Heart Data Unplugged - A Project Report

Dwanith Venkat Girish & Elaine Esther Oruk Opyene

Statistical Machine Learning - STAT-627
December 8, 2024
The American University

1 Table of Contents

Table of contents

1.1 PROJECT OVERVIEW

This project applies statistical machine learning techniques to the 2022 Behavioral Risk Factor Surveillance System (BRFSS) Survey dataset obtained from CDC.

The goal is to identify patterns and make accurate predictions about the risk of heart attacks for individuals based on their relevant demographic, health, and social information.

1.2 RESEARCH QUESTION

1. Can we accurately classify individuals as being at high or low risk for heart attack based on their medical history and social factors?
2. Which individual clinical parameters of individuals contribute to heart attacks?

2 VARIABLES & DESCRIPTIONS

```r
knitr::kable(variable_info, align = "l", col.names = c("Variable", "Description"))
```

Variable	Description
@_STATE	State FIPS code
SEXVAR	Sex of respondent
GENHLTH	General health
EXERANY2	Exercise in past 30 days
SLEPTIM1	How much time do you sleep
CVDINFR4	Ever diagnosed with Heart attack
CVDCRHD4	Ever diagnosed with Angina or Coronary heart disease
CVDSTRK3	Ever diagnosed with a Stroke
ASTHMA3	Ever told had Asthma
CHCCOPD3	(Ever told) you had Chronic Bronchitis
ADDEPEV3	(Ever told) you had a Depressive Disorder
CHCKDNY2	Ever told you have kidney disease?
DIABETE4	(Ever told) you had Diabetes
MARITAL	Marital status
EDUCA	Education level
VETERAN3	Are you a veteran
EMPLOY1	Employment status
INCOME3	Income level
WEIGHT2	Reported weight in pounds
HEIGHT3	Reported height in feet and inches

Variable	Description
LCSNUMCG	On average, how many cigarettes do you smoke each day
AVEDRNK3	Avg alcoholic drinks per day in past 30
LSATISFY	Satisfaction with life
@_IMPRACE	Imputed race/ethnicity value
@_URBSTAT	Urban/rural status
@_PHYS14D	Computed physical health status
@_MENT14D	Computed mental health status
HTM4	Computed height in meters
WTKG3	Computed weight in kilograms
@_BMI5	Computed body mass index

2.1 Response Variable:

The primary response variable for Research Question I is the Risk Group (High Risk vs Low Risk), which is derived by combining information on whether an individual has ever been diagnosed with a heart attack, heart disease, or stroke.

The primary response variable for Research Question II is the (CVDINFR4) variable in response to the question “Ever diagnosed with Heart attack?” which has a binary response (0 = No, 1 = Yes).

3 LITERATURE REVIEW

The article “Artificial Intelligence, Machine Learning, and Cardiovascular Disease” by Mathur et al. (2020) explores how AI and ML have revolutionized cardiovascular medicine, particularly in enhancing diagnostic accuracy, predicting risks, and personalizing treatments. These advancements align with our project’s goal of leveraging AI techniques to classify cardiovascular risk factors using the 2022 CDC BRFSS survey dataset.

3.1 Relevance to the Project

Supervised Learning:

Mathur et al. (2020) highlight the application of supervised learning techniques in cardiovascular medicine, particularly for analyzing clinical variables, imaging data, and outcomes. For instance, Khamis et al. successfully applied supervised learning for the automatic classification of echocardiograms, achieving high accuracy in real-time cardiac function assessments.

Building on these findings, our project uses supervised learning models to predict cardiovascular risk categories and occurrences. Our dataset includes labeled patient data, such as the presence of diabetes, kidney disease, and chronic bronchitis, as well as key demographic factors. By training on this labeled data, our models aim to replicate the high predictive accuracy seen in similar applications while adapting it to our dataset's unique features.

