Module 8: Portfolio Project
Oral and PowerPoint Presentation

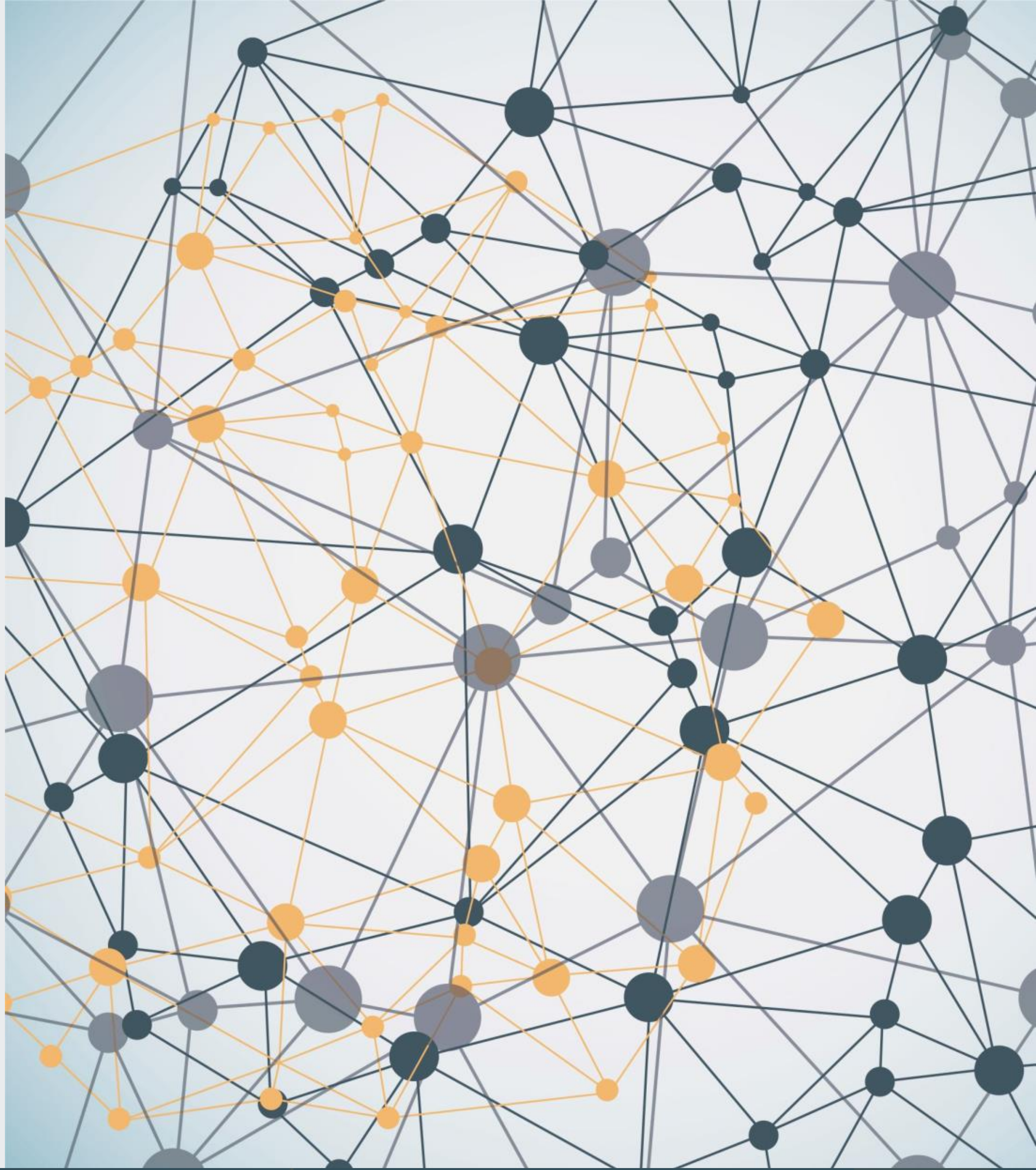
Michelle Goodwin

Colorado State University-Global Campus

MIS 581: Capstone- Business Intelligence and Data Analytics

Dr. Osama Morad

August 4, 2024



Abstract

- Heart disease is a disease that impacts many globally.
- Objectives:
  - Develop predictive models to be used to identify heart disease in patients.
  - Analyze gender as an influencing factor in heart disease.
  - Analyze age impact on heart disease susceptibility.
- Findings:
  - Effective model performance
  - Gender significance in heart disease
  - Age impacts

Introduction

- Heart disease is the leading cause of mortality globally (Di Cesare et al., 2024).
- Role of data analytics in reducing heart disease mortality:
  - Tools for developing predictive models.
  - Identifying individuals at risk more effectively.
  - Discovering contributing factors in heart disease susceptibility.
- Study focus:
  - Develop and evaluate predictive models
  - Analyze gender and age impact
  - Utilize UCI Heart Disease Dataset from Kaggle (Lapp, 2019).

Research Questions and Hypotheses

1. Can a predictive model effectively 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 developing 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

- Predictive models:
  - Desai et al. (2019) created a logistic regression model with 92.58% accuracy and a BPNN model with 85.07% accuracy.
  - Al Reshan et al. (2023) developed a hybrid deep neural network model with 98.56% accuracy.
- Gender-based differences
  - Some studies indicate men are at higher risk (Weidner, 2000); some indicate post-menopausal women are at significant risk (Regtiz-Zagrosek, 2003).
- Age as a significant factor.
  - Farmington Heart Study shows exponential increase after age 50 (Peeters et al., 2002).
  - Dhingra & Vashan (2012) show age as a significant predictor, but interaction with other factors needs exploration.

