Evaluation on the Classification of Cardiovascular Disease Using Logistic Regression

Ben Alexander, Machine Learning and Computational Intelligence Assessment

CONTENTS

I. INTRODUCTION

CARDIOVASCULAR Disease (CVD) is a general term encompassing conditions concerning the heart or blood vessels [1]. In the UK alone, there are around 160,000 deaths due to CVD per year [2]. This collection of diseases desperately needs early diagnosis and measures to be administered for these patients as soon as possible. In recent years, the healthcare industry has been using machine learning methods due to their efficiency in speeding up diagnosis time, and reduction of false positives [3]. Implementing a machine-learning approach to a CVD preliminary screening consultation to be used as an aid for the health professional, could prove revolutionary to this diagnosis procedure. Not only could it increase the accuracy of diagnosis, it could ease pressure on already strained health services around the world. Using a machine learning approach can also pick up on unknown representations of the disease in certain demographics such as females under the age of 55 being misdiagnosed though it is the leading cause of death in the United States for females [4]. Socioeconomic factors too are highly correlated to CVD [5]. Similarly, racial inequality is rife within medical teaching and diagnostic practices due to the misrepresented characteristics of CVD among these cross-sections of society [6]. Hence, an approach to diagnosis that can be abstracted and separated from gender, racial, and socioeconomic bias, could be incredibly important in overturning systemic tendencies of misdiagnosis.

Many machine-learning approaches are black-box methods, giving little to no insight into the inner workings of the classification process. Therefore, this report suggests one white-box approach in the classification of whether a CVD is present in the patient. Although many other approaches were considered, using Logistical Regression seemed the best fit for this dataset and context. Using a white-box method, although potentially less powerful in diagnosis, can be inspected to ensure the misdiagnosis rates for different sections of society are not hindered by the approach. Unfortunately, there are no socioeconomic or ethnicity data in the dataset given, so this cannot be measured, whereas sex is given and will be investigated.

This report will also give a brief introduction to the characteristics, causes, and trajectories of CVDs, an overview of the dataset, data preparation, methods and an evaluation. The objective of this report is to observe the model's performance in this context and always bring the focus back to the context of preliminary screening and the characteristics of CVD.

A. Clinical Background

CVD as a collection of conditions concerning the cardiovascular system, can be grouped into 4 main types: Coronary Heart Disease (CHD), Stroke and Transient Ischaemic Attack (TIA), Peripheral Arterial Disease (PAD), and Aortic Disease (AD). CHD is present when

blood that is oxygen-rich is restricted to the heart. Strokes and TIAs are caused by a cut-off (only temporarily for TIA) in the supply of blood to the brain. PAD occurs due to a blockage of blood flow to a limb from the arteries. Finally, AD is a collection of conditions surrounding the efficacy of the aorta [1].

Although the causes of CVD are not clear, certain risk factors can increase the chance of an individual developing CVD. These include smoking, high cholesterol, diabetes, inactivity, high BMI leading to overweight or obesity, CVD history within the family, ethnicity, age, sex, diet choices, and alcohol. Although there are a lot of risk factors, a lot of these are intertwined and correlated, (for example, overweight/obesity, inactivity and diet, are related [7]). Reducing these risk factors can lead to the prevention of developing CVD, including taking medicine to reduce blood pressure (BP) [1].

Traditional diagnosis of CVD relies on common blood tests to indicate cardiovascular health. These include cholesterol, blood sugar levels, and protein levels among others. Other than these blood tests, a variety of medical imaging, physical activity tests, and cardiac catheterisation can be used to deduce problems in the cardiovascular system to assist diagnosis of the exact type of CVD present in the patient [8].

B. Database

The database consists of $64,511$ entries of patient data, with no identifying features recorded, keeping the data completely and utterly anonymous. The data is arranged in the following way (Table I):

Label	Description
Age	Days
Sex	1: Female, 2: Male
Height	Centimeters (cm)
Weight	Kilograms (kg)
Systolic BP (shown as ap_hi)	millimetres of mercury (mmHg)
Diastolic BP (shown as ap_hi)	millimetres of mercury (mmHg)
Cholesterol	1: Normal, 2: Above Normal, 3: Well Above Normal
Blood Glucose Level	1: Normal, 2: Above Normal, 3: Well Above Normal
Smoker Status	1: Smoker, 0: Non-Smoker
Alcohol Consumption Status	1: Yes, 0: No
Physical Activity	1: Yes, 0: No
CVD Diagnosis	1: Present, 0: Absent

TABLE I
DATASET REPRESENTATION

The problems with this dataset arise from what data has been left out. There is a lack of information in the binary categories of ”Smoker Status”, ”Alcohol Consumption Status”, and ”Physical Activity Status” as this does not describe how often the individuals smoke/drink/exercise. It also does not indicate the time the patient has been smoking, how many units they drink weekly, and how much exercise in time is completed every week. These would be far more beneficial metrics and could improve the model’s performance. Similarly, the categorical data of ”Cholesterol”, and ”Blood Glucose Level”, would be much better indicated in the recorded values, rather than categories. Other metrics that would be beneficial in the classification of CVD would be BMI or BMI-alternative (such as waist-to-hip-ratio [9]), diet scores, and Coronary Artery Calcium (CAC) scores[10]. Although, for this to be a preliminary screening, CAC scores may not be the most suitable as they require a CT scan, and this model is intended to speed up diagnoses [11].

Before any preparation of the data for testing and training the chosen model, it is best to assess the biases of the dataset. As shown in Fig. 1 the age distribution is heavily veered towards people aged 50 and above. This is not a problem as there is a correlation between the risk of developing CVD and age, specifically for those that are over 50 [1]. In Fig. 2, it is clear that there are roughly double the amount of ’female’ entries to ’male’ entries. This however is not a problem, as the model can be separated to accommodate this, and will classify each sex distinctly as well as together to compare.