Big Data Analytics:

Mathur et al. (2020) emphasize the critical role of big data in AI applications for cardiovascular medicine. Large datasets, including clinical variables and patient health records, are essential for training machine learning models and ensuring precision in personalized treatments. However, they also note challenges such as handling missing data, ensuring data privacy, and the computational demands of big data analytics.

Similarly, our project leverages the large-scale 2022 CDC BRFSS survey dataset to classify cardiovascular risks. This dataset provides an extensive range of demographic and health-related variables, offering a rich foundation for machine learning applications. Additionally, our project addresses common big data challenges by employing preprocessing techniques to handle missing data and improve model robustness. This integration of big data allows for scalable and adaptable cardiovascular risk prediction models, with the potential for incorporating additional datasets in future iterations.

4 LOGISTIC REGRESSION

Diabetes, Kidney Disease, Chronic Bronchitis:

Individuals with diabetes (e.g., DIABETE42: 0.958, $p < 0.001$), kidney disease (e.g., CHCKDNY22: 1.041, $p < 0.001$), and chronic bronchitis (e.g., CHCCOPD32: 0.896, $p < 0.001$) have higher odds of being at risk for a heart attack compared to those without these conditions.

Race:

Hispanics (@_IMPRACE5: 0.564, $p < 0.001$) are more likely to be at risk for heart attack compared to the reference group (White, non-Hispanic)

Income:

Individuals in higher income groups particularly those earning 50,000 USD and above (e.g., INCOME39: 0.865, $p < 0.001$) are more likely to be at risk for heart attack compared to lower income categories.

Gender:

Females (SEXVAR: 0.729, $p < 0.001$) have higher odds of being at risk for a heart attack compared to males.

Weight:

Individuals with higher weight (WEIGHT2: 0.00103, $p = 0.0137$) have slightly increased risk of heart attacks.

Alcohol Consumption:

Higher alcohol consumption (AVEDRNK3: 0.107, $p < 0.001$) is associated with increased odds of being at risk for a heart attack.

Sleep: More sleep (SLEPTIM1: -0.041, $p = 0.0004$) is associated with lower odds of being at risk for a heart attack.

Exercise: Exercise (EXERANY2: -0.272, $p < 0.001$) is negatively related to heart attack risk, meaning those who exercise are less likely to be at risk.

Cigarette Smoking: The relationship between smoking and heart attack risk is negative (LCSNUMCG: -0.015, $p < 0.001$), but this result is unexpected and warrants further investigation.

The model is over-predicting "Low" risk for most observations resulting in a low accuracy level of 13.77%. This is likely due to class imbalance as shown below.

The class imbalance in the dataset is quite significant, with a much larger number of "Low" instances compared to "High" instances. We will balance the dataset by oversampling the minority class and undersampling the majority class using the `ROSE` function.

```r
summary(log_model_balanced)
```

```
Call:
glm(formula = RiskGroup ~ DIABETE4 + CHCKDNY2 + CHCCOPD3 + IMPRACE +
    INCOME3 + LCSNUMCG + AVEDRNK3 + SLEPTIM1 + WEIGHT2 + HEIGHT3 +
    EXERANY2 + SEXVAR, family = binomial(link = "logit"), data = balanced_data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.391e+00  2.110e-01  16.069  < 2e-16 ***
DIABETE42   -1.223e+00  1.749e-01  -6.992 2.71e-12 ***
DIABETE43   -8.376e-01  3.589e-02 -23.335  < 2e-16 ***
DIABETE44   -3.528e-01  8.259e-02  -4.272 1.94e-05 ***
DIABETE47   -8.305e-03  3.325e-01  -0.025  0.98008
DIABETE49   -1.120e+01  7.704e+01  -0.145  0.88440
CHCKDNY22   -1.070e+00  5.716e-02 -18.719  < 2e-16 ***
CHCKDNY27   -6.111e-01  2.581e-01  -2.367  0.01792 *
CHCKDNY29   -1.215e+01  1.040e+02  -0.117  0.90700
```

