BMEN619 HEART FAILURE PREDICTION EVALUATION CRITERIA

Veronica Obodozie

11 March 2025

1. INTRODUCTION

According to the World Health Organization, Cardiovascular Diseases (CVDs) are the leading cause of death globally with more than 80% of these deaths attributed to heart attacks and strokes [1]. It is the second leading cause of death in Canada with men having double the chance of having a heart attack than women. [2] Early diagnosis of heart disease based on symptoms and popular risk factors can help with management and increase survival rates. In this area, Machine Learning techniques have been implemented to contribute to diagnostic systems [3].

This assignment describes a proposed methodology, experimental design, and evaluation criterion, for a heart failure prediction model developed using the structured Heart Failure Prediction Dataset [4]. This dataset is open source from Kaggle and consists of freely accessible data from 5 health institutions combined based on 11 similar attributes. The ground truth labels are if the patient has heart disease or not.

The goal of this project is to train a fair, reliable, and reproducible binary classification model using supervised learning to predict if a patient has heart disease or not. To achieve this, multiple traditional machine-learning models will be compared and evaluated based on performance, explainability, and fairness.

2. DATA LOADING AND PREPROCESSING

The dataset was downloaded as a CSV file from the Kaggle platform, this can be loaded as a data frame in Python. No missing data was found during the exploratory data analysis (EDA), and duplicates were removed by the original author. Multiple data types are being handled:

- **Categorical**: Sex, Chest Pain Type, Fasting blood sugar, Resting Electrocardiogram, Exercise Angina, ST slope
- **Numerical**: Age, Resting blood pressure, Serum Cholesterol, Max Heart rate, OldPeak.

The numerical features have a data type that is either integer or float values. For categorical features, some related to binary integer values, and others were object data types with qualitative features. Sex, ChestPainType, RestingECG, ExerciseAngina, and ST_Slope are qualitative features with an object data type. Label encoding shall be applied to update the qualitative categories to numerical.

Data was scaled using the RobustScaler, which uses the interquartile range as a scaling factor to reduce the influence of existing outliers. This ensures the normalization and standardization technique will be applied to both features with normal Gaussian distribution and features like Oldspeak with a distribution skewed to the right.

The dataset was imbalanced with over 70% of the observations being of men. Stratification was the method chosen to combat this which is used in literature for medical datasets of this modality [5], [6].

An important point to consider is outliers in the dataset, only 2 observations were removed due to zero RestingBP and negative OldPeak values which are not possible readings. Other outliers were kept because the difference between anomalies and cases the model should not misclassify was determined. For instance, ignoring younger observations with heart disease.

3. EXPERIMENTAL DESIGN

This section highlights the tools and considerations for reproducibility applied when building this model. It also gives a breakdown of the proposed experimental design and the explainability of the model.

3.1. Tools

Python will be used to develop this detection algorithm and will be heavily reliant on the pandas and sci-kit learn packages for development and metrics. To evaluate the fairness of the model, Aequitas bias and fairness audit [7] will be used with Fairlearn [8] built with sci-kit learn as a base ensuring environment compatibility.

3.2. Reproducibility

The dataset used is open source and is easily accessible on the Kaggle Platform with an account. Code will be hosted on a GitHub repository that is publicly accessible so interested users can run it on platforms like Google Colab or clone the repository to make improvements. A guide on how to use the repository shall be created within the wiki [10].

3.3. Experimental Design

A traditional machine learning approach is taken, and although the feature engineering was performed during the EDA, it was determined that all 11 attributes would be used. Due to the limited number of features and the strong correlation between the features and labels in Figure 1.

