Module 7: Portfolio Project

Final Research Paper

Michelle Goodwin

Colorado State University-Global Campus

MIS 581: Capstone- Business Intelligence and Data Analytics

Dr. Osama Morad

July 28, 2024

Abstract

Heart disease remains the leading cause of mortality worldwide, therefore, research and predictive methods are exceedingly necessary. This study aims to develop and evaluate predictive models for identifying heart disease with additional focus on gender disparities and the impact of age in heart disease susceptibility. Utilizing the UCI Heart Disease Dataset, logistic regression and random forest models were employed to identify patients that are at risk. The logistic regression model performs well with a 75.21% accuracy and an AUC of 0.869, while the random forest model outperforms the logistic regression with an accuracy of 78.51% and an AUC of 0.783. Statistical analysis revels significant relationships between heart disease risk and both gender and age. Females are identified to have a higher average risk compared to males, age was identified as a significant predictor, although contrary to many relevant literature findings, indicating that there are additional significant risk factors that influence heart disease. These findings highlight the effectiveness of predictive models in heart disease risk assessment and emphasize the necessity of incorporating demographic factors to enhance model's predictive accuracy and use in clinical settings. Further research should focus on expanding datasets and refining models to include a broader range of risk factors to improve the model's generalizability and robustness.

Final Research Paper

Introduction

Heart disease is the leading cause of mortality worldwide, requiring research and effective prevention and predictions methods to be performed. Advances in data analytics and machine learning offer a potential tool for developing predictive models that can identify individuals at risk for heart disease and uncover potential factors that contribute to heart disease susceptibility. This study attempts to develop and evaluate predictive models for heart disease risk and to analyze the impact of gender and age on heart disease susceptibility using the UCI Heart Disease Dataset retrieved from Kaggle.

Objectives

The primary objectives of this study are to develop and evaluate predictive models to identify patients at risk for heart disease, examine the relationship between gender and heart disease risk, and investigate the influence of age on heart disease susceptibility. By addressing these objectives, the study aims to provide a comprehensive understanding of how predictive models can be utilized as a tool in clinical setting to improve heart disease diagnosis by identifying patients with heart disease faster and more accurately; therefore, allowing physicians to more effectively prevent complications and decrease patient mortality rates.

Overview of the Study

This study will address three research questions related to heart disease risk prediction: can a predictive model identify patients at risk for heart disease, are females or males more at

risk for developing heart disease, and is age a significant factor in heart disease susceptibility? This study will include a literature review that examines existing research on the efficiency of predictive models in identifying patients at risk for heart disease and explore gender and age as significant contributing factors to heart disease. The research design employs a quantitative approach, utilizing the UCI Heart Disease Dataset to develop and evaluate various predictive models. The methodology includes data collection, preprocessing methods, model development, and performance evaluations. Statistical tests will be conducted to assess the significance of relationships between independent variables (i.e., predictive model accuracy, gender, and age) and the dependent variable (i.e., heart disease risk).

Research Questions and Hypotheses

The hypotheses and research questions that I will address are:

1. Can a predictive model identify patients at risk for heart disease?

   H_0: There is no significant relationship between the predictive model and the accuracy of identifying patients with heart disease.

   H_1: There is a significant relationship between the predictive model and the accuracy of identifying patients with heart disease.

2. Are females or males more at risk for heart disease?

   H_0: There is no significant relationship between gender and the risk of heart disease.

   H_1: There is a significant relationship between gender and the risk of heart disease.

3. Is age a significant factor in heart disease?

H_0: There is no significant relationship between age and heart disease susceptibility.

H_1: There is a significant relationship between age and heart disease susceptibility.

Literature Review

The increasing prevalence of heart disease worldwide necessitates the development of effective predictive models and a comprehensive understanding of risk factors. This literature review examines existing research on the efficacy of predictive models in identifying patients at risk for heart disease and investigates gender and age as significant factors in heart disease susceptibility. The review is structures around three hypotheses, the accuracy of predictive models, gender-based disparities, and age-related sustainability in relation to heart disease.