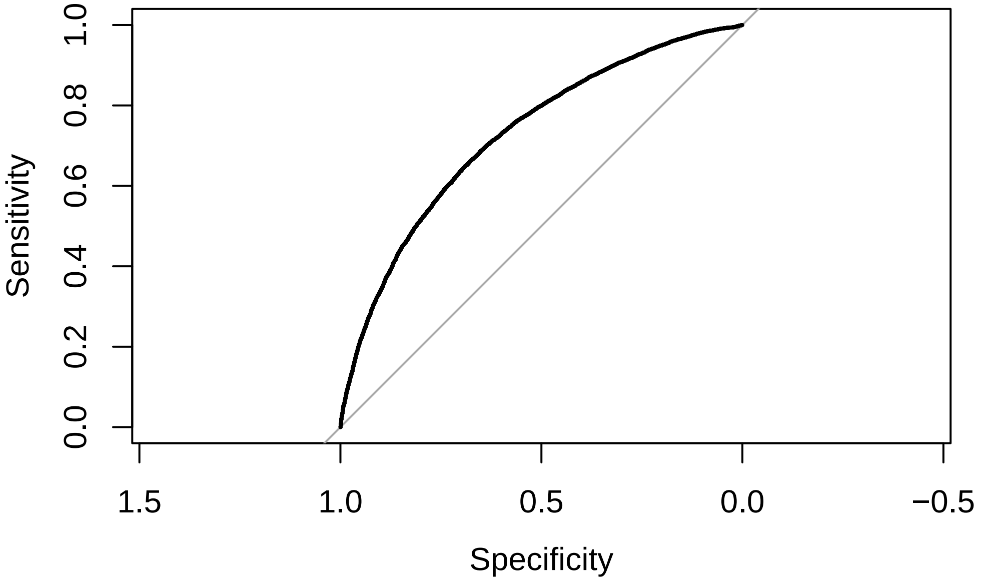
```
CHCCOPD32   -9.689e-01  3.513e-02 -27.581  < 2e-16 ***
CHCCOPD37   -5.401e-01  2.114e-01  -2.554  0.01064 *
IMPRACE2    -6.327e-02  5.495e-02  -1.151  0.24960
IMPRACE3    -9.904e-01  1.644e-01  -6.022 1.72e-09 ***
IMPRACE4     4.501e-02  9.432e-02   0.477  0.63320
IMPRACE5    -5.468e-01  5.901e-02  -9.267  < 2e-16 ***
IMPRACE6    -6.647e-01  9.181e-02  -7.240 4.49e-13 ***
INCOME32     2.462e-01  1.159e-01   2.124  0.03364 *
INCOME33     2.258e-01  1.094e-01   2.064  0.03901 *
INCOME34    -1.664e-01  1.025e-01  -1.623  0.10458
INCOME35    -2.812e-01  9.458e-02  -2.974  0.00294 **
INCOME36    -2.956e-01  9.343e-02  -3.163  0.00156 **
INCOME37    -4.529e-01  9.219e-02  -4.913 8.99e-07 ***
INCOME38    -5.567e-01  9.364e-02  -5.945 2.77e-09 ***
INCOME39    -7.583e-01  9.428e-02  -8.043 8.75e-16 ***
INCOME310   -9.344e-01  1.042e-01  -8.970  < 2e-16 ***
INCOME311   -8.507e-01  1.043e-01  -8.160 3.36e-16 ***
INCOME377   -2.453e-01  1.010e-01  -2.427  0.01521 *
INCOME399   -2.651e-01  9.768e-02  -2.714  0.00665 **
LCSNUMCG     1.632e-02  9.445e-04  17.285  < 2e-16 ***
AVEDRNK3    -7.489e-02  5.085e-03 -14.726  < 2e-16 ***
SLEPTIM1     4.444e-02  7.404e-03   6.003 1.94e-09 ***
WEIGHT2      2.153e-04  2.757e-04   0.781  0.43487
HEIGHT3     -7.076e-04  2.966e-04  -2.386  0.01705 *
EXERANY2     2.737e-01  2.674e-02  10.233  < 2e-16 ***
SEXVAR      -5.904e-01  2.589e-02 -22.807  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 44519  on 32113  degrees of freedom
Residual deviance: 39383  on 32079  degrees of freedom
AIC: 39453

Number of Fisher Scoring iterations: 10
```

