In-Hospital Mortality Prediction

DSEI 103 - Applied Statistics

Nikhita Kannam, Sumaiya Uddin, Shubham Khandale, Allen Lau

Contributions

Nikhita Kannam

EDA, Outlier Cleanup, Logistic Regression, KNN

Sumaiya Uddin

Data Processing, PCA, Decision Tree, Hypothesis Testing

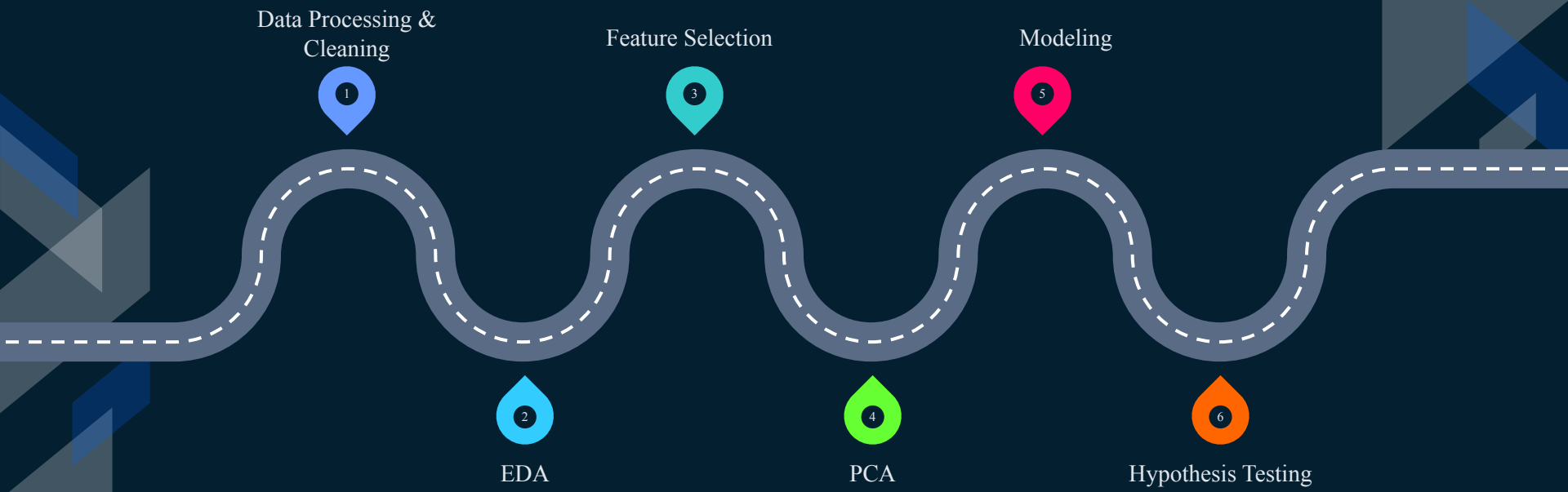
Shubham Khandale

Resampling, LASSO, Random Forest, Lime

Allen Lau

EDA, Logistic Regression, VIF, Decision Tree

Modeling Process



Dataset

- Source: MIMIC-III Database
  - Publicly available database of patient data
  - Admissions to the Intensive Care Unit of the Beth Israel Deaconess Medical Center, in Boston, USA
  - June 1, 2001 – October 31, 2012
  - Kaggle Link
- Response Variable
  - Outcome - Result of Patient Admission to the ICU
    - 0 → Alive
    - 1 → Dead
- Dataset Summary
  - Number of Features → 51
  - Number of Data Points → 1177

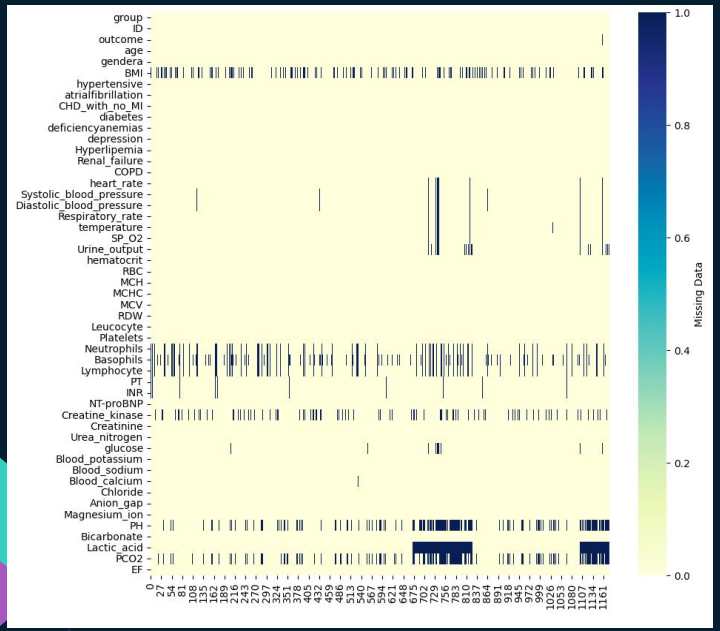
Dataset Features			
group	COPD	MCV	Blood sodium
ID	heart rate	RDW	Blood calcium
outcome	Systolic blood pressure	Leucocyte	Chloride
age	Diastolic blood pressure	Platelets	Anion gap
gendera	Respiratory rate	Neutrophils	Magnesium ion
BMI	temperature	Basophils	PH
hypertensive	SP O2	Lymphocyte	Bicarbonate
atrialfibrillation	Urine output	PT	Lactic acid
CHD with no MI	hematocrit	INR	PCO2
diabetes	RBC	NT-proBNP	EF
deficiencyanemias	MCH	Creatine kinase	Hyperlipemia
depression	MCHC	Creatinine	Renal failure
Urea nitrogen	glucose	Blood potassium	

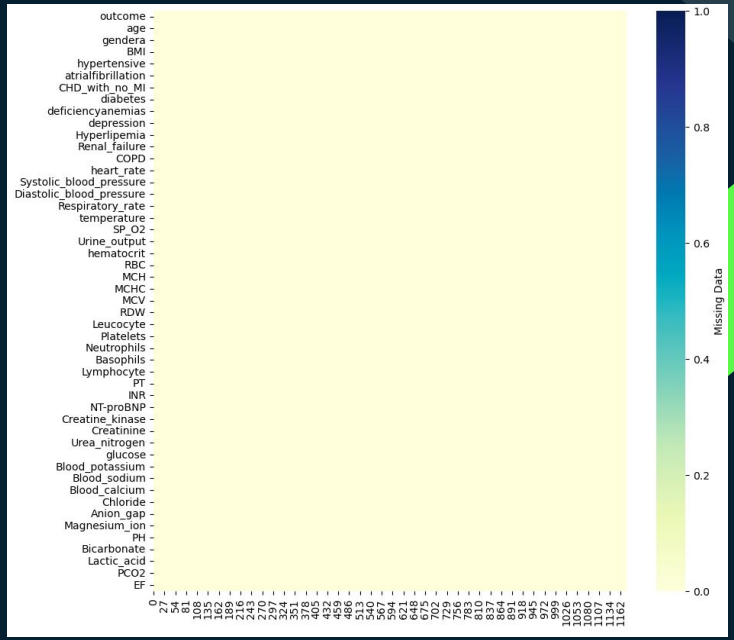
Dataset Null Counts			
PC02	294	PT	20
PH	292	temperature	19
Basophils	259	glucose	18
Lactic Acid	229	Diastolic blood pressure	16
BMI	215	Systolic blood pressure	16
Creatine Kinase	165	SP O2	13
Lymphocyte	145	Respiratory rate	13
Neutrophils	144	heart rate	13
Urine Output	36	Blood calcium	1
INR	20	outcome	1

