Predicting Byssinosis

Raghuram Palaniappan
University of California, Davis
Faculty of Undergraduate Studies
rpalani@ucdavis.edu
https://github.com/ragu0115/PredictingByssinosis

1 Introduction

Byssinosis is a chronic respiratory disease that is caused by prolonged exposure to cotton, flax, or hemp dust in the workplace. The disease can be debilitating and can significantly impact a person's quality of life, making it an important public health concern [1].

Modeling and predicting Byssinosis can have significant implications for both individuals and public health officials. By using statistical models, researchers can identify risk factors and patterns associated with the disease, which can help inform prevention and intervention efforts. For example, models can help identify workplaces with high-risk exposure levels or identify subpopulations that may be more susceptible to the disease.

Furthermore, modeling can also help healthcare professionals and policymakers in resource allocation and decision-making, such as prioritizing prevention and treatment measures or allocating funds to research and education.

2 Data and Methodology

2.1 Data

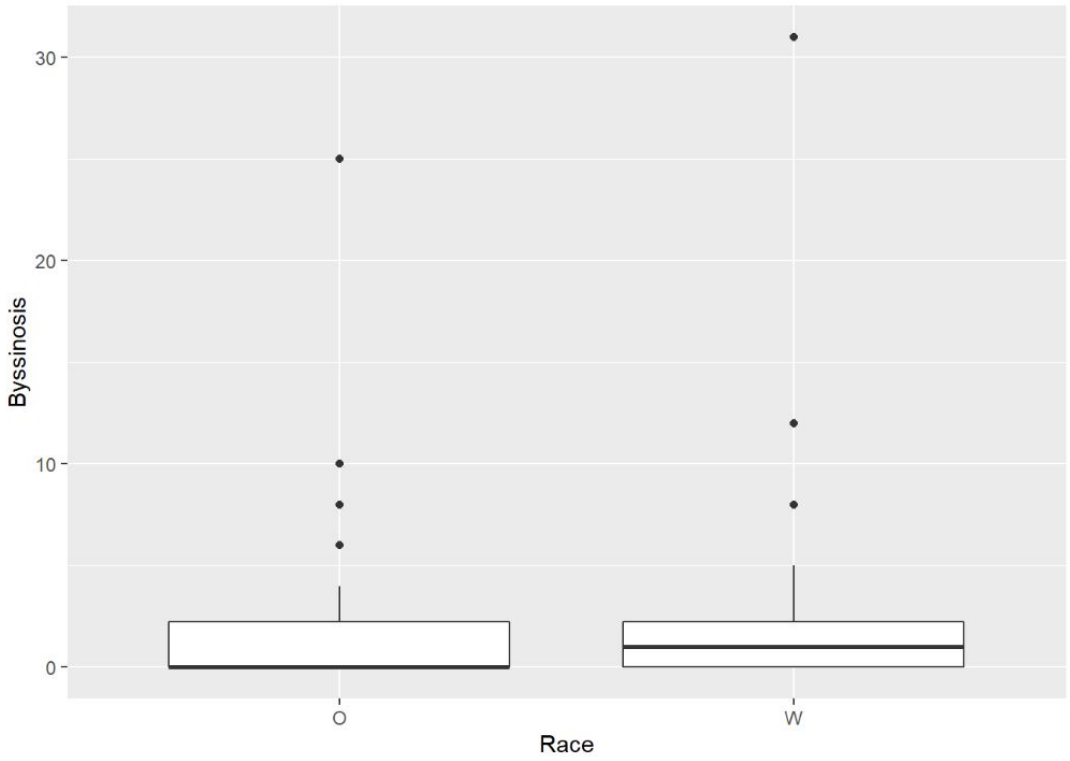
I conducted an analysis of a dataset that was gathered in 1973 from a large cotton textile company located in North Carolina. The objective of the study was to predict the occurrence of Byssinosis using the given explanatory variables. The dataset consists of information on 5,419 workers, including their workplace dust level, duration of employment, smoking habits, gender, race, and Byssinosis status. My primary goal was to investigate the relationships between these factors and the incidence of Byssinosis, and to determine whether exposure to dust at the workplace contributes to the likelihood of developing the disease. To achieve this, I employed appropriate statistical modeling techniques to predict Byssinosis.

