BAN210 Predictive Analysis

ZAA

Chin Hooi Yap

Student ID: 132415217

Table of Contents

[Problem Statement & Dataset 3](#_Toc100878321)

[Data Preprocessing 3](#_Toc100878322)

[Exploratory Data Analysis (EDA) 4](#_Toc100878323)

[Predictive Model Development 6](#_Toc100878324)

[Model Comparison 7](#_Toc100878325)

[Conclusion 7](#_Toc100878326)

[Declaration 7](#_Toc100878327)

[References 7](#_Toc100878328)

# Problem Statement & Dataset

The data given is the breast cancer reoccurrence data given by the Oncology Institute. It includes 201 instance of no-recurrence-events and 85 instances of recurrence-events. There are various parameters such as age, tumor size, Deg-malig (Tumor Histological grade) and etc that can potentially be used as features in our predictive analysis. The objective of this project is to analyze the dataset and develop a predictive model to predict the recurrence of breast cancer.

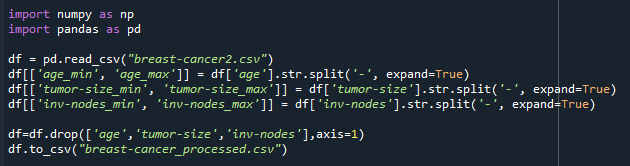
|  |  |  |
| --- | --- | --- |
| **Name** | **Description** | **Values** |
| Class | Recurrence events? (Target Variable) | no-recurrence-events, recurrence-events |
| Age | Age range | 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99 |
| Menopause | Menopause momento | lt40, ge40, premeno |
| Tumor-size | Tumor size excised in mm | 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59 |
| Inv-nodes | A metric of presence | 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39 |
| Node-caps | A metric of presence | yes, no |
| Deg-malig | Tumor Histological grade | 1, 2, 3 |
| Breast | Breast affected | left, right |
| Breastquad | Breast quadrant | left-up, left-low, right-up, right-low, central |
| Irradiat | Radiotherapy | yes, no |

Source: Mackenzie Rivero et al.[[1]](#footnote-1)

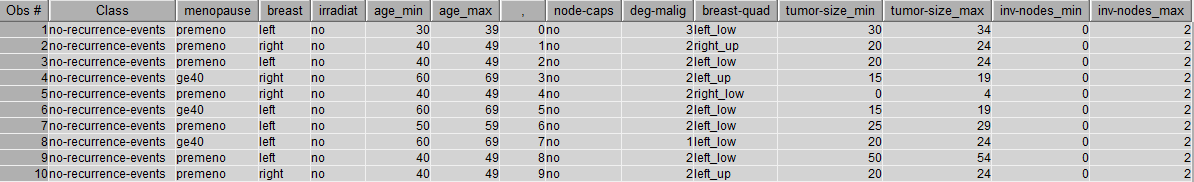
# Data Preprocessing

Before performing any analysis, it is noticeable that there are some inherent issues with the data which will make it difficult for further analysis. For example, for variables such as age and tumor-size, the values are in ranges such as 10-19, 20-29 and it will be difficult to arrange them in order when we are looking at skewness to determine whether we need to perform any transformation. It will also make it hard for any further analysis of interval variables. Hence, Python is used to split the column into two numeric columns consisting of the upper and lower bound of the range, i.e. age to age\_min and age\_max.

Python Code Used:

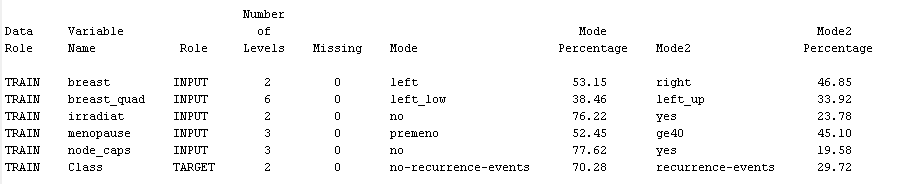


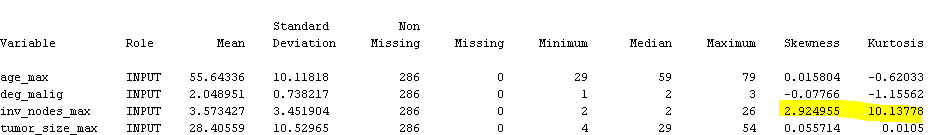
Output:



# Exploratory Data Analysis (EDA)

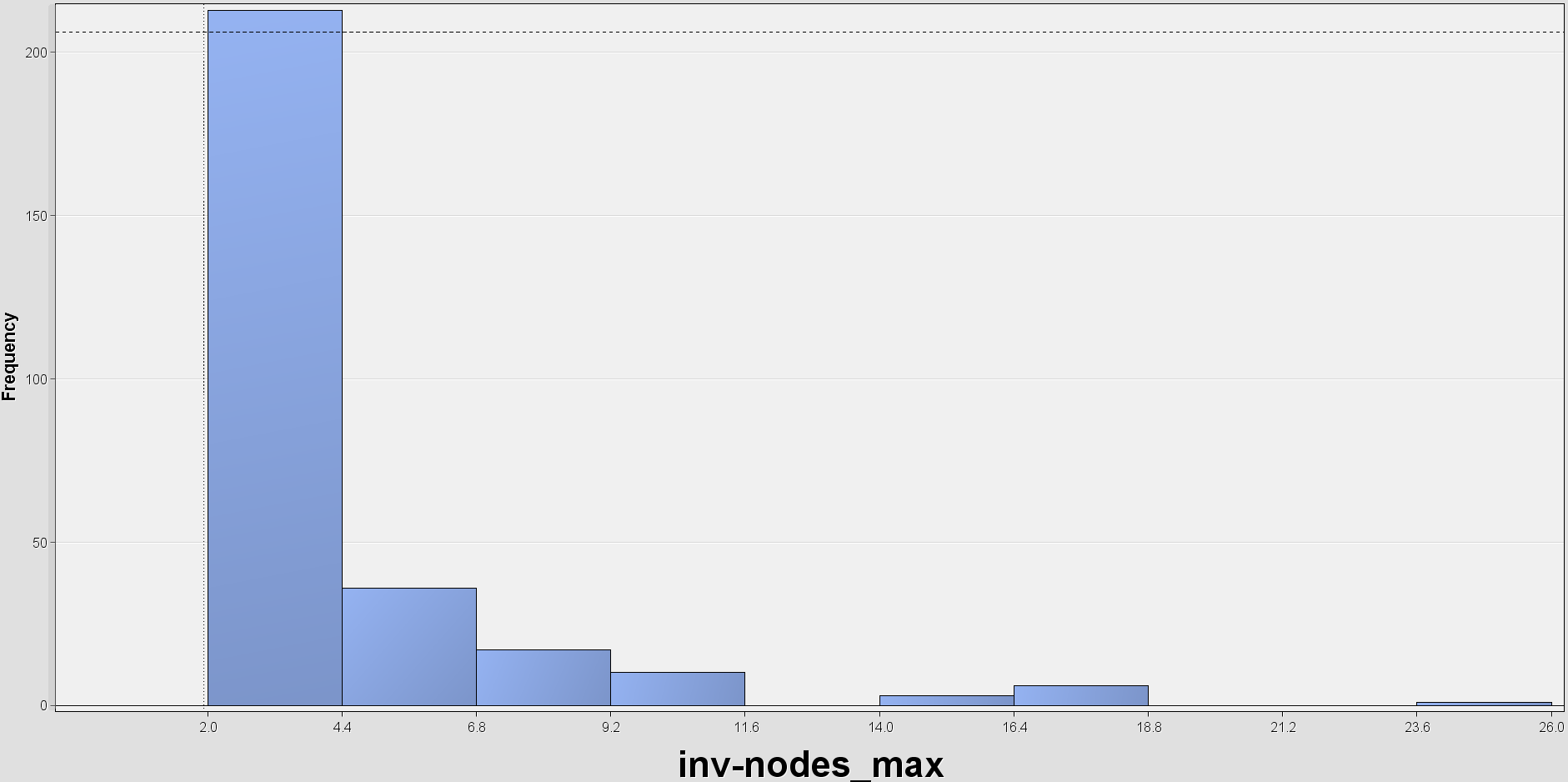
Before developing any predictive model, it is important to perform an exploratory data analysis to investigate the relationship between variables. Firstly, we will start by performing the univariate analysis using StatExplore.





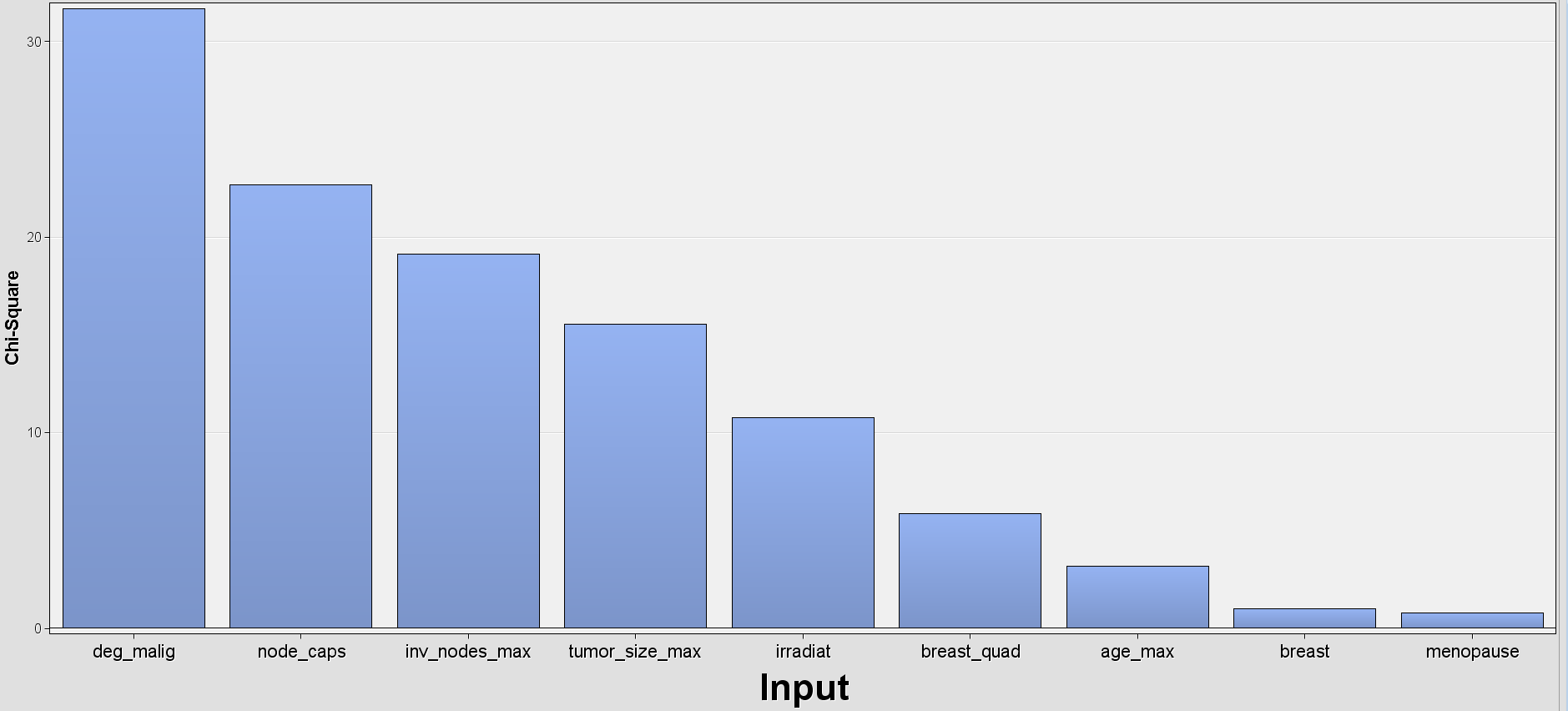
From the output above, we can see that see that inv\_nodes\_max has a very high skewness which would need to be dealt with later with a transformation. This is confirmed with a histogram plot.

