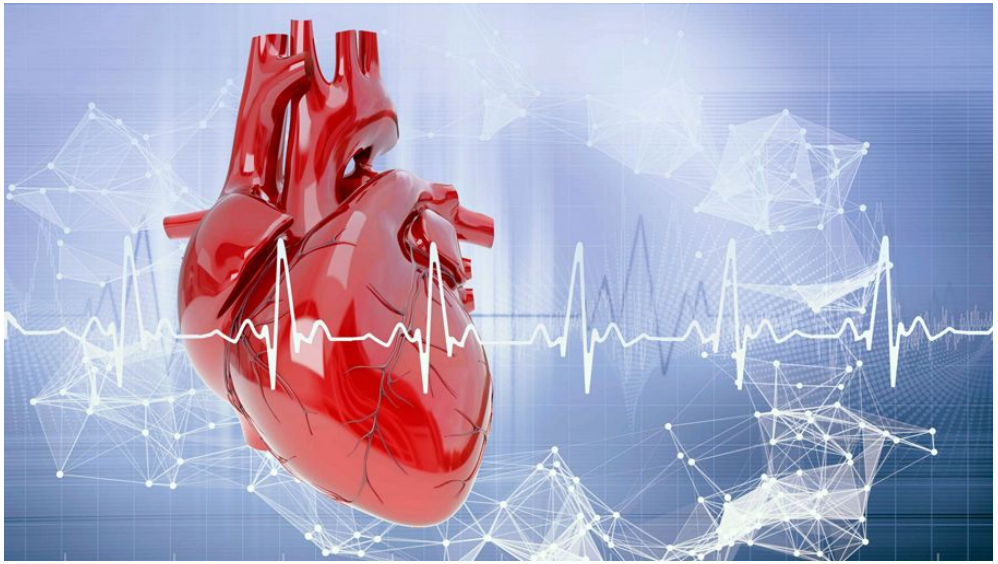
# Predicting Survival of Patients with Heart Failure.



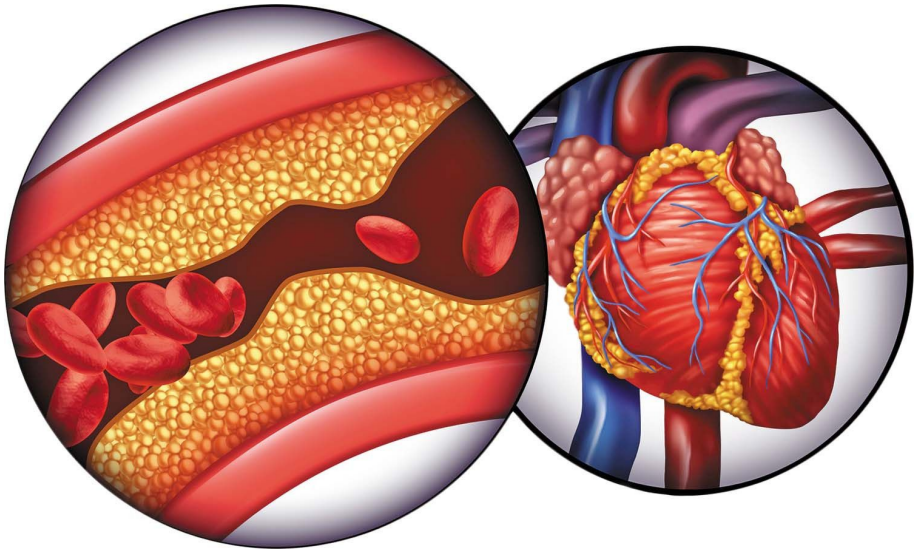
Intended Audience: Healthcare Professionals, Medical Researchers, Students and Educators


DATASCI 207: Applied Machine Learning
Fall 2024
Jasmol Dhesi, Jason Chang, Kent Bourgoing, Sebastian Rosales, Sergey Nam

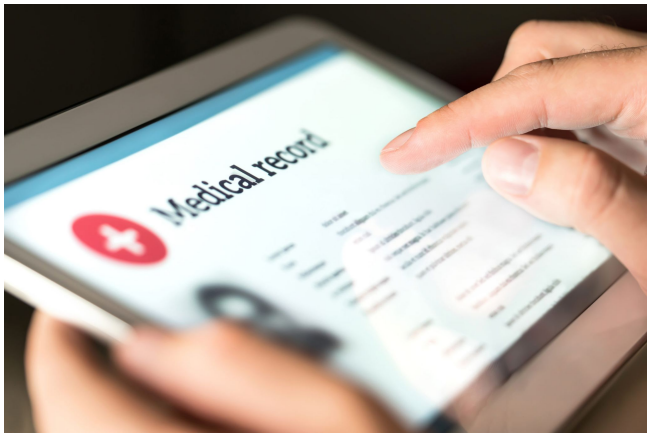

Berkeley
UNIVERSITY OF CALIFORNIA

# Motivation

Cardiovascular diseases, such as heart failure, are responsible approximately **17 million** deaths worldwide each year (WHO, 2021).