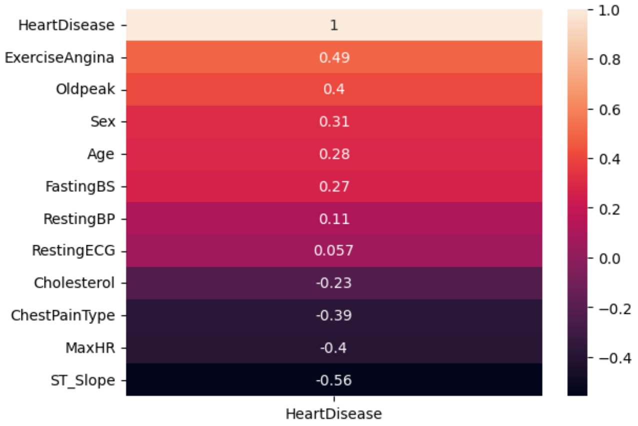

Fig. 1. Correlation of features and HeartDisease label.

The dataset will be split into 80% development and 20% testing with no randomness to ensure reproducibility. And the development set will be split into training and validation using a stratified k-fold cross-validation method to ensure a balanced dataset within each fold, and similar results across folds [11]. Equal feature weighting from data scaling.

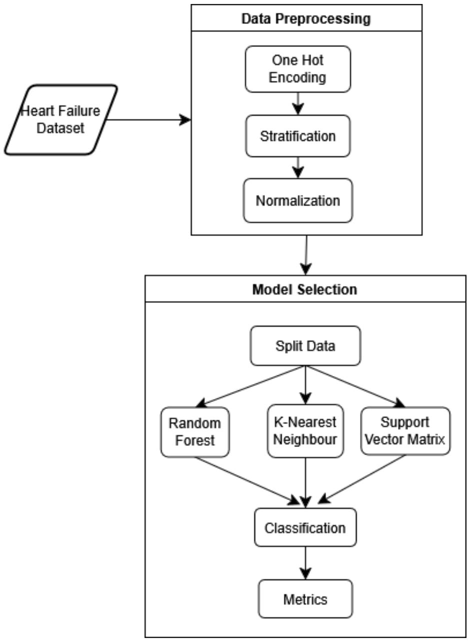

Fig. 2. Proposed Framework for Heart Failure Prediction.

The Support Vector Matric, K-Nearest Neighbour, Linear Classifier, and Random Forest Classifiers will be compared in this project. These models were chosen based on data characteristics, and performance in testing. Figure 2 is the flowchart of the experimental design.

3.4. Responsible

Responsibility refers to the considerations taken to increase the trustworthiness of the tool in healthcare when developing the model. This project follows the fairness, universality, traceability, usability, robustness, and explainability (FUTURE-AI) guideline [12] created to increase the trustworthiness and ethics of healthcare AI tools. Fairness as it relates to equity and robustness in terms of reliable outputs independent of subgroups. This means the performance of the model should be sensitive feature agnostic, and that variability in data like outliers should not affect performance. Explainability is also an important factor for diagnosis tools.

This de-identified dataset came with some sensitive demographic information, age, and sex, and the experimental design was developed to mitigate the exacerbation of existing biases [13], [14]. Although the dataset only considers certain demographics in the United States of America, Switzerland, and Hungary. Heatmaps will be used to display feature importance, and the model limitations will be outlined.

4. EVALUATION CRITERIA

This section focuses on the metrics used to evaluate the performance and fairness of the model. It was chosen based on the nature of the task, and the metrics recommended to measure performance that was used in similar diagnostic performance classification models from literature [3], [6], [12]. The number of heart disease predictions that are right (True Positive), the number of wrong positive predictions (False Positive), the number of correct normal predictions (True Negative), and the number of false normal predictions (False Negative) will be a basis for evaluation.

Recall or sensitivity value is the ratio of the positive values correctly predicted and the total positive outcomes. This is also known as the True Positive Rate and can be stratified across sensitive features to find the equal opportunity [13] fairness metric:

$$Recall = \frac{TP}{TP + FN} \quad (1)$$

Precision describes the reliability of the model, also known as the positive parity.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Specificity, or Negative Predictive Value is the ratio of the True negative values and the total negative outcomes, it describes the reliability of the model.

$$Specificity = \frac{TN}{TN + FN} \quad (3)$$

The F1 score is a weighted average of precision and recall values, and is a popular metric for classification problems [12]:

$$F1\ Score = 2\frac{Precision\ \times Recall}{Precision + Recall} \qquad \textbf{(4)}$$

The confusion matrix for the True Positives and False Negatives of the predicted and target values will be created to capture overall errors. Evaluating the model's explainability will require the use of infidelity and sensitivity metrics. Fairlearn tool helps to determine the statistical parity of the model when evaluating fairness.