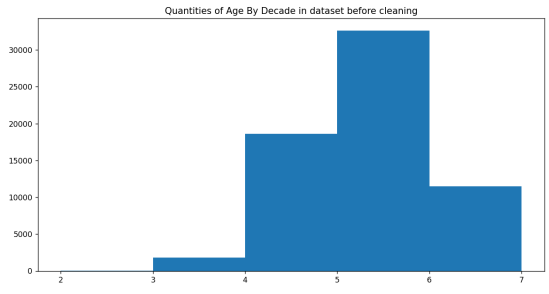

Fig. 1. Quantity of Age by Decade for Uncleaned Dataset

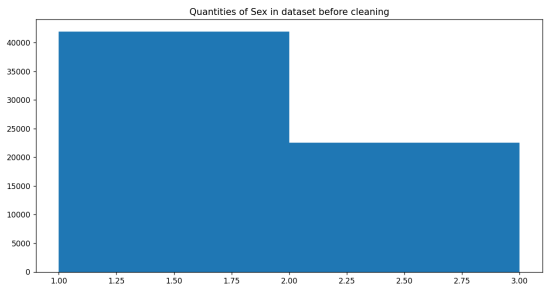

Fig. 2. Quantity of Sex (1: Female, 2: Male) for Uncleaned Dataset

II. Data Preparation

It is important when dealing with a vast dataset like this, to remove any outliers, corrupted values, and nonsensical entries. Outlined below is the process by which the data is prepared and encoded ready to be interpreted and used by the model. This report separates these cleaning stages as the first and second stages, as they remove entries in different ways. The full process can be found in the flowchart Fig. A.1, but will be discussed and explained here.

A. First Stage Data Clean

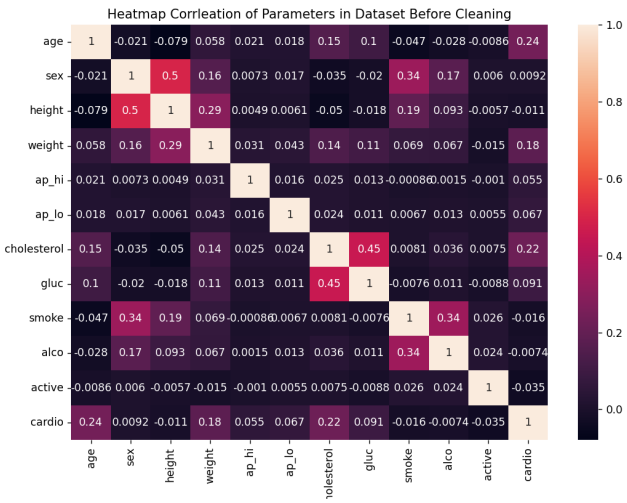

Fig. 3. Correlation between metrics before cleaning

The dataset contains errors in the sex field. This field is sometimes labelled as 'f'/'female'/'m'/'male' (in lower and uppercase) instead of 1 and 0. This can easily be interpreted into either category without losing any data.

After completing this, generating a heat-map to find the correlation between the data (as seen in Fig. 3) shows that there are no strong links between any of the metrics and the CVD ('cardio') classification. Looking into the limits of each of the metrics in the data in Table II, all of the categorical and binary data is within the correct limits, however, some of the integer values seem to be out of the realms of possibility. These values cannot be interpreted to be meaningful data, such as the negative and extremely large BP values.

Metric	Min.	Max.
Age	29.0	64.0
Sex	1	2
Height	55	207
Weight	10.0	200.0
Systolic BP	-150	16020
Diastolic BP	-70	11000
Cholesterol	1	3
Blood Glucose Level	1	3
Smoker Status	0	1
Alcohol Consumption Status	0	1
Physical Activity	0	1
CVD Diagnosis	0	1

TABLE II
MINIMUM AND MAXIMUM VALUES FOR EACH METRIC IN THE DATASET BEFORE CLEANING

On inspection, the data that is of most importance to clean are the metrics 'Height', 'Weight', 'Systolic Blood Pressure', and 'Diastolic Blood Pressure'. After generating the BMI data for the entries. It is possible to discard any values that are out of the realm of possibility using Table III.

Metric	Min.	Max.	References
Height	54.6	272	[12], [13]
Weight	5.9	635.0	[14]
Systolic BP	50	370	[15], [16]
Diastolic BP	20	360	[15], [16]
BMI	12	80	[17], [18]

TABLE III
MINIMUM AND MAXIMUM PHYSICALLY POSSIBLE VALUES

Removing the entries that are out of the range of these values leaves 98.32% (63426) of the dataset, which is higher than human error rate predictions (1%) but not enough to cause concern [19]. However, still leaving in the extremities as shown in the box-plot Fig. 4 leaves many outliers. These outliers are determined by using the $1.5 * IQR$ rule [20]. Although using the $1.5 * IQR$ rule would remove all of these outliers, this may not be beneficial as these values are within the physical bounds shown before. The trade-off between the model being skewed by these data points and those same data points being physically possible is difficult to solve. Hence the second stage of the data clean is introduced.

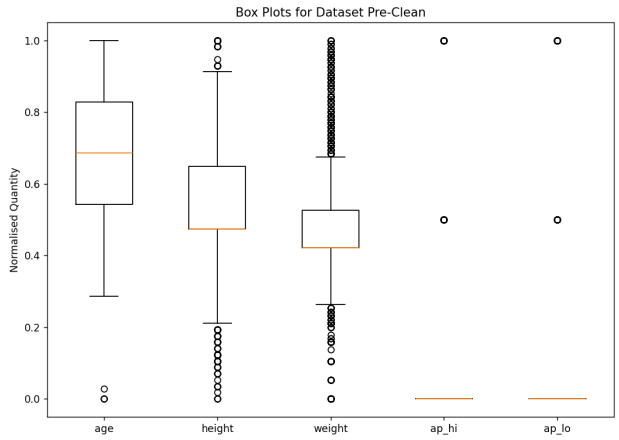

Fig. 4. Box-plots of metrics before cleaning

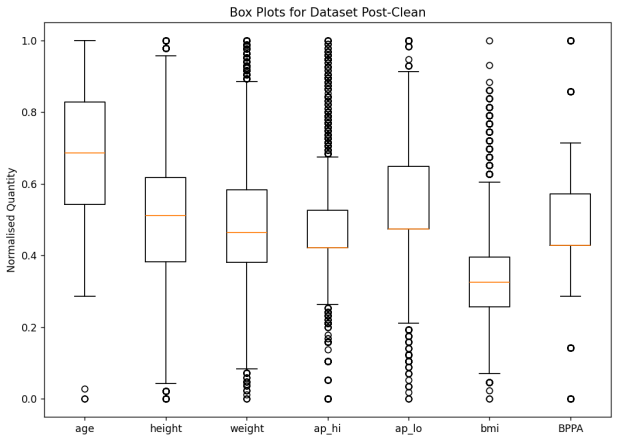

Fig. 5. Box-plots of metrics after cleaning

B. Second Stage Data Clean

Instead of using the $1.5 * IQR$ rule, this method uses the Z-Score of below -3 or $+3$ to exclude outliers as it is less harsh yet still dealing with the absolute extremities present in the data [21]. This is also a trade-off in under-fitting/over-fitting to out-of-dataset data, as these entries could be at the extremes. This method deals with these outliers in decade-sex sub-groupings, to combat the misclassification of an outlier with data of different sexes and age by decades.