**Histogram of inv\_nodes\_max**



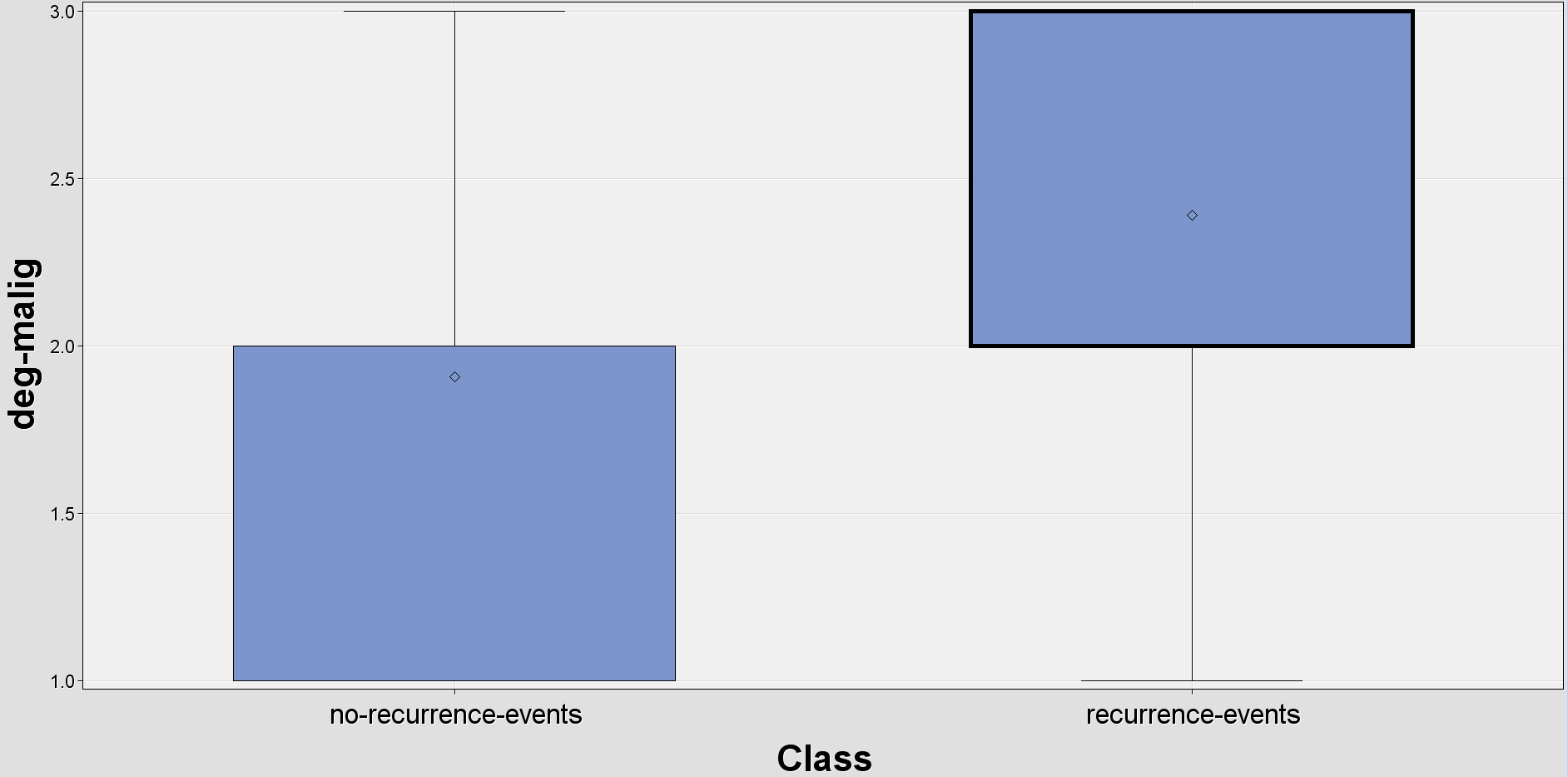
In addition, by looking at the Chi-Square value plot, we will notice the top few variables that are most likely to affect the target variables. In the figure below, we can see that def\_malig, node\_caps, inv\_nodes\_max and tumor\_size\_max are the top 5 variables that we will need to focus on.