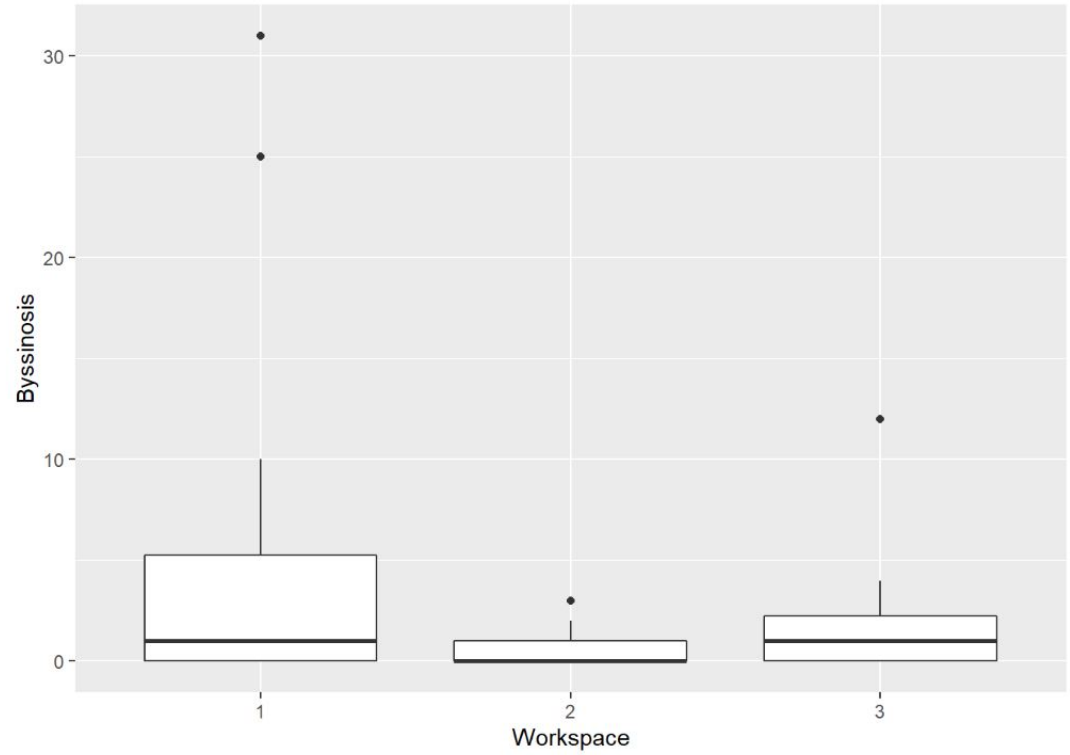
2.2 Methodology

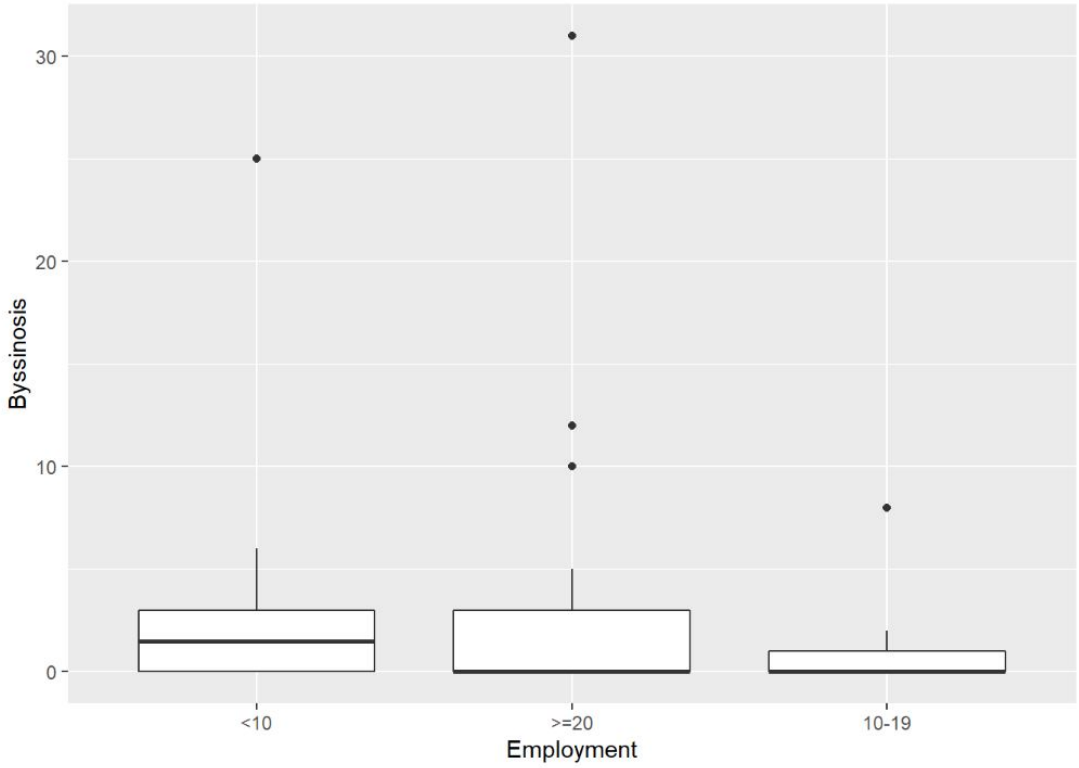
First, boxplots were created to visualize the relationship between the variables found in the dataset. To predict Byssinosis from the data, I employed three different types of models: Logistic Regression, Decision Tree, and Naive Bayes. The dataset was divided into a training set and a test set in an 80/20 ratio, and then fed into three different models. I created a function to predict Byssinosis using the most accurate model.