Predictive Models for Heart Disease Risk

Predictive models use statistical techniques and machine learning algorithms to estimate the likelihood of a patient developing heart disease. Various models have been developed, including logistic regression, neural network models, deep machine learning algorithms, and hybrid models.

Accuracy of Predictive Models

Several studies have evaluated the performance of predictive models in identifying heart disease risk. A study by Desai et al (2019) compared two classification models, using back propagation neural networks (BPNN) and logistic regression to predict heart disease, and determined that linear regression shows higher accuracy with 92.58% compared to BPNN with 85.07% accuracy. In contrast, Al Reshan et al., (2023) found that a hybrid deep neural network

model that combines convolutional neural networks and long short-term memory with additional dense layers outperforms these models with 98.56% accuracy.

Gender-based Differences in Heart Disease Risk

Gender disparities in heart disease prevalence and outcomes have been widely documented. Men traditionally have been considered at a higher risk for heart disease (Weidner, 2000); however, more recent studies suggest that women when post-menopausal are at significant risk (Regitz-Zagrosek, 2003).

Empirical Evidence

Several large cohort studies have investigated gender-based differences in heart disease risk. For instance, the UK Biobank highlighted many differences in risk factors and disease progression between males and females, including blood pressure smoking, diabetes, and smoking (Millett et al., 2018). Additionally, a study by Woodward (2019) found that while men have a higher incidence rate of heart disease, women experience more severe outcomes.

Predictive Models and Gender

Predictive models often incorporate gender as a variable to improve accuracy. Research by Paulus et al. (2015) showed that models accounting for gender differences provide more accurate predictions. However, some studies suggest that there are other risk factors that play a larger role in heart disease prediction (Wilhelmsen et al., 1973).

Age as a Significant Factor in Heart Disease

Age is a well-established risk factor for heart disease, with incidence rates increasing significantly with age. The physiological changes associated with aging, such as arterial stiffness

and increased blood pressure, contribute to higher patient susceptibility to heart disease (Sun, 2015).

Empirical Evidence

Many studies have provided robust evidence linking heart disease risk to age. A study by the Framingham Heart Study cohort found that heart disease increases exponentially with age, especially after age 50 (Peeters et al., 2002).

Integration in Predictive Models

Incorporating age into predictive models is shown to enhance their predictive power. Research by Dhingra & Vashan (2012) demonstrates that age is a significant factor in heart disease prediction. However, the interaction between age and other risk factors, such as hypertension and diabetes, needs further exploration to refine its predictive accuracy.

Research Design

The research design for this study involves a quantitative approach to examine the relationship between predictive accuracy, gender, age, and heart disease risk. The study will utilize the UCI Heart Disease Dataset retrieved from Kaggle, which includes a variety of health indicators and demographic information. The primary goal is to develop and evaluate predictive models for heart disease risk to analyze the influence of gender and age on heart disease susceptibility.

Methodology

The data collection process begins with the retrieval of the UCI Heart Disease dataset from Kaggle, which will be used for this study. This dataset contains health metrics, such as age,

gender, blood pressure, cholesterol levels, and more. For data preprocessing, I will handle missing values by implementing imputation techniques or removing incomplete records. I will normalize numerical features to ensure that they are on a comparable scale. For model development, there will be several machine learning algorithms implemented from logistic regressions to neural networks. These models will be trained and tested with an 60/40 split ratio of training and testing subsets respectively. To determine model performance, the models will be assessed using metrics, such as accuracy, precision, recall, and F1 score. Additionally, cross-validation will be conducted to ensure the models are robust and can be generalized to many healthcare systems.

Methods

The UCI Heart Disease dataset contains mostly quantitative data on patients that were assessed for heart disease. The target variable of the data indicates 0 for a normal heart function and 1 for a fixed defect. The quantitative data in this dataset are age, resting blood pressure (trestbps), cholesterol (chol), maximum heart rate (thalach), and ST depression induced by exercise (oldpeak). The categorical variables in this dataset include gender (sex), chest pain type (cp), fasting blood sugar greater than 120 mg/dl (fbs), resting electrocardiographic results (restecg), slope of the ST segment relative to heart rate (slope), exercise induced angina (exang), thalassemia (thal), and number of vessels colored by fluoroscopy (ca).