**Chi-Square Plot of All Input Variables**



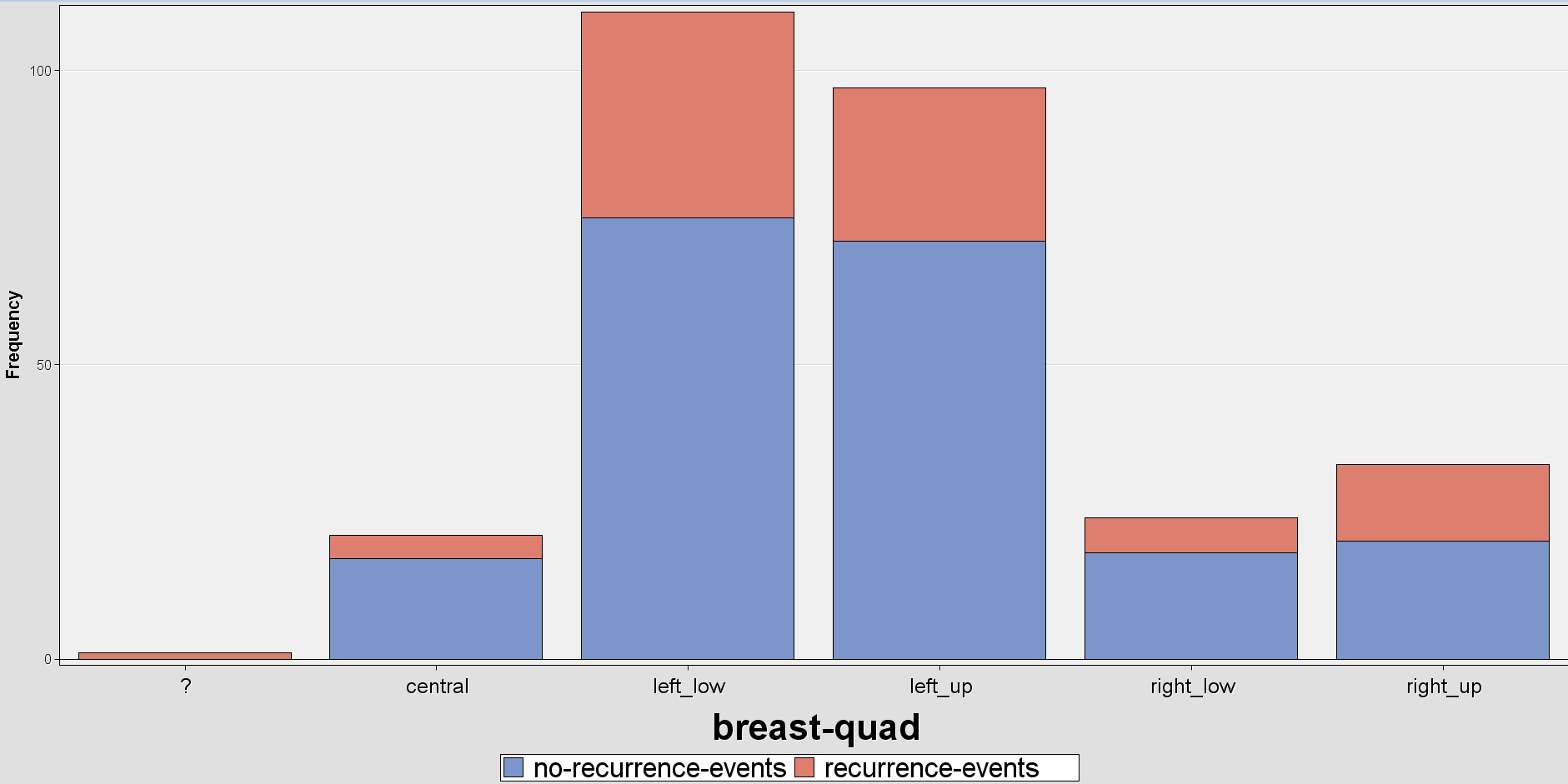
Next, we will look at the some of the variables with respect to the target variable. We can see from the figure below that deg-malig has a very clear relationship with the target variable. No-recurrence group a median of 2 and typically stays between the range of 1 to 2 whereas recurrence group has a median of 3 and typically stays between the range of 2 to 3. This shows that it will be one of the most important features in our model.