# Motivation (cont.)




SPRINGER NATURE Link

Find a journal | Publish with us | Track your research | Search

Home > Medical & Biological Engineering & Computing > Article

**Predicting heart failure in-hospital mortality by integrating longitudinal and category data in electronic health records**

Original Article | Published: 24 March 2023

Volume 61, pages 1857–1873, (2023) Cite this article

Original Investigation | Cardiology

January 10, 2020

**Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims With Electronic Medical Records to Predict Heart Failure Outcomes**

Rishi J. Desai, MS, PhD[1]; Shirley V. Wang, PhD[1]; Muthiah Vaduganathan, MD, MPH[2]; et al

Author Affiliations | Article Information

*JAMA Netw Open.* 2020;3(1):e1918962. doi:10.1001/jamanetworkopen.2019.18962

Machine Learning Website

BMC Part of Springer Nature

BMC Medical Informatics and Decision Making

Home About Articles Submission Guidelines Collections Join The Board Submit manuscript

Research Article | Open access | Published: 03 February 2020

**Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone**

Davide Chicco & Giuseppe Jurman

*BMC Medical Informatics and Decision Making* **20**, Article number: 16 (2020) | Cite this article

# Research Question

**How accurately can machine learning models predict survival outcomes among heart failure patients based on clinical features extracted from electronic medical records?**

- **Secondary Interest:** How does the predictive accuracy of different machine learning algorithms compare in this context?

Berkeley
UNIVERSITY OF CALIFORNIA

# Dataset Overview

- **Source:** 299 heart failure patients, Faisalabad, Pakistan (2015)
- **Collected from:** Faisalabad Institute of Cardiology & Allied Hospital
- **Features:** 13 clinical, demographic, and lifestyle factors
- **Notable Factors:** Age, sex, ejection fraction, serum creatinine, anemia, diabetes, high blood pressure
- **Target:** Survival outcome (died or survived)
- **Follow-up period:** Average of 130 days
- **Demographics:** 105 women, 194 men, ages 40–95

# Features Of Interest

- **Age:** Older age increases risk for cardiovascular events and complicates heart failure management, making it a key mortality predictor.
- **High Blood Pressure:** Elevated blood pressure indicates that the heart is working harder and over long periods of time can lead to chronic heart failure increasing the mortality risk.
- **Ejection Fraction (EF):** Low EF reflects reduced cardiac efficiency and is strongly linked to heart failure severity, essential for assessing patient prognosis.
- **Platelets:** Abnormal platelet levels can indicate cardiovascular complications like thrombosis, impacting survival rates in heart failure.
- **Serum Creatinine:** Elevated levels indicate impaired kidney function, which is closely linked to poor outcomes in heart failure patients.
- **Serum Sodium:** Low sodium (hyponatremia) often marks fluid imbalance and disease severity, impacting mortality risk.
- **Time (Follow-Up Period in Days):** Represents patient monitoring follow-up period, helping assess disease progression and survival trends over time.

# Data Preprocessing and Augmentation

- **No major data quality issues:**
  - No null values
  - No duplicate records
  - No data type issues
- **Preprocessing:**
  - Shuffling records
  - Split data using a 60/20/20 train test and validation split
  - Standardize data using the training dataset
- **Augmentation:**
  - Use Bayesian Information Criterion (BIC) to fit a Gaussian Mixture Model (GMM) with the most optimal amount of clusters
  - Generate 5000 samples from the GMM to augment dataset from 299 to 5299 records
  - Clean up binary data columns and standardize samples with original training dataset

# Project Plan

**Baseline Model** → Majority Classification

**Supervised ML Models**

**Parametric ML Models**

- Logistic Regression
- Neural Network
- Functional API

**Non-Parametric ML Models**

- K-Nearest Neighbors (KNN)
- Decision Tree
- Majority Vote (Ensemble)
- Random Forest
- Bagging
- Gradient Boosting
- Adaptive Boosting (AdaBoost)

