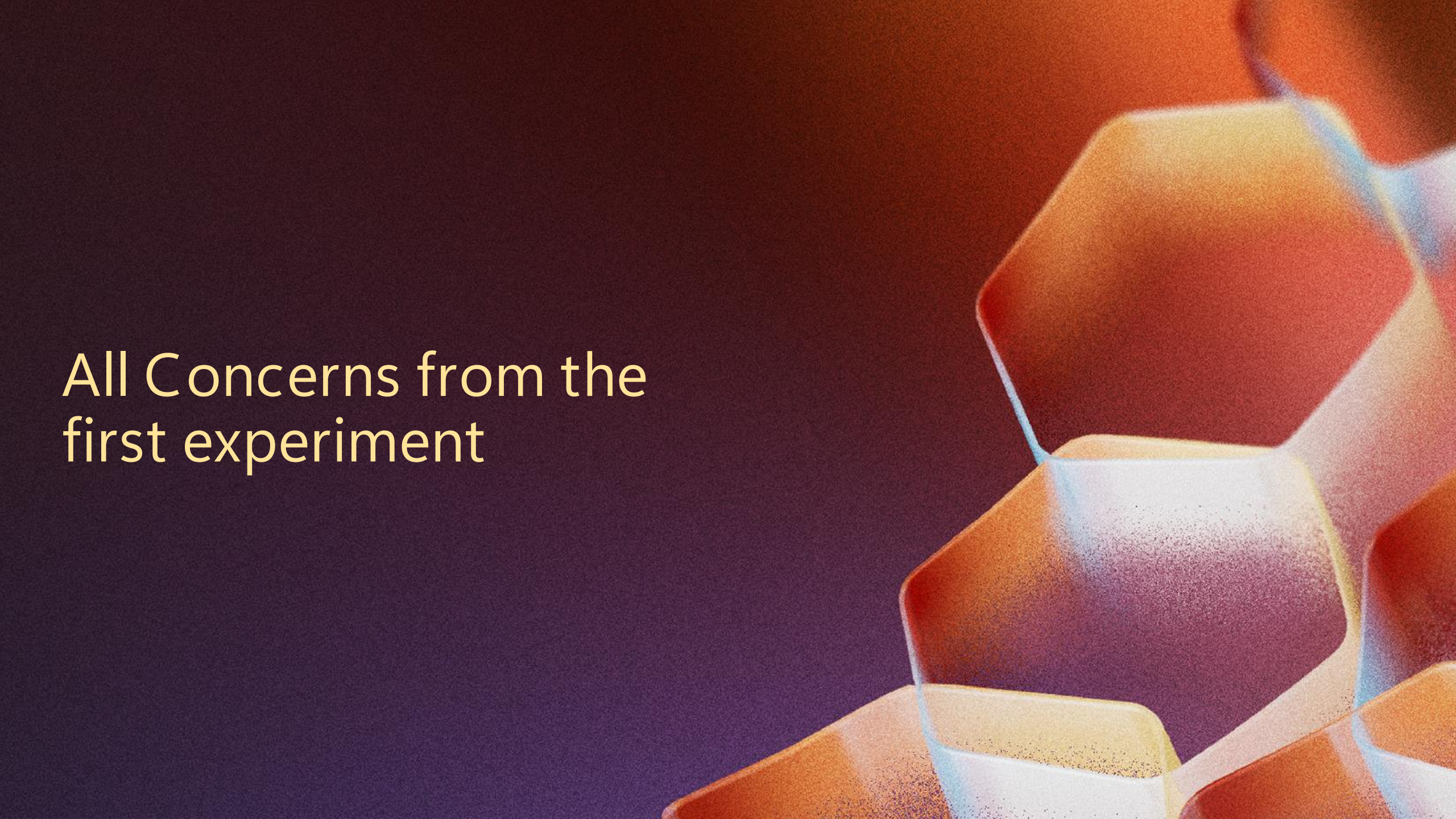
Classification Grouping Heart Disease Patients In The Hospital

By Charunthon Limseelo and Bhagya Saranunt
Advised by Assoc. Prof. Dr. Naruemon Wattanapongsakorn

This presentation is one of the assessment activity of CPE456: Data Mining
Course of Computer Engineering Department (CPE)
Faculty of Engineering, King Mongkut's University of Technology Thonburi

All Concerns from the first experiment



Concern No. 1

Less Ground Truth in the Dataset

Less Ground Truth in the Dataset

- Clustering cardiovascular diseases (CVDs) is challenging due to the complexity of the data and the need for expert knowledge to define meaningful groups. CVDs are often characterized by a wide range of risk factors and clinical presentations, making it difficult to simply apply standard clustering algorithms. Experts are crucial for defining what constitutes a meaningful cluster, as the "ground truth" (the known, correct grouping) may not be immediately apparent from the data alone.