In the new balanced data log model, individuals in very low income groups (earning between 15,000 USD and 20,000) (e.g., INCOME33: 2.258e-01, $p < 0.05$) are also slightly likely to be at risk for heart attack as well as individuals who fall in the racial category of Asian, Non-Hispanic (e.g. IMPRACE3 : -9.904e-01, $p < 0.01$).

The model's accuracy has improved to 66.68%.

- Precision (65.1%): Of all predicted "1s" (high-risk), 65.1% were correct.
- Recall (72.08%): Of all actual "1s" (high-risk), 72.08% were correctly identified.
- F1-Score (68.41%): The balance between precision and recall, reflecting overall effectiveness in predicting high-risk cases.

```
plot(roc_curve)
```



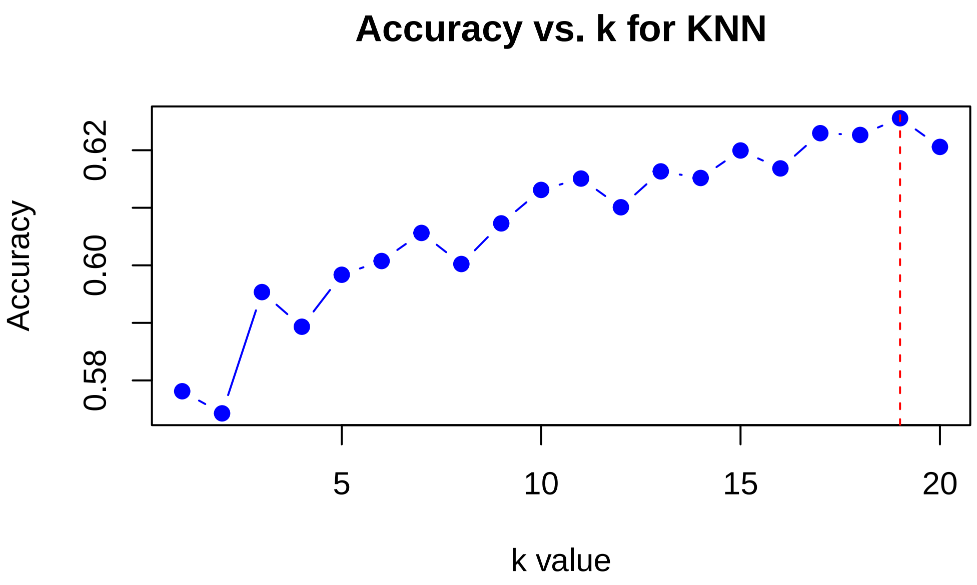
Since our AUC is 0.7268, this suggests that the model is performing reasonably well with a moderate ability to distinguish between the classes, but there's still room for further improvement.

4.1 Comparison of Logistic Regression Model Predictions vs Actual Values

5 KNN METHOD

We split the data by 70% and determined the optimal k for this split that gives us the maximum level of accuracy.

```
plot(1:20, accuracy_values, type = 'b', pch = 19, col = 'blue',
     xlab = "k value", ylab = "Accuracy", main = "Accuracy vs. k for KNN")
abline(v = 19, col = "red", lty = 2)
```



The optimal value of k for the KNN model was found to be 19, which resulted in an accuracy of 62.59%.