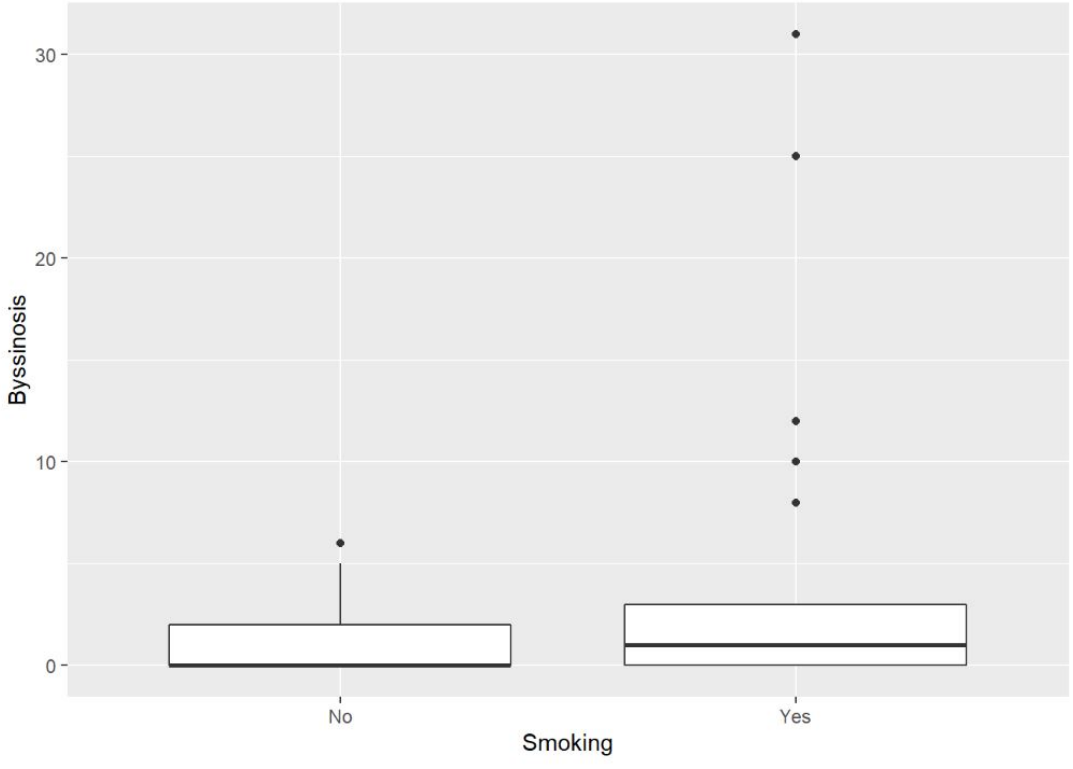
3 Data Visualization

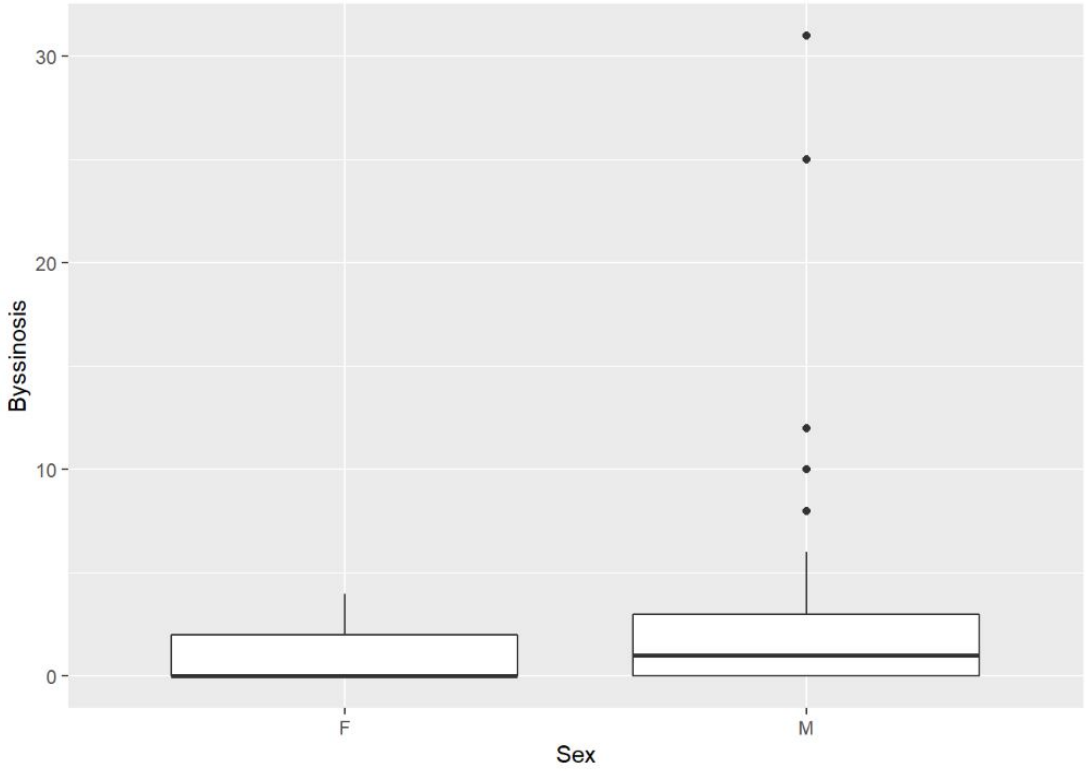
Boxplots examining the relationship between variables in the dataset


30
20
10
0
Byssinosis
O
W
Race


30
20
10
0
Byssinosis
1
2
3
Workspace


30
20
10
0
Byssinosis
<10
>=20
10-19
Employment


30
20
10
0
Byssinosis
No
Yes
Smoking



Variables Race and Sex do not seem to be correlated with Byssinosis due to the lack of difference in the factor level distributions. The other variables had differences in their distributions among factor levels. This means that these other factors outside of Race and Sex will likely contribute more to the final model.