**Boxplot of deg-malig**



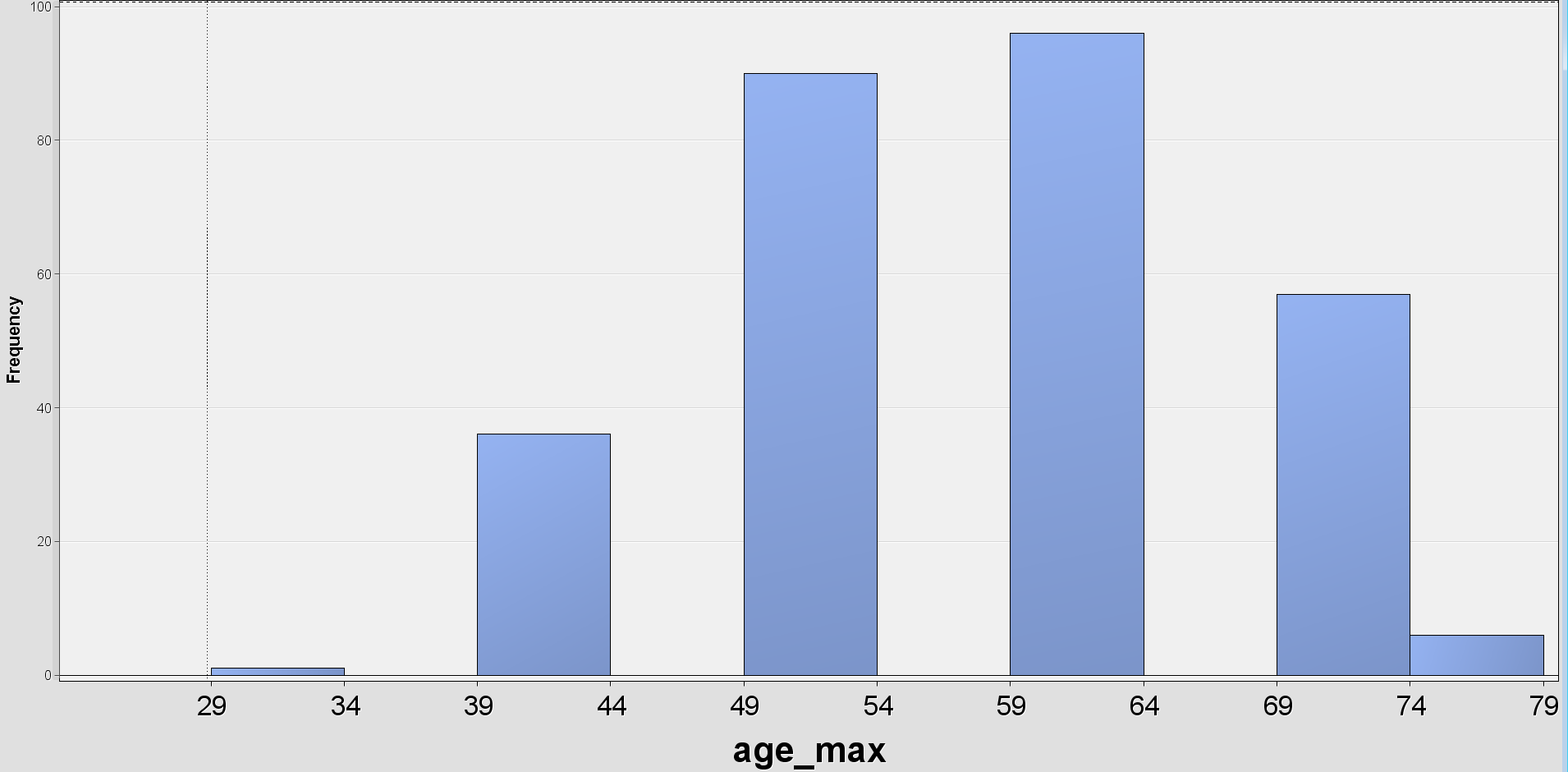
Besides that, we noticed that based on the data available, breast cancer is more commonly happen in the left quadrant of the breast as left\_up and left\_down made up of 72% of the data. This is an important finding as it can be used for future diagnostics of breast cancer.

**Bar plot of breast-quad**



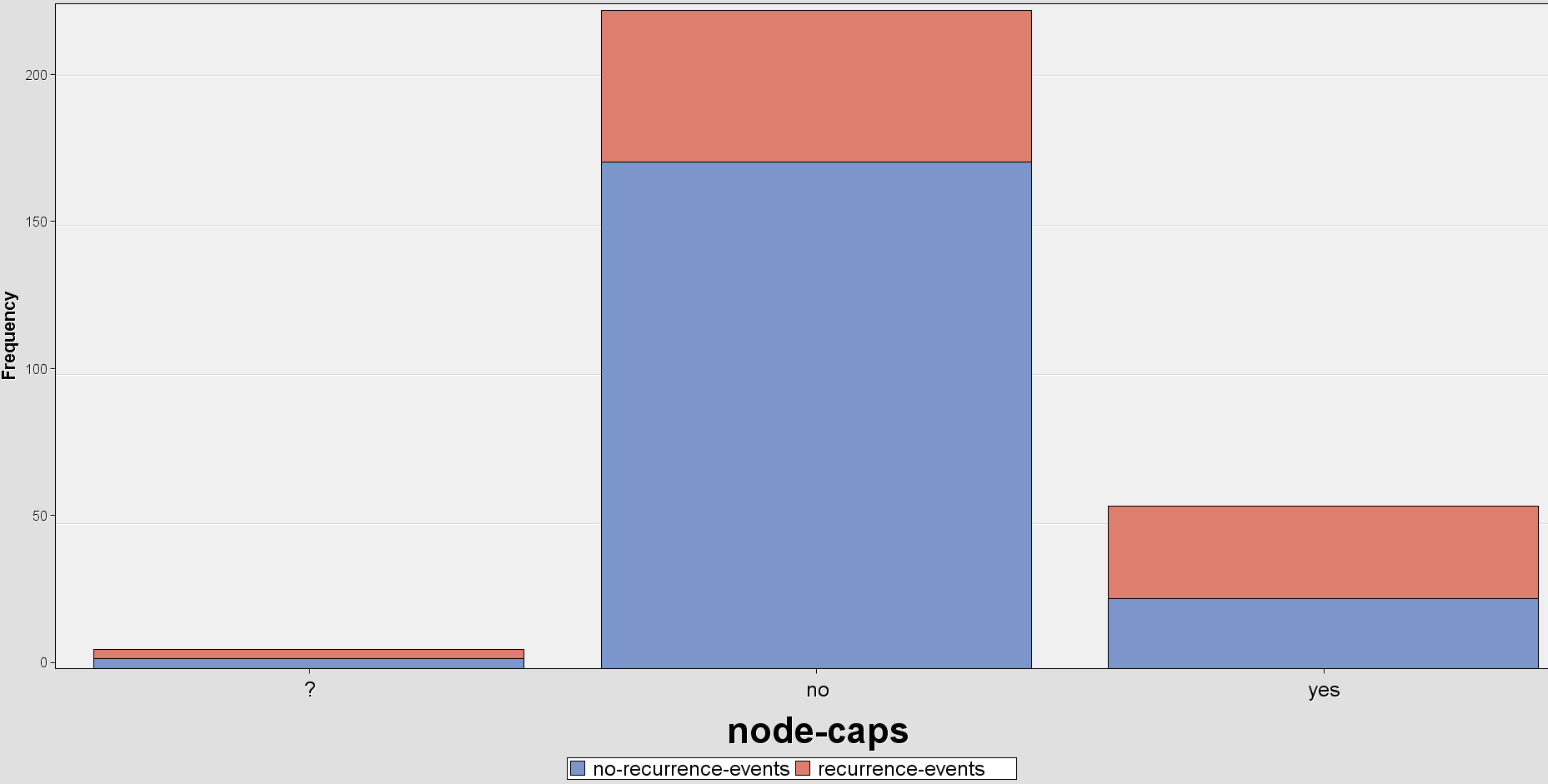
In addition, we also noticed that breast cancer is more commonly happen in the age range of 45-65 based on the figure as shown below.

