Thyroid Disease Analysis [ML Model 80% Accuracy]

Aman Satyendra Yadav

2024-09-13

Introduction

Hi, I am **Aman Yadav**, currently pursuing a Master's degree in **Life Science Informatics** at the Technical University of Deggendorf, Germany. This project was undertaken independently as a personal hobby and is not associated with my university coursework.

In this project, I conducted a comprehensive analysis of the Thyroid Disease Data (available at **https://www.kaggle.com/datasets/jainaru/thyroid-disease-data/data** (https://www.kaggle.com/datasets/jainaru/thyroid-disease-data/data)). This dataset includes 13 clinicopathologic features and aims to predict the recurrence of well-differentiated thyroid cancer. The data was collected over 15 years, with each patient monitored for at least 10 years. The analysis involved data visualization and the development of a machine learning model.

Source

The data was procured from thyroid disease datasets provided by the UCI Machine Learning Repository.

Content The size for the file featured within this Kaggle dataset is shown below — along with a list of attributes, and their description summaries:

Age: The age of the patient at the time of diagnosis or treatment. Gender: The gender of the patient (male or female). Smoking: Whether the patient is a smoker or not. Hx Smoking: Smoking history of the patient (e.g., whether they have ever smoked). Hx Radiotherapy: History of radiotherapy treatment for any condition. Thyroid Function: The status of thyroid function, possibly indicating if there are any abnormalities. Physical Examination: Findings from a physical examination of the patient, which may include palpation of the thyroid gland and surrounding structures. Adenopathy: Presence or absence of enlarged lymph nodes (adenopathy) in the neck region. Pathology: Specific types of thyroid cancer as determined by pathology examination of biopsy samples. Focality: Whether the cancer is unifocal (limited to one location) or multifocal (present in multiple locations). Risk: The risk category of the cancer based on various factors, such as tumor size, extent of spread, and histological type. T: Tumor classification based on its size and extent of invasion into nearby structures. N: Nodal classification indicating the involvement of lymph nodes. M: Metastasis classification indicating the presence or absence of distant metastases. Stage: The overall stage of the cancer, typically determined by combining T, N, and M classifications. Response: Response to treatment, indicating whether the cancer responded positively, negatively, or remained stable after treatment. Recurred: Indicates whether the cancer has recurred after initial treatment.

Data summary

```
data <- read_csv("~/Thyroid_Diff.csv")
```

```
## Rows: 383 Columns: 17
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (16): Gender, Smoking, Hx Smoking, Hx Radiothreapy, Thyroid Function, Ph...
## dbl  (1): Age
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
summary(data)
```

```
##       Age           Gender            Smoking           Hx Smoking
##  Min.   :15.00   Length:383         Length:383         Length:383
##  1st Qu.:29.00   Class :character   Class :character   Class :character
##  Median :37.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :40.87
##  3rd Qu.:51.00
##  Max.   :82.00
##  Hx Radiothreapy    Thyroid Function   Physical Examination  Adenopathy
##  Length:383         Length:383         Length:383           Length:383
##  Class :character   Class :character   Class :character     Class :character
##  Mode  :character   Mode  :character   Mode  :character     Mode  :character
##
##
##
##   Pathology           Focality            Risk               T
##  Length:383         Length:383         Length:383         Length:383
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##       N                  M                Stage             Response
##  Length:383         Length:383         Length:383         Length:383
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    Recurred
##  Length:383
##  Class :character
##  Mode  :character
##
##
##
```

Analysis of Thyroid

Based on AGE

Age is the primary factor in my analysis, and its impact can be better understood through a histogram plot. While thyroid disease can affect individuals of any age group, this dataset provides insights into which age groups are most commonly affected. The histogram allows us to visually observe the distribution of ages and identify the prevalence of thyroid disease across different age ranges. This approach not only highlights patterns within the data but also helps in determining whether specific age groups are more susceptible to thyroid disease recurrence.

```
library(dplyr)
library(ggplot2)


data <- read_csv("~/Thyroid_Diff.csv")
```

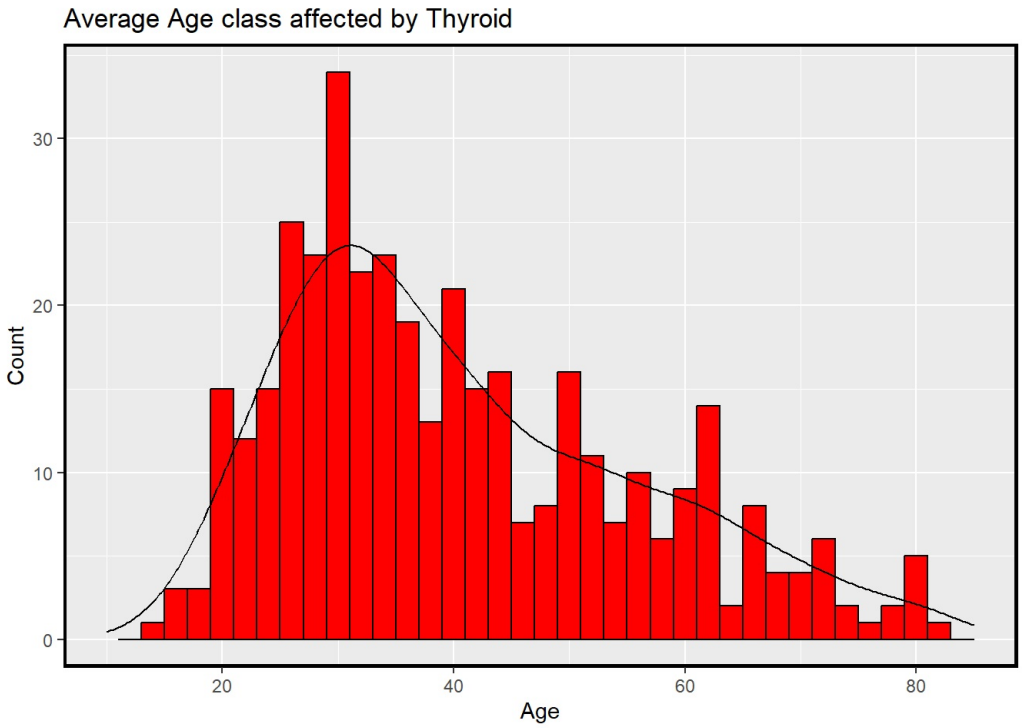
```
## Rows: 383 Columns: 17
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (16): Gender, Smoking, Hx Smoking, Hx Radiothreapy, Thyroid Function, Ph...
## dbl  (1): Age
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#View(data)



data%>%
ggplot(aes(x = Age)) + geom_histogram( binwidth = 2, fill = "red" , color = "black")+
  xlim(10,85) +  geom_density(aes(y = 2 * ..count..), color = "black") +
  labs(title = "Average Age class affected by Thyroid",
  x = "Age",
  y = "Count") + theme(panel.border = element_rect(color = "black", fill = NA , size = 1.5))
```

```
## Warning: The `size` argument of `element_rect()` is deprecated as of ggplot2 3.4.0.
## i Please use the `linewidth` argument instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(count)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_bar()`).
```