Doing this rule to detect outliers leaves the dataset metrics with these box-plots Fig. 5. The box-plots still show outliers but looking at the values in IV, all of these values seem valid without any severe extremities. This second set of box-plots shows the added metrics of ”BMI” and ”Blood Pressure Percentage of Average for Age and Sex Group” (BPPA). BPPA is a metric that seeks to normalise BP for the particular age and sex of the individual. Below Table V shows the average normal BPs for age and sex [22]. Removing these additional outliers leaves the dataset at 95.39% (61535) of the original (Table VI each decade and sex group shows how many were removed).

Now the are extremities removed, the heat-map (shown in Fig 6) expresses the newly made metrics and all of the metrics' relations to the CVD diagnosis. Comparing this to the pre-cleaned heat-map (Fig. 3), there is a lot stronger correlation between the metrics to the CVD diagnosis, hence this should allow for better performance from the model. Note that there are no longer any height and weight metrics as this has been replaced by the more informative BMI metric. Although there are still the Systolic and Diastolic BP metrics as well as the BPPA, this is to catch the entries where one BP is above the average and the other is below, equating to a seemingly ”normal” value for BPPA however it might still be seen as an increased risk factor when the model takes the BP metrics individually. This relationship can also be seen in the pre-cleaned Principle Component Analysis (PCA) of the dataset (Fig. 7) which is not as spread out or

Metric	Min.	Max.
Age	29.0	64.0
Sex	1	2
Height	141.0	188.0
Weight	32.0	116.0
Systolic BP	80	175
Diastolic BP	53	110
Cholesterol	1	3
Blood Glucose Level	1	3
Smoker Status	0	1
Alcohol Consumption Status	0	1
Physical Activity	0	1
BMI	12	55
BPPA	0.7	1.4
CVD Diagnosis	0	1

TABLE IV
MINIMUM AND MAXIMUM VALUES FOR EACH METRIC IN THE DATASET AFTER CLEANING

Sex	Age (years)	Systolic BP	Diastolic BP
Female	18-39	110	68
	40-59	122	74
	60+	139	68
Male	18-39	119	70
	40-59	124	77
	60+	133	69

TABLE V
AVERAGE BP FOR AGE AND SEX [22]

Sex	Decade	Total Entries	% Removed
Female	20s	2	0.00
	30s	1107	2.62
	40s	11428	2.12
	50s	21566	2.74
	60s	7198	3.28
Male	20s	1	0.00
	30s	665	3.46
	40s	6889	3.32
	50s	10517	3.69
	60s	4053	3.80

TABLE VI
AMOUNT OF DECADE AND SEX GROUP REMOVED IN THE Z-SCORE OUTLIER DETECTION

as distinguishable compared to the post-cleaned PCA[1] (Fig. 8). Check the appendix Fig. B.1 to Fig. B.13 for histograms containing the resulting distributions of each metric.

With this cleaned dataset, the data is then stored in training and test sets, of both female and male sets separately, and combined into one. The test sets sit at 20% of the dataset, and each dataset is randomly initialised 10 times to get a mean of the performance of the model, and to escape local minima.

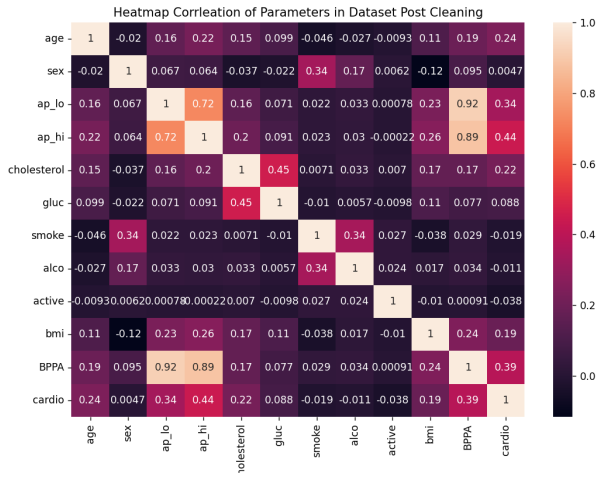

Fig. 6. Correlation between metrics after cleaning

[1]This is not a full representation of the data, as it only shows the 3 most principle axis, however, the data is now more spread out and a split in the CVD diagnosis is more visible

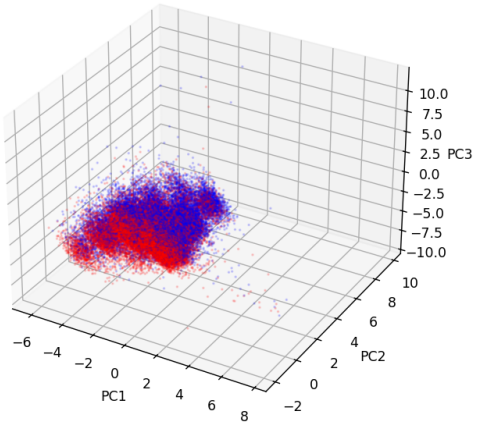

Fig. 7. Principle Component Analysis of Pre-Cleaned Dataset (Key: Blue = CVD Present, Red = CVD Absent)

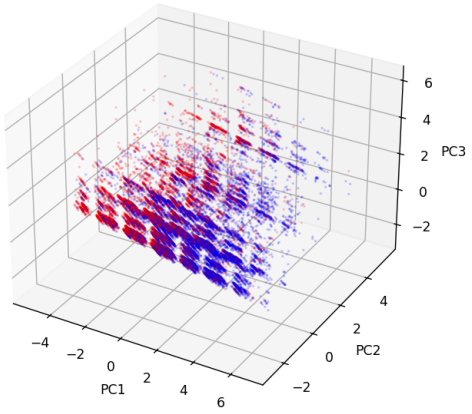

Fig. 8. Principle Component Analysis of Post-Cleaned Dataset (Key: Blue = CVD Present, Red = CVD Absent)

III. METHOD

Logistic regression was chosen, not only for its simplicity and white-box nature but also because it is a supervised learning method. This problem is best suited for either supervised learning or ensemble models, due to the labelling of the CVD diagnosis in the data. Ensemble classifiers work incredibly well and often yield better results compared to supervised learning methods, however, the interpretability of the model decreases with the complexity and number of models in a voting system. This model aims to keep the accuracy of the model high whilst being able to grasp how the model is learning and classifying the data. Not only is this a good application in medical teaching, but it can help large medical bodies find the underlying risk factors of CVD and run public health campaigns combating these.

