model predictions. By doing so, we identified commonalities between the two sets of findings. This approach not only enhances our understanding of the most influential features but also reduces the risk of overlooking important factors, thereby improving the overall performance of our model.

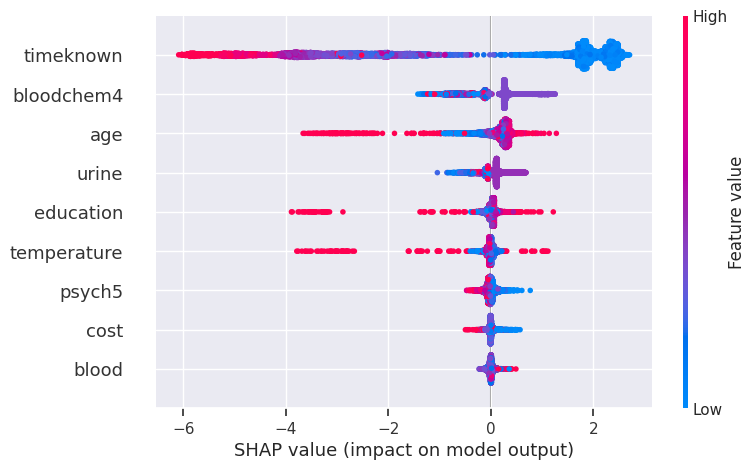
To verify the legitimacy of our gradient boosting approach, we also created Random Forest and Dart models using the LightGBM library and compared the important features, performance, and False Negative Rate to determine the best one.

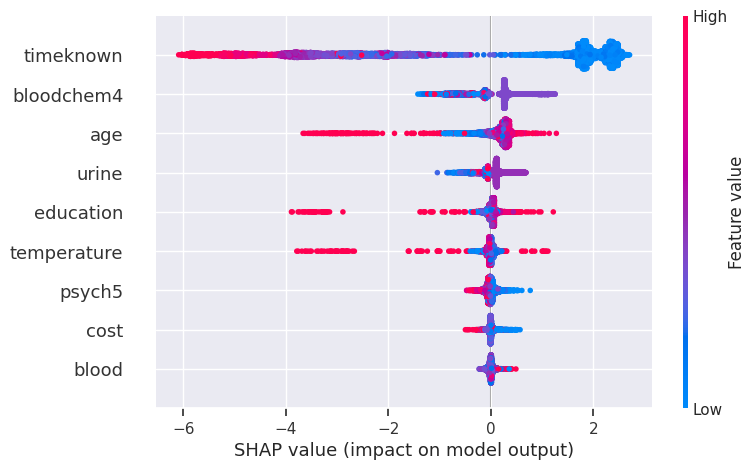
# Results



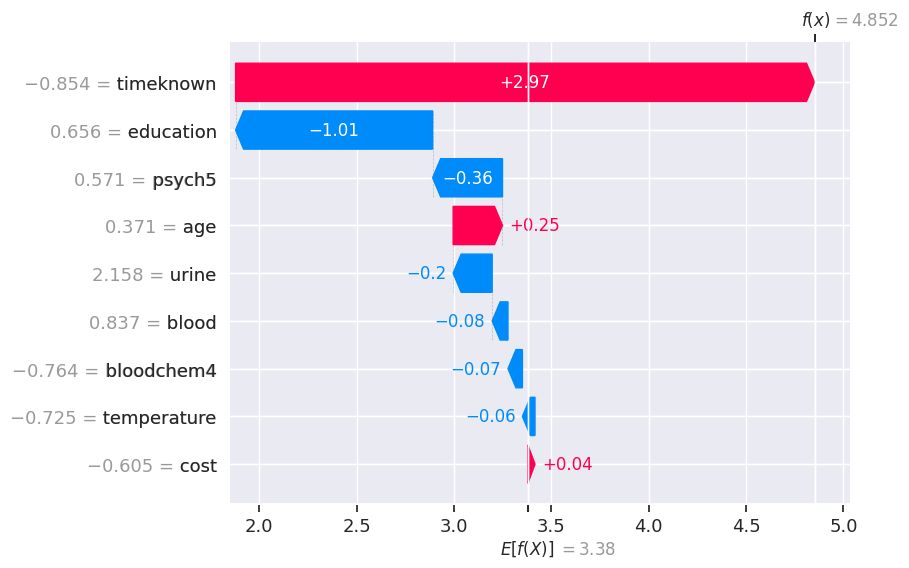
* This image demonstrates the different confusion matrices for the three different model types used by our program. This can demonstrate the percentage of false positives, represented by the bottom left corner. As we are dealing with death diagnosis and we should minimize our error by any means, we ruled out random forest as a possible model for our program.