Summary Statistics

	outcome	age	BMI	heart_rate	Systolic_blood_pressure	Diastolic_blood_pressure	Respiratory_rate	temperature	SP_O2
count	1176.000000	1176.000000	962.000000	1164.000000	1161.000000	1161.000000	1164.000000	1158.000000	1164.000000
mean	0.135204	74.047619	30.188278	84.575848	117.995035	59.534497	20.801511	36.677286	96.272900
std	0.342087	13.437241	9.325997	16.018701	17.367618	10.684681	4.002987	0.607558	2.298002
min	0.000000	19.000000	13.346801	36.000000	75.000000	24.736842	11.137931	33.250000	75.916667
25%	0.000000	65.000000	24.326461	72.371250	105.391304	52.173913	17.925694	36.286045	95.000000
50%	0.000000	77.000000	28.312474	83.610799	116.128205	58.461538	20.372308	36.650794	96.452273
75%	0.000000	85.000000	33.633509	95.907143	128.625000	65.464286	23.391200	37.021991	97.917500
max	1.000000	99.000000	104.970366	135.708333	203.000000	107.000000	40.900000	39.132478	100.000000

Data Processing & Cleaning





The two heatmaps above display the nulls in each column before and after data processing and cleaning. Some methods we used to clean the data set are:

- Dropping columns that are not needed
- Dropping rows that have null values in the data column
- Replacing null values with the median to reduce impact of outliers

Resampling

Response Variable: Outcome

- 0 → Alive
- 1 → Dead

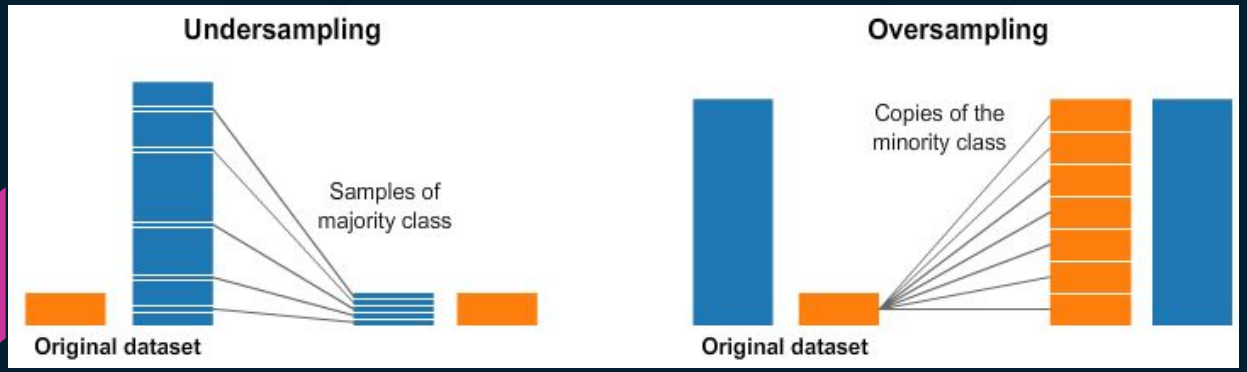
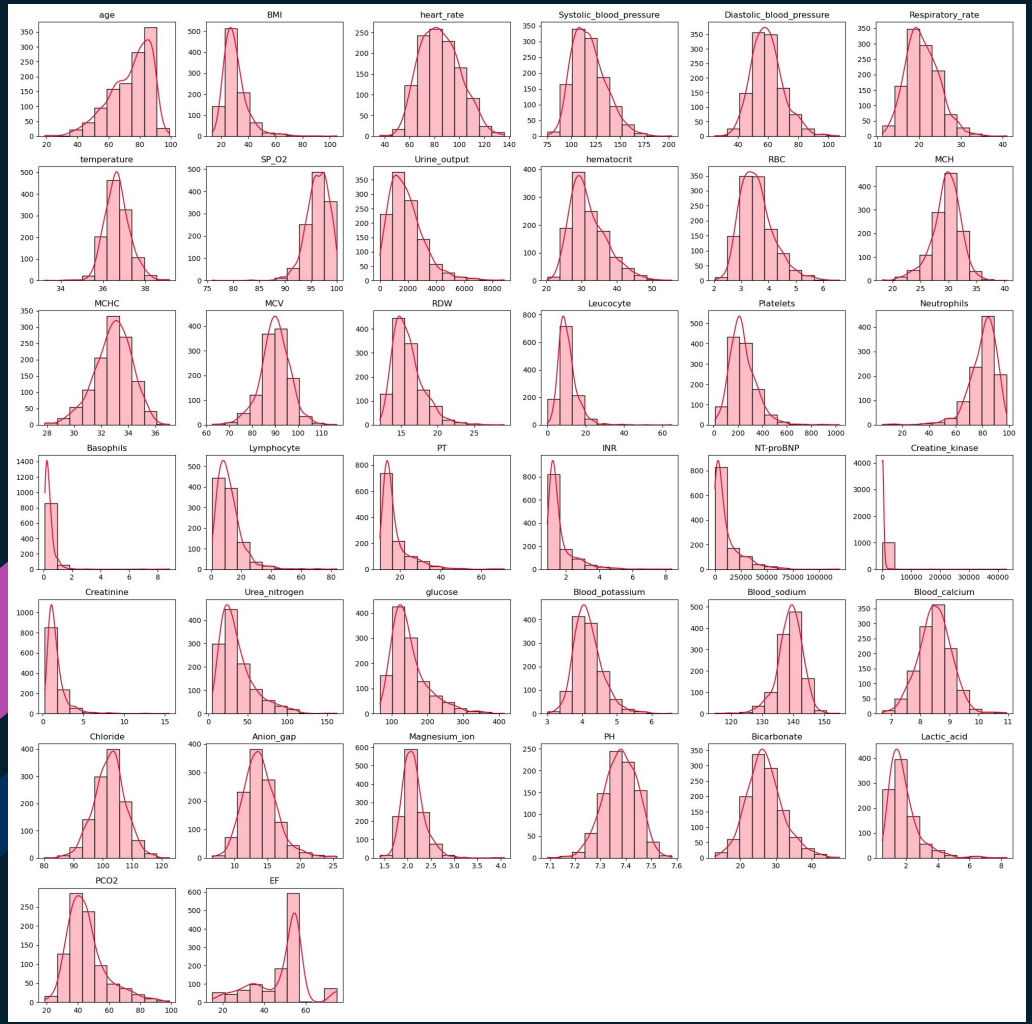


