Predicting Fatal Heart Disease Using ML Models

Team 7b
Antonio Moral, Roozbeh Jafari, Xiang Zhou, Lai Zhang, Kai Guo

Heart failure: A major problem worldwide

Approximately:

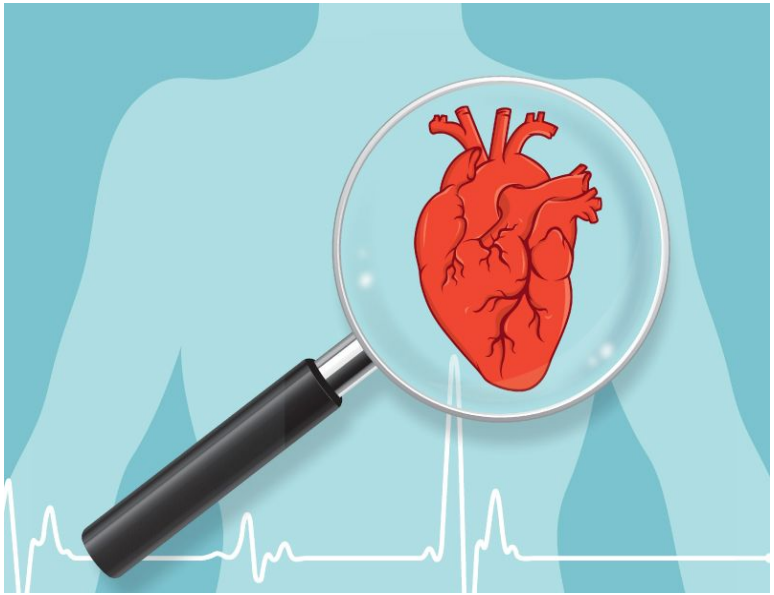
17.9 million

People globally die each year from CVD

Approximately:

85%

Of these deaths are due to heart disease and stroke



Abstract

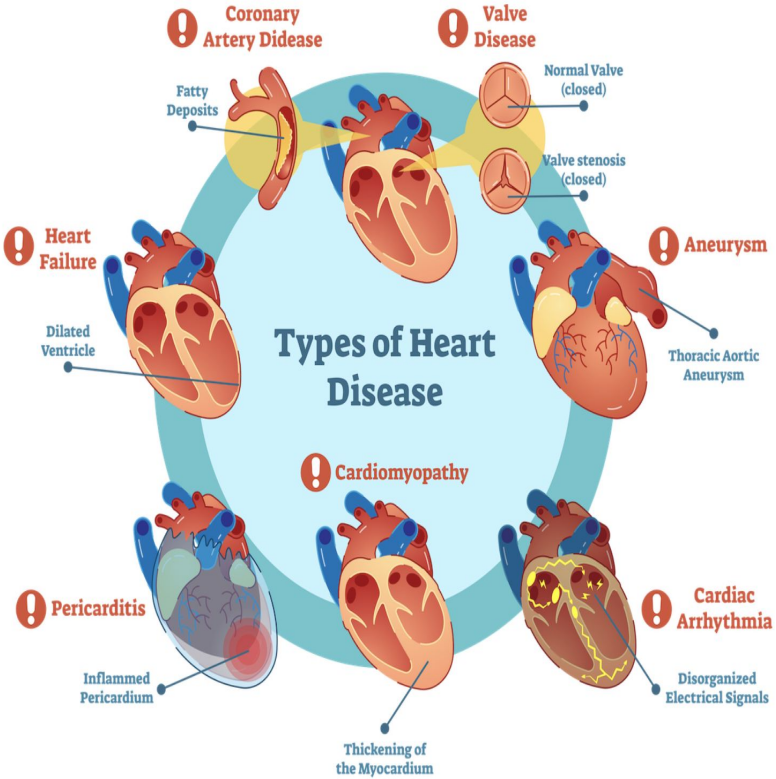
- **Background**
- **Methods**
- **Results**
- **Conclusions**

➔ Number 1 cause of death globally
➔ 31% of all deaths worldwide

➔ EDA
➔ Visualization
➔ Three ML Models

➔ Deducted-feature models
➔ More accurate predictions

➔ New supporting tool for physicians
➔ Focus mainly on serum creatinine and ejection fraction.