In this study, predictive modeling techniques will be utilized to develop and evaluate models for predicting heart disease risk. The primary models considered include logistic regression and a random forest decision tree model. Logistic regression will serve as the baseline model, as a binary classification problem that will be used to predict the likelihood of a patient

having heart disease based on various health indicators and demographic factors. Logistic regression is a simple, easily interpretable, and effective tool. Additionally, random forest ensemble learning methods construct decision trees during the training process and output the classes of the individual trees. This method is known for improving accuracy and handling potential overfitting concerns.

To evaluate the effectiveness of the predictive models and investigate the relationships between the independent variables (i.e., predictive model accuracy, gender, and age) and the dependent variable (i.e., heart disease risk) several statistical tests will be employed. Chi-square tests will be applied to categorical data to assess whether there is a significant association between gender and heart disease risk. T-tests will be used to compare the means of two groups (e.g., heart disease risk between male and female patients) to determine if there is a statistically significant difference between the groups.

Regression analysis will be used to determine the impact of age in heart disease risk and to identify significant predictors among the various health indicators and demographic factors. This analysis will aid in understanding the relationship between the age and heart disease risk, allowing us to understand how these factors may contribute to heart disease.

Limitations

While the UCI Heart Disease dataset is widely used and offers ample information, it does have some limitations. Firstly, it may not capture potential risk factors for heart disease. The dataset includes variables such as age, gender, blood pressure, cholesterol, and more, but it does not include other significant factors like important lifestyle factors, genetic predispositions, or

socioeconomic status of the patient. This omission could lead to an incomplete model that does not fully account for the complexity of heart disease risk. Additionally, the demographic representation within the dataset may limit the generalizability of the findings.

Predictive models developed using the UCI Heart Disease dataset may face challenges related to overfitting, especially if the data is imbalanced. Overfitting could occur if the data does not show a balanced representation of male versus female groups, heart disease cases versus non-heart disease cases, young versus elderly groups, and so on. Using machine learning algorithms will play a crucial role in improving the performance and accuracy of these model predictions. For example, separating the dataset into a training and testing subset will ensure there is no overfitting.

The external validity of the results is another concern. Findings form the UCI Heart Disease dataset may not be applicable to all populations, particularly those outside of the demographic scope of the dataset. For instance, the dataset includes information obtained from Cleveland, Hungary, California, and Switzerland; therefore, the predictive models developed in this study may not be relevant for other geographic regions such as Cairo. To enhance the robustness and applicability of these findings further validation from external datasets will be necessary. The validation process would involve testing the predictive models on different datasets to ensure the results hold true across various populations and settings to determine the models' reliability.

Ethical Considerations

Ensuring the confidentiality and privacy of patient data is essential to this study. To protect individual identities, the dataset has undergone rigorous anonymization processes. This involved removing any personally identifiable information such, as names, social security numbers, and addresses (UCI Machine Learning Repository, n.d.). The anonymization process adheres to strict data protection standards that ensure that the re-identification process of these patients is not possible. Implementing these measures ensure that the privacy of patients comply with HIPPA regulations.

Minimizing bias in predictive models is essential to ethical considerations for this research. Bias can arise from various resources, including how the data is collected, processes, and interpreted. To address this the study considers how variables such as gender and age are incorporated into the models. Ensuring that these variables are represented accurately and fairly is essential to avoid skewed predictions that could disproportionately affect groups. Additionally, this study will emphasize the importance of transparency in reporting model performance across different demographic groups. By providing detailed performance metrics for each group, the study ensures fairness while highlighting and disparities that may need to be addressed in future research (Li et al., 2023).

The ethical integrity of the dataset is another important factor to consider. Although the dataset is publicly available, it is essential that the original collection of data has adhered to strict ethical standards, including informed consent from the participants when the data was obtained (UCI Heart Disease Data Set, 2021). Informed consent ensures that the participants were aware of the nature of the study, the use of their data, and any potential risks involved. It also guarantees that they voluntarily agreed to participate. Upholding these standards ensures that

ethical considerations are at the foundation of each research project and respects the right of the participants.

