Breast Cancer Diagnosis and Prognosis

Final Project Paper

Kendra Castellanos

Kwasi Mensah

Alyssa Moylan

Matt Smith

September 25, 2022

IST 687 – Introduction to Data Science

Syracuse University

**INTRODUCTION**

**Scope/Context.**

For this project we explored the Wisconsin Diagnostic and Prognostic Breast Cancer data sets (UCI, 2007). The diagnostic data set contains 569 records of 32 attributes (18,208 data points) and the prognostic data set contains 198 records of 34 attributes (6,732 data points) (UCI, 2007). All analyses in this project were conducted with the R language in RStudio.

**High-Level Background.**

These data sets were first used by faculty from the University of Wisconsin (General Surgery and Computer Science departments) (UCI, 2007). The diagnostic data set was first used in a study to analyze images of cancer cells in both benign and malignant cases (Street, Wolberg, & Mangasarian, 1993). For the prognostic data set, the same researchers took the features from the diagnostic set, along with two traditional variables used to predict recurrence/nonrecurrence, and analyzed the relationships (Street, Mangasarian, & Wolberg, 1995).

**Business Questions.**

Using the data sets described above, our team set out to answer the following questions:

1. *Which variables are the most determinant of a benign or malignant breast cancer case (in Wisconsin).*
2. *What factors most/least affect a Recur/Non-Recur prognosis?*
3. *Who is at the most risk for occurrence?*
4. *Can we develop a model that accurately predicts a benign/malignant diagnosis or a recur/non-recur prognosis?*

**DATA ACQUISITION, CLEANING AND TRANSFORMATION**

**Data Acquisition Process.**

The data sets for this project were downloaded as a .txt file (written as a comma-delimitated file), and we were able to use the `tidyverse` package (v1.3.2; Wickham et al., 2019) in RStudio to create a data frame from that raw data. The data gathered was not labeled according to the data description files provided from the site, so it needed to be processed to make it easier to deal with.

The only transformation needed for this data was the addition of column names to the data frame (which we named `df\_wdbc` and `df\_wpbc` for the diagnostic and prognostic datasets, respectively. See Appendices A and B, respectively for R code for transformations.

**Data Dictionary.**

The full list of the variables in the diagnosis and prognosis datasets (after cleaning/munging) can be found in Appendix C and D, respectively.

**Demographic Statistics.**

The data for this project did not have any traditional demographic data, such as gender, race, etc.). However, the data can be broken down into summaries of its key categories. For the diagnostic data set, there were 357 benign cases and 212 malignant cases. See Figure 1. R Code for these visualizations can be found in Appendix E.

*Figure 1.*

Chart, treemap chart

Description automatically generated

For the prognostic data set, there were 148 cases where breast cancer did not recur and 46 cases where it did recur. The mean (average) time of recurrence across all subjects in the prognostic data was 46.94 days. See Figure 2.

*Figure -2.*

Chart

Description automatically generated

**Early Observations and Interesting Findings.**

During the team’s initial look into the data sets, it was found that approximately 37.3% of patient’s cancer was identified as malignant and 62%.7 % of the cells were benign. This is interesting because the data sets is heavily skewed towards patients with a benign diagnosis. See Figure 3. R Code for Figure 3 can be found in Appendix N.

*Figure 3.*

Chart, pie chart

Description automatically generated

The variables in the datasets were also analyzed via correlation matrices in order to get an initial feel for the relationships between them as it pertained to the diagnosis and prognoses of the patients involved. See Figures 4 through 7 for diagnosis correlation tables; see Figures 8 through 11 for prognosis correlation tables. For ease of analysis of the team and of reading in this report, the variables for each data set were split into four separate correlation tables. The R Code for these correlation tables can be found in Appendix F (diagnostic) and G (prognostic).

*Table

Description automatically generated with medium confidenceFigure 4. Diagnostic data set correlation table (1 of 4)*

*Figure 5. Diagnostic data set correlation table (2 of 4)*

Calendar

Description automatically generated with low confidence

*Figure 6. Diagnostic data set correlation table (3 of 4)*

Table

Description automatically generated with low confidence

*Figure 7. Diagnostic data set correlation table (4 of 4)*

A picture containing chart

Description automatically generated

*Figure 8. Prognostic data set correlation table (1 of 4)*

Chart

Description automatically generated with medium confidence

*Figure 9. Prognostic data set correlation table (2 of 4)*

Chart

Description automatically generated with medium confidence

*Figure 10. Prognostic data set correlation table (3 of 4)*

Chart

Description automatically generated

*Figure 11. Prognostic data set correlation table (4 of 4)*

Chart

Description automatically generated

Using these tables as a reference to check our work, we then proceeded to single out the factors that had the highest positive correlations (above 0.6) to a negative outcome in the diagnostic (benign versus malignant) and prognostic (recur versus non-recur) data sets. For the diagnostic dataset, those factors are shown in Figure 12. The R Code for the table presented in Figure 12 can be found in Appendix H.