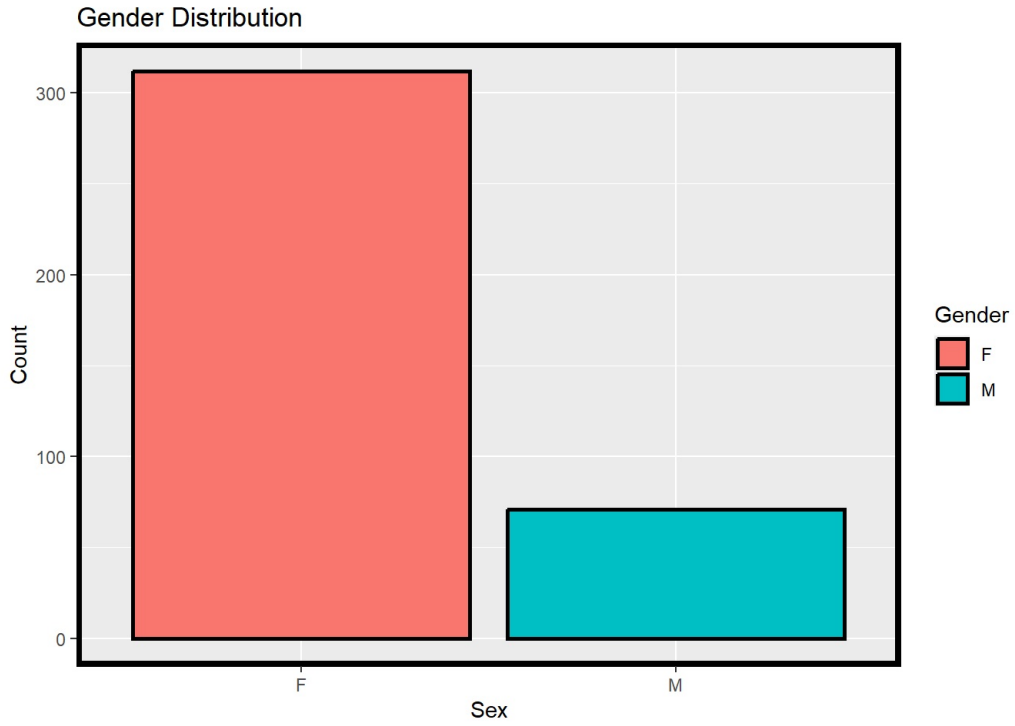
The analysis shows that the **30 to 35** age group is the most common in the thyroid disease dataset, indicating that individuals in this range are more susceptible. While the disease can affect any age, this group appears to be more vulnerable.

Based on Gender.