What data to analyze?

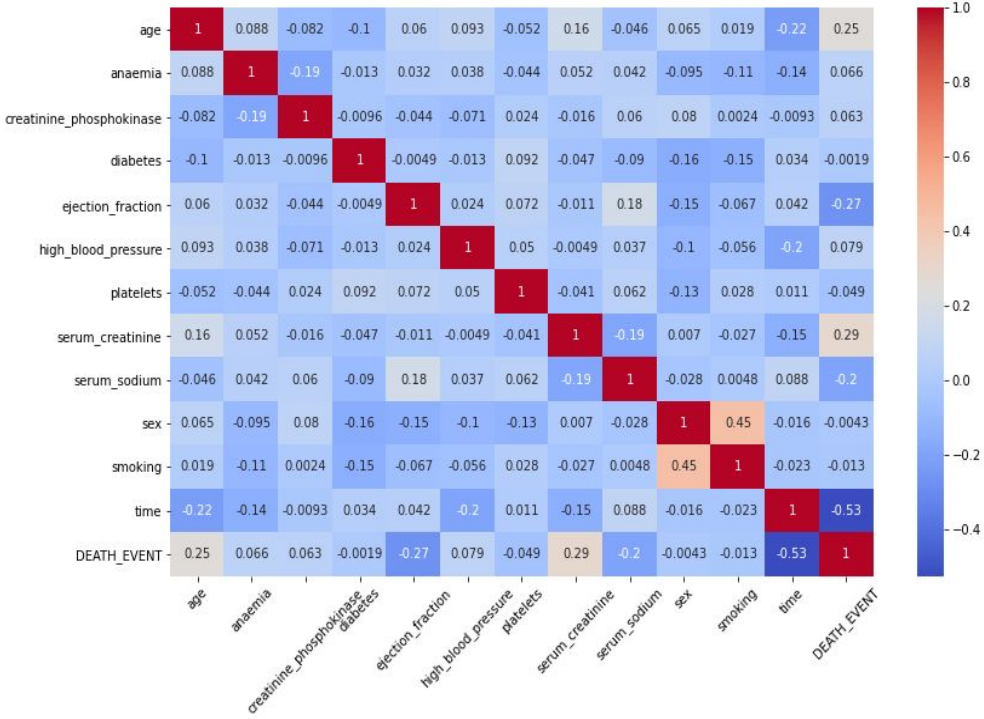
We chose the **Heart Failure Clinical Records Dataset.**

	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
0	75.0	0	582	0	20	1	265000.00	1.9	130	1	0	4	1
1	55.0	0	7861	0	38	0	263358.03	1.1	136	1	0	6	1
2	65.0	0	146	0	20	0	162000.00	1.3	129	1	1	7	1
3	50.0	1	111	0	20	0	210000.00	1.9	137	1	0	7	1
4	65.0	1	160	1	20	0	327000.00	2.7	116	0	0	8	1

```
RangeIndex: 299 entries, 0 to 298
Data columns (total 13 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   age                       299 non-null    float64
 1   anaemia                   299 non-null    int64
 2   creatinine_phosphokinase  299 non-null    int64
 3   diabetes                  299 non-null    int64
 4   ejection_fraction         299 non-null    int64
 5   high_blood_pressure       299 non-null    int64
 6   platelets                 299 non-null    float64
 7   serum_creatinine          299 non-null    float64
 8   serum_sodium              299 non-null    int64
 9   sex                       299 non-null    int64
 10  smoking                   299 non-null    int64
 11  time                      299 non-null    int64
 12  DEATH_EVENT               299 non-null    int64
```