Concern No. 2

There are more factors to consider before clustering

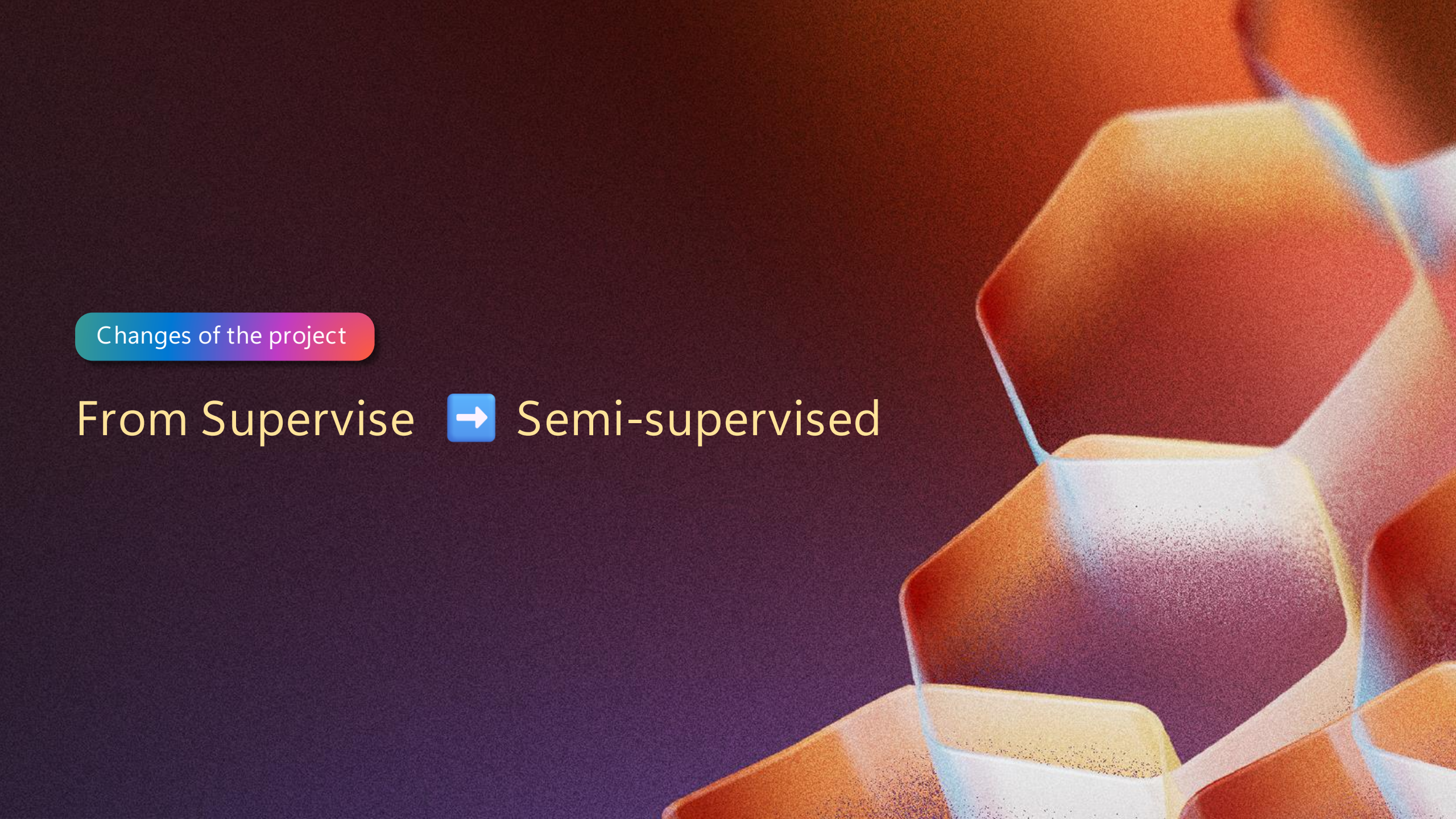
The Main Reason

- Clustering of Cardiovascular Diseases (CVDs) is beneficial when considering imaging data, detailed cardiac biomarkers, and ECG wave data, as it allows for the identification of distinct subgroups of CVDs based on these factors. This helps in developing more targeted diagnostic and treatment approaches.

- Our final approach will be changed from clustering patients by types of disease/root problems to the patients' severity level.