Logistic regression works perfectly when the desired output of the model is binary (which in this case is

1: CVD Present, 0: CVD Absent). Although when the problem presents itself as a non-linear function, there are possibilities of falling into local minima in the loss function. The equation below (Eq. 1) shows the mapping of the vector of inputs $\mathbf{x}$ and the output probability p of classification is rounded to 0 or 1 (adapted from [23]).

$$p(\mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^{n}(\beta_j \mathbf{x}_j))}} \tag{1}$$

β_0 and β_j represent the values in which the model is trained, this will scale the input values associated with each one (with β_0 being the y-intercept) and will adjust these as the model trains. The resulting shape of this curve can be seen in Fig. 9 compared to a more simple linear regression model [23].

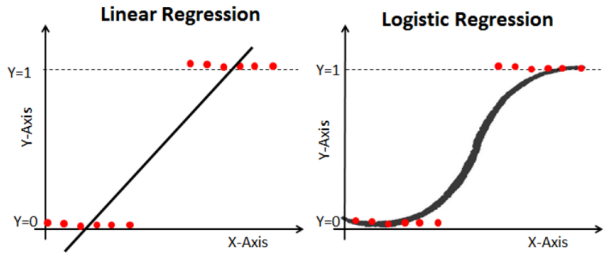

Fig. 9. Logistic regression compared to linear regression, taken from [23]

Because of the potential to fall into local minima, the model was trained and tested 10 times on each female, male, and both sexes datasets. There are not too many parameters associated with this model (sklearn [24]), yet the max iterations parameter was increased to 1000 from the default 100, as it wasn't getting to the desired tolerance value within the default value of max iterations. As was hinted in the previous section the metrics used are as shown in Table VII. This report will showcase the results and analysis of the model in the next section.

x	Metric	Units
$\mathbf{x}_1$	Age	Years (Integer)
$\mathbf{x}_2$	Sex	Binary
$\mathbf{x}_3$	Systolic BP	Integer
$\mathbf{x}_4$	Diastolic BP	Integer
$\mathbf{x}_5$	Cholesterol	Categorical
$\mathbf{x}_6$	Blood Glucose	Categorical
$\mathbf{x}_7$	Smoker Status	Binary
$\mathbf{x}_8$	Alcohol Status	Binary
$\mathbf{x}_9$	Activity Status	Binary
$\mathbf{x}_{10}$	BMI	Float
$\mathbf{x}_{11}$	BPPA	Float

TABLE VII
METRICS USED FOR THE LOGISTIC REGRESSION MODEL

IV. Evaluation

A. Results

The overall average results from running the sexes separately are shown in Table VIII, and the results from the sexes trained together are shown in Table IX. The female equation generated in the first run is outlined in Eq. 2 and the male counterpart is shown in Eq. 3 (note that these have no $\mathbf{x}_2$/sex component). The first run equation of the model with combined sexes is shown in Eq. 4. These serve to show the β coefficients in the sigmoid equation (Eq. 1) of the trained model for that group. The confusion Matrices are shown in Fig. C.1, C.2, and C.3 in the appendix.

Metric	Value in %
Accuracy	73.39
Precision	74.46
Recall	61.47
F1-Score	67.34

TABLE VIII
AVERAGE RESULTS FOR SEXES TRAINED AND TESTED SEPARATELY

Metric	Value in %
Accuracy	73.07
Precision	73.70
Recall	61.22
F1-Score	66.88

TABLE IX
AVERAGE RESULTS FOR SEXES TRAINED AND TESTED TOGETHER

$$\begin{aligned}\mathbf{x} = &-12.586 + 0.0568 * \mathbf{x}_1 + 0.0382 * \mathbf{x}_3 \\ &+ 0.0682 * \mathbf{x}_4 + 0.4901 * \mathbf{x}_5 - 0.1257 * \mathbf{x}_6 \\ &- 0.0808 * \mathbf{x}_7 - 0.1715 * \mathbf{x}_8 - 0.2000 * \mathbf{x}_9 \\ &+ 0.0209 * \mathbf{x}_{10} - 3.1400 * \mathbf{x}_{11}\end{aligned} \tag{2}$$

$$\begin{aligned}\mathbf{x} = &-12.422 + 0.0445 * \mathbf{x}_1 + 0.0281 * \mathbf{x}_3 \\ &+ 0.0697 * \mathbf{x}_4 + 0.4914 * \mathbf{x}_5 - 0.1632 * \mathbf{x}_6 \\ &- 0.1582 * \mathbf{x}_7 - 0.2495 * \mathbf{x}_8 - 0.2973 * \mathbf{x}_9 \\ &+ 0.0430 * \mathbf{x}_{10} - 2.4410 * \mathbf{x}_{11}\end{aligned} \tag{3}$$

$$\begin{aligned}\mathbf{x} = &-12.449 + 0.0511 * \mathbf{x}_1 + 0.0366 * \mathbf{x}_2 \\ &+ 0.0256 * \mathbf{x}_3 + 0.0679 * \mathbf{x}_4 + 0.4883 * \mathbf{x}_5 \\ &- 0.1119 * \mathbf{x}_6 - 0.2032 * \mathbf{x}_7 - 0.1962 * \mathbf{x}_8 \\ &- 0.2480 * \mathbf{x}_9 + 0.0277 * \mathbf{x}_{10} - 2.1140 * \mathbf{x}_{11}\end{aligned} \tag{4}$$

B. Analysis

As can be seen from the models, there is little difference in the efficacy of separating the sexes to not. As can be seen, the precision lies roughly around 73% yet the recall of the models lie around $60 - 63\%$. This is re-iterated by the confusion matrices in the appendix (Fig. C.1, C.2, and C.3). This is especially bad for medical contexts as this means that the diagnosis of individuals who have CVD have roughly 38% chance of being diagnosed as not having CVD. To be used as a preliminary screening, the optimal target would be to have higher recall than precision. This model is good at finding those without CVD, however struggles to diagnose those with CVD. As mentioned previously, the dataset given was not perfect as it had categorical or binary data that could have been better used if it were a more precise metric (such as smoking or alcohol intake). The dataset also does not include diet scores or BMI alternatives that may also have increased the performance of the model [10]. Nor did the dataset include any ethnic detail, which could also increase the effectiveness of the model due to how CVD can affect people of different ethnic backgrounds [25].