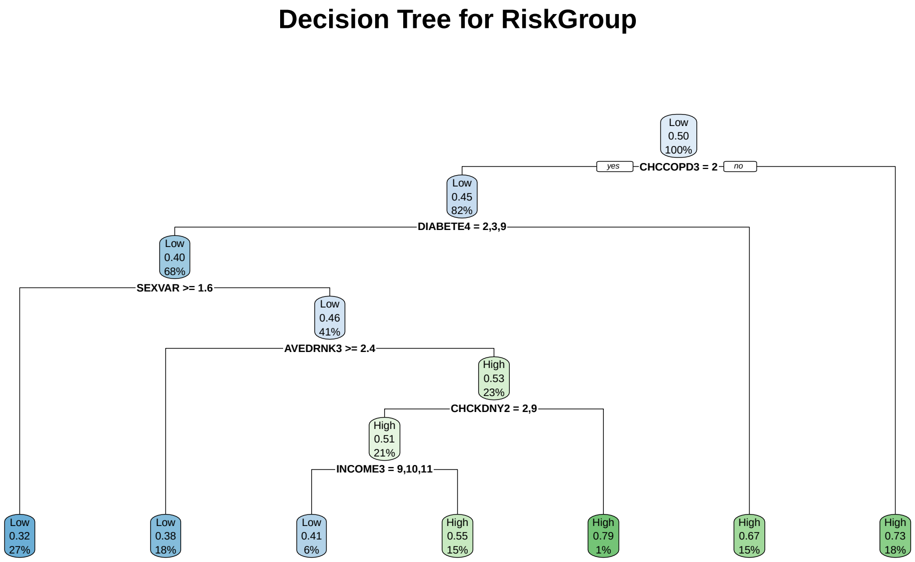
Sensitivity is 71.35%, reflecting a solid performance in predicting the 'Low' risk group while Specificity is 53.80%, indicating a little room for improvement in identifying the 'High' risk group.

However, this model is slightly less accurate than the one for logistic regression (66.68%).

5.1 Comparison of KNN Model Predictions vs Actual Values

6 DECISION TREES

```
rpart.plot(tree_model, main = "Decision Tree for RiskGroup")
```



This Decision Tree model has an accuracy level of 57.53% which is not very good compared to the other models we have previously used i.e. logistic regression and KNN.

This model has an accuracy level of 57.53% which is not very good compared to the other models we have previously used i.e. logistic regression and KNN.