Image Source

Response Variable Counts			
Before Resampling		After Resampling	
0	1017	0	812
1	159	1	812

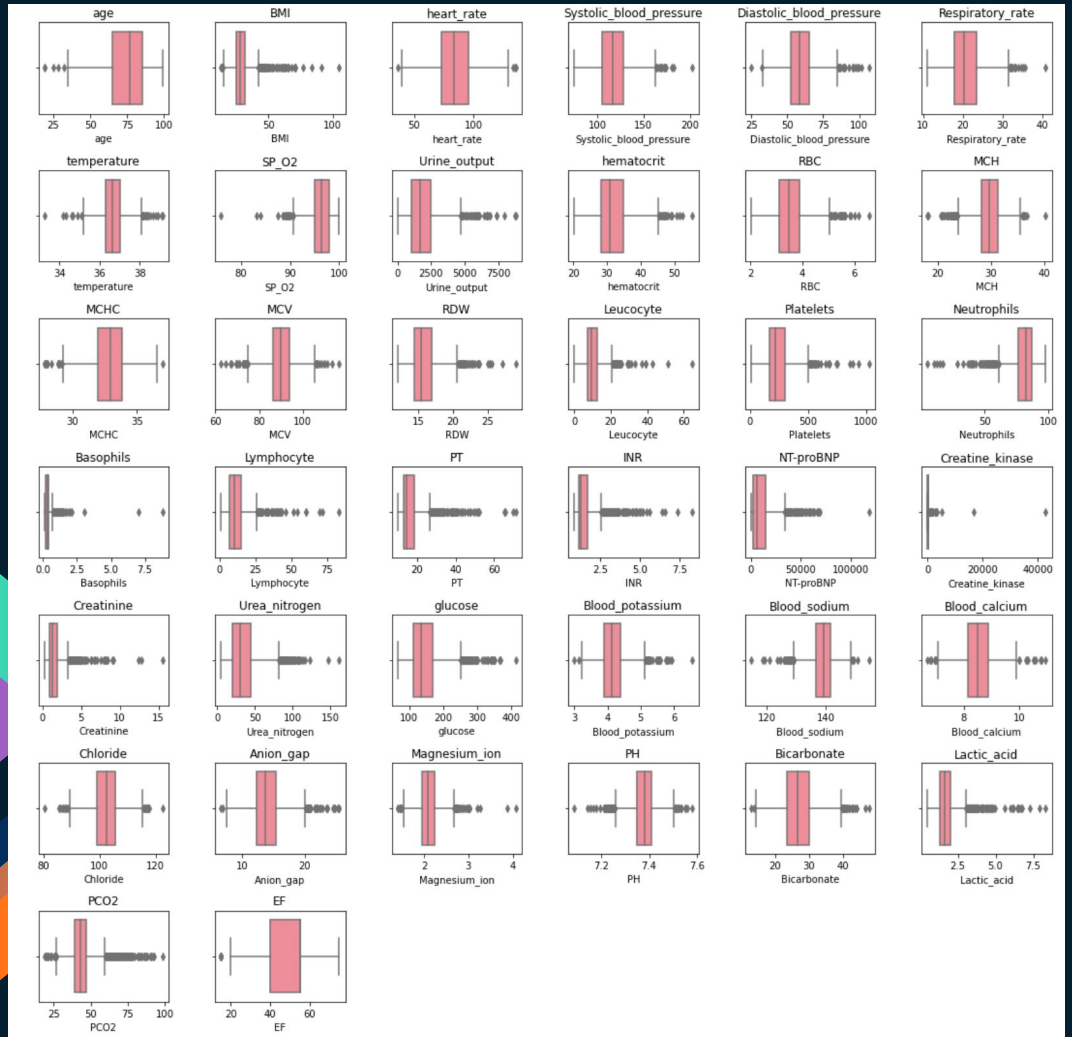
Exploratory Data Analysis



Majority of distributions are right skewed, meaning that most of the data is on the left end. The mean is greater than the median.

Majority of distributions are unimodal.

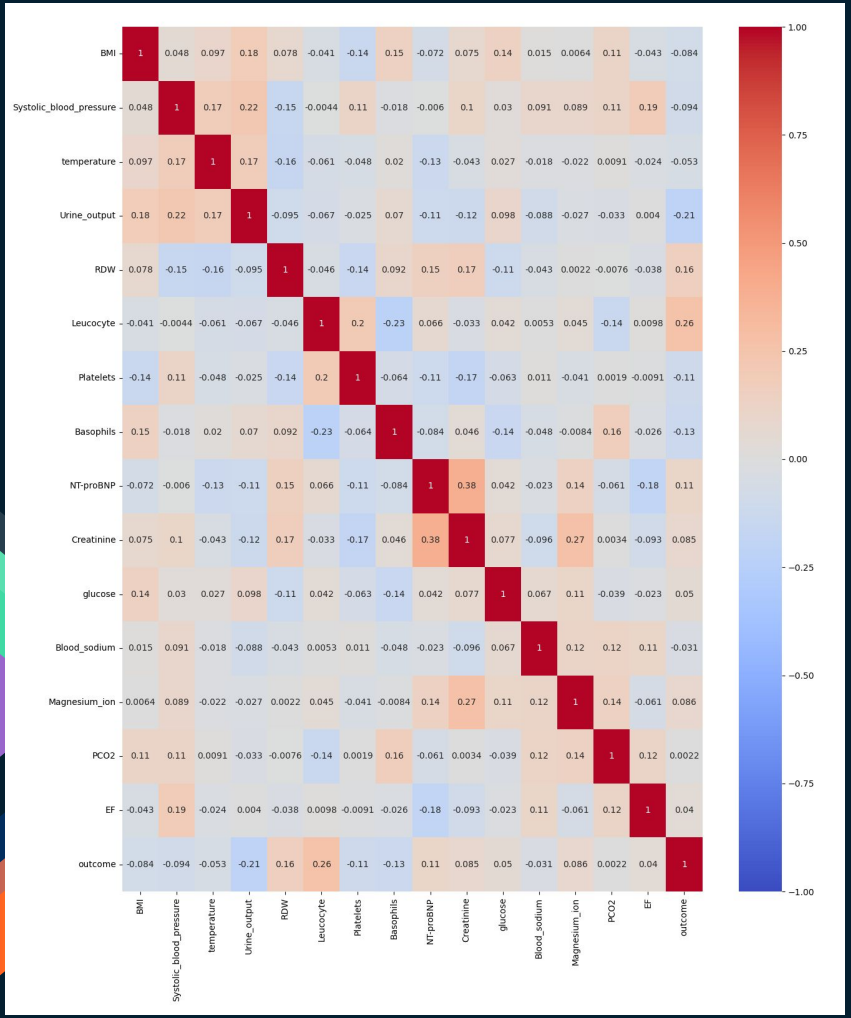
Standard deviations and the scale of the features vary, indicating the need to standardize the data.