In the gender analysis, I found that **females are significantly more susceptible to thyroid disease than males**, with a considerable difference between the two groups.

```
data %>%
  ggplot(aes(x = Gender, fill = Gender)) +
  geom_bar(color = "black", size = 1) +  # Color the borders black and set their thickness
  labs(title = "Gender Distribution",
       x = "Sex",
       y = "Count") +
  theme(panel.border = element_rect(color = "black", fill = NA, size = 2.5))
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Here are more information to justify the data set

Women are more prone to thyroid diseases due to a combination of **hormonal, genetic, and autoimmune factor**s. One significant contributor is the role of **estrogen**, a female hormone that can influence thyroid function by enhancing the immune response. This heightened immune activity may increase the likelihood of autoimmune conditions like Hashimoto's thyroiditis, where the immune system mistakenly attacks the thyroid gland. Estrogen also affects the production of proteins that bind to thyroid hormones, potentially leading to hormonal imbalances.

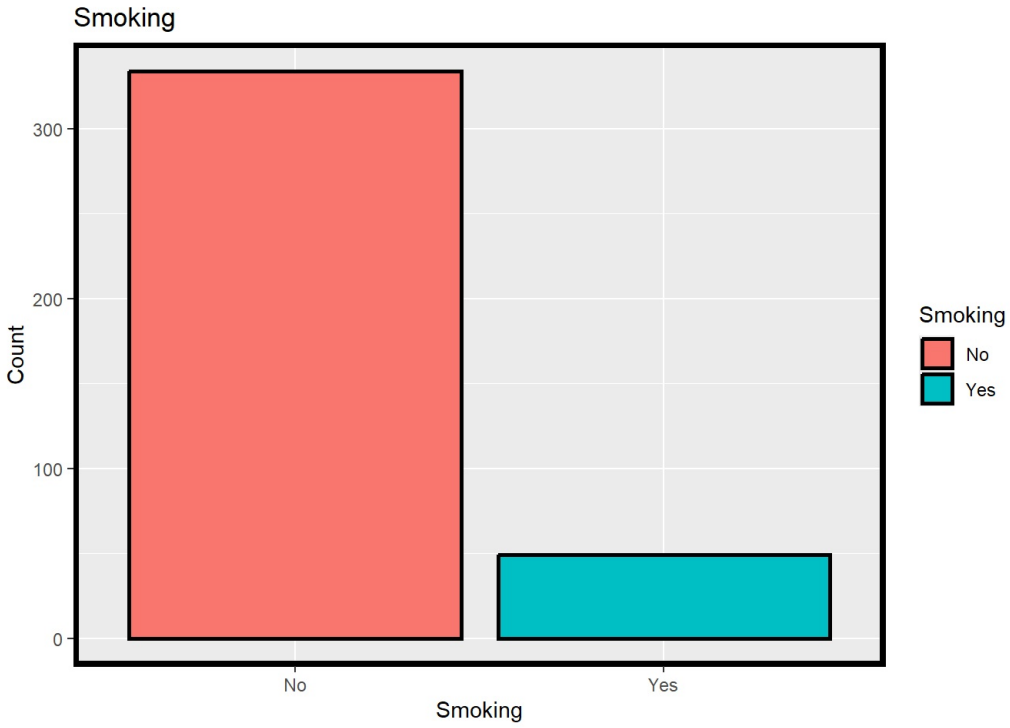
Additionally, women experience fluctuations in hormone levels during life stages like pregnancy, menopause, and puberty, which can further impact thyroid function. Autoimmune disorders, which are a major cause of thyroid issues, are more prevalent in women for reasons that may be linked to evolutionary biology and environmental factors.

The susceptibility of women to thyroid disorders is well-documented, with studies indicating that they are up to 10 times more likely to develop such conditions than men**(Mayo Clinic News Network)(India Today)(Modern Hypothyroid Care | Paloma Health)**.

Smoking habits.

Although smoking is widely recognized as harmful and contributes to many health problems, including lung cancer, the dataset suggests that smoking is not a significant risk factor for thyroid disease. This observation might be due to various factors, such as the limitations of the dataset, which may not fully capture the complex relationship between smoking and thyroid disease. It's also possible that the dataset does not include comprehensive data on other relevant health conditions or lifestyle factors. Therefore, while the dataset indicates no strong association between smoking and thyroid disease, smoking remains a well-established risk factor for numerous other serious health issues.

```
data %>%
  ggplot(aes(x = Smoking, fill = Smoking)) +
  geom_bar(color = "black", size = 1) +  # Color the borders black and set their thickness
  labs(title = "Smoking ",
       x = "Smoking",
       y = "Count") +
  theme(panel.border = element_rect(color = "black", fill = NA, size = 2.5))
```



Smoking and Radiotherapy history.

The analysis of smoking history and radiotherapy in relation to thyroid disease indicates that neither factor is directly associated with the disease. However, these factors can act as risk elements in certain cases, potentially contributing to thyroid disease under specific conditions or in conjunction with other risk factors.

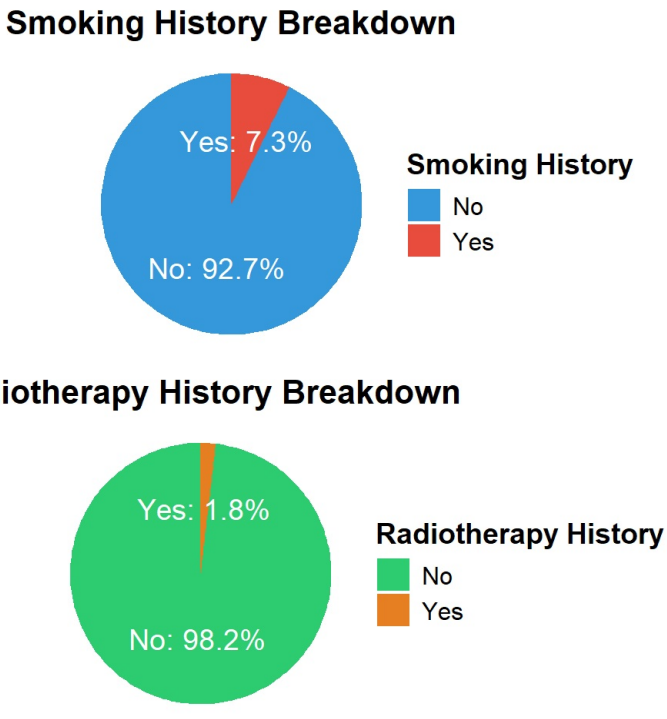
```
value_counts1 <- data %>%
  filter(!is.na(`Hx Radiothreapy`)) %>%
  count(`Hx Radiothreapy`) %>%
  mutate(percentage1 = round(n / sum(n) * 100, 1))
value_counts1
```

```
## # A tibble: 2 × 3
##   `Hx Radiothreapy`     n percentage1
##   <chr>             <int>       <dbl>
## 1 No                  376        98.2
## 2 Yes                   7         1.8
```

```
value_counts2 <- data %>%
  filter(!is.na(`Hx Smoking`)) %>%
  count(`Hx Smoking`) %>%
  mutate(percentage2 = round(n / sum(n) * 100, 1))
value_counts2
```