*Figure 12 Highest correlation values - Diagnostic*

*Graphical user interface, application

Description automatically generated*

*All correlation values were significant (p < 0.05).*

The values shown in the “malignant coded” column above are the correlation values. The team decided on using only the top ten highest correlation values to focus on for the rest of this project. Below is the table (Figure 13) with the top ten highest correlation values for a “recur/non-recur” outcome in the prognosis data set. The R Code for this table is shown in Appendix I.

*Figure 13. Highest correlation values - Prognostic*

Graphical user interface, text, application

Description automatically generated

*All correlation values were significant (p < 0.05).*

From here, we could now move to developing the models for these datasets.

**MODELING TECHNIQUES**

**Support Vector Machine.**

We chose a Support Vector Machine (SVM) to start as our first attempt at modeling these two datasets. For the diagnosis data set, the SVM model created had a balanced accuracy of approximately 97%, with two false negatives (a benign prediction for malignant diagnosis), and a model significance of p < .001. The prognosis data set was not as accurate with a balanced accuracy of approximately 72%, and a model significance of p > 0.05 (~0.08). See Figures 14 and 15 for the diagnosis and prognosis data set model summaries, respectively. R Code for these summaries can be found in Appendices J and K, respectively. Scaling the independent variables in both datasets produced the same exact results as shown below, not providing any additional accuracy.

*Figure 14. Confusion matrix for SVM - Diagnostic*

Text

Description automatically generated

*Figure 15. Confusion matrix for SVM - Prognostic*

Text

Description automatically generated

**Decision Tree.**

In order to see if we could improve our accuracy, we looked at creating a decision tree model for both datasets. Again, for the diagnosis data set, the Decision Tree Classifier model created a relatively accurate model with a balanced accuracy of approximately 90%, with seven false negatives (a benign prediction for malignant diagnosis) and three false positives (a malignant prediction for a benign diagnosis), and a model significance of *p* < .001. The prognosis data set, again, was not as accurate with a balanced accuracy of approximately 65%, and a model significance of *p* > 0.1 (~0.6). See Figures 16 and 17 for the diagnosis and Figures 18 and 19 for the prognosis data set model summaries, respectively. R Code for these summaries can be found in Appendices L and M, respectively. Scaling the independent variables in both datasets produced the same exact results as shown below, not providing any additional accuracy.

*Figure 16. Confusion matrix for Decision Tree Classifier - Diagnostic*

*Text

Description automatically generated*

*Figure 17. Decision Tree Classifier plot - Diagnostic*

Timeline

Description automatically generated

*Figure 18. Confusion matrix for Decision Tree Classifier - Prognostic*

*Text

Description automatically generated*

*Figure 19. Decision Tree Classifier plot - Prognostic*

Timeline

Description automatically generated

**INTERPRETATION OF RESULTS**

**SVM Results.**

From the SVM model on the diagnostic data set, we can say for certain that an accurate model (98.23%) can be produced in order to predict breast cancer in patients (at least with the data provided); answering business question #4. However, the same cannot be said for the prognostic data provided, which also did not produce a statistically significant model (*p > 0.*05). We believe that this is possibly due to an initial cancer diagnosis being easier to recognize, but not easily predicted to return in the patient or not. It is also possible that the independent variables chosen for the prognosis dataset were not statistically significant or relevant to the prediction of a cancer recurrence. Further investigation and data gathering may be required (related to business question #2). Based on this, we cannot determine the answer for our business question #3 at this time.

**Decision Tree Results.**

The Decision Tree Classifier model created for both datasets yielded the same results as the SVM; higher accuracies in the diagnosis data model versus the prognosis data model. Exploring this method of modeling also provided insight into the most important factors influencing a malignant/benign cancer diagnosis (see Figure 20 and 21). This insight can help us answer business question #3 and shows us the factors that must be watched closely when monitoring the status of a cancer patient.