**Unsupervised ML Models**

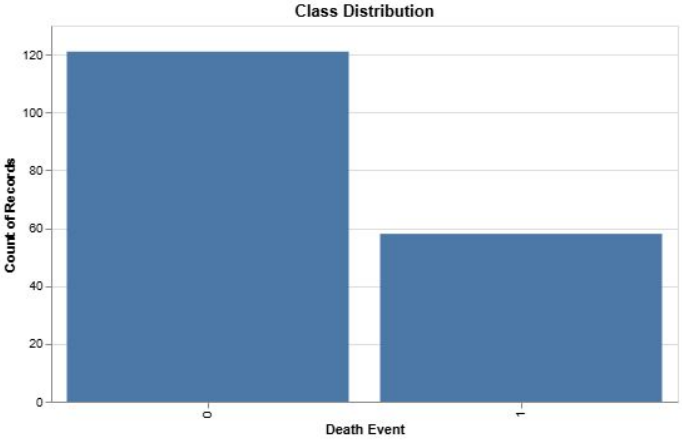
- K-Means Clustering
- Agglomerative Hierarchical Clustering
- Density-Based Spatial Clustering (DBSCAN)
- Gaussian Mixture Model (Data Augmentation)
- PCA/SVD*

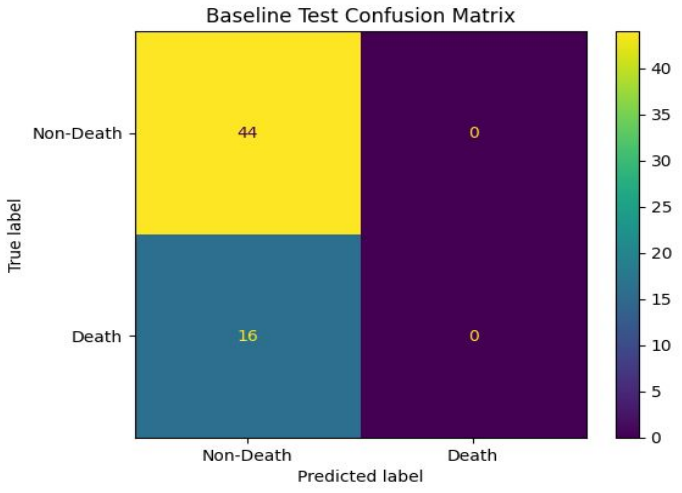
# Model Performance Metrics

- **Chosen Metric**: Accuracy
  - Reason: Simple and suitable as our dataset is balanced between death and survival cases.
- **Alternative Metrics**
  - Precision: Useful if false positives (predicting death when survival occurs) should be minimized.
  - Recall: Essential if catching all true death cases is critical, even with some false positives.
  - F1 Score: Balances precision and recall, ideal for imbalanced datasets.
  - AUC-ROC: Measures model's ability to distinguish classes, helpful if class imbalance is present.
  - Loss: Helpful in understanding the amount of errors in the test set.

# Baseline Model

- Model: Majority Class Classifier
- Within the training dataset we have a total of 179 records
  - 121 non-deaths
  - 58 deaths
- Training Accuracy: **67.59%**
- Training Loss: **11.94**
- Validation Accuracy: **63.33%**
- Validation Loss: **13.51**
- Test Accuracy: **73.33%**
- Test Loss: **9.82**