```
## # A tibble: 2 × 3
##   `Hx Smoking`     n percentage2
##   <chr>        <int>       <dbl>
## 1 No             355        92.7
## 2 Yes             28         7.3
```

```
pichart1 <- value_counts1 %>%
  ggplot(aes(x = "", y = n, fill = `Hx Radiothreapy`)) +
  geom_bar(stat = "identity", width = 1) +
  geom_text(aes(label = paste0(`Hx Radiothreapy`, ": ", percentage1, "%")),
            position = position_stack(vjust = 0.5), size = 5, color = "white") +
  scale_fill_manual(values = c("#2ecc71", "#e67e22")) +
  coord_polar("y", start = 0) +
  theme_void() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
    legend.position = "right",
    legend.title = element_text(size = 14, face = "bold"),
    legend.text = element_text(size = 12)
  ) +
  labs(fill = "Radiotherapy History", title = "Radiotherapy History Breakdown")




pichart2 <- value_counts2 %>%
  ggplot(aes(x = "", y = n, fill = `Hx Smoking`)) +
  geom_bar(stat = "identity", width = 1) +
  geom_text(aes(label = paste0(`Hx Smoking`, ": ", percentage2, "%")),
            position = position_stack(vjust = 0.5), size = 5, color = "white") +
  scale_fill_manual(values = c("#3498db", "#e74c3c")) +
  coord_polar("y", start = 0) +
  theme_void() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
    legend.position = "right",
    legend.title = element_text(size = 14, face = "bold"),
    legend.text = element_text(size = 12)
  ) +
  labs(fill = "Smoking History", title = "Smoking History Breakdown")

grid.arrange(pichart2,pichart1, ncol =1 )
```



Above 90% people didnt had any smoking or radiotherapy history

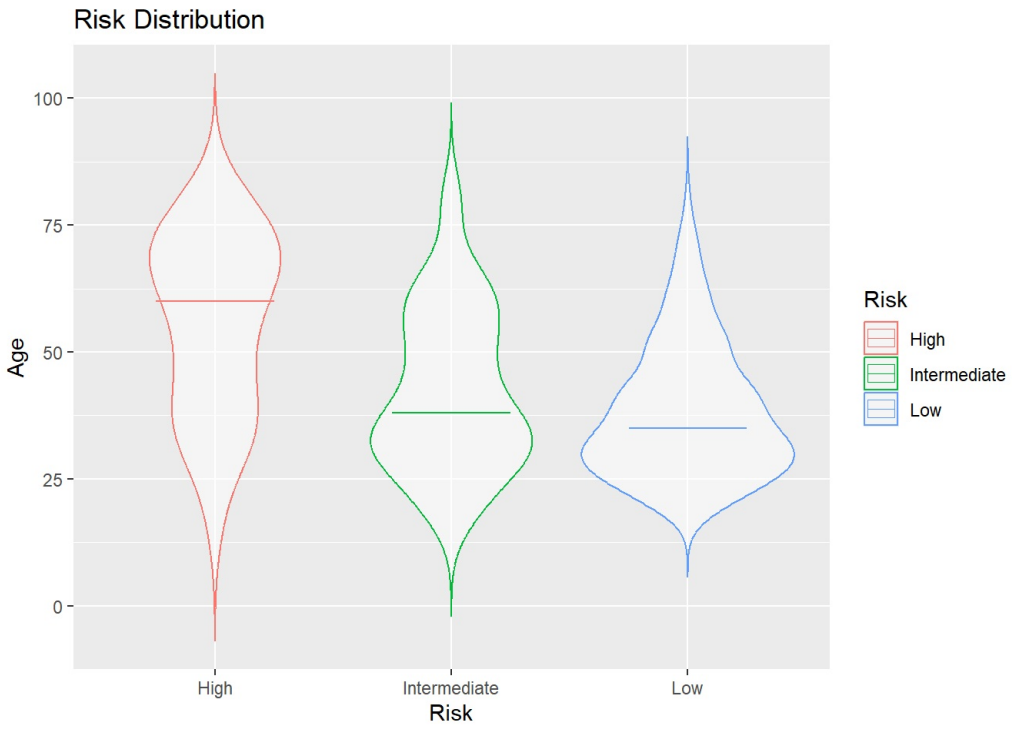
Risk Distribution

It's well-established that older individuals are generally more susceptible to a range of health issues, and this trend is evident in the analysis of thyroid disease risk as well. Using a violin plot (or diamond plot) to examine the risk distribution across different age groups in the dataset, it was observed that the risk of thyroid disease increases with age. This analysis supports the common understanding that age is a significant risk factor for thyroid disease, with older age groups showing higher risk levels.

```
violonplot <- data%>%
  ggplot(aes(x = Risk, y = Age, color = Risk)) +
  geom_violin(trim = FALSE, alpha = 0.5) +
  stat_summary(fun = median, geom = "crossbar", aes(color = Risk), width = 0.5, size = 0.2)+
  labs(title = "Risk Distribution")
  theme(panel.border = element_rect(color = "black", fill = NA, size = 2.5))+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## List of 2
##  $ panel.border:List of 5
##   ..$ fill         : logi NA
##   ..$ colour       : chr "black"
##   ..$ linewidth    : num 2.5
##   ..$ linetype     : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_rect" "element"
##  $ axis.text.x :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 1
##   ..$ vjust        : NULL
##   ..$ angle        : num 45
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

```
violonplot
```



Physical Examination

Findings from a **physical examination** of the patient, which may include palpation of the thyroid gland and surrounding structures.