**Histogram of age\_max**

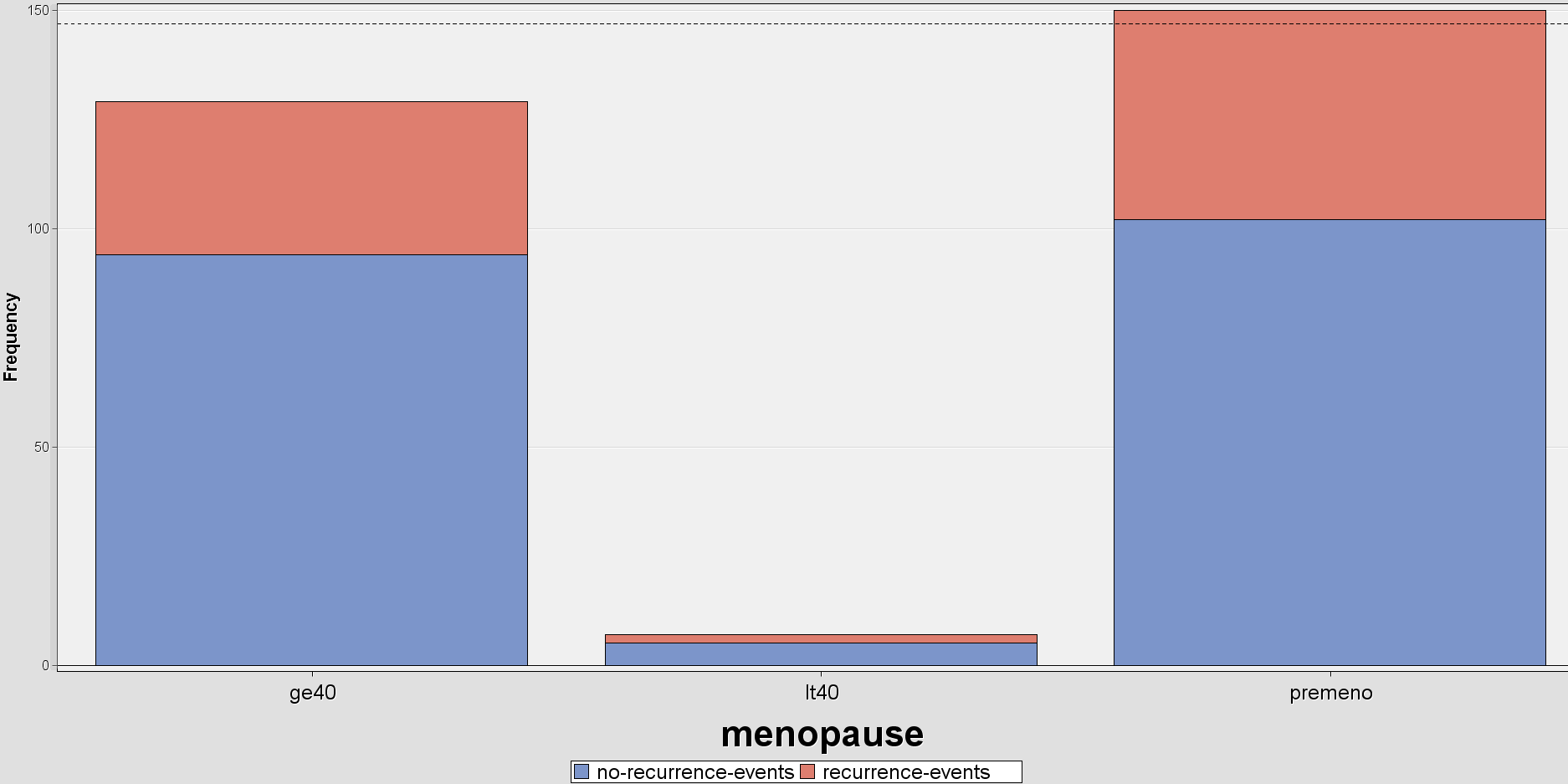


Furthermore, 78% of the breast cancer patients does not have node-caps and 98% of them are either greater than or equal to 40 or at premeno stage.