Changes of the project

From Supervise ➡️ Semi-supervised



From Supervise to Semi-supervised

The most suitable way for clustering as changing from classifying heart disease type to heart disease severity. As it has less ground truth, we are now decided to be on Semi-supervised training instead.

More Reference Added

The Severity Prediction of the Binary and Multi-Class Cardiovascular Disease – A Machine Learning-Based Fusion Approach

THE SEVERITY PREDICTION OF THE BINARY AND MULTI-CLASS CARDIOVASCULAR DISEASE - A MACHINE LEARNING-BASED FUSION APPROACH

Hafsa Binte Kibria
Department of Electrical & Computer Engineering
Rajshahi University of Engineering & Technology,
Rajshahi-6204,Bangladesh
hafsabintekibria@gmail.com

Abdul Matin
Department of Electrical & Computer Engineering
Rajshahi University of Engineering & Technology,
Rajshahi-6204,Bangladesh
ammuaj.cseruet@gmail.com

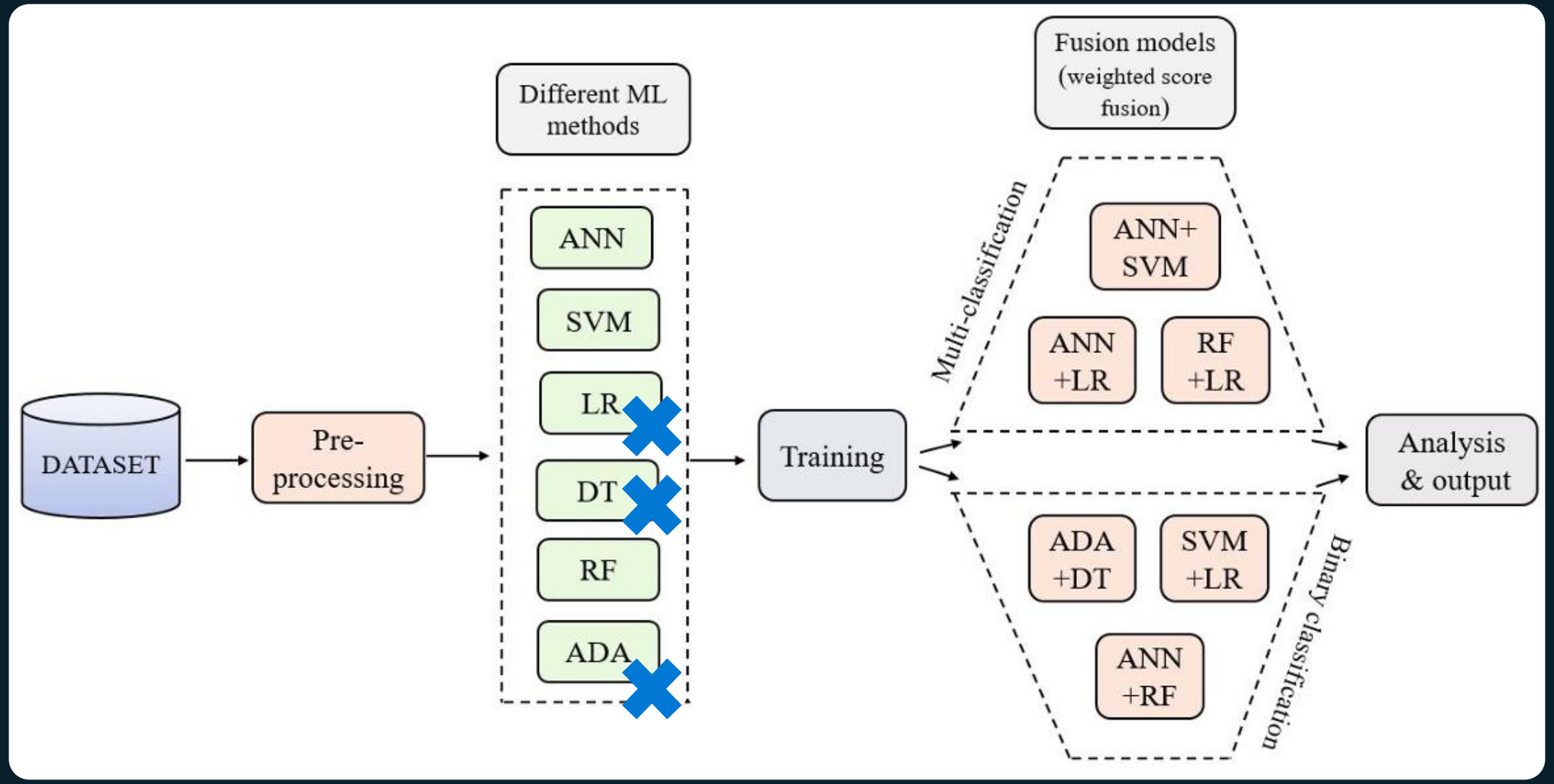
March 10, 2022

ABSTRACT

In today's world, a massive amount of data is available in almost every sector. This data has become an asset as we can use this enormous amount of data to find information. Mainly health care industry contains many data consisting of patient and disease-related information. By using the machine learning technique, we can look for hidden data patterns to predict various diseases. Recently CVDs, or cardiovascular disease, have become a leading cause of death around the world. The number of death due to CVDs is frightening. That is why many researchers are trying their best to