6.1 Comparison of Tree Model Predictions vs Actual Values

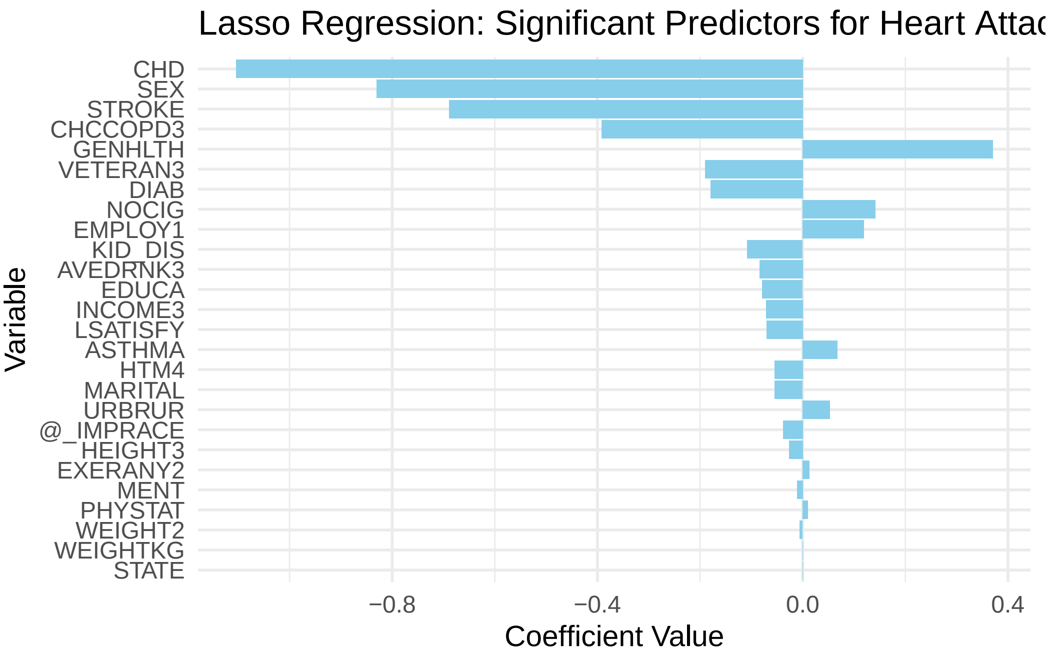
7 RESEARCH QUESTION REFORMATTED

Can we accurately predict individuals getting a heart attack based on their medical history and social factors?

- Numeric variables: WEIGHT2, HEIGHT3, NOCIG, AVEDRNK3, HTM4, WEIGHTKG, BMI
- Categorical variables: (all variables apart from the above)
- Dependent variable: HA (heart attack, 0 = no, 1 = yes)

#Lasso regression

```
lassoplot
```



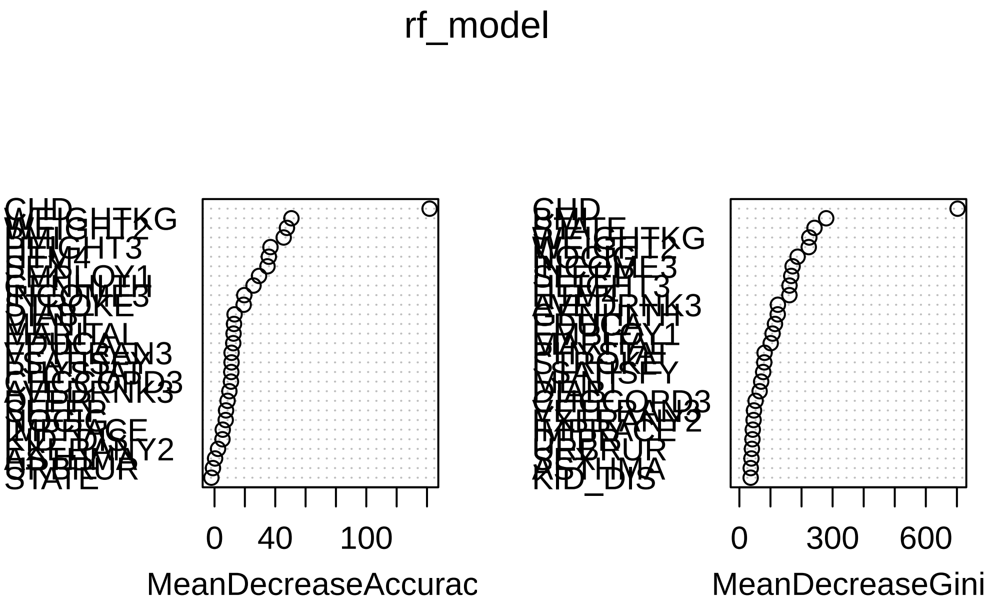
##Interpretation

The Lasso regression analysis on the CDC dataset identified negative coefficients for “CHD” (Coronary Heart Disease), “SEX” (male), “STROKE”, and “CHCCOPD3” (Chronic Bronchitis) as significant predictors of heart attacks. These results underscore that these conditions and male gender are associated with an increased likelihood of heart attacks in the studied population.

#Spline Regression

#Random Forest

```
varImpPlot(rf_model)
```



In our Random Forest model, “WEIGHTKG”, “HEIGHT3”, and “CHD” were identified as key predictors, significantly influencing accuracy and Gini impurity, highlighting their importance in heart attack risk assessments.