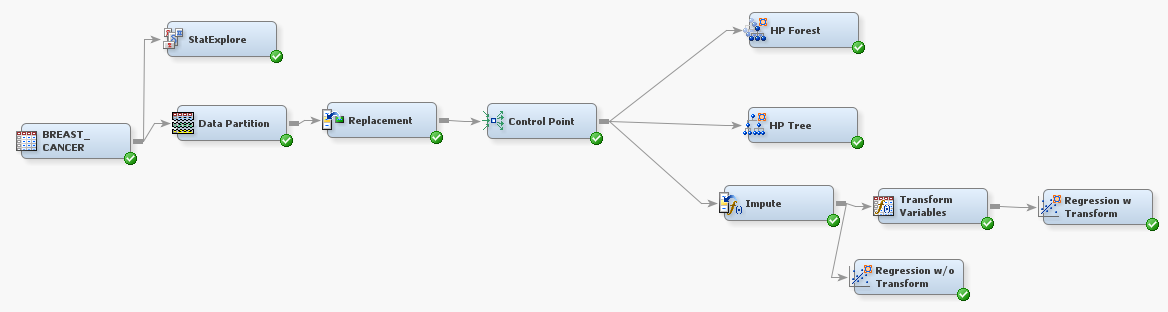
**Bar plot of node-caps**



**Bar plot of menopause**



# Predictive Model Development

As for predictive model development, we developed a few different models using the process as shown in the diagram below. Firstly, the data is split into 70% training and 30% validation. Then we replaced the “?” data with “\_UNKNOWN\_” in the replacement step. After that the data is used to develop three different models. The algorithms we chose are Random Forest, Decision Tree and Logistic Regressions.

**Random Forest**

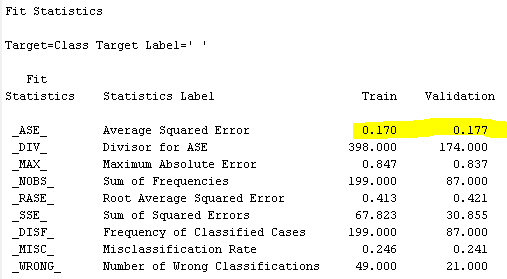
A maximum of 20 trees was used in the training of the random forest after discovering that after 20 trees, the misclassification did not improve any further

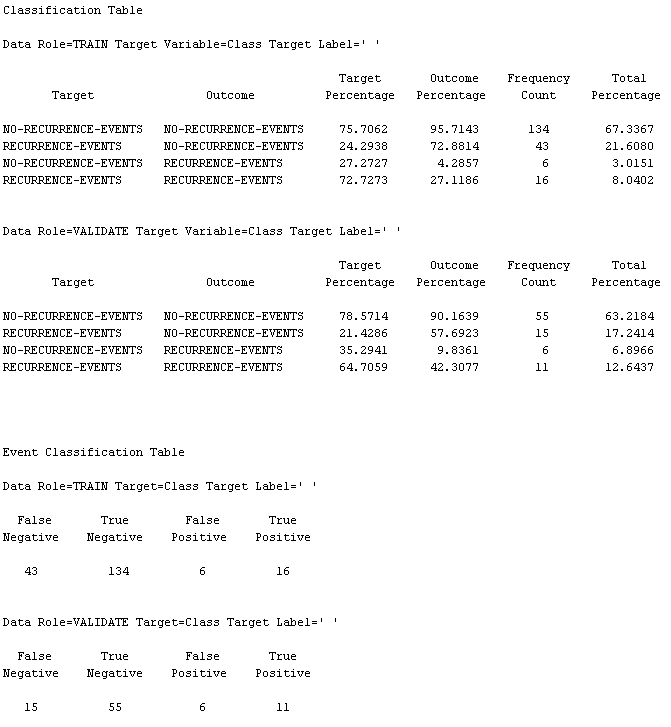
**Misclassification did not improve further after 20 trees**



**Fit Performance of Random Forest**

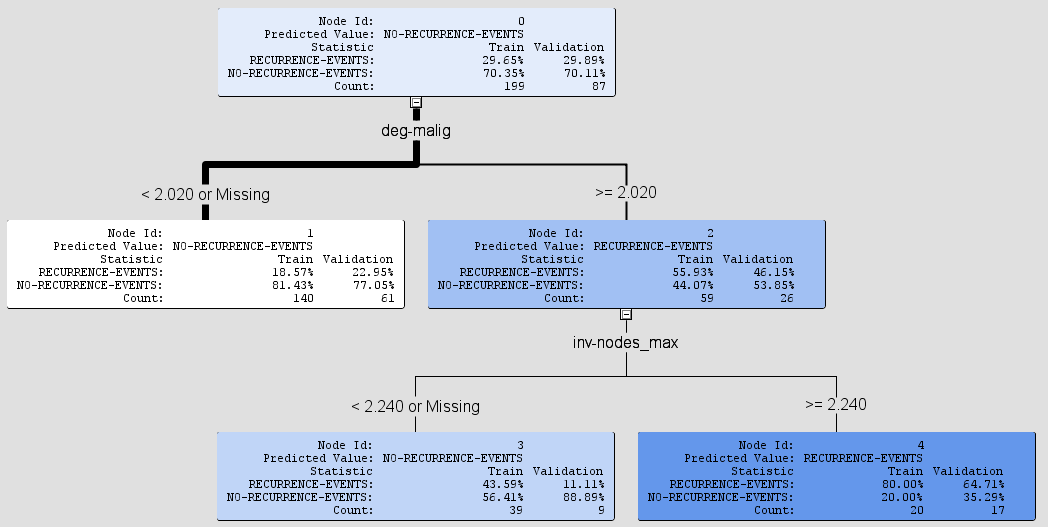
We can see that the difference between the training and validation Average Squared Error, hence, we can conclude that there is no overfitting or underfitting.