design a predictive model that can save many lives using the data mining model. In this research, some fusion models have been constructed to diagnose CVDs along with its severity. Machine learning(ML) algorithms like artificial neural network, SVM, logistic regression, decision tree, random forest, and AdaBoost have been applied to the heart disease dataset to predict disease. Randomoversampler was implemented because of the class imbalance in multiclass classification. To improve the performance of classification, a weighted score fusion approach was taken. At first, the models were trained. After training, two algorithms' decision was combined using a weighted sum rule. A total of three fusion models have been developed from the six ML algorithms. The results were promising in the performance parameter. The proposed approach has been experimented with different test training ratios for binary and multiclass classification problems, and for both of them, the fusion models performed well. The highest accuracy for multiclass classification was found as 75%, and it was 95% for binary. The code can be found in : github/hafsakibria/Weightedscorefusion

https://arxiv.org/pdf/2203.04921

Proposed Architecture



Feature Engineering Analysis

- The Null values have been handled.
- The data are NOT Standardizing, Normalizing, or Scaling

Binary-class Data

- 1918 entries with labeled target

Multiclass Data

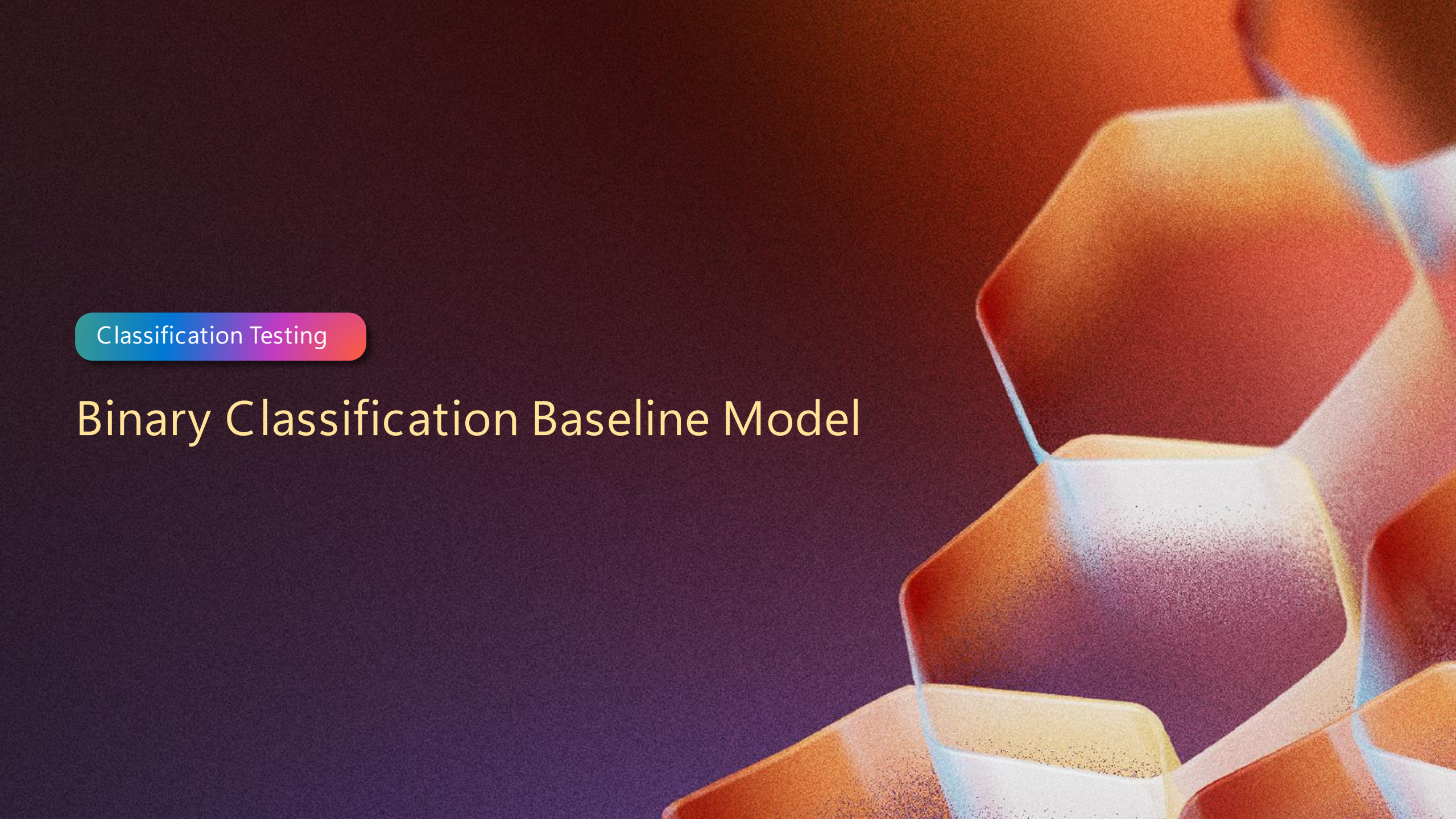
- 587 labeled data from Europe hospital
- 777 unlabeled data from India hospital