This study aims to advance the understanding of heart disease risk factors by developing and evaluating predictive models and analyzing the impact of gender and age on heart disease. Through data preprocessing, model development, and evaluation, this research will provide valuable insight into the efficiency of predictive models and the significance of demographic factors in heart disease. Future work should focus on expanding the dataset, refining the predictive models, and exploring additional risk factors to further enhance the predictive accuracy and generalizability.

Findings

Predictive Model

The logistic regression model in figure 1 is evaluated in figure 2 to provide insight into the predictive abilities of identifying patients at risk for heart disease. The model was evaluated using a variety of metrics, indicating that the independent variables sex, chest pain type (cp), maximum heart rate achieved (thalach), ST depression induced by exercise relative to rest (oldpeak), and number of major vessels colored by fluoroscopy (ca), significantly influence heart disease risk. The confusion matrix shows that the model achieves an accuracy of 75.21% with a 95% confidence interval of 66.54% to 82.60%. The sensitivity (recall) is 65.45% and specificity is 83.33%, demonstrating that the model is more effective in correctly identifying true negatives than true positives. The positive predictive value (precision) is 76.60%, and the F1 score is 70.59%, indicating a balance performance between precision and recall. The area under the ROC

curve (AUC) is 0.869, highlights the model's strong discriminatory power. These results support rejecting the null hypothesis, indicating a significant relationship between the predictive model and the accuracy of identifying patient with heart disease.