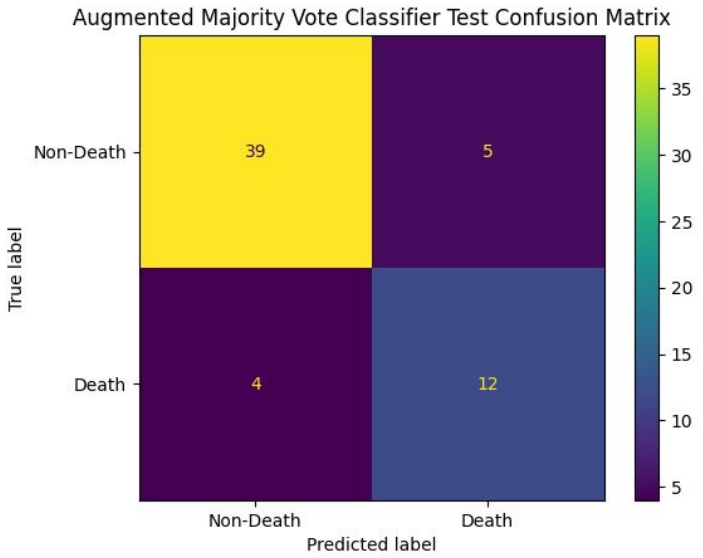
# Model Results

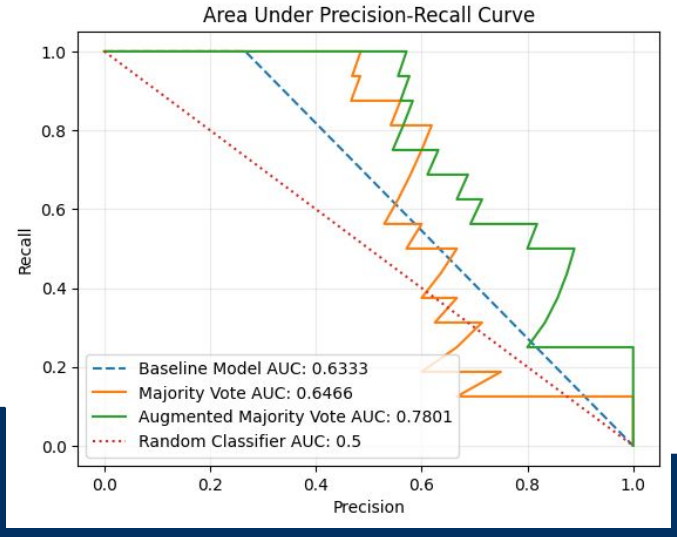
| | Model | Train Accuracy | Validation Accuracy | Test Accuracy | Test Loss | Test Recall | Test Precision |
|---|---|---|---|---|---|---|---|
| 1 | Augmented Gradient Boosting | 0.883568 | 0.883333 | 0.850000 | 0.272829 | 0.6250 | 0.769231 |
| 2 | Augmented Voting Classifier | 0.966017 | 0.833333 | 0.850000 | 0.333343 | 0.7500 | 0.705882 |
| 3 | Augmented Decision Tree | 0.778336 | 0.783333 | 0.800000 | 0.335736 | 0.4375 | 0.700000 |
| 4 | Random Forest | 0.960894 | 0.866667 | 0.866667 | 0.340767 | 0.6875 | 0.785714 |
| 5 | Augmented Logistic Regression | 0.770033 | 0.833333 | 0.866667 | 0.351654 | 0.6250 | 0.833333 |
| 6 | Augmented Sequential Neural Network | 0.818112 | 0.883333 | 0.833333 | 0.352741 | 0.5625 | 0.750000 |
| 7 | Logistic Regression | 0.837989 | 0.800000 | 0.816667 | 0.365822 | 0.5625 | 0.692308 |
| 8 | Augmented Random Forest | 0.793976 | 0.850000 | 0.816667 | 0.391763 | 0.4375 | 0.777778 |
| 9 | Voting Classifier | 0.960894 | 0.766667 | 0.783333 | 0.402732 | 0.4375 | 0.636364 |
| 10 | Augmented KNN | 0.837807 | 0.850000 | 0.766667 | 0.418960 | 0.3750 | 0.600000 |
| 11 | AdaBoost | 0.865922 | 0.850000 | 0.850000 | 0.520810 | 0.6875 | 0.733333 |
| 12 | Gradient Boosting | 1.000000 | 0.850000 | 0.816667 | 0.548312 | 0.5625 | 0.692308 |
| 13 | Augmented AdaBoost | 0.863294 | 0.866667 | 0.833333 | 0.610284 | 0.5000 | 0.800000 |
| 14 | Augmented Bagging Classifier | 0.983008 | 0.883333 | 0.866667 | 0.884097 | 0.6250 | 0.833333 |
| 15 | Bagging Classifier | 0.983240 | 0.833333 | 0.800000 | 0.926520 | 0.5000 | 0.666667 |
| 16 | KNN | 0.832402 | 0.783333 | 0.800000 | 0.987466 | 0.4375 | 0.700000 |
| 17 | Sequential Neural Network | 0.977654 | 0.816667 | 0.783333 | 1.033662 | 0.6250 | 0.588235 |
| 18 | Augmented Functional Neural Network | 0.789342 | 0.816667 | 0.900000 | 3.604365 | 0.7500 | 0.857143 |
| 19 | Functional Neural Network | 0.849162 | 0.833333 | 0.833333 | 6.007276 | 0.7500 | 0.666667 |
| 20 | Decision Tree | 0.944134 | 0.816667 | 0.783333 | 6.671077 | 0.5625 | 0.600000 |
| 21 | Baseline | 0.675978 | 0.633333 | 0.733333 | 9.824363 | 0.0000 | 0.000000 |

| | Test Accuracy | Validation Accuracy | Test Accuracy | Test Loss | Test Recall | Test Precision |
|---|---|---|---|---|---|---|
| Mean | 88.85% | 83.67% | 82.42% | 1.3574 | 56.25% | 72.44% |
| SD | 7.66% | 3.32% | 3.17% | 0.0317 | 10.92% | 8.63% |

# Augmented Voting Classifier Model

- Model: Augmented Majority Vote
- Training Accuracy: 0.9660
- Training Loss: 0.2201
- Validation Accuracy: 0.8333
- Validation Loss: 0.3735
- Test Accuracy: 0.8500
- Test Recall: 0.7500
- Test Precision: 0.7059