The boxplots for each variable is used to identify the outliers, and gain insights on the distributions.

Extreme outliers or erroneous measurements are suspected to be incorrect were removed.

For example, Respiratory rate has one extreme outlier, which was removed.



Top 5 Correlated Features		
Feature 1	**Feature 2**	**Corr. Coeff.**
Urine Output	Systolic blood pressure	0.22195
Basophils	Leucocyte	0.225909
Leucocyte	outcome	0.262817
Magnesium ion	Creatinine	0.266649
NT-proBNP	Creatinine	0.384566

Bottom 5 Correlated Features		
Feature 1	**Feature 2**	**Corr. Coeff.**
Platelets	PCO2	0.001942
Magnesium_ion	RDW	0.002165
outcome	PCO2	0.002196
Creatinine	PCO2	0.003369
EF	Urine_output	0.004009

Feature Selection

Standard Scaler

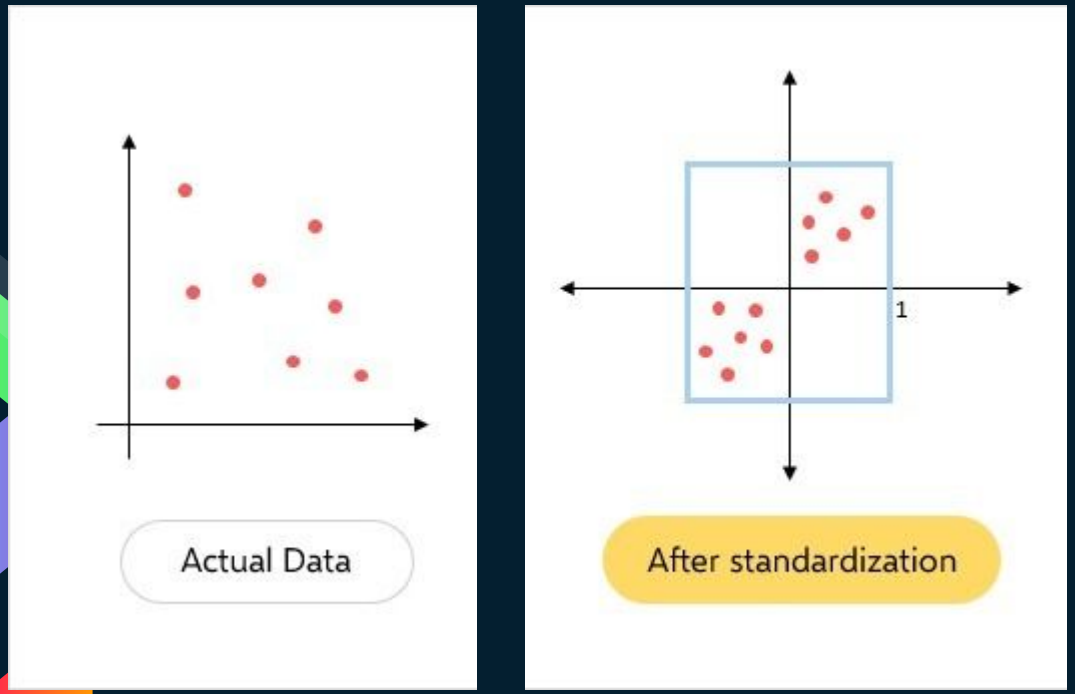
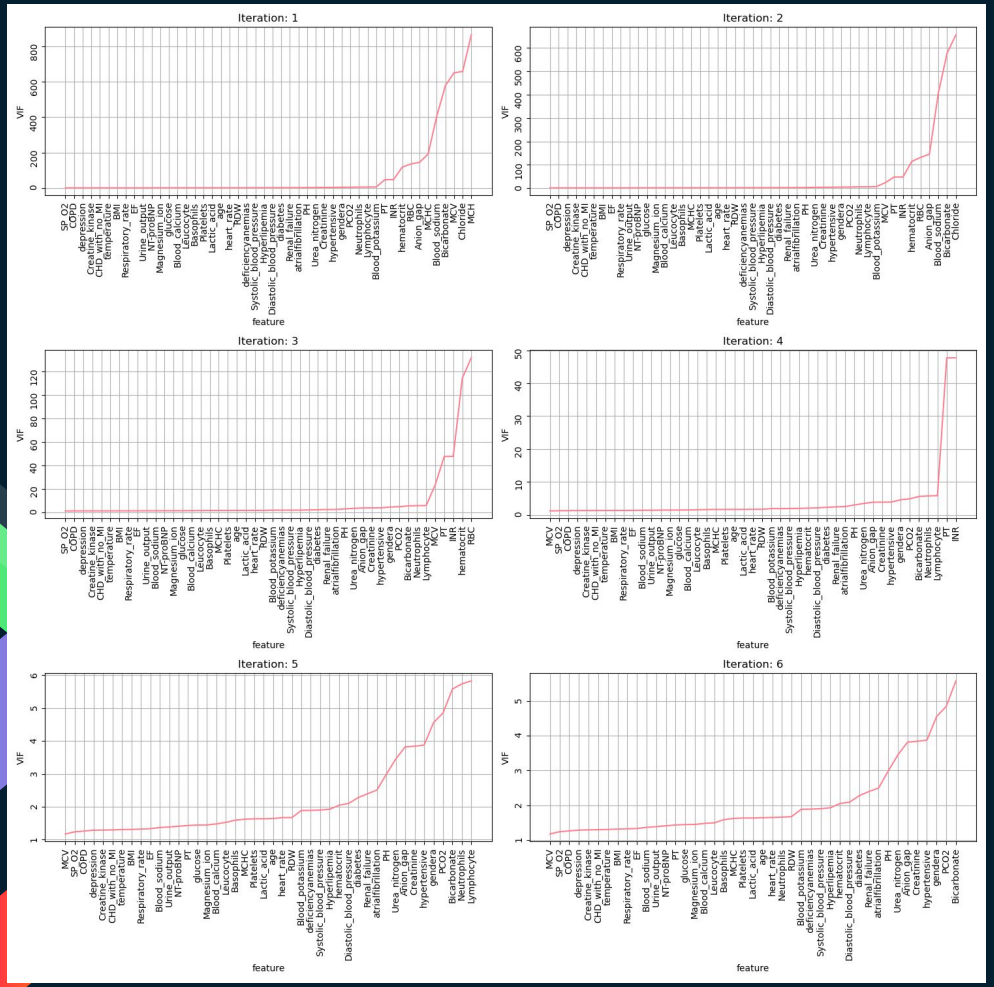