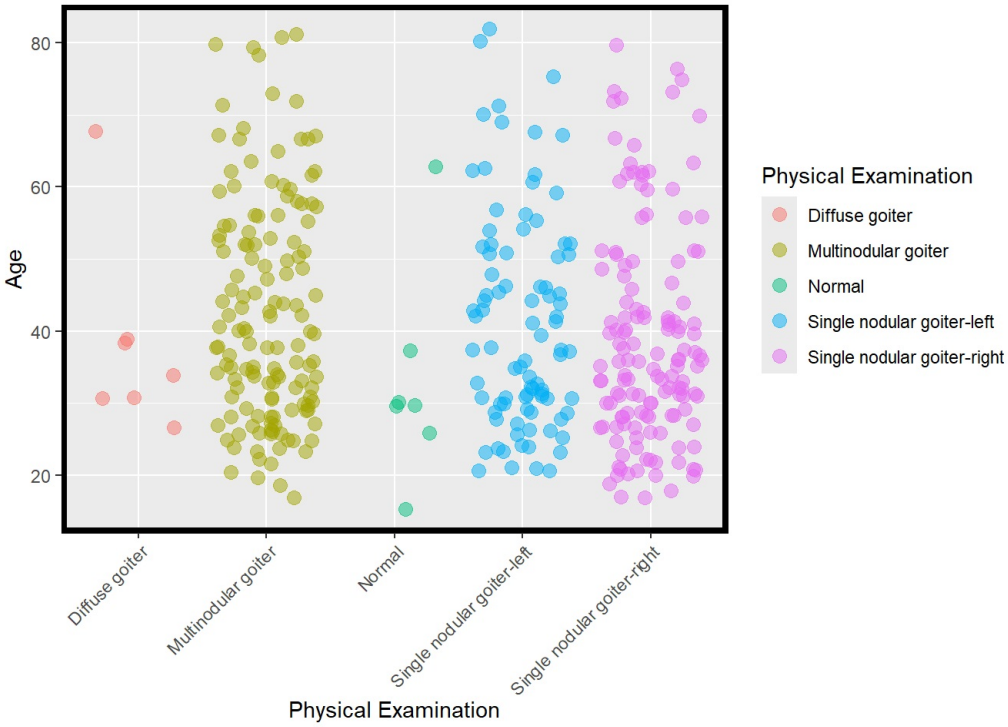
Palpation of the thyroid gland: The doctor uses their hands to feel (palpate) the thyroid gland located in the front of the neck. This is done to check for:

Size: An enlarged thyroid may indicate a condition such as goiter.

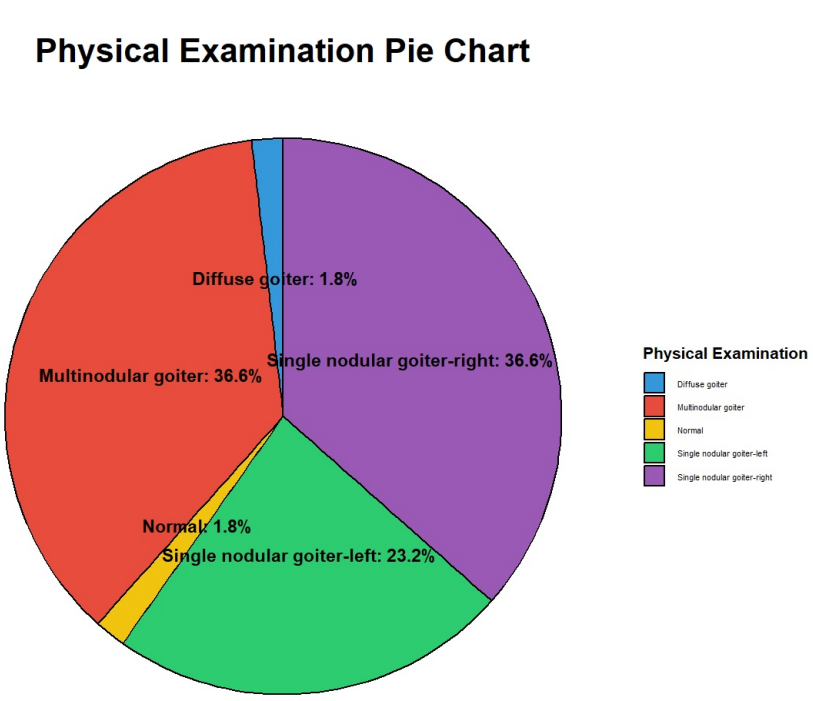
Texture: The provider feels for any abnormalities like lumps or nodules, which could suggest thyroid disease or, in some cases, thyroid cancer.

Tenderness: Pain or discomfort during palpation may point to inflammation, such as in thyroiditis.

```
jitter <- data%>%
  ggplot(aes(x = `Physical Examination`, y = Age, color = `Physical Examination`)) +
  geom_jitter(size = 3, alpha = 0.5) +
  theme(panel.border = element_rect(color = "black", fill = NA, size = 2.5))+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

jitter
```




```
Valuecount3 <- data %>%
  filter(!is.na(`Physical Examination`)) %>%
  count(`Physical Examination`) %>%
  mutate(percentage1 = round(n / sum(n) * 100, 1))

pichart3 <- Valuecount3 %>%
  ggplot(aes(x = "", y = n, fill = `Physical Examination`)) +
  geom_bar(stat = "identity", width = 1, color = "black") +
  coord_polar("y", start = 0) +
  geom_text(aes(label = paste0(`Physical Examination`, ": ", percentage1, "%")),
            position = position_stack(vjust = 0.5),
            size = 3,
            color = "black",
            fontface = "bold") +
  scale_fill_manual(values = c("#3498db", "#e74c3c", "#f1c40f", "#2ecc71", "#9b59b6")) +
  theme_void() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
    legend.position = "right",
    legend.title = element_text(size = 8, face = "bold"),
    legend.text = element_text(size = 4),
    legend.key.size = unit(0.4, 'cm')
  ) +
  labs(fill = "Physical Examination", title = "Physical Examination Pie Chart")
pichart3
```