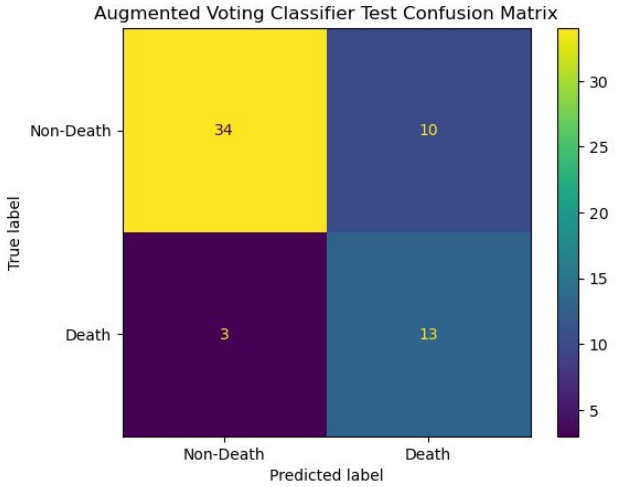



Berkeley
UNIVERSITY OF CALIFORNIA

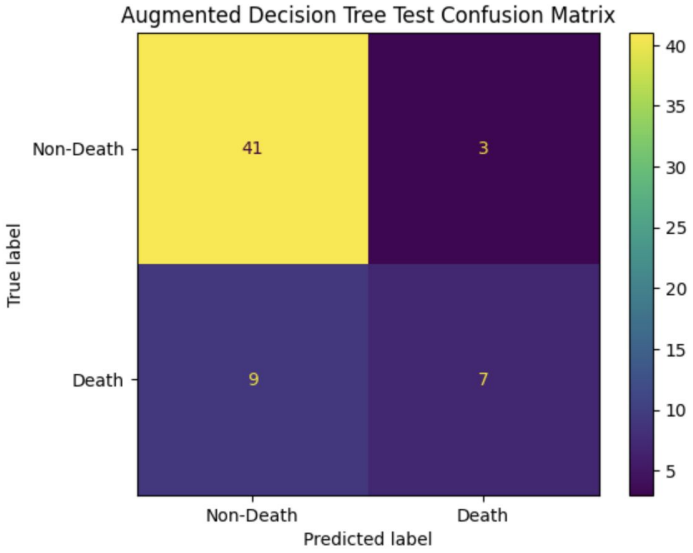
# Augmented Voting Classifier Model

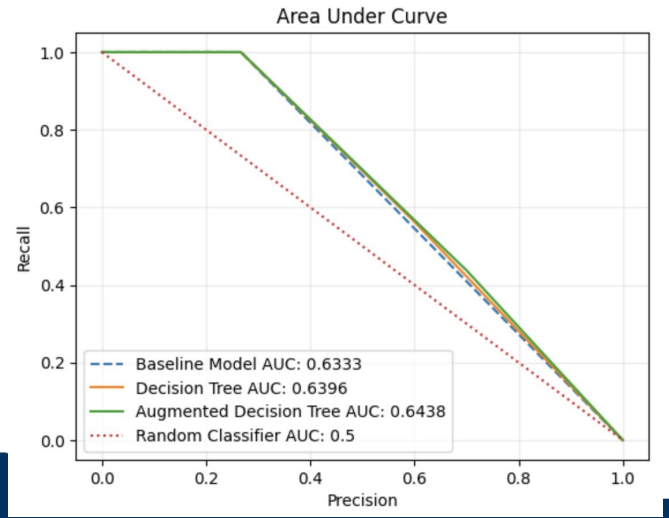
- Improved
  - **Probability-Based Voting (Soft):** Instead of using hard voting, soft voting helped make a weighted decision. This improved performance by leveraging the confidence of each estimator.
  - **Decision Boundary**: Lowering the threshold to 0.25 shifts our focus to improving recall. Reducing false negatives is imperative when working with mortality prediction.
  - **Cross-Validation**: Tested different values of nearest neighbors and verified that k=5 provided the best balance of performance metrics. (accuracy, precision, and recall)



# Augmented Decision Tree Model

- Model: Augmented Decision Tree
- Training Accuracy: 0.7783
- Training Loss: 0.4677
- Validation Accuracy: 0.7833
- Validation Loss: 0.4098
- Test Accuracy: 0.8000
- Test Recall: 0.4375
- Test Precision: 0.7000