Looking at the model's (containing both sexes) misclassifications against age by decade and sex (Fig. 10 and 11), it is clear to see that there is no difference in the misdiagnosis against sex, and the misdiagnosis percentages against age tend to be those falling into an older age category. This is good as it works against the misdiagnosis rates in medical practitioners for younger (under 55s) females discussed earlier in the report [4]. On the other hand, there is still a bias in the misclassification that will need further exploration in the future. To see the misclassifications for the single-sex models for age by decade, see the appendix Fig. D.1 and D.2.

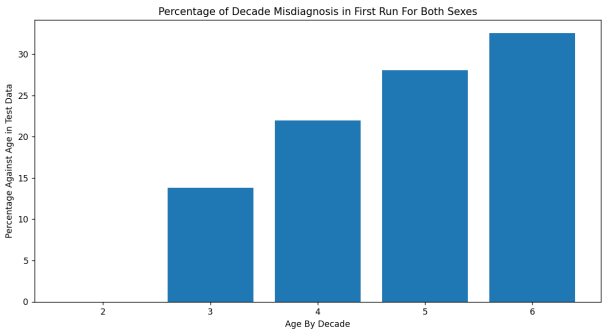

Fig. 10. Misdiagnosis Percentages of Age by Decade for Both Sex Model

The dataset does not provide the model with a clear distinction as to whether or not an individual has CVD

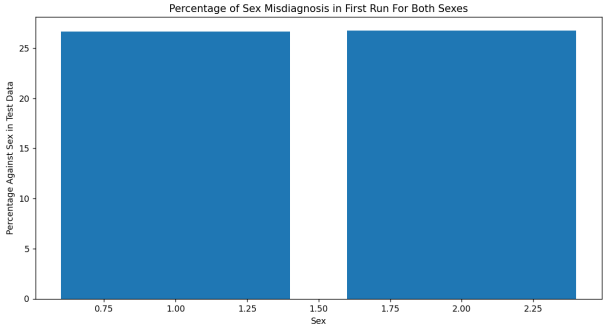

Fig. 11. Misdiagnosis Percentages of Sex for Both Sex Model

(interpreted from PCA). Logistic regression may not be the best solution in terms of accurate diagnosis, however, using it prevents over-fitting to the dataset. Without any more metrics added to the dataset, the performance could be increased, or the dataset could be cleaned further, but this comes as a trade-off of becoming over-fitted to the training set.

Using a logistic regression model allows for the scale of the dataset for training to be increased in the future and become better adapted to when new entries arise. Logistic regression is also fast to both test and predict results, as it only adjusts the coefficients of the input entries for training, and for testing, it is just dropping values into an equation. Therefore compared to other methods where the model needs to be present to predict outcomes, only the equation needs to be distributed to medical practitioners. This has the opportunity save lots of memory and processing power on the end-user's device compared to other methods.

V. Conclusion

Although this method did not appear to have great success in terms of accuracy and recall metrics, the reasons for using this method are still valid in comparison to other machine learning methods. Utilizing the method to aid medical professionals in coming to a diagnosis on a preliminary screening could be beneficial due to its lightweight nature, white-box design, and little processing power. Although the accuracy is not too high and will misdiagnose some people, using this as one of many tools in diagnosis should be considered but not relied upon.

This report has discussed the shortcomings of the model and the dataset and what measures would be needed to improve the performance whilst sticking to the model of logistic regression. Having a bigger sample size and more metrics per entry that are related to CVD would allow the model to make better classification decisions, however, it is clear that the anonymity of the data collected is kept and cannot lead to any identifiable traits.

Hopefully, this report has proved interesting, and this model could not just help medical experts in their diagnoses but allow public health organisations to focus on the highest risk factors for CVDs. Using this as a tool with extra patient data that is corrlated to CVD (such as ethnicity [25] or socioeconomic factors [5]) could give insight into how CVD affects different cross-sections of society and could prove to be useful in medical teaching applications.

References

[1] "Cardiovascular disease," NHS, https://www.nhs.uk/conditions/cardiovascular-disease/:~:text=Cardiovascular%20disease%20(CVD)%20is%20a,increased%20risk%20of%20blood%20clots.

[2] "BHF UK CVD Factsheet," British heart foundation, https://www.bhf.org.uk/-/media/files/for-professionals/research/heart-statistics/bhf-cvd-statistics-uk-factsheet.pdf?rev=5c76af77f68e4c43b19f957890005bbe&hash=D31DB43089AAD361320212D15D4B70FB.

[3] P. Singh, N. Singh, K. K. Singh, and A. Singh, Chapter 5 - Diagnosing of disease using machine learning, https://doi.org/10.1016/B978-0-12-821229-5.00003-3.

[4] R. P. Bullock-Palmer, L. J. Shaw, and M. Gulati, "Emerging misunderstood presentations of cardiovascular disease in young women," WILEY, https://onlinelibrary.wiley.com/doi/full/10.1002/clc.23165.

[5] W. M. Schultz, H. M. Kelli, J. C. Lisko, Et al., "Socioeconomic Status and Cardiovascular Outcomes: Challenges and Interventions," PubMed, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5958918/

[6] Z. Javed, M. H. Maqsood, T. Yahya, and Et al, "Race, Racism, and Cardiovascular Health: Applying a Social Determinants of Health Framework to Racial/Ethnic Disparities in Cardiovascular Disease," AHA Journals, https://www.ahajournals.org/doi/full/10.1161/CIRCOUTCOMES.121.007917.

[7] "Obesity - Causes," NHS, https://www.nhs.uk/conditions/obesity/causes/:~:text=If%20you%20are%20not%20active,by%20the%20body%20as%20fat.

[8] "Cardiovascular disease," Cleveland Clinic, https://my.clevelandclinic.org/health/diseases/21493-cardiovascular-disease.