4 Results

4.1 Logistic Regression

To assess the predictive value of the explanatory variables in the logistic regression model, a train-test split was implemented, with the model being trained on the train dataset and tested on the test dataset. The significance of each variable was determined through the computation of z-values, which reflect the degree to which the variable contributes to the outcome of interest. The workplace dust explanatory variables yielded z-values of -8.597 and -12.26, indicating a strong rejection of the null hypothesis and highlighting the significant impact of workplace dust on the incidence of Byssinosis. However, despite the model's accuracy of approximately 54%, further improvements are necessary to enhance the model's predictive power.

```
##
## Call:
## glm(formula = cbind(Byssinosis, Non.Byssinosis) ~ Workspace +
##     Employment + Smoking + Sex + Race, family = binomial(link = "logit"),
##     data = train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4195  -0.8313  -0.1337   0.3589   1.4292
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -2.363981   0.292736  -8.075 6.72e-16 ***
## Workspace2      -2.848580   0.331341  -8.597  < 2e-16 ***
## Workspace3      -2.771386   0.226046 -12.260  < 2e-16 ***
## Employment>=20   0.702328   0.244534   2.872  0.00408 **
## Employment10-19  0.501969   0.278906   1.800  0.07190 .
## SmokingYes       0.752412   0.235080   3.201  0.00137 **
## SexM            -0.202745   0.262157  -0.773  0.43930
## RaceW            0.008068   0.221504   0.036  0.97094
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 303.970  on 53  degrees of freedom
## Residual deviance:  31.812  on 46  degrees of freedom
## AIC: 138.05
##
## Number of Fisher Scoring iterations: 5
```

4.2 Decision Tree Model

The testing on the Decision Tree model yielded an approximate 46% accuracy. The Decision tree was not a strong model for the data.

4.3 Naive Bayes Model

The testing on the Naive Bayes model yielded an approximate 62% accuracy. The Naive Bayes was the best model for the data.

4.4 Python Model Function

I created a R function to make predictions. This function allowed us to predict Byssinosis using our best Naive Bayes model.

```r
# Prediction Function for Byssinosis using Tree Model
predict_Byssinosis <- function(workspace, employment, smoking, sex, race) {
  # Create a data frame with the input variables
  input_data <- data.frame(Workspace = workspace,
                           Employment = employment,
                           Smoking = smoking,
                           Sex = sex,
                           Race = race)
  # Predict the value of Byssinosis using the tree model
  prediction <- predict(tree_model, newdata = input_data, type = "class")
  # Return the predicted value of Byssinosis
  return(prediction)
}
```

5 Conclusion

The study aimed to predict the occurrence of Byssinosis by investigating the relationships between workplace dust exposure, smoking habits, gender, race, and Byssinosis incidence. Three models, including Logistic Regression, Decision Tree, and Naive Bayes, were employed to predict Byssinosis from the dataset. The workplace dust explanatory variables were found to be an essential contributor due to their high z-values of -8.597 and -12.26 in the Logistic Regression model, indicating a strong rejection of the null hypothesis. However, the model's accuracy of approximately 54% highlights the need for further improvements to enhance its predictive power. The Decision Tree model yielded an accuracy of approximately 46% and was not a strong model for the data. In contrast, the Naive Bayes model produced the highest accuracy of approximately 62% and proved to be the best model for the dataset. A R function was developed to predict Byssinosis using the most accurate model. The findings suggest that workplace dustiness is a crucial risk factor for Byssinosis incidence and can potentially inform prevention and intervention efforts to mitigate the disease's impact on public health.

References

[1] National Academies of Sciences, Engineering, and Medicine. (2019). Fostering Healthy Mental, Emotional, and Behavioral Development in Children and Youth: A National Agenda. National Academies Press (US).