8 CHALLENGES AND ETHICAL CONSIDERATIONS

Data Privacy Concerns:

The integrity and privacy of patient data are paramount in any healthcare-related ML project. Since our dataset includes sensitive health and demographic information, it is essential to ensure compliance with data protection and privacy regulations.

While we were not working directly with medical records or real-time patient data and our informants were kept anonymous, the principles of data privacy still apply. If our project were to scale or incorporate additional data sources, it would be important to ensure that privacy concerns are addressed.

Data Integrity Issues (Poor Data Selection, Biases):

Mathur et al. (2020) highlights that AI-based systems are only as good as the data they are trained on. Issues like poor data selection, selection bias, and unintentional biases can lead to inaccurate or discriminatory predictions (White House Reports, 2016; Petersen et al., 2020).

For instance, our data source predominantly represented low-risk individuals, which caused the ML models to perform poorly for high-risk groups. We addressed this by oversampling the minority group to balance the dataset for training.

Reproducibility of ML Models:

ML models, particularly those trained on complex, high-dimensional datasets, require reproducibility to ensure that they can be trusted in clinical settings (Petersen et al., 2020).

Similarly, in our project, we used standardized methods for data pre-processing and model validation, ensuring that our ML model's results can be reliably reproduced across different environments or with new data.

8.1 Key Predictors of Heart Attack:

Our project tested multiple machine learning models for cardiovascular risk prediction, with Logistic Regression performing the best at 66.68% accuracy, followed by KNN (62.59%) and Decision Tree (57.53%).

Our study applied Lasso regression, spline, and random forests to the CDC dataset, identifying significant predictors of heart attack risk: 'CHD' (Coronary Heart Disease), 'SEX' (male), 'STROKE', and 'CHCCOPD3' (Chronic Bronchitis). These findings address our first research question by demonstrating that machine learning can accurately classify individuals as being at high or low risk for a heart attack based on their medical history and social factors. Regarding our second question, the significant predictors—CHD, male gender, history of stroke, and chronic bronchitis—highlight which individual clinical parameters most contribute to heart attacks.

Splines confirmed non-linear relationships for numeric predictors like WEIGHT2 and BMI.

Lasso and Splines achieved the highest accuracy (94.1%), demonstrating the effectiveness of variable selection and non-linear modeling. Random Forest closely followed (93.92%) and provided robust predictions with feature importance insights.

Predictive accuracy across models indicates consistency in identifying important health related factors. Random Forest proved useful for mixed data types, while Lasso and Splines excelled in variable focused and relationship specific modeling.

9 CONCLUSIONS AND RECOMMENDATIONS

These findings show potential but also highlight areas for improvement, like balancing the dataset, refining features, clubbing variables and eventually scaling numeric variables yielded sound results. While the results are promising, the models need to be more accurate and reliable for practical use, showing just how important quality data and thoughtful model design are in healthcare AI.

10 APPENDIX

I, Elaine Esther Oruk Opyene have contributed to this project in terms of data processing, classification tasks which include Logistic regression, K nearest neighbors and decision trees, and towards the write-up of the report.

I, Dwanith Venkat Girish, have contributed to this project in terms of finding the dataset, conducting prediction tasks which include lasso regression, spline regression and RandomForest methods, and as well as towards the write-up of the report.

10.1 REFERENCES

- Mathur P, Srivastava S, Xu X, Mehta JL. Artficial Intelligence, Machine Learning, and Cardiovascular Disease. Clinical Medicine Insights: Cardiology. 2020;14. doi:10.1177/1179546820927404
- Khamis H, Zurakhov G, Azar V, Raz A, Friedman Z, Adam D. Automatic api cal view classification of echocardiograms using a discriminative learning dic tionary. Med Image Anal. 2017;36:15-21
- Petersen SE, Abdulkareem M, Leiner T. Artificial intelligence will transform cardiac imaging-opportunities and challenges. Front Cardiovasc Med. 2019;6:133.