[9] K. Miller, "Researchers Say Waist-to-Hip Ratio Should Replace BMI—Here's Why," Verywell Health, https://www.verywellhealth.com/could-waist-hip-ratio-replace-bmi-6745714.

[10] S. Kim, Y. Chang, J. Cho, and Et al, "Life's simple 7 cardiovascular health metrics and progression of ...," Aha Journals, https://www.ahajournals.org/doi/10.1161/ATVBAHA.118.311821.

[11] "Coronary artery calcium scoring:," HCA Healthcare UK, https://www.hcahealthcare.co.uk/our-services/tests/ct-coronary-calcium-scoring:~:text=This%20can%20be%20easily%20detected,problems%20and%20prevent%20heart%20attacks.

[12] "World's shortest man: All you need to know about Chandra Bahadur Dangi," Guinness World Records, https://www.guinnessworldrecords.com/news/2012/2/worlds-shortest-man-all-you-need-to-know-about-chandra-bahadur-dangi/.

[13] "Robert Wadlow: Tallest man ever," Guinness World Records, https://www.guinnessworldrecords.com/records/hall-of-fame/robert-wadlow-tallest-man-ever.

[14] "World's heaviest and lightest people", https://www.topendsports.com/testing/records/weight.htmgoogle_vignette.

[15] H. Barcroft and O. G. Edholm, "On the vasodilatation in human skeletal muscle during post-haemorrhagic fainting," U.S. National Library of Medicine, https://pubmed.ncbi.nlm.nih.gov/16991676/.

[16] J. A. Narloch and M. E. Brandstater, "Influence of breathing technique on arterial blood pressure during heavy weight lifting," PubMed, https://pubmed.ncbi.nlm.nih.gov/7741618/:~:text=The%20highest%20pressure%20recorded%20in,maximal%20lifting%20with%20slow%20exhalation.

[17] "The limits of human starvation," Field Exchange , https://www.ennonline.net/fex/15/limits.
[18] T Yoshizawa, K Ishikawa, H Nagasawa, Et al, "A Fatal Case of Super-super Obesity (BMI 80) in a Patient with a Necrotic Soft Tissue Infection", PubMed, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5995720/:∼:text=The%20patient%20ultimately%20died%20of,treated%20before%20lethal%20complications%20occur.
[19] H. Laurila, "Manual Data Entry Errors," Beamex Blog for Calibration Professionals, https://blog.beamex.com/manual-data-entry-errors:∼:text=This%20is%20despite%20the%20fact,data%20entry%20is %20about%201%20%25.
[20] "Identifying outliers: IQR method: Stat 200," PennState: Statistics Online Courses, https://online.stat.psu.edu/stat200/lesson/3/3.2.
[21] K. Feldman, Z-score: A handy tool for detecting outliers in Data, https://www.isixsigma.com/dictionary/z-score/:∼:text=Outlier%20detection,3%20is%20considered%20an%20outlier.
[22] "Healthy blood pressure by age and Gender (chart)," Baptist Health, https://www.baptisthealth.com/blog/heart-care/healthy-blood-pressure-by-age-and-gender-chart.
[23] A. Saini, "A beginner's guide to logistic regression," Analytics Vidhya, https://www.analyticsvidhya.com/blog/2021/08/conceptual-understanding-of-logistic-regression-for-data-science-beginners/.
[24] "sklearn.linear_model.logisticregression," scikit-learn, https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.
[25] C. Razieh, F. Zaccardi, J. Miksza, and E. al, "Differences in the risk of cardiovascular disease across ethnic groups: UK Biobank Observational Study," Nutrition, Metabolism and Cardiovascular Diseases, https://www.sciencedirect.com/science/article/pii/S0939475322003295:∼:text=Differences%20between%20ethnic%20groups%20were%20found%20across%20all%20these%20factors.&text=Higher%20CVD%20risk%20in%20South,independent%20of%20all%20measured%20factors.&text=Black%20individuals%20generally%20had%20similar,risk%20compared%20to%20white%20Europeans.

APPENDIX

A. Flowchart of Data Cleaning

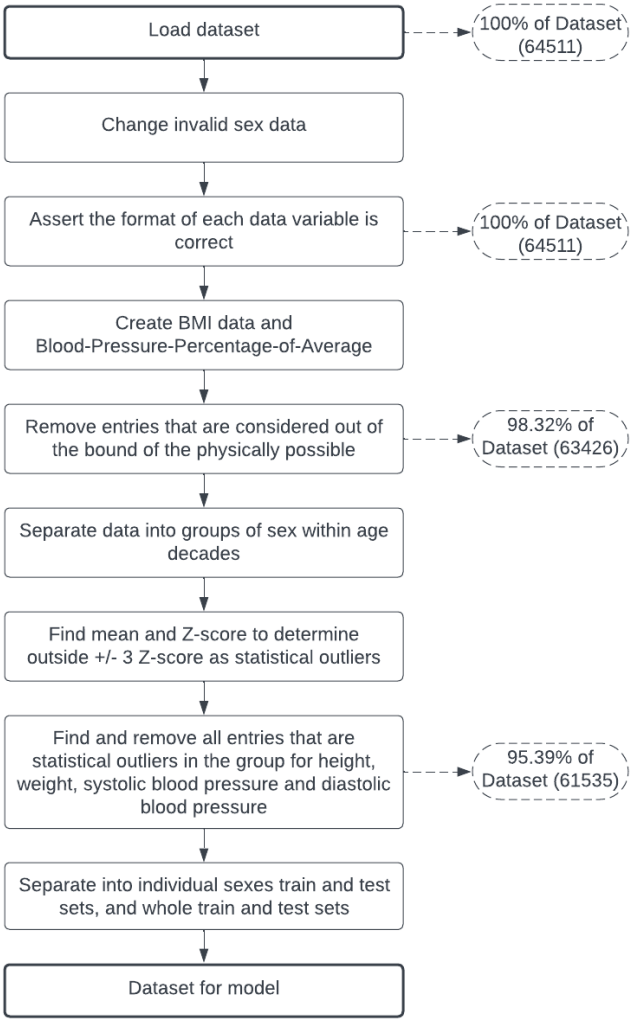

Fig. A.1. The process undertaken to clean the data, including how much of the dataset is left after the invalid entries have been removed