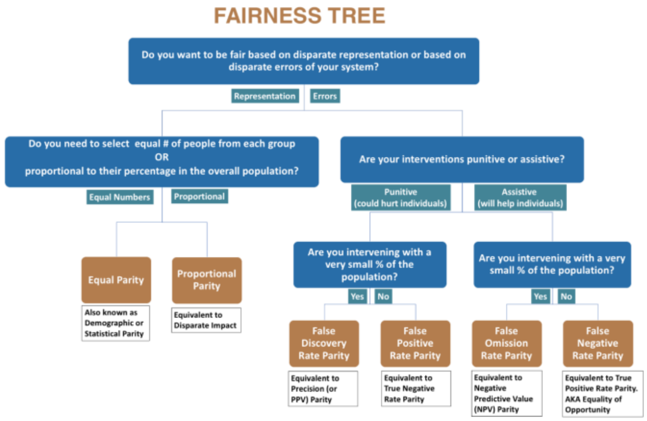
FAIRNESS TREE
Do you want to be fair based on disparate representation or based on disparate errors of your system?
Representation
Errors
Do you need to select equal # of people from each group OR proportional to their percentage in the overall population?
Equal Numbers
Proportional
Are your interventions punitive or assistive?
Punitive (could hurt individuals)
Assistive (will help individuals)
Equal Parity
Also known as Demographic or Statistical Parity
Proportional Parity
Equivalent to Disparate Impact
Are you intervening with a very small % of the population?
Yes
No
Are you intervening with a very small % of the population?
Yes
No
False Discovery Rate Parity
Equivalent to Precision (or PPV) Parity
False Positive Rate Parity
Equivalent to True Negative Rate Parity
False Omission Rate Parity
Equivalent to Negative Predictive Value (NPV) Parity
False Negative Rate Parity
Equivalent to True Positive Rate Parity. AKA Equality of Opportunity

Fig. 3. Aequitas Fairness Metric Decision Tree [7].

Demographic parity requires equal decision rates between subgroups and will be applied specifically to the sex feature due to the heavy imbalance in the data. A General Demographic Parity [13] shall be applied to the Age feature as this is a continuous attribute.

$$P(\hat{y} = 1 \mid z = M) = P(\hat{y} = 1 \mid z = F) \qquad \textbf{(5)}$$

5. CONCLUSION

Based on the Heart failure prediction dataset obtained from Kaggle, a reliable model will be created using Python as the primary algorithm for development. To achieve this, the data will be preprocessed using robust scaling to reduce the outlier effect, and categorical data ne-hot encoded to numeric values, The methodology outlined in this assignment shall be implemented and improvements shall be made during testing and results outlined in the final project. For optimum results model hyperparameters will be tuned.

6. REFERENCES

[1] "Cardiovascular diseases," World Health Organization, Jun. 2021. Accessed: Mar. 06, 2025. [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases

[2] P. H. A. of Canada, "Heart Disease in Canada." Accessed: Mar. 06, 2025. [Online]. Available: https://www.canada.ca/en/public-health/services/publications/diseases-conditions/heart-disease-canada.html

[3] D. K. Plati *et al.*, "Machine Learning Techniques for Predicting and Managing Heart Failure," in *Predicting Heart Failure*, John Wiley & Sons, Ltd, 2022, pp. 189–226. doi: 10.1002/9781119813040.ch9.