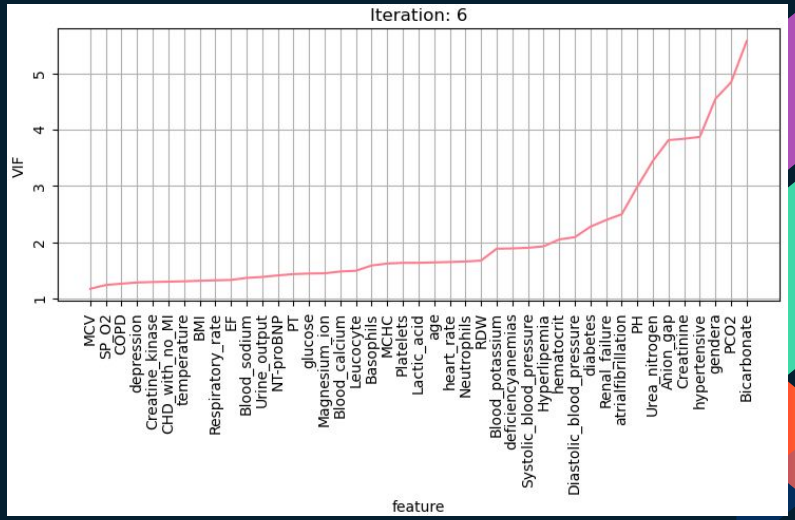
Image Source

Standardizing the features is required due to the differences in standard deviations and magnitudes of the values.

Accomplished by subtracting the mean and dividing by the standard deviation.

Results in a more stable model





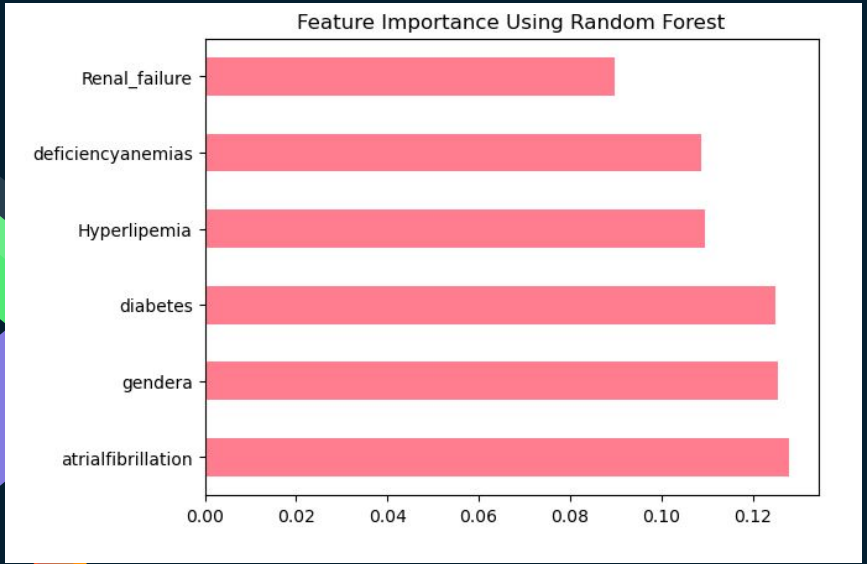
Variance Inflation Factor (VIF) is used to remove features whose variance can be explained by other features.

VIF > 5 is the conservative threshold for a feature to be considered highly correlated with the others

Reduce multicollinearity in the dataset.

Random Forest & LASSO Regression Feature Selection

Random Forest



LASSO Regression (α = 0.01)

Weight[?]	Feature
+0.119	Systolic_blood_pressure
+0.083	Platelets
+0.069	glucose
+0.057	RDW
+0.049	Leucocyte
+0.044	Basophils
+0.040	BMI
+0.034	Magnesium_ion
+0.034	temperature
+0.026	EF
+0.021	outcome
+0.013	Urine_output
+0.012	NT-proBNP
+0.011	Blood_sodium
+0.005	PCO2
+0.003	Creatinine

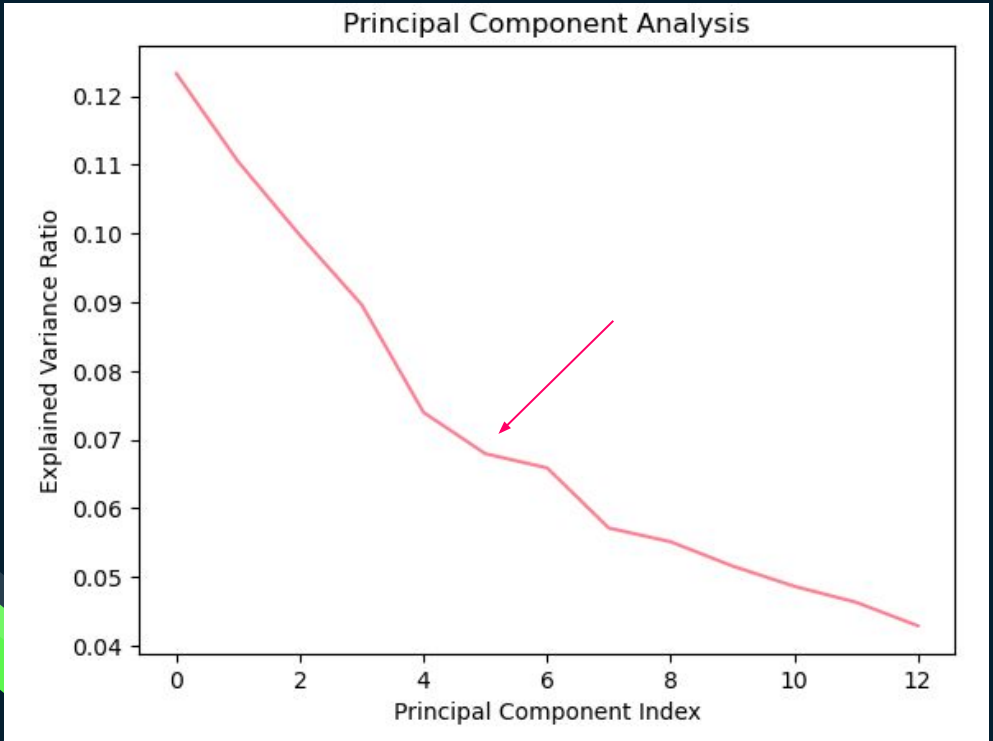
Selected Features

Categorical
gender
hypertensive
atrialfibrillation
CHD_with_no_MI
diabetes
deficiencyanemias
depression
Hyperlipemia
Renal_failure
COPD