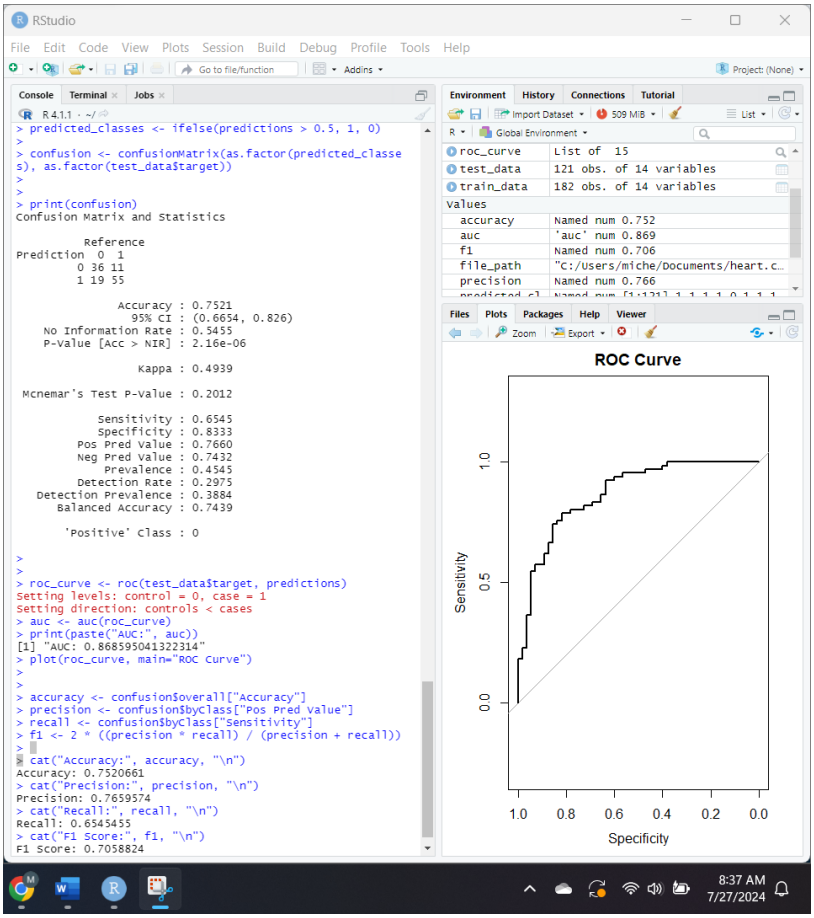
Figure 1

Logistic Regression Predictive Model

```
> logistic_model<- glm(target ~ ., data=train_data, family=
binomial)
> summary(logistic_model)

Call:
glm(formula = target ~ ., family = binomial, data = train_d
ata)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.97434  -0.31355   0.09571   0.50702   2.83972

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.810906   3.631425   1.325 0.185238
age         -0.039371   0.032795  -1.201 0.229942
sex         -2.115277   0.653947  -3.235 0.001218 **
cp           1.008265   0.265517   3.797 0.000146 ***
trestbps    -0.020821   0.015256  -1.365 0.172332
chol        -0.006005   0.004959  -1.211 0.225952
fbs          0.005500   0.729685   0.008 0.993986
restecg     -0.248127   0.506798  -0.490 0.624418
thalach      0.033756   0.014739   2.290 0.022006 *
exang       -0.817871   0.638215  -1.281 0.200019
oldpeak     -0.677221   0.313164  -2.163 0.030579 *
slope        0.082241   0.525914   0.156 0.875735
ca          -0.760241   0.238664  -3.185 0.001445 **
thal        -0.536856   0.391637  -1.371 0.170438
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 250.90  on 181  degrees of freedom
Residual deviance: 113.45  on 168  degrees of freedom
AIC: 141.45

Number of Fisher Scoring iterations: 6
```

Figure 2

Logistic Regression Performance Metrics



The results from the random forest model in figure 3 provides insight into how a predictive model can accurately identify patients at risk for heart disease. The model, trained with 500 trees, yielded an accuracy of approximately 78.51% on the testing data. The 95% confidence interval for accuracy ranged from 70.12% to 85.46%, indicating robust performance. The model's sensitivity and specificity were 76.36% and 80.30%, respectively. This suggests that it performs well in correctly identifying patients with and without heart disease. The area under the ROC curve (AUC) was 0.783, further supporting the model's strong predictive power. Additionally, the positive predictive value (precision) and the F1 score are both 76.36%, which shows that the model maintains a balanced performance between precision and recall. These metrics demonstrate that the random forest model has a significant relationship with the accuracy

of identifying patients with heart disease, allowing the null hypothesis to be rejected. Therefore, this predictive model effectively identifies patients at risk for heart disease in this dataset.