Used Features For Analysis – Attributes/Columns

- age --> Age of the patient (28-77 in years).
- gender --> Gender of the patient: typically 1 = Male, 0 = Female.
- chestpain --> Type of chest pain experienced (encoded):• 1 = Typical Angina• 2 = Atypical Angina• 3 = Non-Anginal Pain• 4 = Asymptomatic
- restingBP --> Resting Blood Pressure (in mm Hg) when the patient is seated and relaxed.
- serumcholestrol --> Serum cholesterol level (in mg/dl).
- fastingbloodsugar --> Whether the patient's fasting blood sugar is > 120 mg/dl:• 1 = Yes• 0 = No
- restingrelectro --> Results of resting electrocardiographic test (ECG) (encoded):• 0 = Normal• 1 = ST-T wave abnormality• 2 = Possible left ventricular hypertrophy
- maxheartrate --> Maximum heart rate achieved during exercise test.
- exerciseangia --> Whether the patient experienced exercise-induced angina:• 1 = Yes• 0 = No
- oldpeak --> ST depression induced by exercise relative to rest. (Numeric value showing how much the ST segment dropped below baseline after exercise.)
- slope --> Slope of the peak exercise ST segment (encoded):• 1 = Upsloping• 2 = Flat• 3 = Downsloping
- target --> Diagnosis of heart disease:• 0 = No heart disease• 1 = Presence of heart disease

Classification Testing

Binary Classification Baseline Model



Model Accuracy

Support Vector Machine

```
SVM Multi-class Clasification
accuracy: 0.7013888888888888
[[ 72 172]
 [  0 332]]
              precision    recall  f1-score   support

         0.0       1.00      0.30      0.46       244
         1.0       0.66      1.00      0.79       332

    accuracy                           0.70       576
   macro avg       0.83      0.65      0.62       576
weighted avg       0.80      0.70      0.65       576

roc_auc_RF 0.6475409836065573
```

- Accuracy: 0.7013888

Random Forest

```
Random Forrest Binary Clasification
accuracy: 0.9079861111111112
[[222  34]
 [ 19 301]]
              precision    recall  f1-score   support

         0.0       0.92      0.87      0.89       256
         1.0       0.90      0.94      0.92       320

    accuracy                           0.91       576
   macro avg       0.91      0.90      0.91       576
weighted avg       0.91      0.91      0.91       576

roc_auc_RF 0.90390625
```

- Accuracy: 0.907986111

Artificial Neural Network

```
ANN
accuracy:  0.8385416666666666
[[184  72]
 [ 21 299]]
              precision    recall  f1-score   support

         0.0       0.90      0.72      0.80       256
         1.0       0.81      0.93      0.87       320

    accuracy                           0.84       576
   macro avg       0.85      0.83      0.83       576
weighted avg       0.85      0.84      0.84       576

roc_auc_ann 0.8265625
```