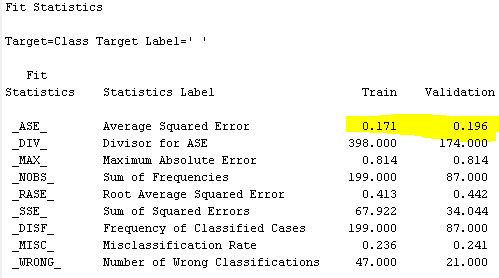
**Decision Tree**

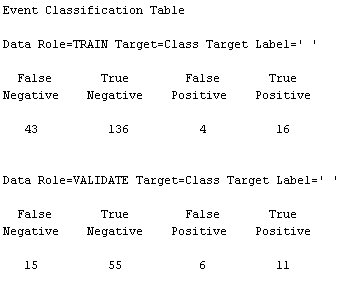
10-fold cross validation is used for this decision tree algorithm. Besides hyperparameters such as maximum depth and leaf size are set to 10 and 5 respectively.



**Fit Performance of Decision Tree**

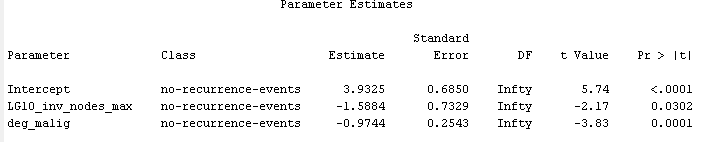
We can see that the difference between the training and validation Average Squared Error, hence, we can conclude that there is no overfitting or underfitting.



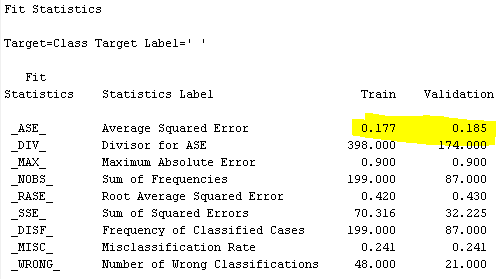


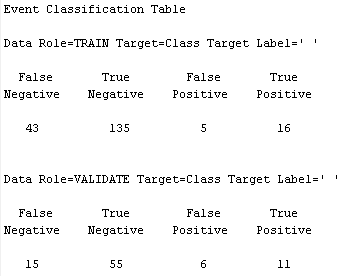
**Logistic Regression**

For logistic regression, additional imputation and transformation steps are needed before fitting the model. It is worth noting that the performance of the logistic regression improved significantly after going through the log transformation of the inv\_nodes\_max which was found to be skewed in the EDA previously. The Average Squared Error improved from 0.2 to 0.185



**Fit Statistics of Logistic Regression with Transformation**





# Model Comparison

The evaluation metric we used to compare the model is Averaged Squared Error. The following table summarizes the performance of all the models developed in this project. You can see that Random Forest has the best performance among all three models, likely due to its ensemble effect of using multiple trees. We can also see a huge improvement by using transformation on skewed variables for Logistic Regression. Hence, we would highly recommend using the Random Forest to develop a predictive model for this problem.

|  |  |
| --- | --- |
| **Model** | **Averaged Squared Error (Validation)** |
| Random Forest | 0.177 |
| Decision Tree | 0.196 |
| Logistic Regression without Transformation | 0.2 |
| Logistic Regression with Transformation | 0.185 |

# Conclusion

Through EDA, we found that variable inv\_nodes\_max is skewed and transformation is applied to it when developing the Logistic Regression model. In addition, deg-malig has a very clear relationship with breast cancer recurrence and this is proven by all models using it as the most important feature in the later stage. Some other interesting findings include breast cancer commonly happens in the left quadrant of the breast, and among women of age 45-65.

We have developed predictive models for classification with three different algorithms and compared their performance using Average Squared Error (ASE) as the evaluation metrics for this project. Random Forest has the best ASE of 0.177 probably due to the ensemble effect of using multiple decision trees in its decision making. Besides, we have also proven that logistic regression, though, is a simple algorithm, if optimized properly, for example, applying a transformation to skewed variables, can have comparable performance (ASE of 0.185) with complex algorithms like Random Forest.

# 

# Declaration

I, Chin Hooi Yap (mention your name), declare that the attached assignment is my own work in accordance with the Seneca Academic Policy. I have not copied any part of this assignment, manually or electronically, from any other source including websites, unless specified as references. I have not distributed my work to other students.

# 

# References

Mackenzie Rivero, Alexander, et al. “Machine Learning for the Evolutionary Analysis of Breast Cancer.” *Journal of Science and Research: Revista Ciencia e Investigación*, vol. 3, no. CITT2017, 2018, pp. 44–49., https://doi.org/10.26910/issn.2528-8083vol3isscitt2017.2018pp44-49.

1. https://dialnet.unirioja.es/descarga/articulo/7349569.pdf [↑](#footnote-ref-1)