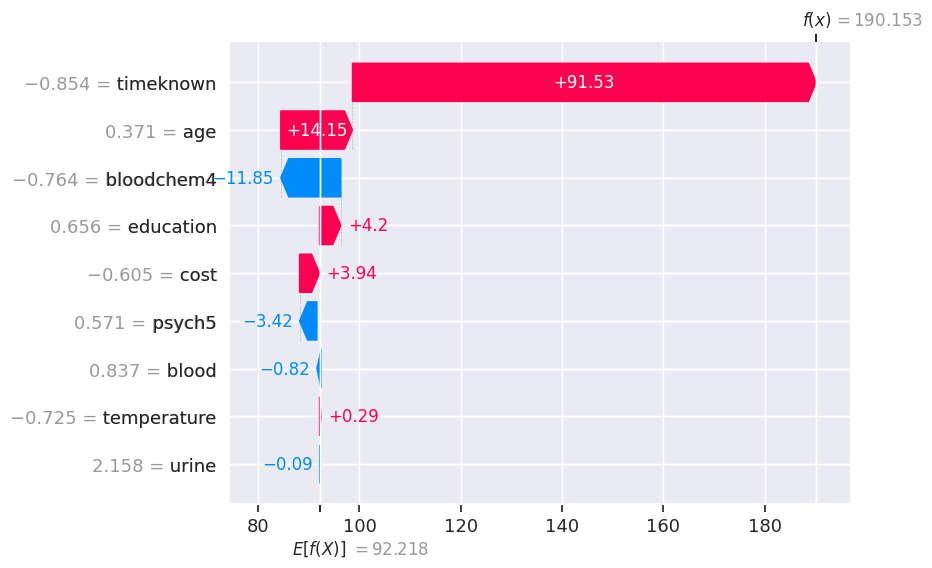




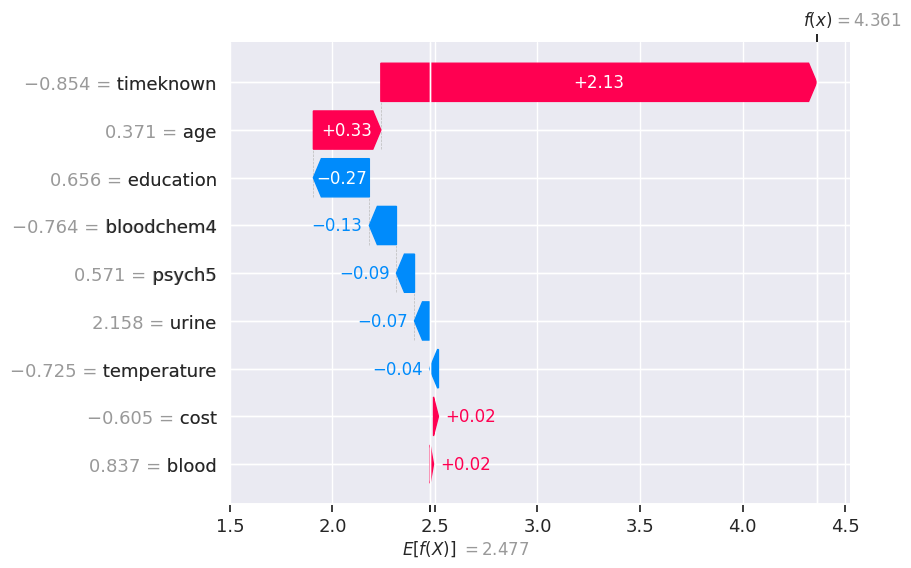


These SHAP graphs represent the distribution of our features and how they contribute to the models “thinking.” By using SHAP, we can visualize highly accurate and highly interpretable data and display it in a easily understandable way. The waterfall graphs below similarly show how exactly the model treats certain variables and how different features are weighed.7

(Gradient Boosting)