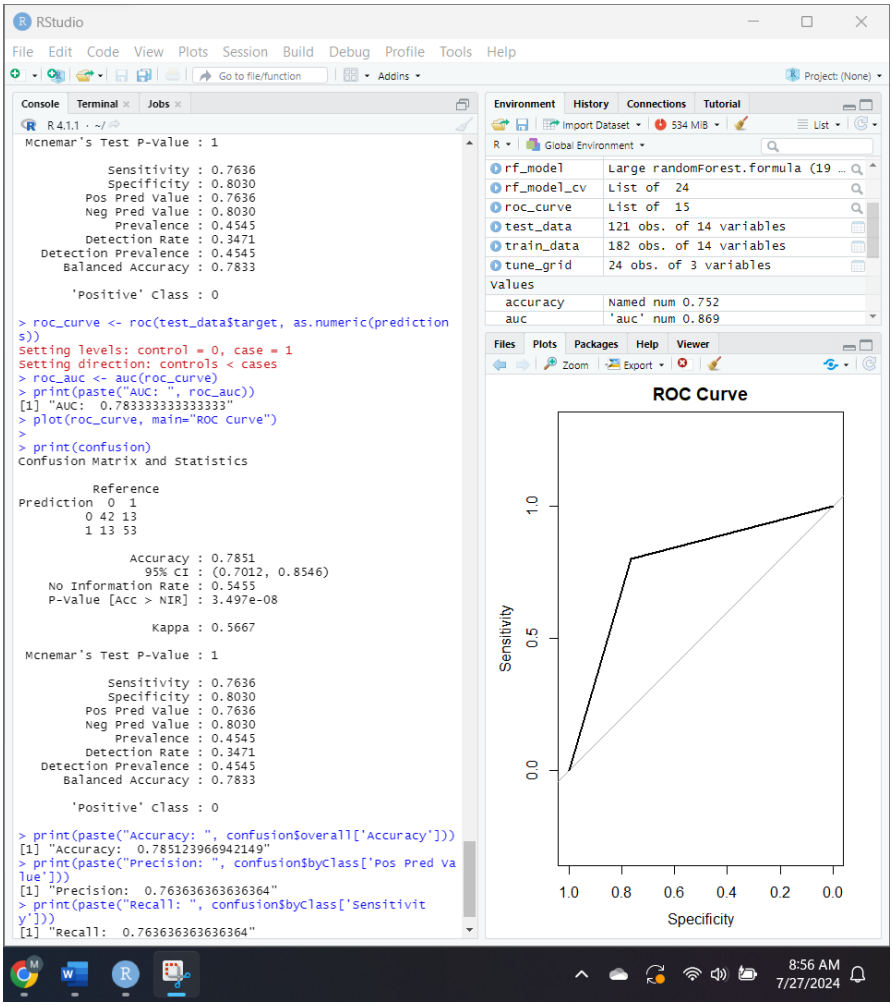
Figure 3

Random Forest Decision Tree Model

```
> rf_model <- randomForest(target ~ ., data=train_data, ntr
ee=500, mtry=3, nodesize=5)
> predictions <- predict(rf_model, test_data)
> predictions <- factor(predictions, levels = levels(test_d
ata$target))
> confusion <- confusionMatrix(predictions, test_data$targe
t)
> print(confusion)
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 42 13
         1 13 53
```

Figure 4

Random Forest Decision Tree Predictive Model Performance Metrics



The logistic regression model, as evaluated through cross-validation in figure 5, achieved an RMSE of 0.3463, an R-squared of 0.5182, and an MAE of 0.2357. These metrics indicate moderate predictive performance with a good balance between accuracy and complexity. The logistic regression model's accuracy on the test data was 75.21%, with a sensitivity of 65.45% and specificity of 83.33%. The model's AUC of 0.869 suggests strong discriminatory power, with a precision of 76.60% and an F1 score of 70.59%. This performance underscores the model's effectiveness in identifying heart disease risk, confirming that the model provides a significant relationship with identifying patients at risk.

The random forest model, cross-validated with 10 folds, demonstrated an RMSE of 0.3539, R-squared of 0.5175, and MAE of 0.2863 with an optimal mtry of 2. The model achieved an accuracy of 78.51% on the test data, with a sensitivity of 76.36% and specificity of 80.30%. The AUC of 0.783 reflects solid predictive capability, and the precision and F1 score are both 76.36%, indicating balanced performance. Compared to logistic regression, the random forest model has a slightly higher accuracy and F1 score but a marginally lower AUC, suggesting it also performs effectively in identifying patients at risk for heart disease.

Both models exhibit robust performance in predicting heart disease risk. The logistic regression model shows slightly higher AUC and a balanced performance between precision and recall, while the random forest model has higher accuracy and maintains strong performance metrics. The cross-validation results suggest that both models are competitive, though the logistic regression model has a slight edge in terms of AUC, indicating potentially better discriminatory ability. Overall, both models provide significant insights into heart disease risk and are effective tools for identifying patients at risk.

Figure 5

Cross-validation of the Logistic Regression Model and Random Forest Model

```
> print("Logistic Regression Model Results:")
[1] "Logistic Regression Model Results:"
> print(logistic_model)
Generalized Linear Model

303 samples
 13 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 273, 273, 273, 272, 273, 272, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.3463067  0.5181505  0.2356525

>
> print("Random Forest Model Results:")
[1] "Random Forest Model Results:"
> print(random_forest_model)
Random Forest

303 samples
 13 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 272, 273, 272, 273, 273, 273, ...
Resampling results across tuning parameters:

  mtry  RMSE       Rsquared   MAE
   2    0.3538667  0.5175181  0.2862720
   7    0.3623235  0.4846184  0.2713840
  13    0.3658807  0.4748271  0.2677444

RMSE was used to select the optimal model using
 the smallest value.
The final value used for the model was mtry = 2.
>
> results <- resamples(list(Logistic_Regression = logistic_
model,
+                          Random_Forest = random_forest_m
odel))
> print("Model Comparison:")
[1] "Model Comparison:"
> print(results)

Call:
resamples.default(x = list(Logistic_Regression
 = logistic_model, Random_Forest = random_forest_model))

Models: Logistic_Regression, Random_Forest
Number of resamples: 10
Performance metrics: MAE, RMSE, Rsquared
Time estimates for: everything, final model fit
```