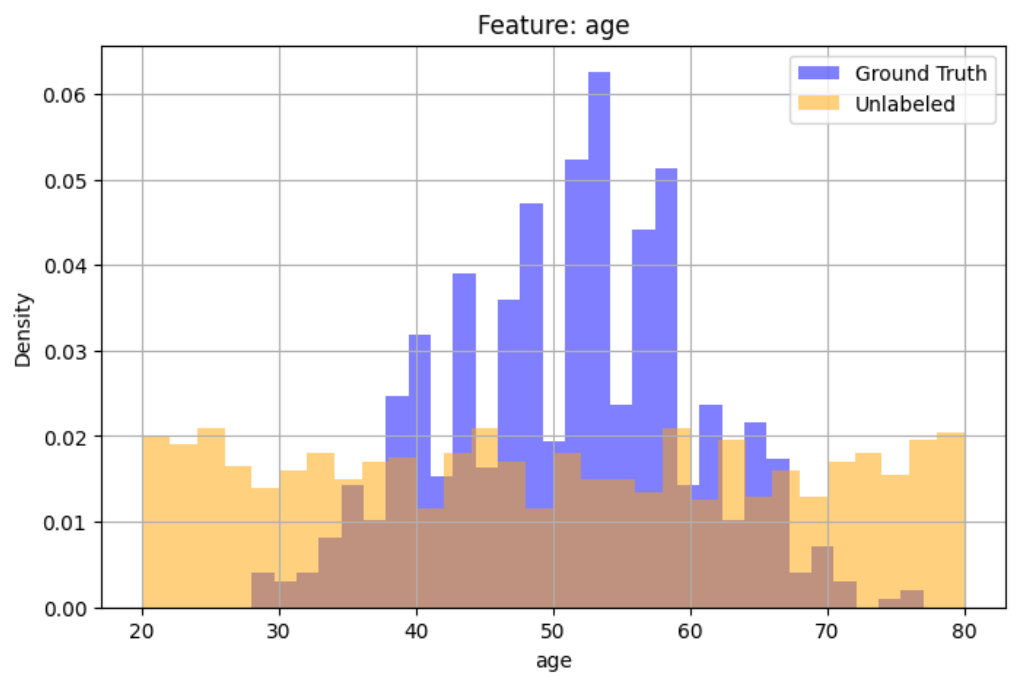
- Accuracy: 0.838541666

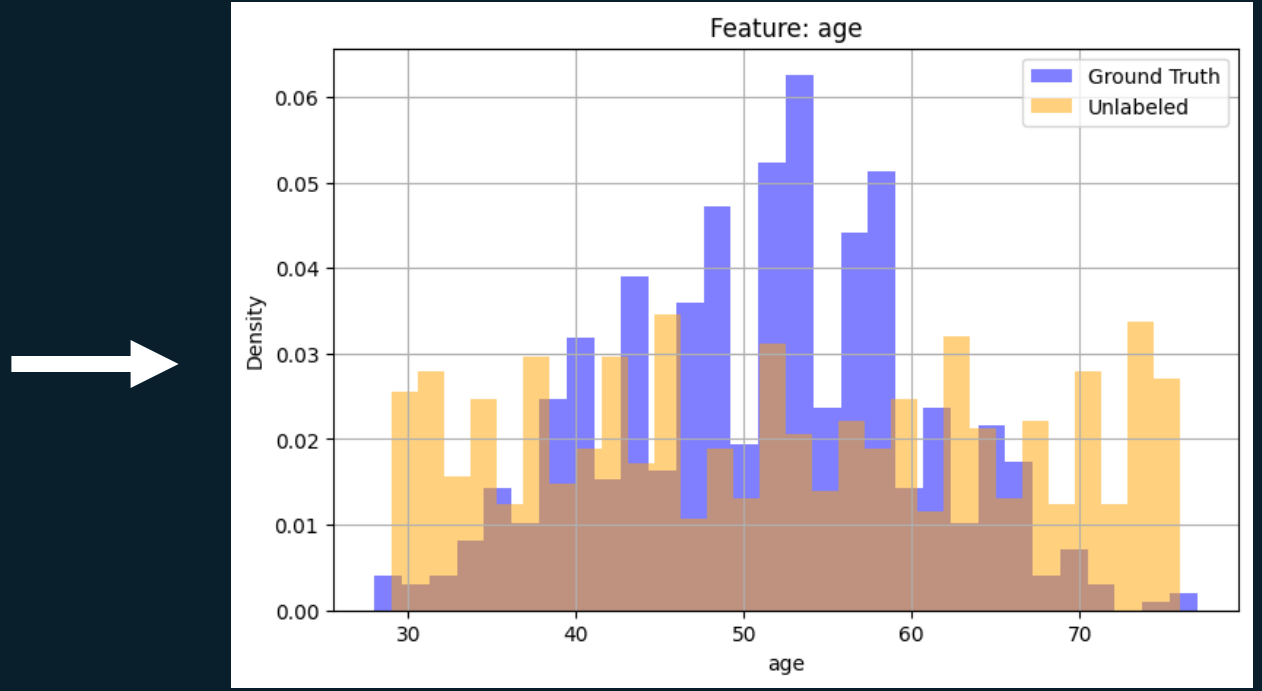
https://colab.research.google.com/drive/16W6xTJWUu6tzYtVBvq7xXw6psNiWdl-3?usp=sharing