*Figure 20. Most important diagnosis factors (Decision Tree Classifier)*

*Text

Description automatically generated*

*Figure 21. Most important prognosis factors (Decision Tree Classifier)*

**Table

Description automatically generated**

**APPENDICES**

**APPENDIX A.**

# R code for cleaned diagnostic data set:

#install.packages("tidyverse", repos = "http://cran.us.r-project.org")

library(tidyverse)

fixed\_col\_names\_wdbc = c("id", "diagnosis", "radius\_mean", "texture\_mean", "perimeter\_mean", "area\_mean", "smoothness\_mean", "compactness\_mean", "concavity\_mean", "concave\_points\_mean", "symmetry\_mean", "fractal\_dimension\_mean", "radius\_SE", "texture\_SE", "perimeter\_SE", "area\_SE", "smoothness\_SE", "compactness\_SE", "concavity\_SE", "concave\_points\_SE", "symmetry\_SE", "fractal\_dimension\_SE", "radius\_worst", "texture\_worst", "perimeter\_worst", "area\_worst", "smoothness\_worst", "compactness\_worst", "concavity\_worst", "concave\_points\_worst", "symmetry\_worst", "fractal\_dimension\_worst”)

df\_wdbc = read\_csv("https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.data", col\_names = fixed\_col\_names\_wdbc)

**APPENDIX B.**

# R code for cleaned prognostic data set:

library(tidyverse)

fixed\_col\_names\_wpbc = c("id", "outcome", "time", "radius\_mean", "texture\_mean", "perimeter\_mean", "area\_mean", "smoothness\_mean", "compactness\_mean", "concavity\_mean", "concave\_points\_mean", "symmetry\_mean", "fractal\_dimension\_mean", "radius\_SE", "texture\_SE", "perimeter\_SE", "area\_SE", "smoothness\_SE", "compactness\_SE", "concavity\_SE", "concave\_points\_SE", "symmetry\_SE", "fractal\_dimension\_SE", "radius\_worst", "texture\_worst", "perimeter\_worst", "area\_worst", "smoothness\_worst", "compactness\_worst", "concavity\_worst", "concave\_points\_worst", "symmetry\_worst", "fractal\_dimension\_worst", "tumor\_size", "lymph\_node\_status")

df\_wpbc = read\_csv("https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wpbc.data", col\_names = fixed\_col\_names\_wpbc)

df\_wpbc <- df\_wpbc %>% filter(lymph\_node\_status != "?")

**APPENDIX C:**

**Data Dictionary**

Features (Diagnostic Data Set):

1) id (unique identifier for individuals)

2) diagnosis (M = Malignant; B = Benign)

3) radius\_mean (mean of distances from center to points on the perimeter

4) texture\_mean (standard deviation of gray-scale values)

5) perimeter\_mean

6) area\_mean

7) smoothness\_mean (local variation in radius lengths)

8) compactness\_mean (perimeter^2 / area - 1.0)

9) concavity\_mean (severity of concave portions of the contour)

10) concave\_points\_mean (number of concave portions of the contour)

11) symmetry\_mean

12) fractal\_dimension\_mean ("coastline approximation" - 1)

13) radius\_SE (mean of distances from center to points on the perimeter

14) texture\_SE (standard deviation of gray-scale values)

15) perimeter\_SE

16) area\_SE

17) smoothness\_SE (local variation in radius lengths)

18) compactness\_SE (perimeter^2 / area - 1.0)

19) concavity\_SE (severity of concave portions of the contour)

20) concave points\_SE (number of concave portions of the contour)

21) symmetry\_SE

22) fractal\_dimension\_SE ("coastline approximation" - 1)

23) radius\_worst (mean of distances from center to points on the perimeter); (mean of the three largest values)

24) texture\_worst (standard deviation of gray-scale values)

25) perimeter\_worst; (mean of the three largest values)

26) area\_worst; (mean of the three largest values)

27) smoothness\_worst (local variation in radius lengths); (mean of the three largest values)

28) compactness\_worst (perimeter^2 / area - 1.0); (mean of the three largest values)

29) concavity\_worst (severity of concave portions of the contour); (mean of the three largest values)

30) concave\_points\_worst (number of concave portions of the contour); (mean of the three largest values)

31) symmetry\_worst; (mean of the three largest values)

32) fractal\_dimension\_worst ("coastline approximation" - 1); (mean of the three largest values)

**APPENDIX D:**

Features (Prognostic Data Set):