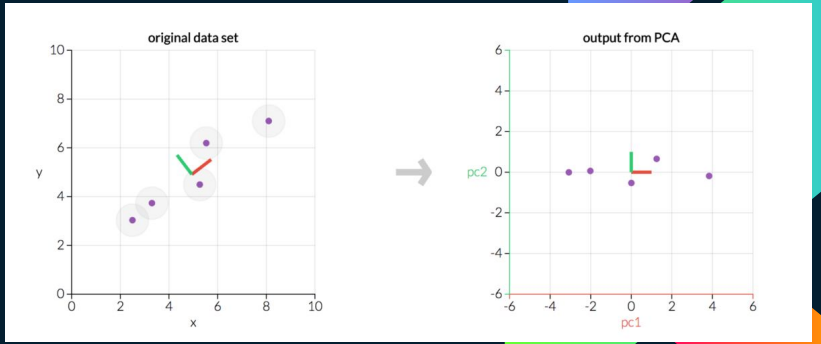
Continuous	
BMI	NT-proBNP
Systolic_blood_pressure	Creatinine
temperature	glucose
Urine_output	Blood_sodium
RDW	Magnesium_ion
Leucocyte	PCO2
Platelets	EF
Basophils	outcome

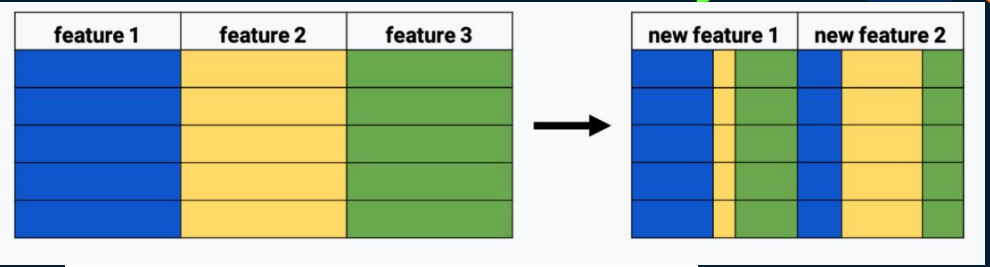
Number of features reduced from 51 to 26

Principal Component Analysis

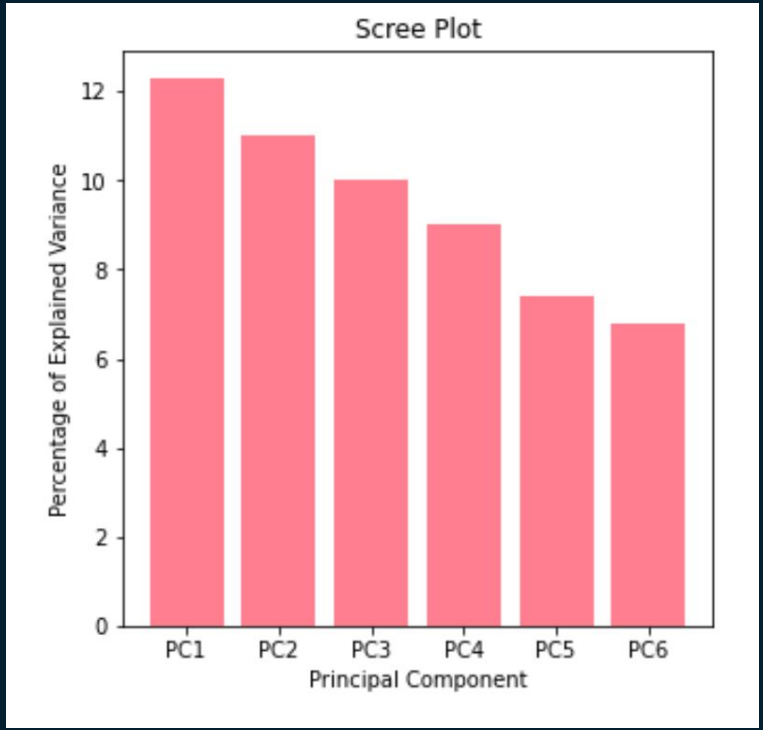


The Explained Variance Ratio is graphed to determine the elbow. According to the graph, we should keep 6 components.





https://campus.datacamp.com/courses/dimensionality-reduction-in-python/feature-extraction?ex=1

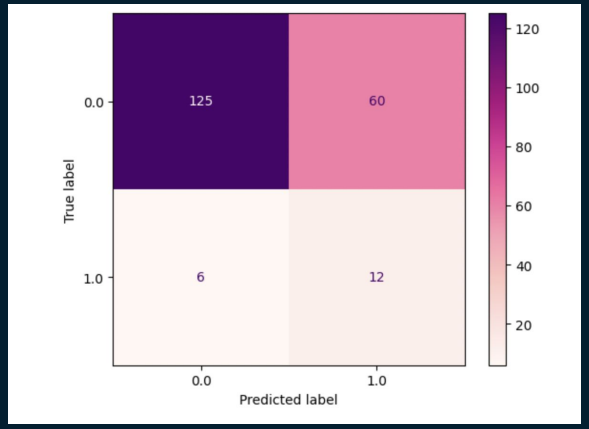


The scree plot is used to determine the number of factors to retain in an exploratory factor analysis (FA) or principal components to keep in a principal component analysis (PCA).

	PC1	PC2	PC3	PC4	PC5	PC6
0	-0.966972	1.782877	-0.952053	-1.405088	-1.621522	-1.099229
1	-2.012096	3.499352	0.531030	2.334477	-1.753157	-0.820873
2	-1.483876	0.197878	0.131047	-0.317830	1.396916	1.403093
3	-0.950295	-1.398700	0.352502	1.029159	0.076503	-0.760649
4	-0.860199	1.807029	-1.555054	-0.279860	1.367071	-0.265036

Modeling

Logistic Regression



Input: PCA components resulting from numerical features determined by Lasso feature selection

The accuracy of the logistic model is 67%

The precision of the model on class 0 is 0.95 and on class 1 is 0.17. Out of all the people that the model predicted would die, about 17% actually did.

```
Accuracy Score:  0.6748768472906403
Confusion Matrix:
 [[125  60]
 [  6  12]]
Classification Report:
              precision    recall  f1-score   support

         0.0       0.95      0.68      0.79       185
         1.0       0.17      0.67      0.27        18

    accuracy                           0.67       203
   macro avg       0.56      0.67      0.53       203
weighted avg       0.88      0.67      0.74       203
```

The model performs worse on identifying class 1 because we have resampled, resulting in duplicate data points for class 1