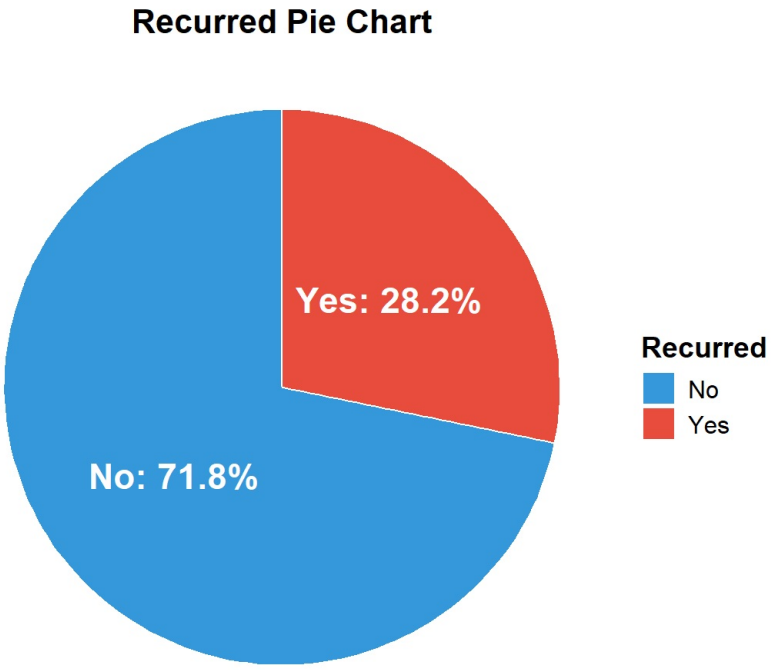


From the plot above, it is evident that the presence of **goiter** is a common finding among patients suffering from thyroid disease. This condition appears consistently across all age groups, **indicating that goiter may be a significant clinical sign regardless of the patient's age**. Although other factors may contribute to the disease, goiter remains a prevalent physical manifestation observed during diagnosis

Recurrance

Even though our analysis shows that the treatment outcomes for most thyroid disease patients were **excellent**, the risk of **recurrence** remains a significant concern. Thyroid disease can resurface after a period of remission, making recurrence one of the most critical factors to consider in diagnosis and long-term management. Monitoring for recurrence is essential, as it can impact both **treatment plans and patient prognosis**. The plot below provides a clear insight into the prevalence of recurrence within our sample population, highlighting how common this issue is among patients despite initially successful treatments.

```
value_counts <- data %>%
  count(Recurred) %>%
  mutate(percentage = round(n / sum(n) * 100, 1))


pichart3 <- value_counts %>%
  ggplot(aes(x = "", y = n, fill = Recurred)) +
  geom_bar(stat = "identity", width = 1, color = "white") +
  coord_polar("y", start = 0) +
  geom_text(aes(label = paste0(Recurred, ": ", percentage, "%")),
            position = position_stack(vjust = 0.5),
            size = 6,
            color = "white",
            fontface = "bold") +
  scale_fill_manual(values = c("#3498db", "#e74c3c")) +
  theme_void() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
    legend.position = "right",
    legend.title = element_text(size = 14, face = "bold"),
    legend.text = element_text(size = 12)
  ) +
  labs(fill = "Recurred", title = "Recurred Pie Chart")

pichart3
```



Machine Learning Model.

In this machine learning model, I used three independent variables—Age, Gender, and Risk—to predict recurrence. I chose a decision tree model due to the categorical nature of the dataset, which is well-suited for handling categorical variables and providing interpretable results.

Necesary libraries

```
install.packages("rpart")
```

```
## Installing package into 'C:/Users/Aman Yadav/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)
```

```
## package 'rpart' successfully unpacked and MD5 sums checked
```

```
## Warning: cannot remove prior installation of package 'rpart'
```

```
## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying
## C:\Users\Aman
## Yadav\AppData\Local\R\win-library\4.3\00LOCK\rpart\libs\x64\rpart.dll to
## C:\Users\Aman Yadav\AppData\Local\R\win-library\4.3\rpart\libs\x64\rpart.dll:
## Permission denied
```

```
## Warning: restored 'rpart'
```

```
##
## The downloaded binary packages are in
##  C:\Users\Aman Yadav\AppData\Local\Temp\RtmpeCGhSi\downloaded_packages
```

```
install.packages("rpart.plot")
```

```
## Installing package into 'C:/Users/Aman Yadav/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)
```

```
## package 'rpart.plot' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\Aman Yadav\AppData\Local\Temp\RtmpeCGhSi\downloaded_packages
```

```
install.packages("caret")
```

```
## Installing package into 'C:/Users/Aman Yadav/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)
```

```
## package 'caret' successfully unpacked and MD5 sums checked
```

```
## Warning: cannot remove prior installation of package 'caret'
```

```
## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying
## C:\Users\Aman
## Yadav\AppData\Local\R\win-library\4.3\00LOCK\caret\libs\x64\caret.dll to
## C:\Users\Aman Yadav\AppData\Local\R\win-library\4.3\caret\libs\x64\caret.dll:
## Permission denied
```

```
## Warning: restored 'caret'
```

```
##
## The downloaded binary packages are in
##  C:\Users\Aman Yadav\AppData\Local\Temp\RtmpeCGhSi\downloaded_packages
```

```
install.packages("nnet")
```

```
## Installing package into 'C:/Users/Aman Yadav/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)
```

```
## package 'nnet' successfully unpacked and MD5 sums checked
```

```
## Warning: cannot remove prior installation of package 'nnet'
```

```
## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying
## C:\Users\Aman
## Yadav\AppData\Local\R\win-library\4.3\00LOCK\nnet\libs\x64\nnet.dll to
## C:\Users\Aman Yadav\AppData\Local\R\win-library\4.3\nnet\libs\x64\nnet.dll:
## Permission denied
```

```
## Warning: restored 'nnet'
```

```
##
## The downloaded binary packages are in
##  C:\Users\Aman Yadav\AppData\Local\Temp\RtmpeCGhSi\downloaded_packages
```

```
library(rpart)
library(rpart.plot)
library(caret)
```

```
## Loading required package: lattice
```

```
library(nnet)
```