	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
count	299.000000	299.000000	299.000000	299.000000	299.000000	299.000000	299.000000	299.00000	299.000000	299.000000	299.00000	299.000000	299.00000
mean	60.833893	0.431438	581.839465	0.418060	38.083612	0.351171	263358.029264	1.39388	136.625418	0.648829	0.32107	130.260870	0.32107
std	11.894809	0.496107	970.287881	0.494067	11.834841	0.478136	97804.236869	1.03451	4.412477	0.478136	0.46767	77.614208	0.46767
min	40.000000	0.000000	23.000000	0.000000	14.000000	0.000000	25100.000000	0.50000	113.000000	0.000000	0.00000	4.000000	0.00000
25%	51.000000	0.000000	116.500000	0.000000	30.000000	0.000000	212500.000000	0.90000	134.000000	0.000000	0.00000	73.000000	0.00000
50%	60.000000	0.000000	250.000000	0.000000	38.000000	0.000000	262000.000000	1.10000	137.000000	1.000000	0.00000	115.000000	0.00000
75%	70.000000	1.000000	582.000000	1.000000	45.000000	1.000000	303500.000000	1.40000	140.000000	1.000000	1.00000	203.000000	1.00000
max	95.000000	1.000000	7861.000000	1.000000	80.000000	1.000000	850000.000000	9.40000	148.000000	1.000000	1.00000	285.000000	1.00000