GridSearchCV

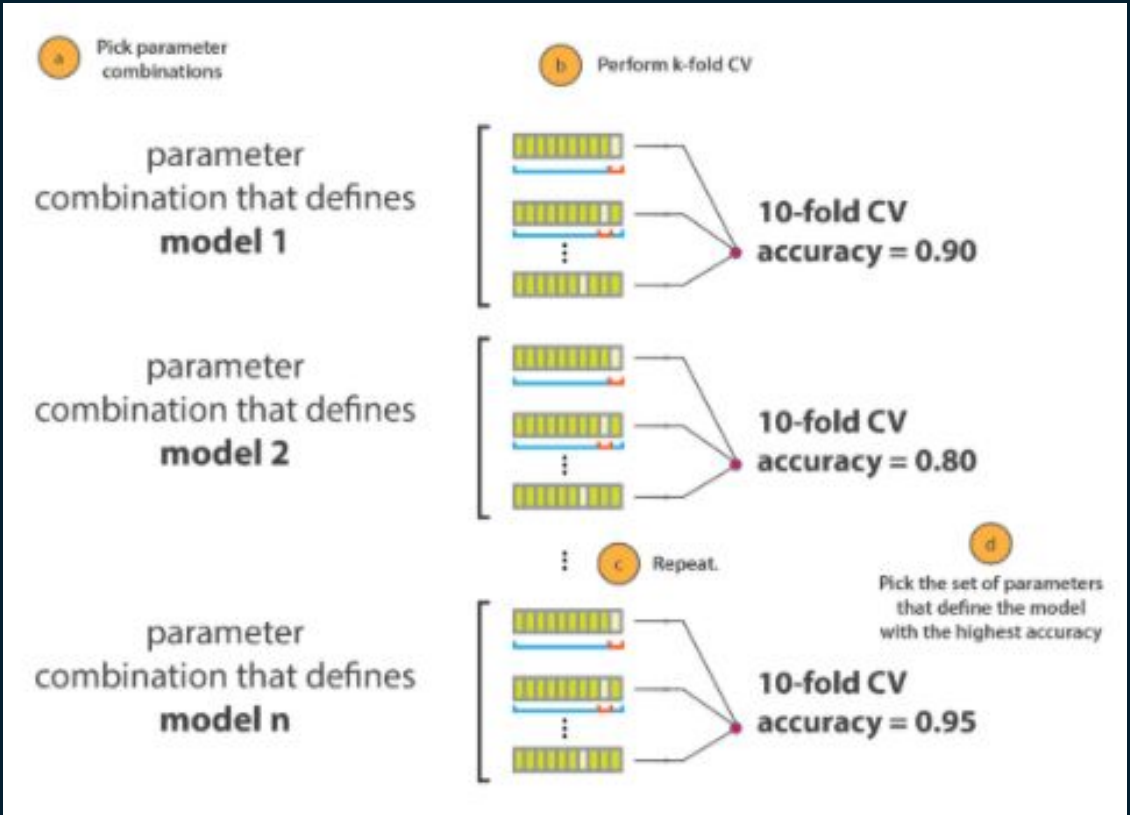
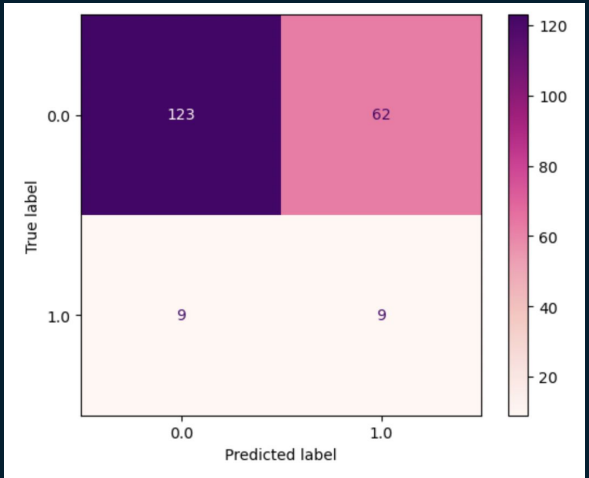


Image Source

Decision Tree



Hyperparameters: Information Gain, Tree Depth = 12

Input: Categorical features selected from the random forest feature selection

The accuracy of the Decision Tree model is 65%

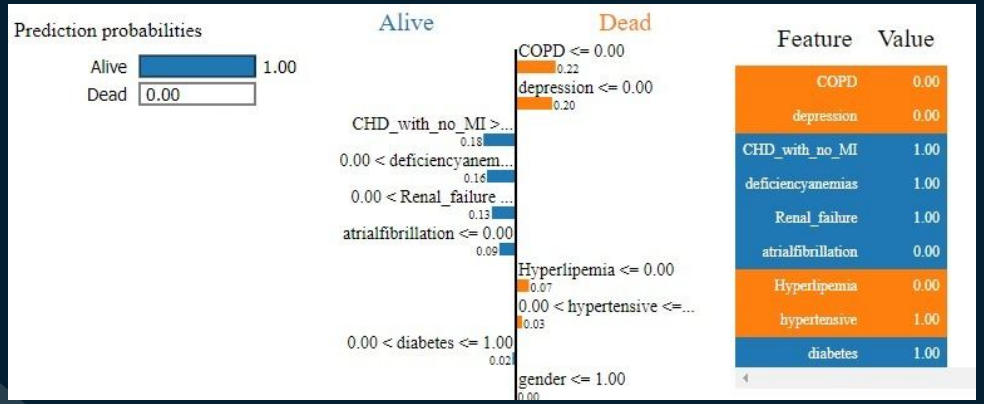
The precision of the model on class 0 is 0.93 and on class 1 is 0.13. Out of all the people that the model predicted would die, about 13% actually did.

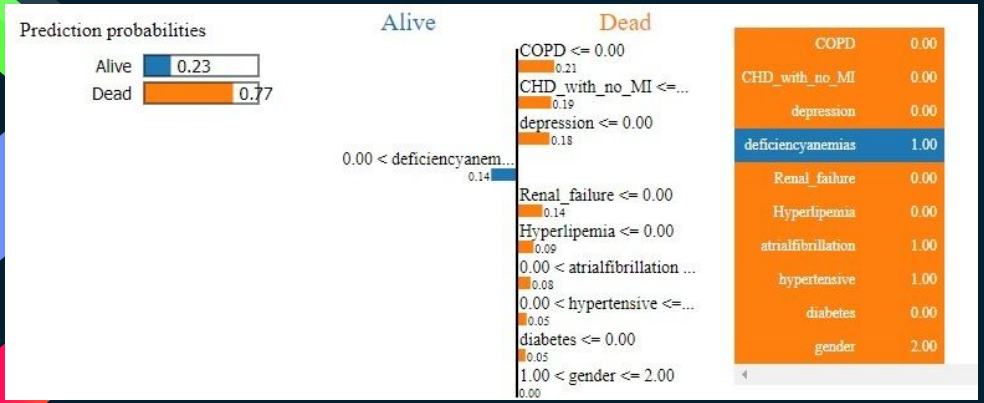
Classification Report:

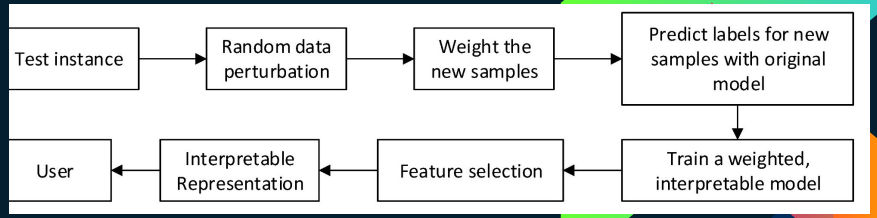
	precision	recall	f1-score	support
0.0	0.93	0.66	0.78	185
1.0	0.13	0.50	0.20	18
accuracy			0.65	203
macro avg	0.53	0.58	0.49	203
weighted avg	0.86	0.65	0.73	203

Decision tree does worse at identifying the actual positive class 1 labeled data points when compared to logistic regression.