Gender Disparities

The Chi-Square test in figure 6 and the Two Sample t-test in figure 7 indicate a significant relationship between gender and risk of heart disease, allowing the null hypothesis to be rejected that there is no significant relationship between gender and heart disease risk. The t-test reveals a statistically significant difference in mean values of heart disease risk between males and females, with a p-value of $2.44 * 10^{-7}$, indicating a highly significant result. The mean heart disease risk for females in higher (1.75) compared to males (1.449), with a 95% confidence interval suggesting that the rue difference in means lies between 0.19 and 0.41. Additionally, the

Chi-square statistic of 22.717 and p-value of $1.877 * 10^{-6,}$ further confirms the presence of a disparity. These findings suggest that, on average, females in this sample have a higher risk of heart disease compared to males. Therefore, the evidence strongly supports a relationship between gender and heart disease risk, with females being more at risk. These findings align with the existing literature on gender disparities. However, other literature suggests that men are more at risk, it could be that females with heart disease are underreported or that the specific characteristics of this dataset include more post-menopausal women. This requires further evaluation for more accurate and comprehensive understanding of gender disparities.

Figure 6

Chi-Square Test

```
> library(dplyr)

Attaching package: ‘dplyr’

The following object is masked from ‘package:randomForest’:

    combine

The following objects are masked from ‘package:stats’:

    filter, lag

The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union

>
> heart_data$sex <- factor(heart_data$sex, levels = c(0,
 1), labels = c("Female", "Male"))
> heart_data$target <- factor(heart_data$target)
>
> contingency_table <- table(heart_data$sex, heart_data$tar
get)
>
> chi_sq_test <- chisq.test(contingency_table)
>
> print(chi_sq_test)

        Pearson's Chi-squared test with Yates' continuity
        correction

data:  contingency_table
X-squared = 22.717, df = 1, p-value = 1.877e-06
```

Figure 7

T-test

```
> heart_data$target <- as.numeric(heart_data$target)
> heart_data$sex <- as.factor(heart_data$sex)
>
> t_test <- t.test(target ~ sex, data = heart_data)
>
> print(t_test)

        Welch Two Sample t-test

data:  target by sex
t = 5.3372, df = 209.95, p-value = 2.44e-07
alternative hypothesis: true difference in means between gr
oup Female and group Male is not equal to 0
95 percent confidence interval:
 0.1896497 0.4117996
sample estimates:
mean in group Female   mean in group Male
            1.750000             1.449275
```

Age and Heart Disease Susceptibility

The logistic regression analysis in Figure 8 reveals that age is a statistically significant factor influencing the risk of heart disease. The model, which predicts heart disease probability based on age, shows a negative coefficient for age (-0.05235), indicating that with each additional year of age, the log odds of developing heart disease decrease. This finding is supported by a highly significant p-value (0.000122), which is much lower than the threshold of 0.05. The model's fit is robust, as shown by the reduction in deviance from the null model and a low Akaike Information Criterion (AIC)value. The null deviance of 417.64 decreases to a residual deviance of 401.86 when age is included, signifying an improved fit. These results indicate that the null hypothesis (H_0) can be rejected, confirming there is a significant relationship between age and heart disease susceptibility. However, these results are the opposite of what precious research may suggest, with heart disease risk increasing age. Therefore, this indicates that there may be more variables that need to be included to capture the heart disease risk more accurately.

Figure 8

Logistic Regression with Age as a Predictor for Heart Disease

```
> logistic_model <- glm(target ~ age, data = heart_data, fa
mily = binomial)
> summary(logistic_model)

Call:
glm(formula = target ~ age, family = binomial, data = heart
_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7125  -1.1773   0.8296   1.0685   1.5947

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.03623    0.75639   4.014 5.97e-05 ***
age         -0.05235    0.01363  -3.841 0.000122 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 417.64  on 302  degrees of freedom
Residual deviance: 401.86  on 301  degrees of freedom
AIC: 405.86

Number of Fisher Scoring iterations: 4
```