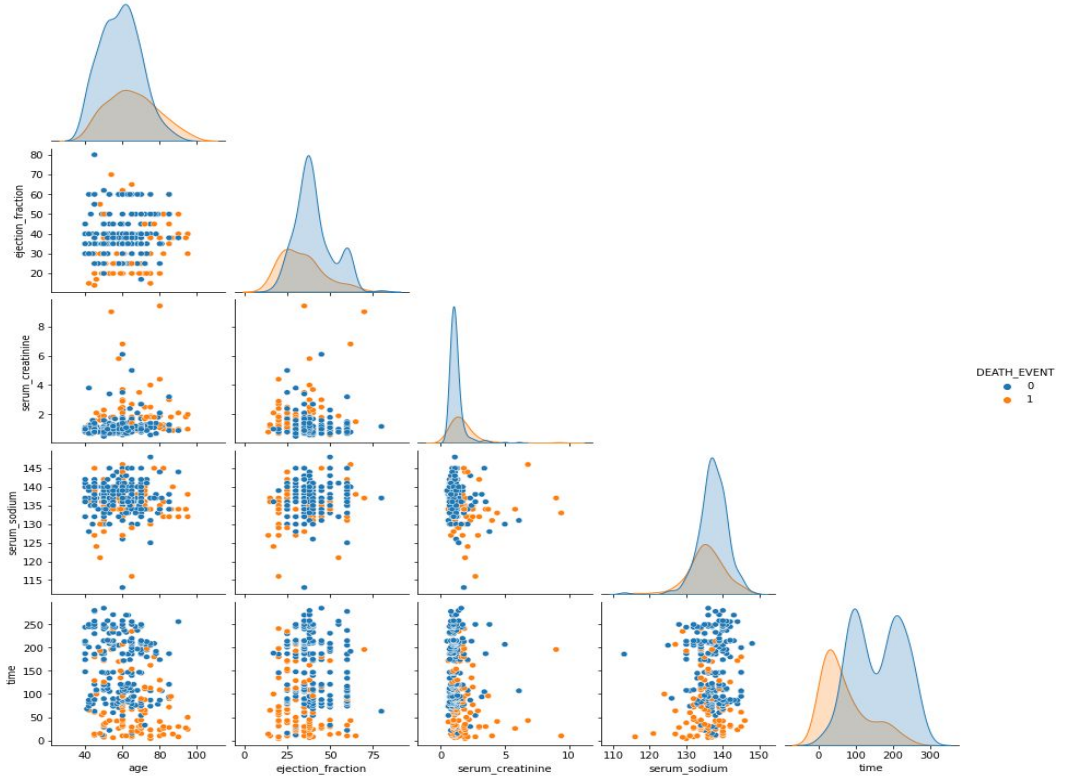
Correlation of Features

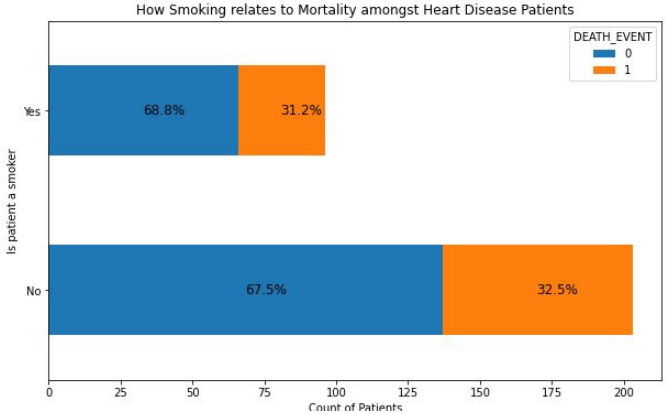


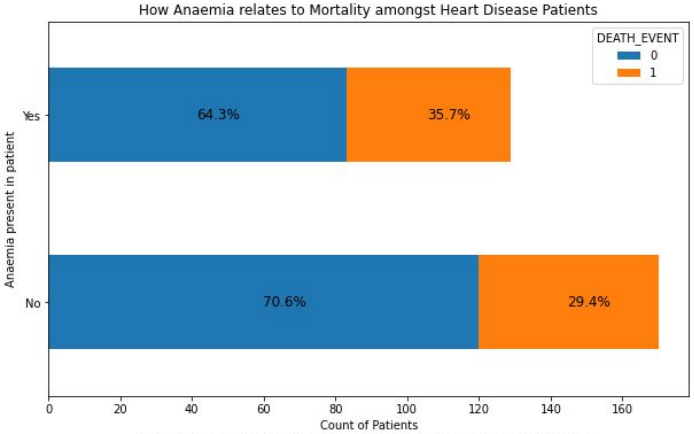
Understanding the features most correlated to fatal heart disease.

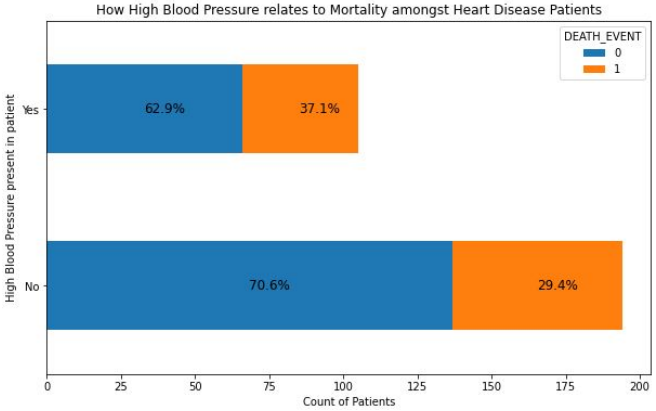
Findings:

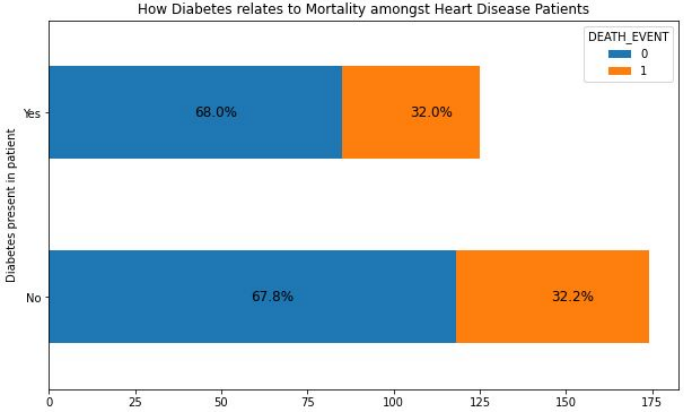
- We can see some initial tendencies such as age being a factor on whether a patient survives.
- Heart failure patients with lower ejection fraction are more likely to die
- The higher the serum creatinine value, patients are more likely to die
- Patients with lower serum sodium value are more likely to die
- Most death events occur during a short period of follow-up time