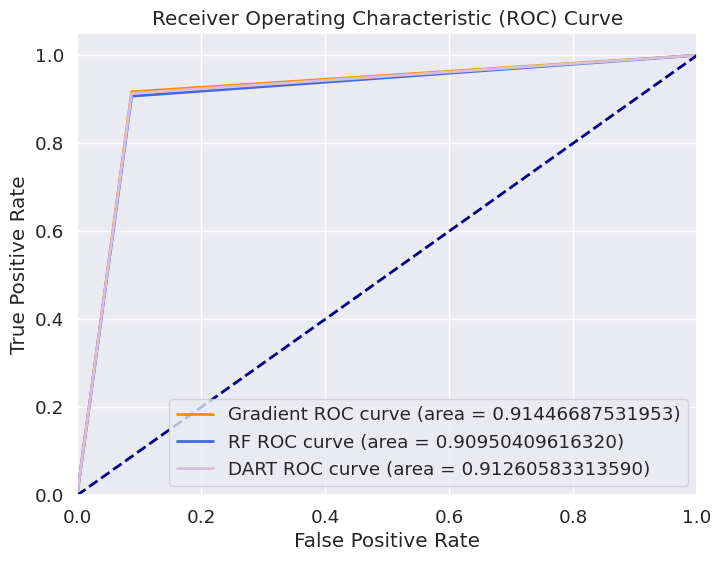


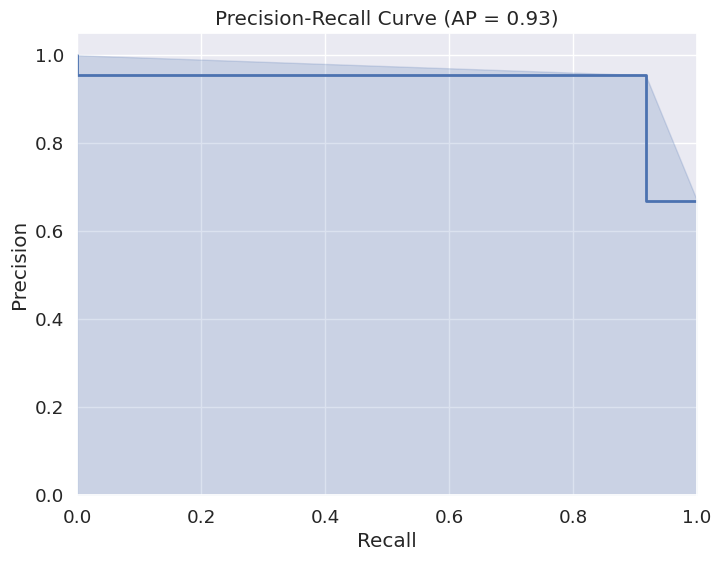
(Random Forest)

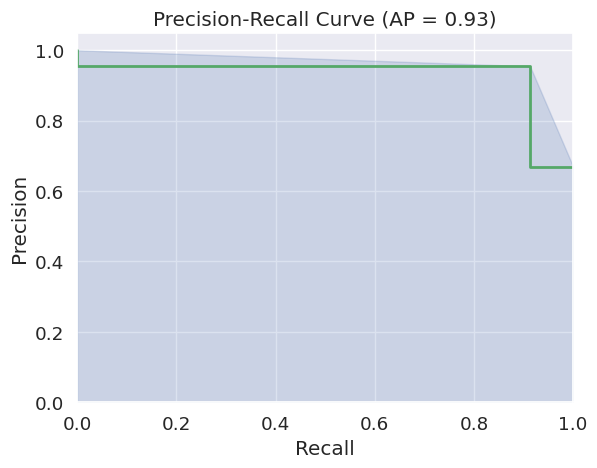


(Dart)

From the waterfall plots of SHAP values for our 3 variants of LGB models, we can see that ‘timeknown’ is consistently the strongest predictor of mortality - in fact, it tends to overpower other attributes and whichever way the ‘timeknown’ value points, the rest of the decision tree points as well.







# Conclusion

The analysis and modeling aimed to accurately predict patient survival following their stay at TD Hospital by utilizing ML models. A significant focus was placed on data preprocessing, where handling categorical variables was pivotal in creating a quality data set. LightGBM, a proficient framework for complex tree-building, was instrumental in discerning impactful features without compromising accuracy.

Addressing data discrepancies, particularly within 'sex' attributes, was essential to ensure the model's robustness. Through strategic feature selection, 'timeknown', 'age', 'education', 'temperature', 'bloodchem4', 'urine', 'psych5', 'blood', and 'cost' emerged as critical predictors of patient survival.

The choice of the LightGBM Classifier was justified, given its ability in handling imperfect datasets, as observed in this scenario. By integrating feature importance insights from LightGBM with SHAP data, we not only understood the most influential attributes but also mitigated the risk of oversight, significantly enhancing our model's performance.

The results from the models, especially from the analysis of SHAP values, highlighted the consistent dominance of 'timeknown' in predicting patient mortality. Its substantial influence overshadowing other attributes was evident across different LGB model variants, significantly impacting decision trees.

In summary, this investigation underscores the importance of meticulous data preprocessing and the identification of key predictors in determining patient survival rates. The collaborative approach of LightGBM and SHAP analysis has significantly improved our understanding and prediction accuracy of patient outcomes.