1) id (unique identifier for individuals)

2) outcome (R = recur, N = nonrecur)

3) time (recurrence time if field 2 = R, disease-free time if field 2 = N)

4) radius\_mean (mean of distances from center to points on the perimeter

5) texture\_mean (standard deviation of gray-scale values)

6) perimeter\_mean

7) area\_mean

8) smoothness\_mean (local variation in radius lengths)

9) compactness\_mean (perimeter^2 / area - 1.0)

10) concavity\_mean (severity of concave portions of the contour)

11) concave\_points\_mean (number of concave portions of the contour)

12) symmetry\_mean

13) fractal\_dimension\_mean ("coastline approximation" - 1)

14) radius\_SE (mean of distances from center to points on the perimeter

15) texture\_SE (standard deviation of gray-scale values)

16) perimeter\_SE

17) area\_SE

18) smoothness\_SE (local variation in radius lengths)

19) compactness\_SE (perimeter^2 / area - 1.0)

20) concavity\_SE (severity of concave portions of the contour)

21) concave points\_SE (number of concave portions of the contour)

22) symmetry\_SE

23) fractal\_dimension\_SE ("coastline approximation" - 1)

24) radius\_worst (mean of distances from center to points on the perimeter); (mean of the three largest values)

25) texture\_worst (standard deviation of gray-scale values)

26) perimeter\_worst; (mean of the three largest values)

27) area\_worst; (mean of the three largest values)

28) smoothness\_worst (local variation in radius lengths); (mean of the three largest values)

29) compactness\_worst (perimeter^2 / area - 1.0); (mean of the three largest values)

30) concavity\_worst (severity of concave portions of the contour); (mean of the three largest values)

31) concave\_points\_worst (number of concave portions of the contour); (mean of the three largest values)

32) symmetry\_worst; (mean of the three largest values)

33) fractal\_dimension\_worst ("coastline approximation" - 1); (mean of the three largest values)

34) tumor\_size - diameter of the excised tumor in centimeters

35) lymph\_node\_status - number of positive axillary lymph nodes observed at time of surgery

**APPENDIX E**:

library(ggplot2)

# R Code for diagnoses descriptive stats bar chart

wdbc\_diag\_freqs\_plot <-

ggplot(df\_wdbc, aes(diagnosis, fill=diagnosis)) +

geom\_bar(stat = 'count') +

stat\_count(geom = 'text',

colour = 'black',

size = 5,

aes(label = ..count..),

position = position\_stack(vjust = 0.5)) +

scale\_fill\_manual(values = c('lightgreen',

'pink')) +

ggtitle("Distribution of Benign and Malignant Diagnosis",

"Wisconsin Diagnosis Breast Cancer dataset")\

wdbc\_diag\_freqs\_plot

# R Code for prognosis descriptive stats bar chart

wpbc\_diag\_freqs\_plot <-

ggplot(df\_wpbc, aes(outcome, fill=outcome)) +

geom\_bar(stat = 'count') +

stat\_count(geom = 'text',

colour = 'black',

size = 5,

aes(label = ..count..),

position = position\_stack(vjust = 0.5)) +

scale\_fill\_manual(values = c('lightgreen',

'pink')) +

ggtitle("Distribution of Recur and Non-recur Prognosis",

"Wisconsin Prognosis Breast Cancer dataset")

wpbc\_diag\_freqs\_plot

**APPENDIX F:**

# Changing .$diagnosis variable values from "M/B" to "1/0", respectively

df\_wdbc <-

df\_wdbc %>%

mutate(diagnosis = replace(diagnosis, diagnosis == "M", 1)) %>%

mutate(diagnosis = replace(diagnosis, diagnosis == "B", 0))

# Change .$diagnosis to numeric values

df\_wdbc$diagnosis <- as.numeric(df\_wdbc$diagnosis)

# Need to create seperate dataframes for each corrplot(?):

df\_wdbc\_corrplotdata\_1 <- data.frame(c(df\_wdbc[2], df\_wdbc[3:11]))

corrplot(cor(df\_wdbc\_corrplotdata\_1), method = 'number', number.cex=0.65)

# Now for the rest of the factors:

df\_wdbc\_corrplotdata\_2 <- data.frame(c(df\_wdbc[2], df\_wdbc[12:20]))

df\_wdbc\_corrplotdata\_3 <- data.frame(c(df\_wdbc[2], df\_wdbc[21:29]))

df\_wdbc\_corrplotdata\_4 <- data.frame(c(df\_wdbc[2], df\_wdbc[29:32]))

corrplot(cor(df\_wdbc\_corrplotdata\_2), method = 'number', number.cex=0.65)

corrplot(cor(df\_wdbc\_corrplotdata\_3), method = 'number', number.cex=0.65)

corrplot(cor(df\_wdbc\_corrplotdata\_4), method = 'number', number.cex=0.65)

**APPENDIX G:**

df\_wpbc <-

df\_wpbc %>%

filter(lymph\_node\_status != "?")