Classification Testing

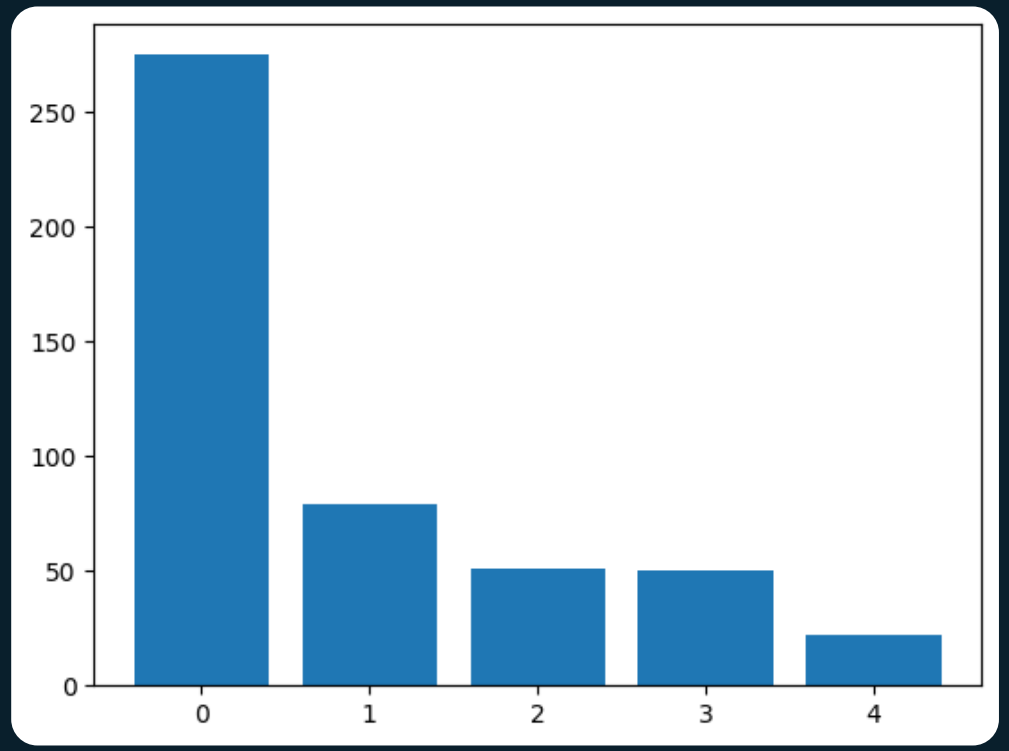
Multiclass Classification Based Model

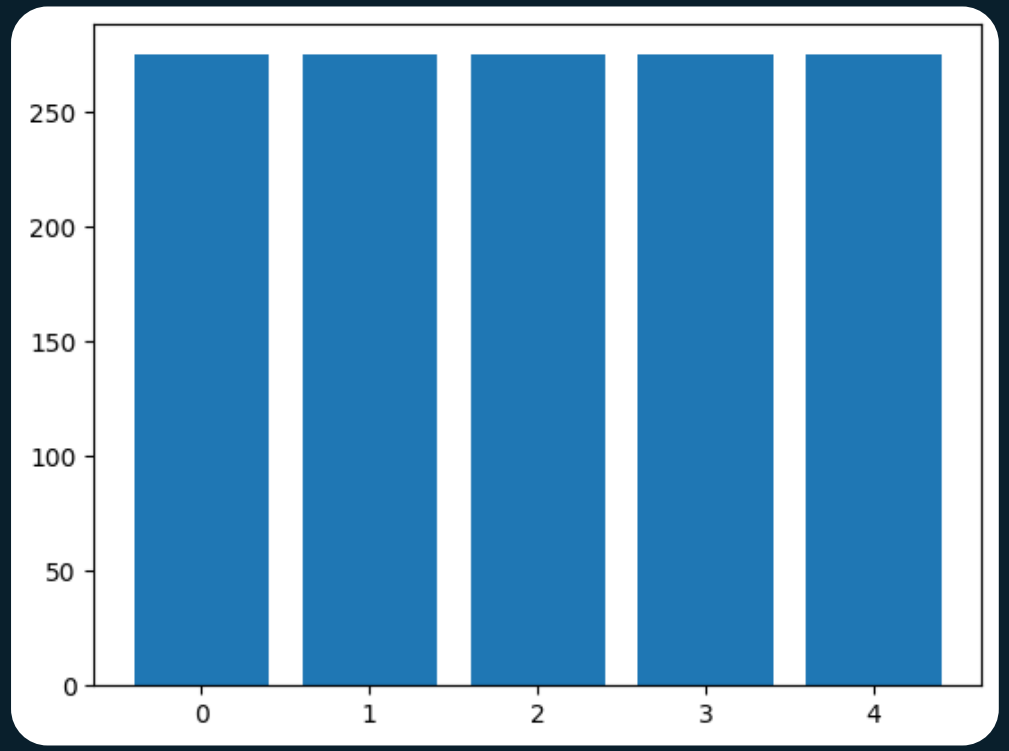
Unlabeled Data – Data Cleaning





Up-sampling






```
Class=0, n=275 (57.652%)
Class=1, n=79 (16.562%)
Class=2, n=51 (10.692%)
Class=3, n=50 (10.482%)
Class=4, n=22 (4.612%)
```

```
The number of classes before fit Counter({0: 275,
1: 79, 2: 51, 3: 50, 4: 22})
The number of classes after fit Counter({0: 275,
1: 275, 2: 275, 4: 275, 3: 275})
Class=0, n=275 (20.000%)
Class=1, n=275 (20.000%)
Class=2, n=275 (20.000%)
Class=3, n=275 (20.000%)
Class=4, n=275 (20.000%)
```

Model Accuracy

Support Vector Machine

```
SVM Multi-class Clasification
accuracy: 0.5333333333333333
[[57 12  3  4  1]
 [ 6  1  5  0  1]
 [ 2  1  4  1  3]
 [ 3  3  2  1  4]
 [ 1  0  3  1  1]]
              precision    recall  f1-score   support

           0       0.83      0.74      0.78        77
           1       0.06      0.08      0.07        13
           2       0.24      0.36      0.29        11
           3       0.14      0.08      0.10        13
           4       0.10      0.17      0.12         6

    accuracy                           0.53       120
   macro avg       0.27      0.28      0.27       120
weighted avg       0.58      0.53      0.55       120
```

- Accuracy: 0.5333333

Random Forest

```
Random Forrest Multiclass Clasification
accuracy: 0.6166666666666667
[[65  6  4  1  1]
 [ 3  3  4  3  0]
 [ 3  1  6  0  1]
 [ 5  2  3  0  3]
 [ 2  1  1  2  0]]
              precision    recall  f1-score   support

           0       0.83      0.84      0.84        77
           1       0.23      0.23      0.23        13
           2       0.33      0.55      0.41        11