Seting X and Y variable.

here our X variable will be Gender + Age + Risk Y variable is Recurrence. **rpart()** Function: Builds the decision tree model.

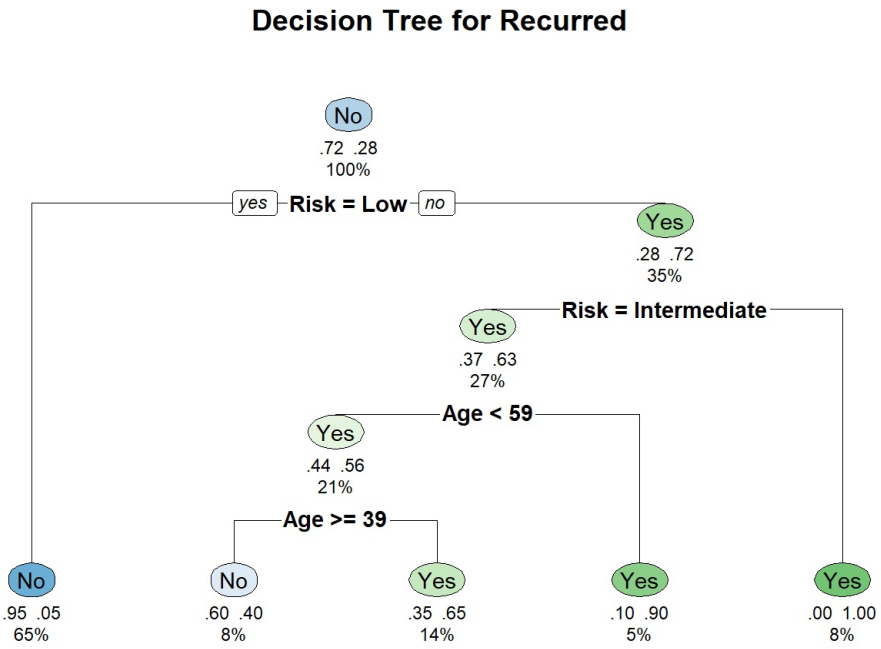
```
set.seed(123)  # for reproducibility
tree_model <- rpart(Recurred ~ Gender + Age + Risk , data = data, method = "class")
summary(tree_model)
```

```
## Call:
## rpart(formula = Recurred ~ Gender + Age + Risk, data = data,
##     method = "class")
##   n= 383
##
##           CP nsplit rel error    xerror       xstd
## 1 0.53703704      0 1.0000000 1.0000000 0.08153707
## 2 0.01851852      1 0.4629630 0.4629630 0.06104976
## 3 0.01000000      4 0.4074074 0.4814815 0.06207154
```

```
##
## Variable importance
##   Risk    Age Gender
##     76     15      9
##
## Node number 1: 383 observations,    complexity param=0.537037
##   predicted class=No   expected loss=0.2819843  P(node) =1
##     class counts:   275   108
##    probabilities: 0.718 0.282
##   left son=2 (249 obs) right son=3 (134 obs)
##   Primary splits:
##       Risk   splits as  RRL,      improve=77.80025, (0 missing)
##       Gender splits as  LR,       improve=16.70460, (0 missing)
##       Age    < 57.5 to the left,  improve=16.47216, (0 missing)
##   Surrogate splits:
##       Gender splits as  LR,       agree=0.695, adj=0.127, (0 split)
##       Age    < 57.5 to the left,  agree=0.692, adj=0.119, (0 split)
##
## Node number 2: 249 observations
##   predicted class=No   expected loss=0.04819277  P(node) =0.6501305
##     class counts:   237    12
##    probabilities: 0.952 0.048
##
## Node number 3: 134 observations,    complexity param=0.01851852
##   predicted class=Yes  expected loss=0.2835821  P(node) =0.3498695
##     class counts:    38    96
##    probabilities: 0.284 0.716
##   left son=6 (102 obs) right son=7 (32 obs)
##   Primary splits:
##       Risk   splits as  RL-,      improve=6.761487, (0 missing)
##       Age    < 58.5 to the left,  improve=5.385627, (0 missing)
##       Gender splits as  LR,       improve=2.030589, (0 missing)
##   Surrogate splits:
##       Age < 63.5 to the left,  agree=0.799, adj=0.156, (0 split)
##
## Node number 6: 102 observations,    complexity param=0.01851852
##   predicted class=Yes  expected loss=0.372549  P(node) =0.2663185
##     class counts:    38    64
##    probabilities: 0.373 0.627
##   left son=12 (82 obs) right son=13 (20 obs)
##   Primary splits:
##       Age    < 58.5 to the left,  improve=3.696031, (0 missing)
##       Gender splits as  LR,       improve=1.394352, (0 missing)
##
## Node number 7: 32 observations
##   predicted class=Yes  expected loss=0  P(node) =0.08355091
##     class counts:     0    32
##    probabilities: 0.000 1.000
##
## Node number 12: 82 observations,    complexity param=0.01851852
##   predicted class=Yes  expected loss=0.4390244  P(node) =0.2140992
##     class counts:    36    46
##    probabilities: 0.439 0.561
##   left son=24 (30 obs) right son=25 (52 obs)
##   Primary splits:
##       Age    < 39   to the right, improve=2.4517820, (0 missing)
##       Gender splits as  LR,       improve=0.6306811, (0 missing)
##
## Node number 13: 20 observations
##   predicted class=Yes  expected loss=0.1  P(node) =0.05221932
##     class counts:     2    18
##    probabilities: 0.100 0.900
##
## Node number 24: 30 observations
##   predicted class=No   expected loss=0.4  P(node) =0.07832898
##     class counts:    18    12
##    probabilities: 0.600 0.400
##
## Node number 25: 52 observations
##   predicted class=Yes  expected loss=0.3461538  P(node) =0.1357702
##     class counts:    18    34
##    probabilities: 0.346 0.654
```