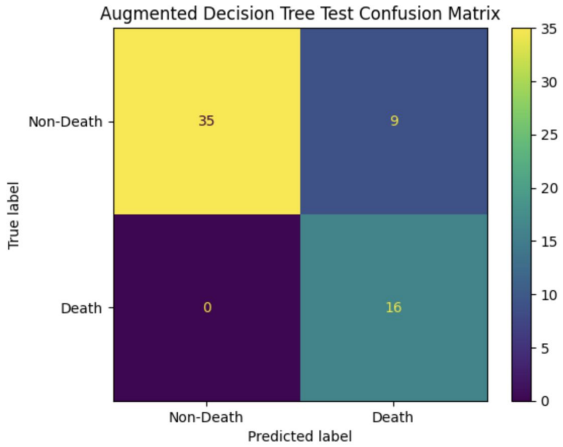

Berkeley
UNIVERSITY OF CALIFORNIA

# Augmented Decision Tree Model

```
dt_model = DecisionTreeClassifier(max_depth=5, random_state=1234)
```

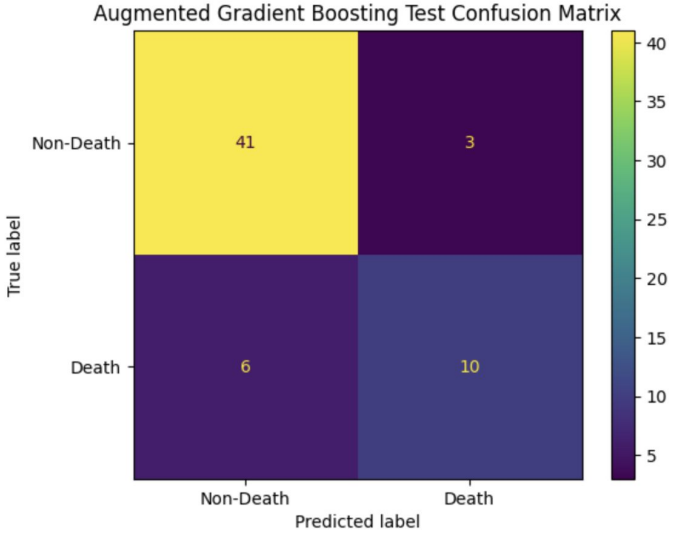
- Hyperparameter Tuning
  - **Max Depth:** The maximum depth of the tree determines the number of levels it can expand. The model with `max_depth=5` achieved the best performance by striking a balance between simplicity and complexity, avoiding overfitting while capturing sufficient patterns to generalize effectively to unseen data. Smaller values led to underfitting, while larger values caused overfitting.
  - **Decision Boundary:** When predicting death events for heart patients, a threshold of 0.3 is more suitable in situations where missing an at-risk patient could result in fatal consequences. However, this approach increases the likelihood of false positives, whic must be carefully assessed based on the available resources and the potential impact of unnecessary interventions.

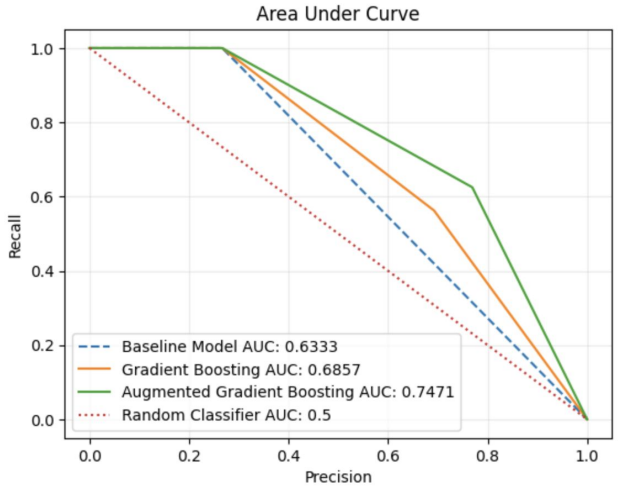
```
new_aug_dt_test_pred = [1 if i > .3 else 0 for i in(aug_dt_model.predict_proba(X_test))[:,1]]
```



# Augmented Gradient Boosting Model

- Model: Augmented Gradient Boosting
- Training Accuracy: 0.8836
- Training Loss: 0.3068
- Validation Accuracy: 0.8833
- Validation Loss: 0.2926
- Test Accuracy: 0.85
- Test Recall: 0.6250
- Test Precision: 0.7692



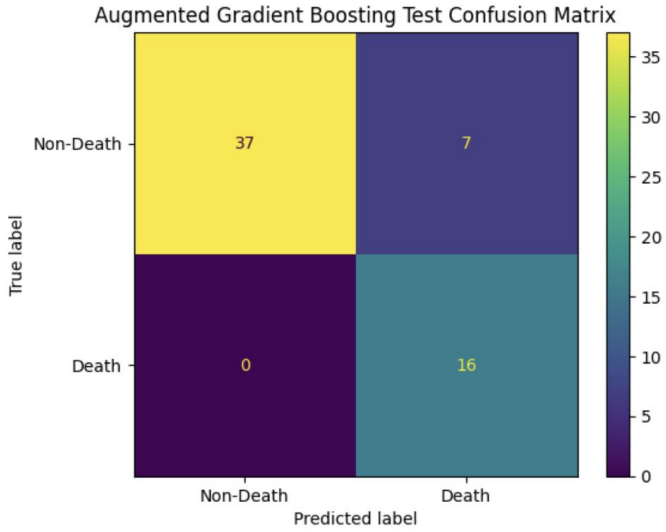




Berkeley
UNIVERSITY OF CALIFORNIA

# Augmented Gradient Boosting Model

```
gb_model = GradientBoostingClassifier(max_depth = 4, n_estimators=200,random_state=1234)
```