B. Dataset Post Clean Quantities

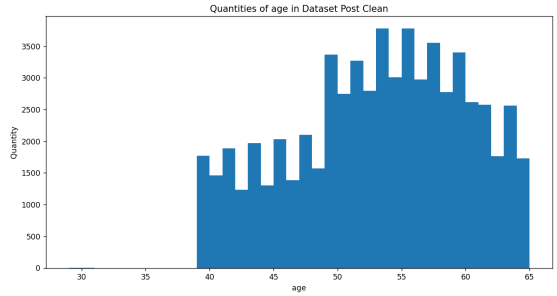

Fig. B.1. Quantity of Age in Years for Cleaned Dataset

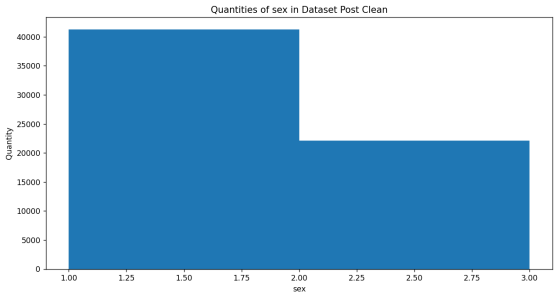

Fig. B.2. Quantity of Sex (1: Female, 2: Male) for Cleaned Dataset

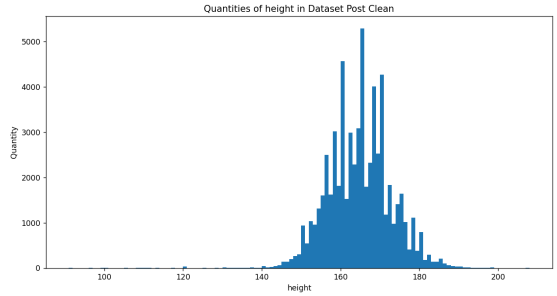

Fig. B.3. Quantity of Height in cm for Cleaned Dataset

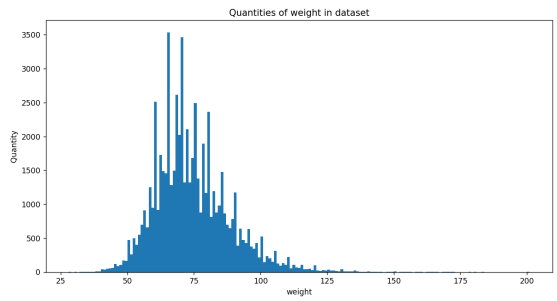

Fig. B.4. Quantity of Weight in kg for Cleaned Dataset

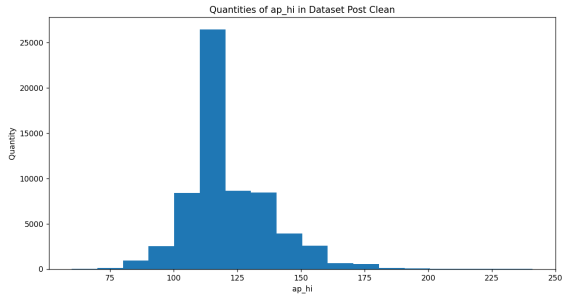

Fig. B.5. Quantity of Systolic Blood Pressure in mmHg for Cleaned Dataset

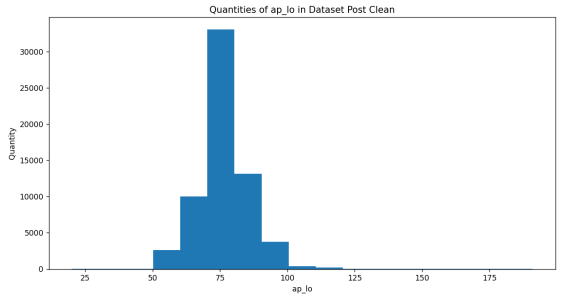

Fig. B.6. Quantity of Diastolic Blood Pressure in mmHg for Cleaned Dataset

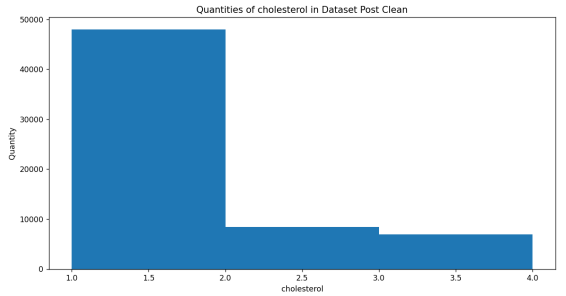

Fig. B.7. Quantity of Cholesterol (1: Normal, 2: Above Normal, 3: Well Above Normal) for Cleaned Dataset

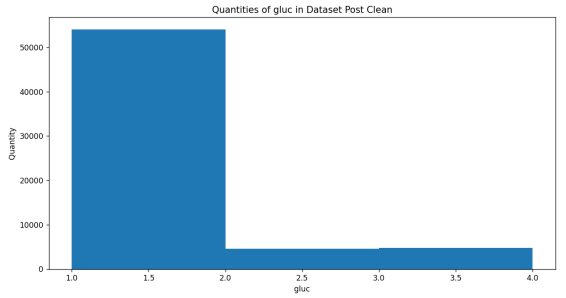

Fig. B.8. Quantity of Blood Glucose (1: Normal, 2: Above Normal, 3: Well Above Normal) for Cleaned Dataset

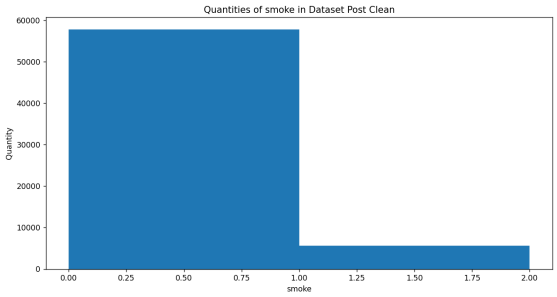

Fig. B.9. Quantity of Smoke Status (1: Smoker, 0: Non-smoker) for Cleaned Dataset

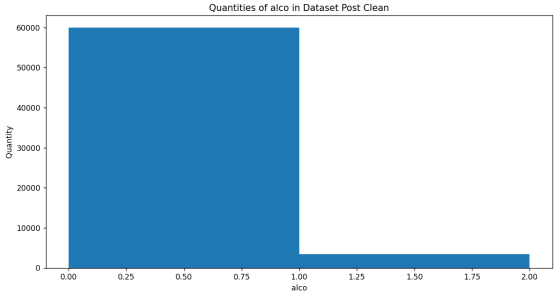

Fig. B.10. Quantity of Alcohol Consumption Status (1: Yes, 0: No) for Cleaned Dataset