[4] fedesoriano, "Heart Failure Prediction Dataset." Kaggle, https://www.kaggle.com/fedesoriano/heart-failure-prediction, Sep. 2021. Accessed: Feb. 08, 2025. [Online]. Available: https://www.kaggle.com/fedesoriano/heart-failure-prediction

[5] D. Ang, K. Naineni, and J. Ho, "Healthcare Data Handling with Machine Learning Systems: A Framework," in *2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE)*, Jul. 2023, pp. 1331–1334. doi: 10.1109/CSCE60160.2023.00223.

[6] D. K. Lutfi and G. F. Shidik, "Improvement Heart Failure Prediction Using Binary Preprocessing," in *2023 International Conference on Advanced Mechatronics, Intelligent Manufacture and Industrial Automation (ICAMIMIA)*, Nov. 2023, pp. 236–241. doi: 10.1109/ICAMIMIA60881.2023.10427846.

[7] P. Saleiro *et al.*, "Aequitas: A Bias and Fairness Audit Toolkit," Apr. 29, 2019, *arXiv*: arXiv:1811.05577. doi: 10.48550/arXiv.1811.05577.

[8] H. Weerts, M. Dudík, R. Edgar, A. Jalali, R. Lutz, and M. Madaio, "Fairlearn: Assessing and Improving Fairness of AI Systems," *J. Mach. Learn. Res.*, vol. 24, no. 257, pp. 1–8, 2023.

[9] R. K. E. Bellamy *et al.*, "AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias," Oct. 03, 2018, *arXiv*: arXiv:1810.01943. doi: 10.48550/arXiv.1810.01943.

[10] P. Vandewalle, J. Kovacevic, and M. Vetterli, "Reproducible research in signal processing," *IEEE Signal Process. Mag.*, vol. 26, no. 3, pp. 37–47, May 2009, doi: 10.1109/MSP.2009.932122.

[11] M. T r, V. K. V, D. K. V, O. Geman, M. Margala, and M. Guduri, "The stratified K-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification," *Healthc. Anal.*, vol. 4, p. 100247, Dec. 2023, doi: 10.1016/j.health.2023.100247.

[12] K. Lekadir *et al.*, "FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare," *BMJ*, vol. 388, p. e081554, Feb. 2025, doi: 10.1136/bmj-2024-081554.

[13] Q. Feng, M. Du, N. Zou, and X. Hu, "Fair Machine Learning in Healthcare: A Review," Feb. 01, 2024, *arXiv*: arXiv:2206.14397. doi: 10.48550/arXiv.2206.14397.

[14] J. E. Alderman *et al.*, "Tackling algorithmic bias and promoting transparency in health datasets: the STANDING Together consensus recommendations," *Lancet Digit. Health*, vol. 7, no. 1, pp. e64–e88, Jan. 2025, doi: 10.1016/S2589-7500(24)00224-3.

[15] Fahad Rehman and Muhammad Aammar Tufail, "Heart Disease Prediction Using 9 Models," Kaggle. Accessed: Mar. 01, 2025. [Online]. Available: https://kaggle.com/code/fahadrehman07/heart-disease-prediction-using-9-models

APPENDIX: DATA DESCRIPTION

This is a support document detailing the exploratory data analysis findings for the dataset. Heart disease is one of the leading causes of death globally and this dataset is a comma-separated version (CSV) file obtained from Kaggle containing 11 features to predict heart diseases and 1 common label. This file combined 5 heart datasets across 11 common features [4]. EDA is performed with the help of similar notebooks [15].

There are a total of 918 observations, with duplicate data removed by the author before dataset publication and no missing data found when inspected using Python. In creating this label, data on the type of heart disease is lost.

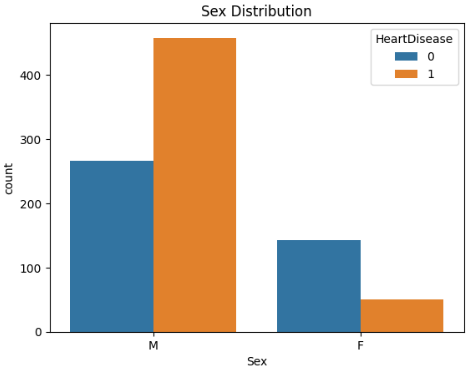

Fig. 4. Stratified histogram of label by Sex category.