- Hyperparameter Tuning
  - **Max Depth:** The choice of `max_depth=4` limits the complexity of individual trees, preventing overfitting while allowing the model to capture meaningful patterns in the data.
  - **Number Estimators:** The choice of `n_estimators=200` strikes a balance between improving the model's ability to learn complex patterns and avoiding overfitting or excessive training time. It provides sufficient iterations for boosting to refine predictions without overcomplicating the model.
  - **Decision Boundary:** Reducing the threshold from 0.5 to 0.25 was chosen to increase recall by capturing all true positive cases ("Death") while tolerating more false positives ("Non-Death" incorrectly classified as "Death"). This approach is suitable in contexts like predicting death events for heart patients, where missing true positive cases could have severe consequences, even at the expense of slightly lower precision.




```
new_aug_gb_test_pred = [1 if i > .25 else 0 for i in(aug_gb_model.predict_proba(X_test))[:,1]]
```

# Clustering Approach

Objective:

- Identify patterns and structures in the data
- Anomaly detection
- Data Simplification

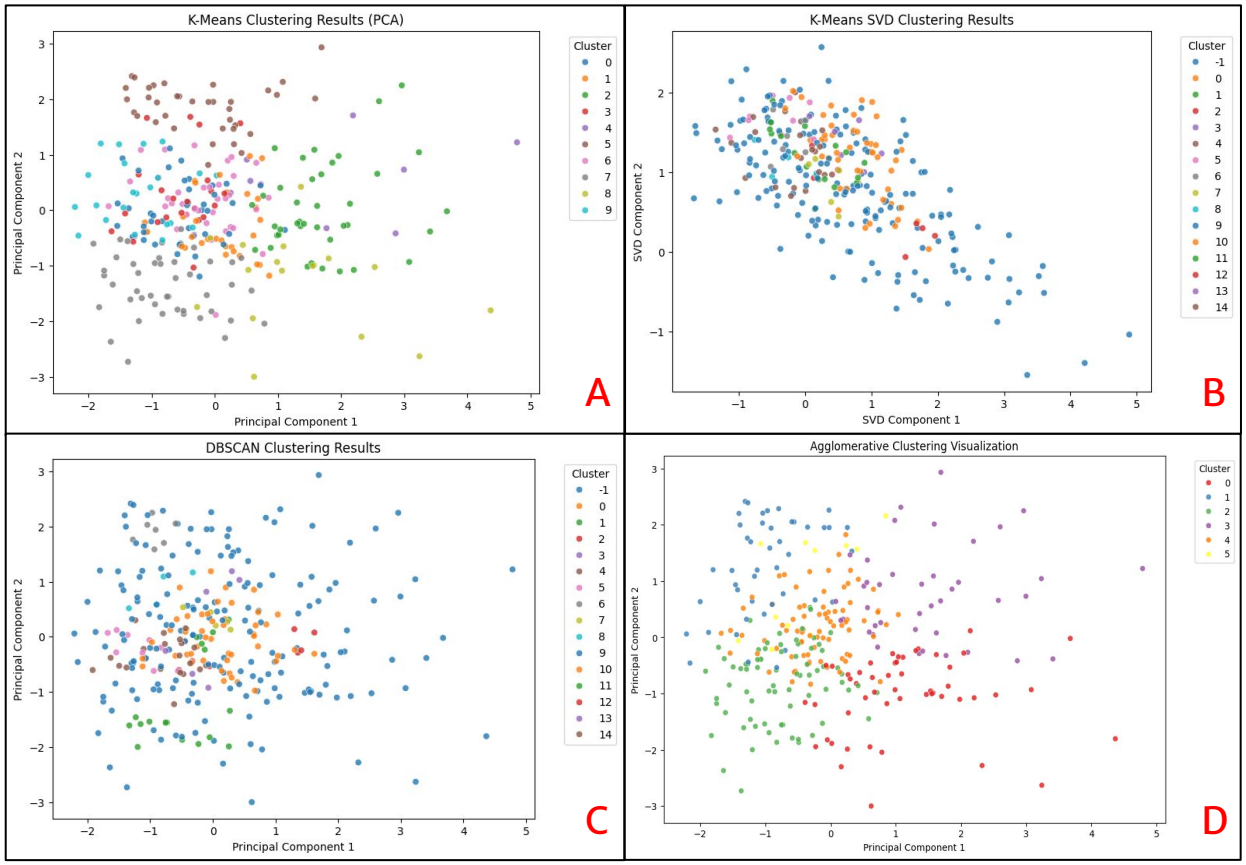
Algorithms Used:

- K-Means, Density-Based Clustering, Agglomerative Clustering

Dimensionality Reduction Techniques:

- Principal Component Analysis (PCA), Singular Value Decomposition (SVD)

# Clustering Results (PCA)



# Clustering Scoring (Silhouette Scores)

| Method | Parameters | Silhouette Score (Higher is Better), [-1,1] |
|---|---|---|
| K Means | K = 2 | 0.1866 |
| K Means (SVD) | K = 5 | 0.2255 |
| DBSCAN | Epsilon = 0.7, min_samples = 3 | -0.2212 |
| Agglomerative | K = 6 | 0.1688 |

Silhouette score was used as the metric to evaluate the clustering algorithms. K-Values were decided by selecting the k values associated with optimal silhouette scores.

- A silhouette score measures how well-separated and compact your clusters are in a dataset. It ranges from -1 to 1, with higher values indicating better-defined clusters.

The k-means algorithm with SVD applied had the highest silhouette score

# Clustering Conclusions

- Agglomerative Clustering had the highest silhouette score HOWEVER it wasn't indicative of the data being good for clustering.
- A desirable silhouette score is >= 0.5 for defined clusters; no algorithms yielded sufficient results.
- As shown in the reduced plots, clusters have significant overlap preventing us from drawing meaningful insights.
- Unsupervised Learning was not useful in this context.

It is challenging to determine whether a heart failure patient will die or survive simply by clustering the data, as the groups are not easily separable and individuals from different outcomes often share overlapping clinical features.

# Conclusion

- Machine learning models predicted mortality risk in HF patients with an average accuracy of 85% (SD 5%).
- Non-parametric models outperformed parametric models, with a lower average test loss of 0.98 (SD 1.66) vs. 2.25 (SD 2.70).
- Augmented models performed better than non-augmented models, with a test loss of 0.99 (SD 1.77) vs. 1.72 (SD 2.31).
- Augmented Gradient Boosting was the top model, with accuracy ≥ 0.85, test loss ~0.27, and <3% difference across training, validation, and test accuracies.

# Conclusion

## Potential Biases

- **Augmented Data Validity:** The GMM-based data augmentation may not fully capture the true distribution, potentially affecting result reliability.
- **Generalization Bias:** Limited to two hospitals in Pakistan, with an age bias (40–95 years) and gender imbalance (194 males vs. 105 females), restricting broader applicability.

# Conclusion

- Models **did not exceed 90%** accuracy across training, validation, and testing datasets.
- Predictions should **complement**, not replace, healthcare professionals' clinical judgment

Future Work:

- Explore advanced machine learning algorithms beyond DATASCI 207 for improved accuracy.
- Focus on hyperparameter tuning for flexible models like Sequential Neural Networks and Functional API.
- Conduct a deeper bias and fairness analysis.

# References


Data Set Link from UCI ML Repo:
https://archive.ics.uci.edu/dataset/519/heart+failure+clinical+records

Davide Chicco, Giuseppe Jurman: "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone". BMC Medical Informatics and Decision Making 20, 16 (2020). https://doi.org/10.1186/s12911-020-1023-5

World Health Organization. (2021, June 11). Cardiovascular diseases (cvds). World Health Organization.
https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)#:~:text=Key%20facts%201%20Cardiovascular%20diseases%20%28CVDs%29%20are%20the,2019%2C%2038%25%20were%20caused%20by%20CVDs.%20More%20items

Mayo Clinic. (2023, April 20). Heart failure.
https://www.mayoclinic.org/diseases-conditions/heart-failure/symptoms-causes/syc-20373142

# GitHub Repository

https://github.com/JasmolSD/207_007_final_project

# Individual Contributions

**Jasmol Dhesi**: K-Means, DBSCAN, Agglomerative Clustering, SVD, PCA

**Jason Chang**: KNN, Majority Vote, Bagging

**Kent Bourgoing:** Slides, Functional API Neural Network, AdaBoost, and model result statistics

**Sebastian Rosales**: Data preprocessing, Data augmentation with Gaussian Mixture Model, Baseline Model, Logistic Regression, Sequential Neural Network, Decision Boundary exploration

**Sergey Nam**: Decision Tree, Random Forest, Gradient Descent, Zoom meetings