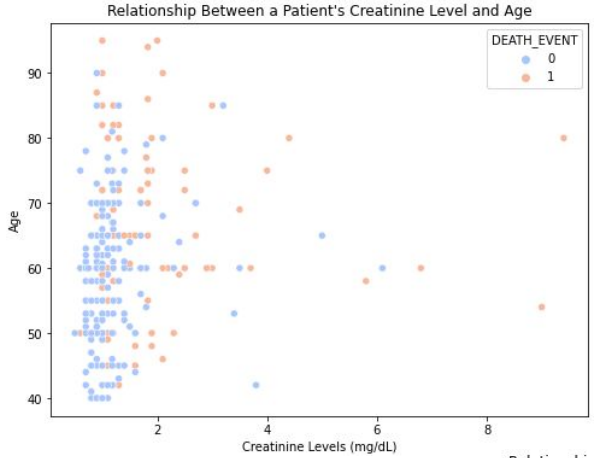


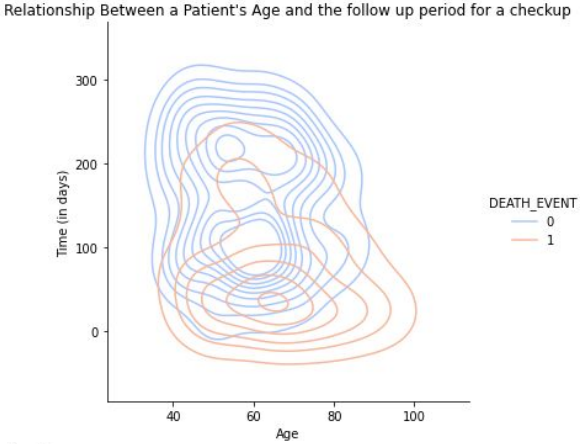


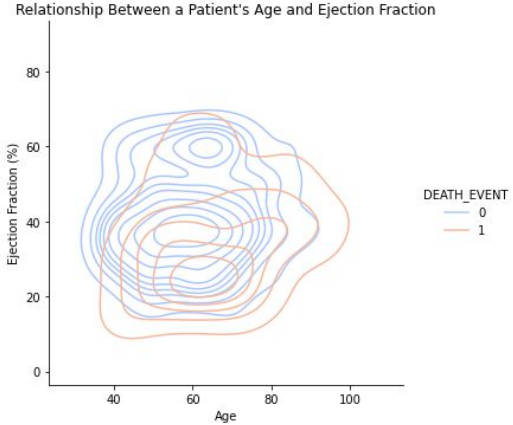




Understanding the binary attributes in the dataset







Examining other relationships between different attributes and a patients age.

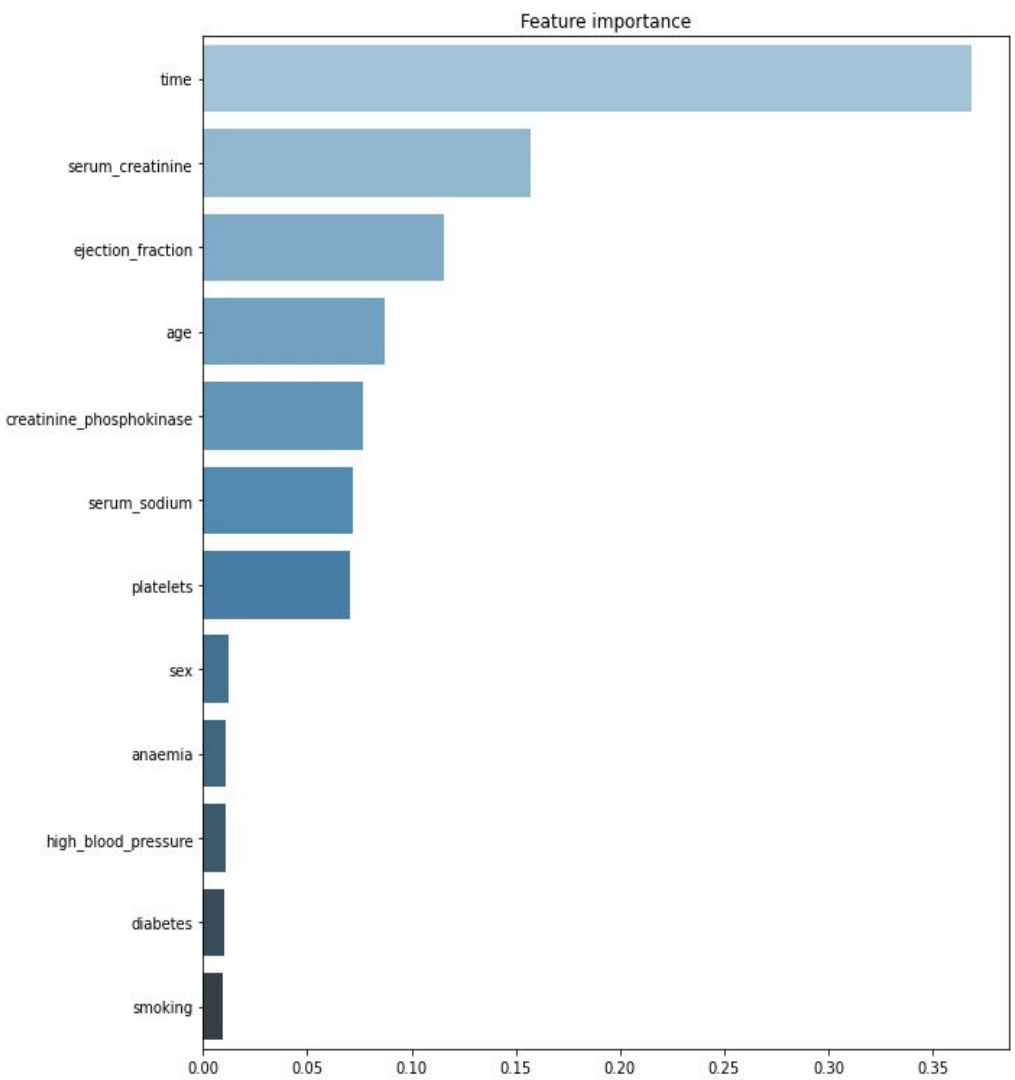
Creating Machine Learning Models to Predict Fatal Heart Disease