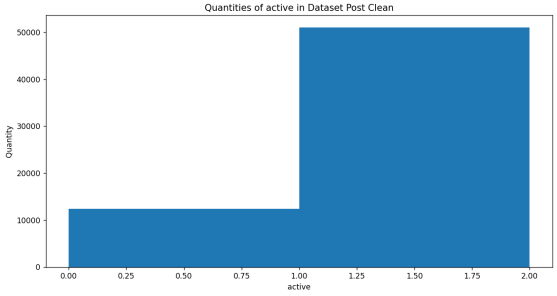

Fig. B.11. Quantity of Activity Status (1: Active, 0: Inactive) for Cleaned Dataset

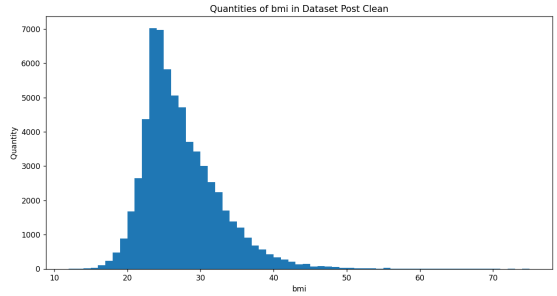

Fig. B.12. Quantity of BMI in kg/m^2 for Cleaned Dataset

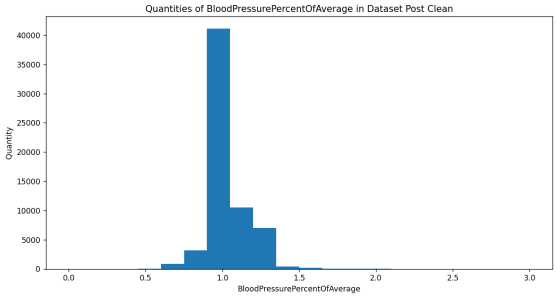

Fig. B.13. Quantity of Blood Pressure Percentage of Average for Age and Sex Group % for Cleaned Dataset

C. Confusion Matrices for First Runs of Models

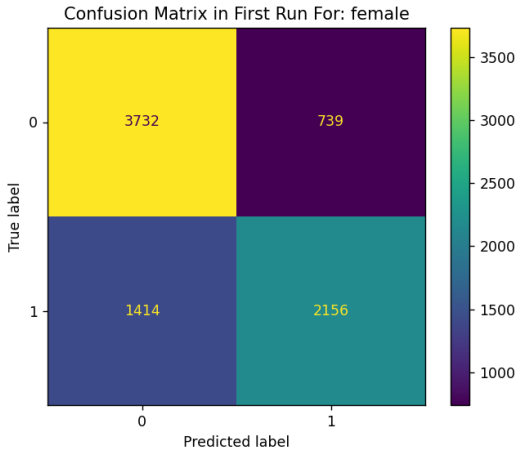

Fig. C.1. Confusion Matrix for First Run of Female Only Set

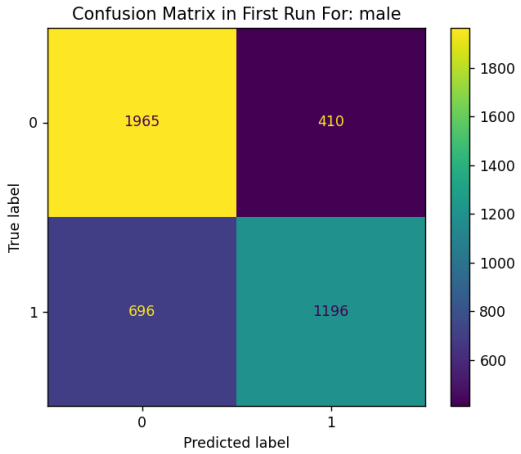

Fig. C.2. Confusion Matrix for First Run of Male Only Set

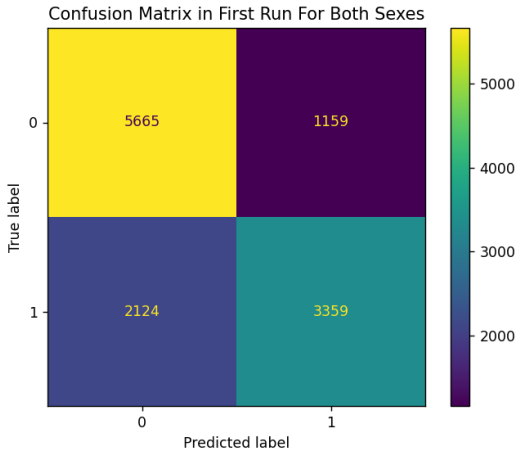

Fig. C.3. Confusion Matrix for First Run of Both Sexes Set

D. Percentage Misdiagnosises for Age by Decade for Single Sex Models

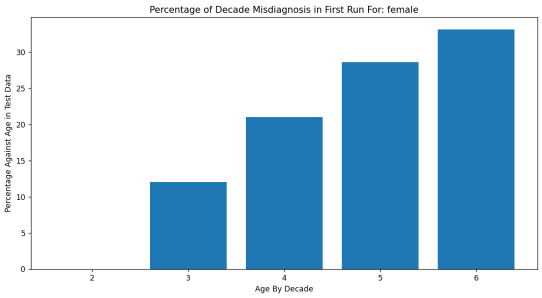

Fig. D.1. Misdiagnosis Percentages of Age by Decade for Female Model

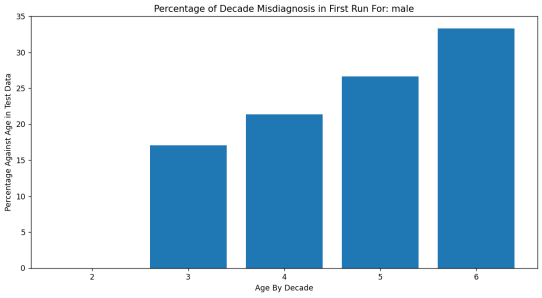

Fig. D.2. Misdiagnosis Percentages of Age by Decade for Male Model