We can also visualize our decision tree accordding to the following variables. The code below mentions the same

```
rpart.plot(tree_model, extra = 104, under = TRUE, faclen = 0, main = "Decision Tree for Recurred")
```



Splitin the Data.

It is very important to split the data in test and training set in 8:2 ratio before training the model. The testing dataset will be used latter to find the accuracy of our Machine learning model.

```
set.seed(123)
train_index <- createDataPartition(data$Recurred, p = 0.8, list = FALSE)
train_data <- data[train_index, ]
test_data <- data[-train_index, ]
test_data$Recurred <- as.factor(test_data$Recurred)
```

**train_index <- createDataPartition(data
$Recurred, p = 0.8, list = FALSE) * * thisfunctionwillcreateanindexforsplittingthedataintotraining(80testing(20 * * test_data$Recurred <-
as.factor(test_data$Recurred) make sure that variable of test set are considered as a factor for our prefiction.

Model Building

After preparing the Data we are ready to build the model by putting everything together.

```
tree_model <- rpart(Recurred ~ Gender + Age + Risk, data = train_data, method = "class")
predictions <- predict(tree_model, newdata = test_data, type = "class")


confusionMatrix(predictions, test_data$Recurred)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction No Yes
##        No  48   5
##        Yes  7  16
##
##                Accuracy : 0.8421
##                  95% CI : (0.7404, 0.9157)
##     No Information Rate : 0.7237
##     P-Value [Acc > NIR] : 0.01146
##
##                   Kappa : 0.6165
##
##  Mcnemar's Test P-Value : 0.77283
##
##             Sensitivity : 0.8727
##             Specificity : 0.7619
##          Pos Pred Value : 0.9057
##          Neg Pred Value : 0.6957
##              Prevalence : 0.7237
##          Detection Rate : 0.6316
##    Detection Prevalence : 0.6974
##       Balanced Accuracy : 0.8173
##
##        'Positive' Class : No
##
```

The confusion matrix provides a summary of the classification performance of the model. In this case, the confusion matrix shows how the model's predictions compare to the actual values for the test data:

True Negatives (No predicted as No): 48 False Positives (Yes predicted as No): 5 False Negatives (No predicted as Yes): 7 True Positives (Yes predicted as Yes): 16 **Accuracy**: The model correctly predicted the outcome for approximately 84.21% of the cases. This metric provides a general measure of how well the model performs overall.

95% Confidence Interval for Accuracy: The accuracy estimate is between 74.04% and 91.57%. This range reflects the uncertainty in the accuracy estimate due to sample variability.

No Information Rate (NIR): This value of 72.37% represents the accuracy that could be achieved by always predicting the most frequent class. The model's accuracy is significantly higher than this baseline.

P-Value [Acc > NIR]: The p-value of 0.01146 indicates that the model's accuracy is statistically significantly better than what would be achieved by random guessing.

Kappa: The kappa statistic of 0.6165 measures the agreement between the predicted and actual classifications, accounting for chance. A value of 0.6165 suggests substantial agreement beyond what would be expected by chance.

Mcnemar's Test P-Value: The p-value of 0.77283 for McNemar's test assesses whether the proportion of errors is significantly different between the two classes. A high p-value suggests that the error rates for the two classes are not significantly different.

Sensitivity: The sensitivity of 87.27% indicates the model's ability to correctly identify positive cases (Yes). In other words, it successfully identifies 87.27% of the actual positive cases.

Specificity: The specificity of 76.19% measures how well the model identifies negative cases (No). It correctly identifies 76.19% of the actual negative cases.

Positive Predictive Value (PPV): The PPV of 90.57% indicates the proportion of positive predictions that are actually correct. This high value suggests that when the model predicts a positive outcome, it is usually correct.

Negative Predictive Value (NPV): The NPV of 69.57% shows the proportion of negative predictions that are accurate. This indicates that when the model predicts a negative outcome, it is correct 69.57% of the time.

Prevalence: The prevalence of 72.37% represents the proportion of actual positive cases in the dataset. This high prevalence indicates that positive cases are common in the test set.

Detection Rate: The detection rate of 63.16% measures the proportion of actual positives that are correctly identified. It shows that the model detects 63.16% of the actual positive cases.

Detection Prevalence: The detection prevalence of 69.74% represents the proportion of predicted positives out of all predictions. This indicates that the model predicts 69.74% of cases as positive.

Balanced Accuracy: The balanced accuracy of 81.73% is the average of sensitivity and specificity. This metric provides a balanced measure of performance, considering both positive and negative cases.

'Positive' Class: In this analysis, "No" is considered the positive class. This means the model's performance metrics are reported with respect to predicting "No" as the positive outcome.

#END