           3       0.00      0.00      0.00        13
           4       0.00      0.00      0.00         6

    accuracy                           0.62       120
   macro avg       0.28      0.32      0.30       120
weighted avg       0.59      0.62      0.60       120
```

- Accuracy: 0.616666666

Artificial Neural Network

```
ANN Multi-class Clasification
accuracy: 0.5916666666666667
[[59  8  9  1  0]
 [ 5  3  5  0  0]
 [ 2  1  6  2  0]
 [ 4  2  4  3  0]
 [ 1  1  3  1  0]]
              precision    recall  f1-score   support

           0       0.83      0.77      0.80        77
           1       0.20      0.23      0.21        13
           2       0.22      0.55      0.32        11
           3       0.43      0.23      0.30        13
           4       0.00      0.00      0.00         6

    accuracy                           0.59       120
   macro avg       0.34      0.35      0.33       120
weighted avg       0.62      0.59      0.60       120
```

- Accuracy: 0.59166666

https://colab.research.google.com/drive/1alfoDE1z-FLbFyaNUpL9GRiepKtO1Mlf?usp=sharing#scrollTo=8vgpzKPfyuGH

Comparing The Accuracy for All Models

Binary Classification Based

- SVM: 0.7013888
- **Random Forest: 0.907986111**
- ANN: 0.838541666

Multiclass Classification Based

- SVM: 0.5333333
- **Random Forest: 0.616666666**
- ANN: 0.59166666

Can be more optimized

Learnings and next steps

- Comparing accuracy and other metrics that can be used to apply on the project, and select two main models that need to be used for final project

- Finalizing the final version of the conference paper format of the report.

Thank You | Q&A

Contact us by

Charunthon Limseelo
boat.charunthon@gmail.com or visit official website

Bhagya Saranunt
parksaranunt@gmail.com