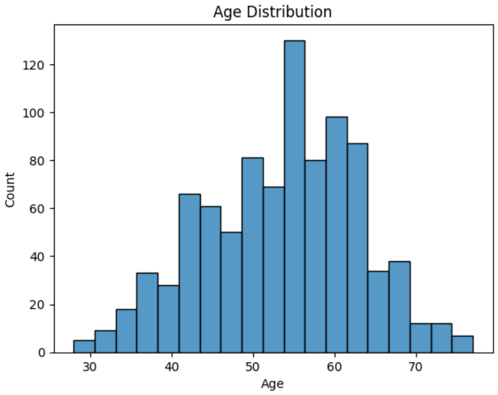

Fig. 5. Histogram shows distribution of Age feature values.

The dataset was imbalanced, across the sensitive features of age and sex with more observations from Male and patients over 40 as seen in Figures 4 and 5. Categorical qualitative features were used to stratify the label data to better understand their characteristics. Figure 6 shows how more observations with asymptomatic (ASY) chest pain type are likely to have heart disease, compared to the opposite behaviour for atypical (ATA). A similar distribution is seen in Figure 7, with more observations having exercise-induced angina positive prediction.

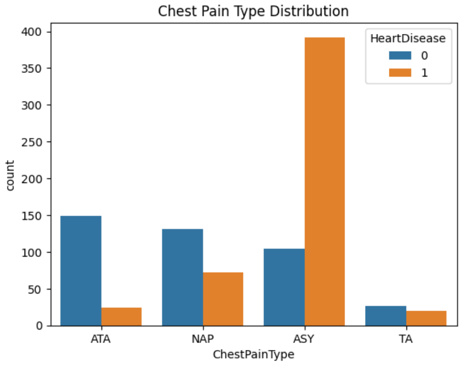

Fig. 6. Stratified histogram of label by ChestPainType.

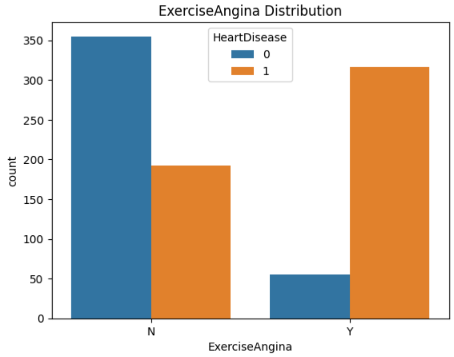

Fig. 7. Stratified histogram of label by ExerciseAngina.

Table 1: Statistical Analysis Results for Numerical Features

Feature	mean	std	min	50%	max
Age	53.51	9.433	28	54	77
RestingBP	132.4	18.51	0	130	200
Cholesterol	198.8	109.4	0	223	603
FastingBS	0.2331	0.423	0	0	1
MaxHR	136.8	25.46	60	138	202
Oldpeak	0.8874	1.067	-2.6	0.6	6.2
HeartDisease	0.5534	0.497	0	1	1

A detailed statistical analysis was performed for features with numerical values, Table 1 shows some of the results. Using the Z-score, some outliers were found within the dataset, however, to avoid the false negative errors they will be left in the dataset during processing. An example case is observation 76 of a 32-year-old Male having an ASY chest pain type and a cholesterol above the fiftieth percentile having heart disease. This was classified as an outlier due to the z-score of age, but the correlations from other features would be useful to have within the dataset.

There were a total of 19 outliers found within the numerical values based on the Z score: 8 from restingBP, 1 from the MaxHR, 3 from Cholesterol, and 7 from Oldpeak. Outlier values were inspected with observation rows: 324 with OldPeak value of -2.6 and 449 with RestingBP value of 0 will be removed during preprocessing.