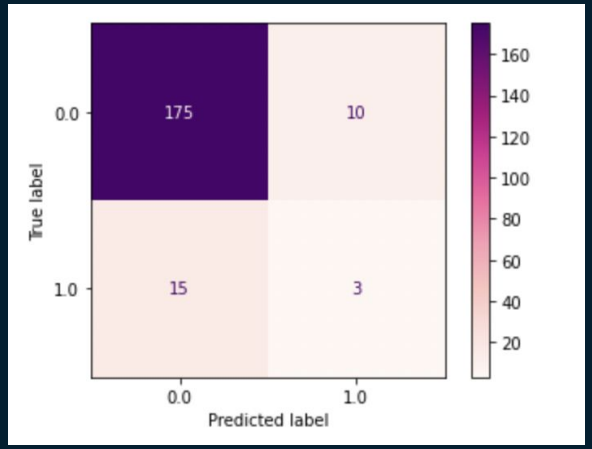
LIME (Local Interpretable Model-Agnostic Explanations)







KNN



Hyperparameters: Euclidean Distance, 2 Neighbors

Input: Both continuous and categorical variables were used for the KNN model.

The KNN model was the best performing model with an accuracy of 88%.

The precision of the model on class 0 is 0.94 and on class 1 is 0.23. Out of all the people that the model predicted would die, about 23% actually did.

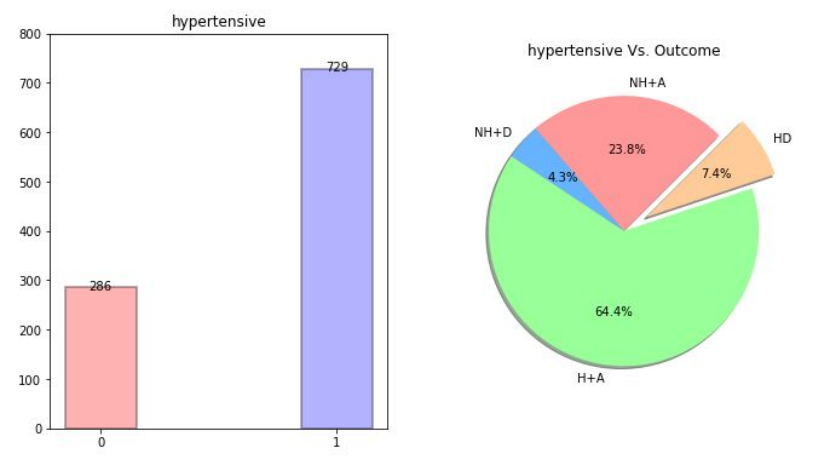
```
Accuracy Score:  0.8768472906403941
Classification Report:
              precision    recall  f1-score   support

         0.0       0.92      0.95      0.93       185
         1.0       0.23      0.17      0.19        18

    accuracy                           0.88       203
   macro avg       0.58      0.56      0.56       203
weighted avg       0.86      0.88      0.87       203
```

This model performs better due to the inclusion of both categorical and numerical variables.

Modeling Conclusions

- Unbalanced dataset resulted in modeling difficulties
  - Upsampling of label 1 results in duplicate data points, so models do not do well with label 1 patients with different attributes
  - If model misclassified one label 1 datapoint, then it will misclassify all potential duplicates
- KNN model accuracy is higher than logistic and decision tree models
  - Additional features being trained on (categorical, continuous)
  - Important features in both sets of features
- Further Analysis
  - Transform continuous features into categorical
  - Obtain data to represent categorical features as continuous
  - Obtain more data points for label 1

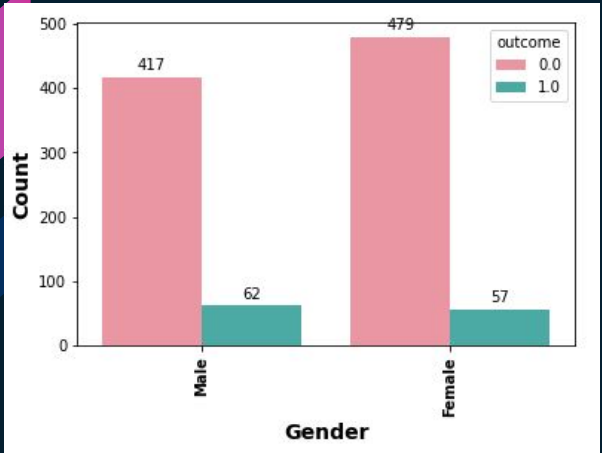
Hypothesis Testing

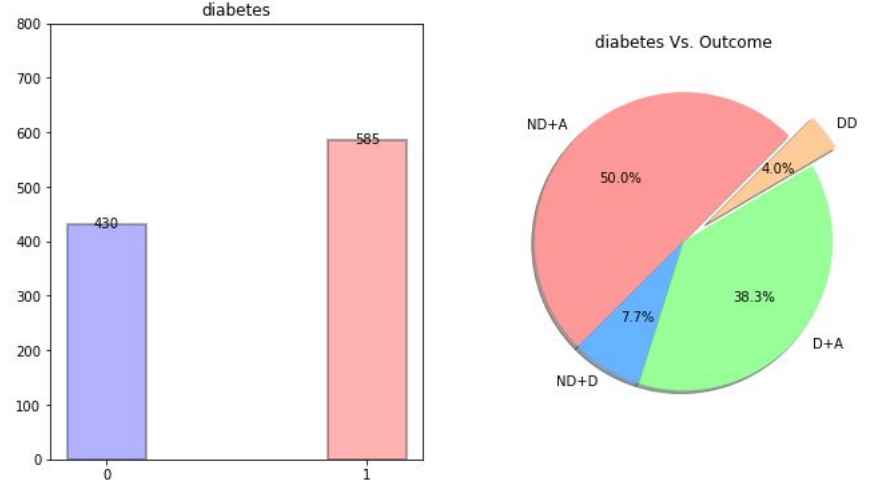


P Value for Hypertensive Vs outcome: 0.0019

P Value for Diabetes Vs outcome: 0.0667

P Value for Gender Vs outcome: 0.5595





Thank You!

Questions?