Conclusion

The analysis of logistic regression and random forest models in predicting heart disease risk has yielded insightful findings. Both models demonstrated significant capabilities in identifying patients at risk for heart disease, with logistic regression achieving an accuracy of 75.21% and an AUC of 0.869, while the random forest model reached an accuracy of 78.51% and an AUC of 0.783. These metrics indicate that both models are effective in distinguishing between patients with and without heart disease, though logistic regression slightly outperforms random forest in terms of discriminatory ability. The statistical tests confirm a significant relationship between gender and heart disease risk, with females showing higher average risk levels. This aligns with existing literature but also suggests potential underreporting or dataset-specific characteristics influencing these results. Furthermore, age was identified as a significant

predictor in the logistic regression model, contrary to previous findings, highlighting the complexity of heart disease risk factors.

Recommendations

Based on the results, it is recommended that both the logistic regression and random forest models be utilized in conjunction to leverage their respective strengths. While logistic regression provides strong discriminatory power, the random forest model's higher accuracy makes it valuable for practical applications. Further research should explore the potential for including additional variables or refining current predictors to better capture heart disease risk, particularly regarding age and gender disparities. Additionally, a more comprehensive dataset, possibly including diverse populations and more extensive demographic information, is recommended to validate the findings and ensure robust, generalizable results. Future studies should also investigate the potential underreporting of heart disease risk in females and examine post-menopausal status more closely to understand its impact on risk levels.

References

Al Reshan, M. S., Amin, S., Zeb, M. A., Sulaiman, A., Alshahrani, H., & Shaikh, A. (2023). A robust heart disease prediction system using hybrid deep neural networks. IEEE Access.

Desai, S. D., Giraddi, S., Narayankar, P., Pudakalakatti, N. R., & Sulegaon, S. (2019). Back-propagation neural network versus logistic regression in heart disease classification. In Advanced Computing and Communication Technologies: Proceedings of the 11th ICACCT 2018 (pp. 133-144). Springer Singapore.

Dhingra, R., & Vasan, R. S. (2012). Age as a risk factor. Medical Clinics, 96(1), 87-91.

Gudadhe, M., Wankhade, K., & Dongre, S. (2010, September). Decision support system for heart disease based on support vector machine and artificial neural network. In 2010 International Conference on Computer and Communication Technology (ICCCT) (pp. 741-745). IEEE.

Lapp, D. (2019, June 6). *Heart disease dataset*. Kaggle. https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset?resource=download

Li, F., Wu, P., Ong, H. H., Peterson, J. F., Wei, W. Q., & Zhao, J. (2023). Evaluating and mitigating bias in machine learning models for cardiovascular disease prediction. Journal of biomedical informatics, 138, 104294.

Millett, E. R., Peters, S. A., & Woodward, M. (2018). Sex differences in risk factors for myocardial infarction: cohort study of UK Biobank participants. bmj, 363.

Paulus, J. K., Shah, N. D., & Kent, D. M. (2015). All else being equal, men and women are still not the same: using risk models to understand gender disparities in care. Circulation: Cardiovascular Quality and Outcomes, 8(3), 317-320.

Peeters, A., Mamun, A. A., Willekens, F., & Bonneux, L. (2002). A cardiovascular life history. European heart journal, 23(6), 458-466.

Regitz-Zagrosek, V. (2003). Cardiovascular disease in postmenopausal women. *Climacteric*, *6*, 13.

Sun, Z. (2015). Aging, arterial stiffness, and hypertension. Hypertension, 65(2), 252-256.

UCI Heart disease data set. (2021, January 1). Kaggle. https://www.kaggle.com/datasets/lourenswalters/uci-heart-disease-data-set

UCI Machine Learning Repository. (n.d.). https://archive.ics.uci.edu/dataset/45/heart+disease

Weidner, G. (2000). Why do men get more heart disease than women? An international perspective. Journal of American College Health, 48(6), 291-294.

Wilhelmsen, L., Wedel, H., & Tibblin, G. (1973). Multivariate analysis of risk factors for coronary heart disease. Circulation, 48(5), 950-958.

Woodward, M. (2019). Cardiovascular disease and the female disadvantage. International journal of environmental research and public health, 16(7), 1165.