[2] Statology. "How to Split Data into Training Testing Sets in R (with Examples)." Statology, 19 Mar. 2021, www.statology.org/train-test-split-r/.

[3] Dunn, Tiffany. "Categorical Stats." Tiffany Dunn, 25 July 2020, https://tiffanydunn.tech/project/categorical-stats/categorical_stats/.

Code Appendix

The code appendix may be bugged on the format. For all the code, it is best to look at the Github Repository linked at the top of the report;

```
# Utilized Tiffany Dunn's Code as Inspiration to set up original model and data (Linked in refere

# Set up Libraries and load up data
library(ggplot2)
library(caret)
library(randomForest)

# Read in Data
df <- read.csv("Byssinosis.csv")
df$Workspace <- as.factor(df$Workspace) # Converts variable to categorical

# Data Visualizations
ggplot(df, aes(x=Race, y=Byssinosis)) +
  geom_boxplot()
ggplot(df, aes(x=Workspace, y=Byssinosis)) +
  geom_boxplot()
ggplot(df, aes(x=Employment, y=Byssinosis)) +
  geom_boxplot()
ggplot(df, aes(x=Smoking, y=Byssinosis)) +
  geom_boxplot()
ggplot(df, aes(x=Sex, y=Byssinosis)) +
  geom_boxplot()

# Modeling the Data
set.seed(123) # set seed for reproducibility
# Create Train Test Data Split
```

```
train_index <- createDataPartition(df$Byssinosis, p = 0.8, list = FALSE)
train_data <- df[train_index, ]
test_data <- df[-train_index, ]
# Logistic Regression
model <- glm(cbind(Byssinosis, Non.Byssinosis) ~ Workspace + Employment + Smoking + Sex + Race, d
summary(model)
# Workplace Dust Explanatory variables give out extremely high t statistic meaning that Workplace
# Test Accuracy of the Model by using the Test Data Set
glm_pred <- predict(model, newdata = test_data, type = "response")
glm_pred_class <- ifelse(glm_pred > 0.5, 1, 0)
glm_accuracy <- mean(glm_pred_class == test_data$Byssinosis)
glm_accuracy

library(rpart)
# Modeling the Data
set.seed(77) # set seed for reproducibility
# Create Train Test Data Split
train_index <- createDataPartition(df$Byssinosis, p = 0.8, list = FALSE)
train_data <- df[train_index, ]
test_data <- df[-train_index, ]
# Fit decision tree model
tree_model <- rpart(Byssinosis ~ Workspace + Employment + Smoking + Sex + Race, data = train_data
summary(tree_model)
# Predict on test data and calculate accuracy
tree_pred <- predict(tree_model, newdata = test_data, type = "class")
tree_accuracy <- mean(tree_pred == test_data$Byssinosis)
tree_accuracy

# Load required library
library(e1071)
# Create Train Test Data Split
set.seed(100) # set seed for reproducibility
train_index <- createDataPartition(df$Byssinosis, p = 0.8, list = FALSE)
train_data <- df[train_index, ]
test_data <- df[-train_index, ]
# Fit Naive Bayes model
nb_model <- naiveBayes(Byssinosis ~ Workspace + Employment + Smoking + Sex + Race, data = train_d
# Make predictions on test data and calculate accuracy
nb_pred <- predict(nb_model, newdata = test_data)
nb_accuracy <- mean(nb_pred == test_data$Byssinosis)
nb_accuracy

# Prediction Function for Byssinosis using Tree Model
predict_Byssinosis <- function(workspace, employment, smoking, sex, race) {
  # Create a data frame with the input variables
  input_data <- data.frame(Workspace = workspace,
                           Employment = employment,
                           Smoking = smoking,
                           Sex = sex,
                           Race = race)
  # Predict the value of Byssinosis using the tree model
  prediction <- predict(tree_model, newdata = input_data, type = "class")
  # Return the predicted value of Byssinosis
  return(prediction)
}
```