Research Design and Methodology

Quantitative approach using UCI Heart Disease dataset.

Model development:

- Logistic regression and Random Forest models.
- Training/testing split : 60/40
- Evaluation metrics: Accuracy, Precision Recall, F1 score.
- Cross-Validation for robustness

Statistical tests:

- Chi-square test for gender data.
- T-test for mean comparison of male/female data.
- Regression analysis for age impact.

Model Evaluation and Performance

Logistic Regression:

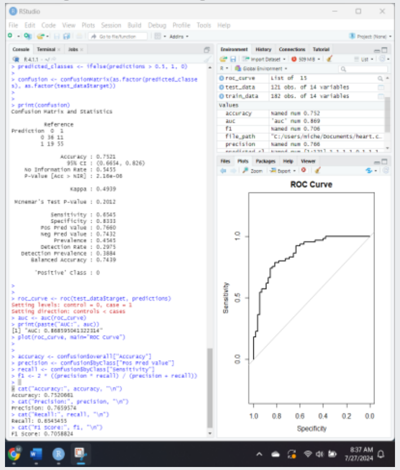
Key influential variables: Sex, chest pain type (cp), maximum heart rate (thalach), ST depression (oldpeak), and major vessels (ca).

Metrics:

Accuracy (75.21%), sensitivity/recall (65.45%), specificity (83.33%), precision (76.60%), F1 score (70.59%), AUC (0.869).

Figure 1

Logistic regression



Random :

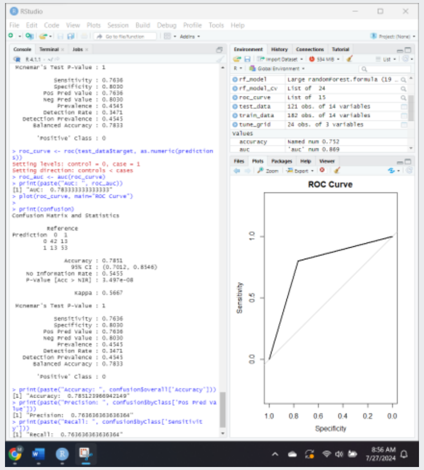
Trained with 500 trees

Metrics:

Accuracy (78.51%), sensitivity/recall (76.36%), specificity (80.30%), precision (76.36%), F1 score (76.36%), AUC (0.783).

Figure 2

Random Forest



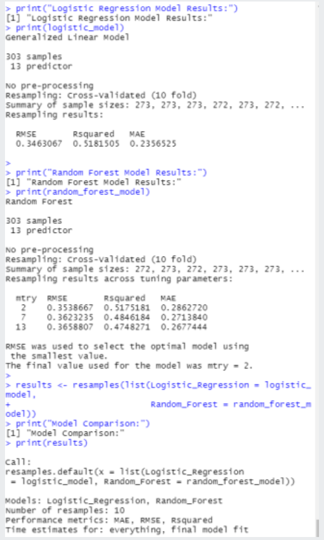
Cross-Validation:

Logistic regression: RMSE (0.3463), R-squared (0.5182), MAE (0.2357).

Random Forest: RMSE (0.3539), R-squared (0.5175), MAE (0.2863).

Figure 3

Cross-Validation



```
> 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
```

Findings on Gender

Chi-square Test

Chi-square statistic: 22.717

P-value: 1.877* 10^{-6}

Significant relationship between gender and heart disease.

Figure 4

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
```

- Two Sample T-test
  - Mean Heart risk: females (1.75), Males (1.449)
  - P-Value: 2.44*10^{-7}
  - 95% confidence interval: 0.19 to 0.41
  - Statistically significant difference in means

Figure 5

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
```

Findings on Age

Logistic regression of age and heart disease:

Age coefficient (-0.05235), p-value (0.000122), Reduction in deviance (417.64 to 401.86)

Figure 6

Logistic Regression with Age

```
> 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

Predictive Models

- Both models are effective in identifying heart disease risk.
- Logistic regression outperforms random forest in discriminatory ability (AUC).

Gender and Heart Disease

- Significant relationship
- Females show higher average risk level.

Age as a Predictor

- Significant logistic regression model
- Contradictory to previous findings, suggesting a complexity of risk factors.

Recommendations

1

Utilize multiple models:

- Logistic regression model shows strong discriminatory power, while random forest shows higher accuracy.

2

Further research:

- Include additional variables and perform further research on age and gender disparities.

3

Utilize other datasets:

- Include data on more diverse populations with extensive demographic information

4

Investigate Underreporting

- Post-menopausal impact on female heart disease risk.

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.
- Di Cesare, M., Perel, P., Taylor, S., Kabudula, C., Bixby, H., Gaziano, T. A., ... & Pinto, F. J. (2024). The heart of the world. Global heart, 19(1).
- Lapp, D. (2019, June 6). Heart disease dataset. Kaggle.
- 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.
- Weidner, G. (2000). Why do men get more heart disease than women? An international perspective. Journal of American College Health, 48(6), 291-294.