# Changing .$outcome variable values from "R/N" to "1/0", respectively

df\_wpbc <-

df\_wpbc %>%

mutate(outcome = replace(outcome, outcome == "R", 1)) %>%

mutate(outcome = replace(outcome, outcome == "N", 0))

# Change .$outcome and .$lymph\_node\_status to numeric values

df\_wpbc$outcome <- as.numeric(df\_wpbc$outcome)

df\_wpbc$lymph\_node\_status <- as.numeric(df\_wpbc$lymph\_node\_status)

# Need to create seperate dataframes for each corrplot(?):

df\_wpbc\_corrplotdata\_1 <- data.frame(c(df\_wpbc[2], df\_wpbc[3:11]))

corrplot(cor(df\_wpbc\_corrplotdata\_1), method = 'number', number.cex=0.65)

# Now for the rest of the factors:

df\_wpbc\_corrplotdata\_2 <- data.frame(c(df\_wpbc[2], df\_wpbc[12:20]))

df\_wpbc\_corrplotdata\_3 <- data.frame(c(df\_wpbc[2], df\_wpbc[21:29]))

df\_wpbc\_corrplotdata\_4 <- data.frame(c(df\_wpbc[2], df\_wpbc[29:35]))

corrplot(cor(df\_wpbc\_corrplotdata\_2), method = 'number', number.cex=0.65)

corrplot(cor(df\_wpbc\_corrplotdata\_3), method = 'number', number.cex=0.65)

corrplot(cor(df\_wpbc\_corrplotdata\_4), method = 'number', number.cex=0.65)

**APPENDIX G:**

df\_wdbc$malignant\_coded <- ifelse(df\_wdbc$diagnosis == "M", 1, 0)

df\_wdbc\_diag\_coded <-

df\_wdbc %>%

select(-2) %>%

select(id, malignant\_coded, everything())

df\_wdbc\_diag\_coded <-

df\_wdbc\_diag\_coded %>%

select(-1)

df\_corr\_table <- data.frame(cor(df\_wdbc\_diag\_coded))

df\_corr\_table <-

df\_corr\_table %>%

slice(-1)

df\_corr\_table <- df\_corr\_table %>% filter(malignant\_coded > .6) %>% arrange(desc(malignant\_coded))

df\_corr\_table

**APPENDIX H:**

df\_wdbc$malignant\_coded <- ifelse(df\_wdbc$diagnosis == "M", 1, 0)

df\_wdbc\_diag\_coded <-

df\_wdbc %>%

select(-2) %>%

select(id, malignant\_coded, everything())

df\_wdbc\_diag\_coded <-

df\_wdbc\_diag\_coded %>%

select(-1)

df\_wdbc\_corr\_table <- data.frame(cor(df\_wdbc\_diag\_coded))

df\_wdbc\_corr\_table <-

df\_wdbc\_corr\_table %>%

slice(-1)

df\_wdbc\_corr\_table <- df\_wdbc\_corr\_table %>% filter(malignant\_coded > .6) %>% arrange(desc(malignant\_coded))

df\_wdbc\_corr\_table

**APPENDIX I:**

df\_wpbc$outcome\_coded <- ifelse(df\_wpbc$outcome == "R", 1, 0)

df\_wpbc\_outc\_coded <-

df\_wpbc %>%

select(-2) %>%

select(id, outcome\_coded, everything())

df\_wpbc\_outc\_coded

df\_wpbc\_outc\_coded <-

df\_wpbc\_outc\_coded %>%

select(-1)

df\_wpbc\_outc\_coded$lymph\_node\_status <- as.numeric(df\_wpbc\_outc\_coded$lymph\_node\_status)

df\_wpbc\_corr\_table <- data.frame(cor(df\_wpbc\_outc\_coded))

df\_wpbc\_corr\_table <-

df\_wpbc\_corr\_table %>%

slice(-1)