Feature Importance Ranking

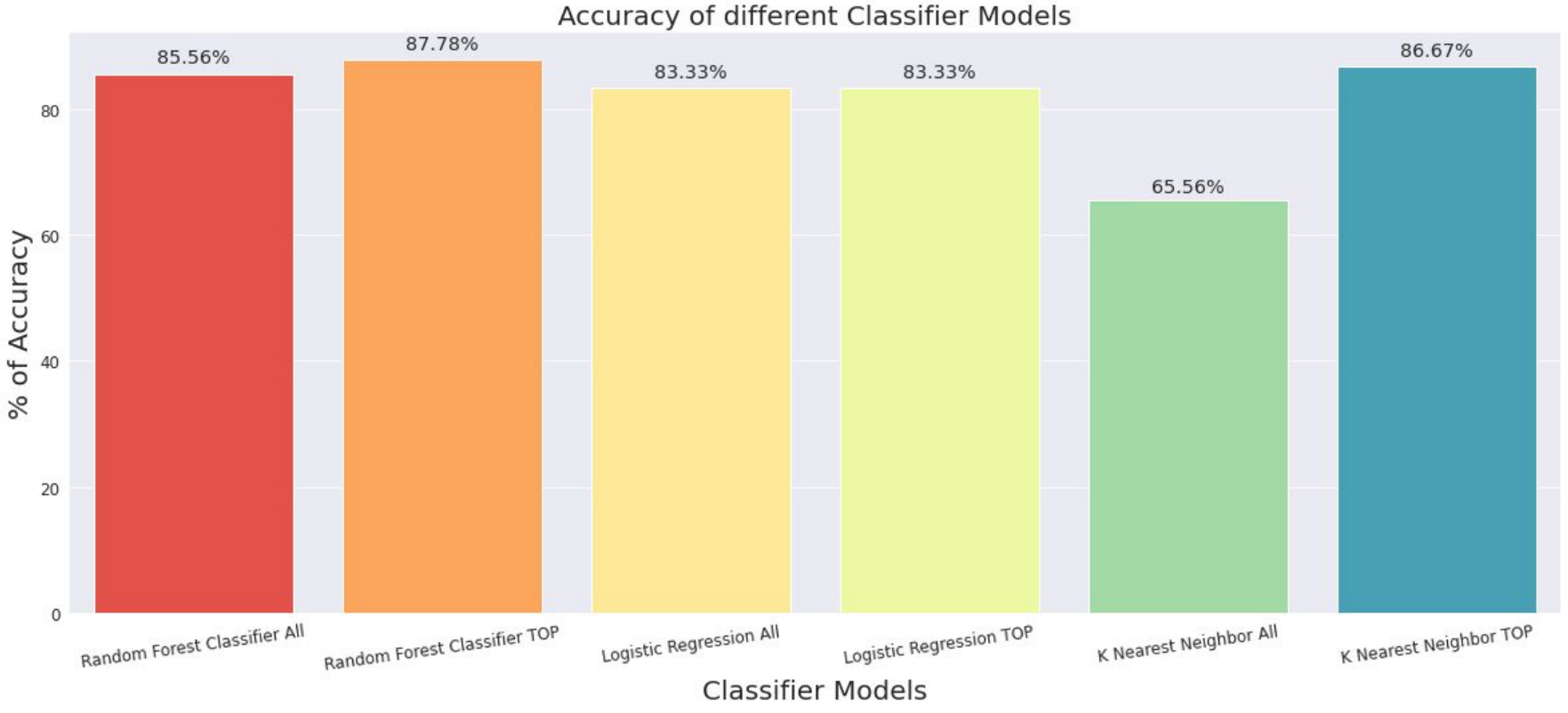
- Random Forest Classifier

Top features:

- time
- serum_creatinine
- ejection_fraction
- age

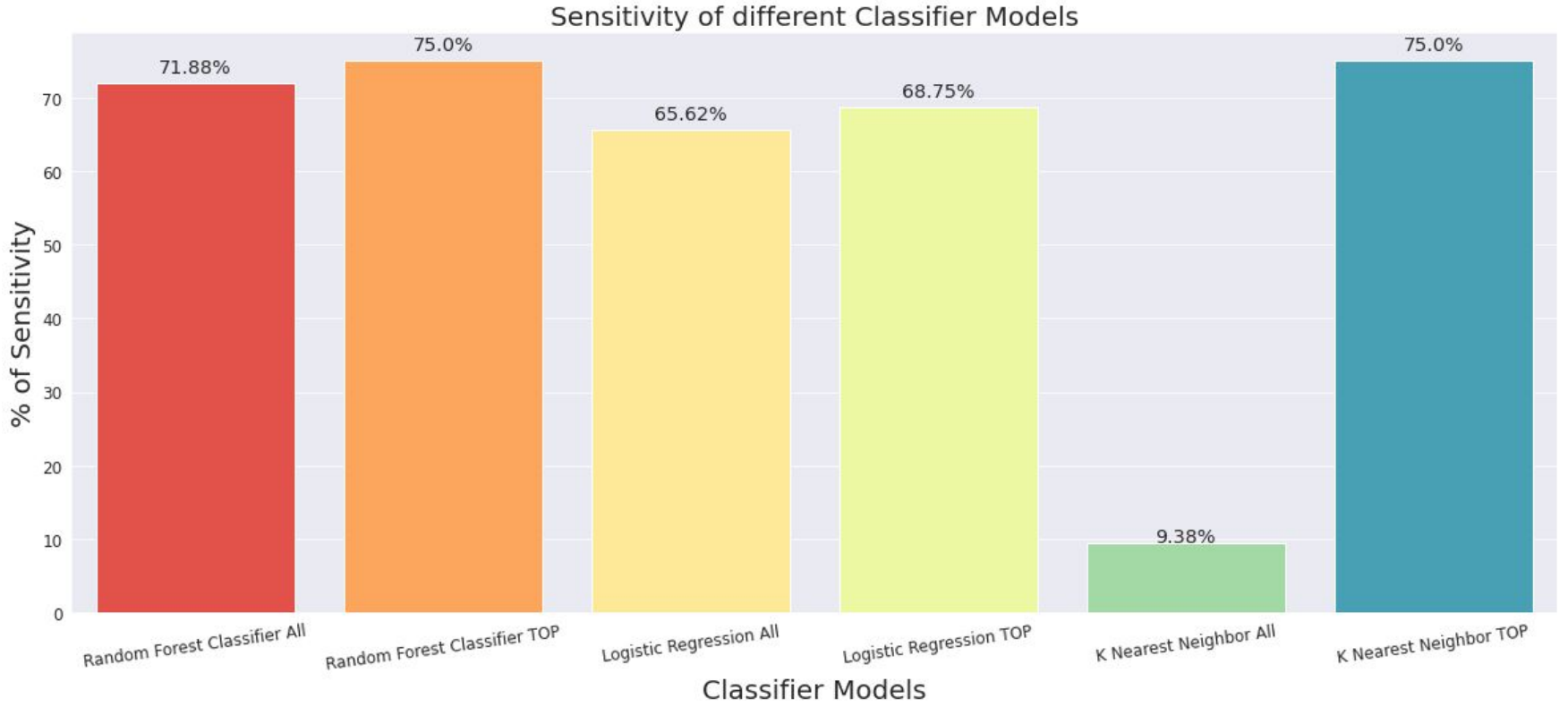


Accuracy of different classifier models



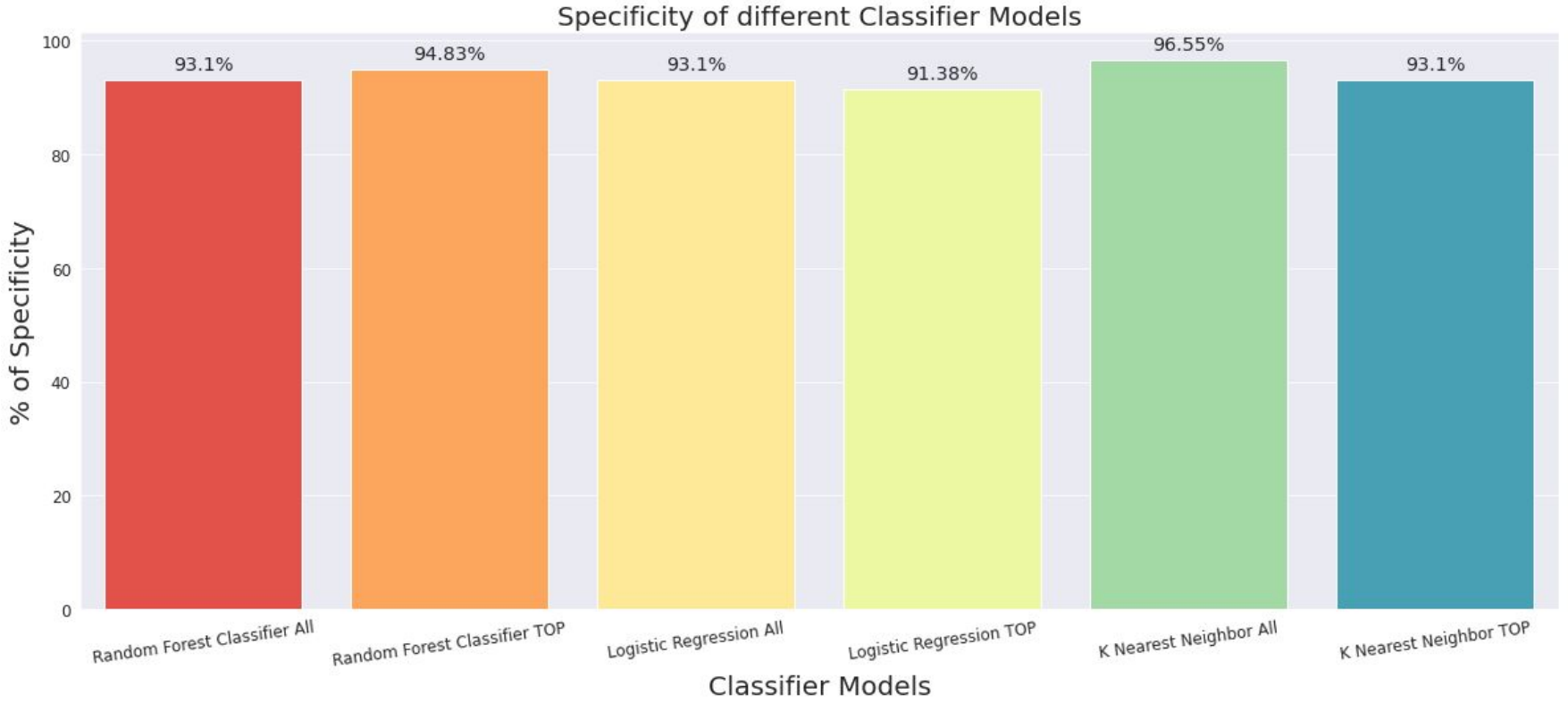
Sensitivity of different classifier models

- TP/P
- How many actually -died patients the model correctly identifies as died?



Specificity of different classifier models

- TN/N
- How many patients survived does the model correctly confirmed as survived?



Conclusion:

This discovery has the potential to impact on clinical practice, becoming a new supporting tool for physicians when predicting if a heart failure patient will survive or not.