df\_wpbc\_corr\_table <- df\_wpbc\_corr\_table %>% arrange(desc(outcome\_coded))

df\_wpbc\_corr\_table

**APPENDIX J:**

**set.seed(1234)**

df\_wdbc\_diag\_coded$malignant\_coded <- as.factor(df\_wdbc\_diag\_coded$malignant\_coded)

trainList <- createDataPartition(

y = df\_wdbc\_diag\_coded$malignant\_coded,

p = .80,

list = F)

trainSet <- df\_wdbc\_diag\_coded[trainList,]

testSet <- df\_wdbc\_diag\_coded[-trainList,]

svm\_model\_wdbc <- train(malignant\_coded ~ .,

data = df\_wdbc\_diag\_coded,

method = "svmRadial")

preds <- predict(svm\_model\_wdbc, testSet)

confusionMatrix(preds, testSet$malignant\_coded, positive = "1")

**APPENDIX K:**

set.seed(1234)

df\_wpbc\_outc\_coded$outcome\_coded <- as.factor(df\_wpbc\_outc\_coded$outcome\_coded)

trainList <- createDataPartition(

y = df\_wpbc\_outc\_coded$outcome\_coded,

p = .80,

list = F)

trainSet <- df\_wpbc\_outc\_coded[trainList,]

testSet <- df\_wpbc\_outc\_coded[-trainList,]

svm\_model\_wpbc <- train(outcome\_coded ~ .,

data = df\_wpbc\_outc\_coded,

method = "svmRadial")

preds <- predict(svm\_model\_wpbc, testSet)

confusionMatrix(preds, testSet$outcome\_coded, positive = "1")

**APPENDIX L:**

set.seed(1234)

df\_wdbc\_tree <- df\_wdbc[, c(-1, -33)]

df\_wdbc\_tree

trainList <- createDataPartition(

y = df\_wdbc\_tree$diagnosis,

p = .80,

list = F)

trainSet <- df\_wdbc\_tree[trainList,]

testSet <- df\_wdbc\_tree[-trainList,]

dt\_model\_wdbc <- rpart(diagnosis ~ ., data = trainSet)

prp(dt\_model\_wdbc, extra = 104)

rpart.plot(dt\_model\_wdbc, extra = 104)

varImp(dt\_model\_wdbc)

predictions <- predict(dt\_model\_wdbc, testSet, type="class")

confusionMatrix(predictions, as.factor(testSet$diagnosis), positive = "M")

**APPENDIX M:**

set.seed(1234)

df\_wpbc\_tree <- df\_wpbc[, c(-1, -36)]

df\_wdbc\_tree

trainList <- createDataPartition(

y = df\_wpbc\_tree$outcome,

p = .80,

list = F

)

trainSet <- df\_wpbc\_tree[trainList,]

testSet <- df\_wpbc\_tree[-trainList,]

dt\_model\_wpbc <- rpart(outcome ~ ., data = trainSet)

prp(dt\_model\_wpbc, extra = 104)

rpart.plot(dt\_model\_wpbc, extra = 104)

varImp(dt\_model\_wpbc)

predictions <- predict(dt\_model\_wpbc, testSet, type="class")

confusionMatrix(predictions, as.factor(testSet$outcome), positive = "R")

**APPENDIX N:**

diagnosis\_data <- table(df\_wdbc$diagnosis)

diagnosis\_data\_table <- prop.table(diagnosis\_data)\*100

Diagnosis\_df <- as.data.frame(diagnosis\_data\_table)

Plabels <- sprint(“%s - %3.2f%s”, diagnosis\_df[,1], diagnosis\_data\_table, “%”)

Colors <- terrain.colors(2)

Pie(diagnosis\_data\_table, labels = plabels, clockwise = T, col = colors, radius = 1, cex = 1, main = “Frequency of Breast Cancer Diagnosis”)

**REFERENCES**

Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993). Biomedical image processing and biomedical visualization. doi:10.1117/12.148698

Street, W. N., Mangasarian, O. L., & Wolberg, W. H. (1995). An Inductive Learning Approach to Prognostic Prediction. Machine Learning Proceedings 1995, 522–530. doi:10.1016/b978-1-55860-377-6.50071-2

UCI (2007). UCI machine learning repository: Breast cancer wisconsin (Diagnostic) dataset. Retrieved from <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29>.

UCI (2007). UCI machine learning repository: Breast cancer wisconsin (Prognostic) dataset. Retrieved from <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Prognostic%29>.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686.