BAN210- Final Assessment

## Instructions

The project is posted on BB on April 11st with the deadline of April 15th at midnight. The late penalty will be 10% per day.

Unlike the workshops, the instructions in the final assessment are minimal (this is an assessment!). Your task is to design and implement necessary components of Predictive Modeling data pipeline to get the job done. When needed, refer to the workshops for relevant instructions and ideas, or research methods in the SAS Enterprise Miner book or online.

Also, unlike the workshops, your task is not to simply paste results from Enterprise Miner. You need to **EXPLAIN** what the results indicate, why they are important or nontrivial, and how they can be used as a PA-based insight.

Note: This is not a group work task.

## Overview

In this assessment, you will be implementing, testing, and documenting Predictive Analytics models for **one** of following data sets:

**Dataset 1:** Breast Cancer Data Set ([**Download link**](https://archive.ics.uci.edu/ml/datasets/Breast%2BCancer)). This is one of three domains provided by the Oncology Institute that has repeatedly appeared in the machine learning literature. This data set includes 201 instances of one class and 85 instances of another class. The instances are described by 9 attributes, some of which are numeric, and some are nominal.

**Attribute Information:**

1. Class: no-recurrence-events, recurrence-events

2. age: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99.

3. menopause: lt40, ge40, premeno.

4. tumor-size: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54,

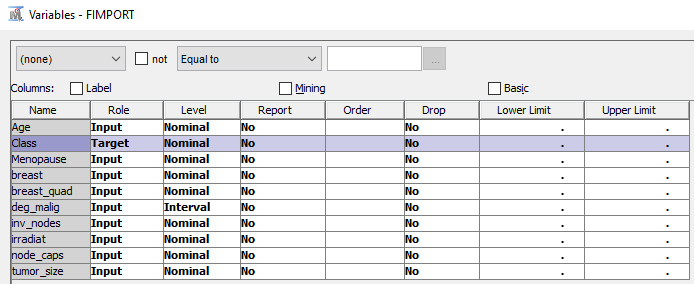
55-59.

5. inv-nodes: 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.

1. node-caps: yes, no.
2. deg-malig: 1, 2, 3.
3. breast: left, right.
4. breast-quad: left-up, left-low, right-up, right-low, central.
5. irradiat: yes, no.

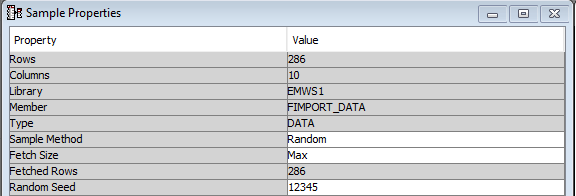
**SELECTING TARGET VARIABLE**

Choose *Class* as the **Target**.****

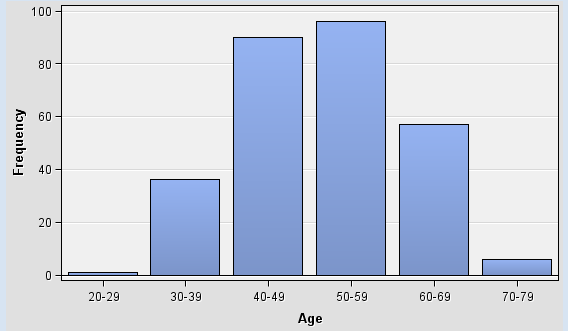
* frequency plot of Class

****

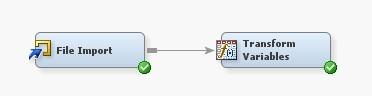
* There are 285 rows and 10 Columns imported

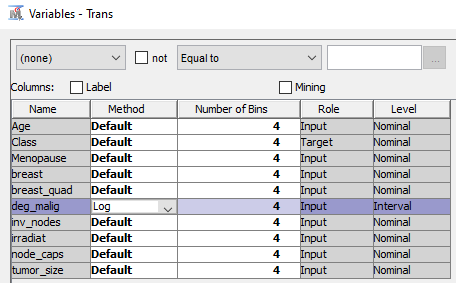


* Age is suitable for linear regression as it shows a U-shaped curve.

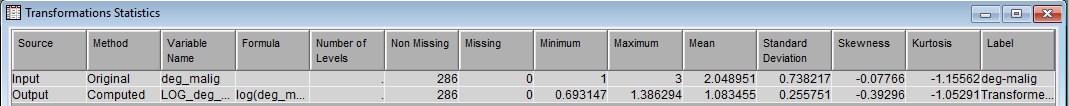


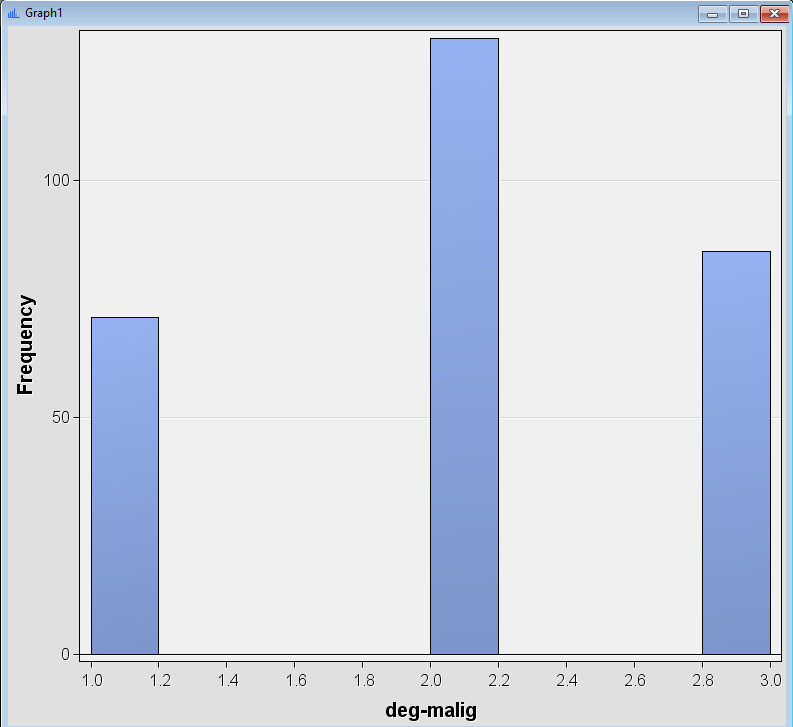
In **Transform** node, choose breast\_quad variable since all of the variables are Nominal.



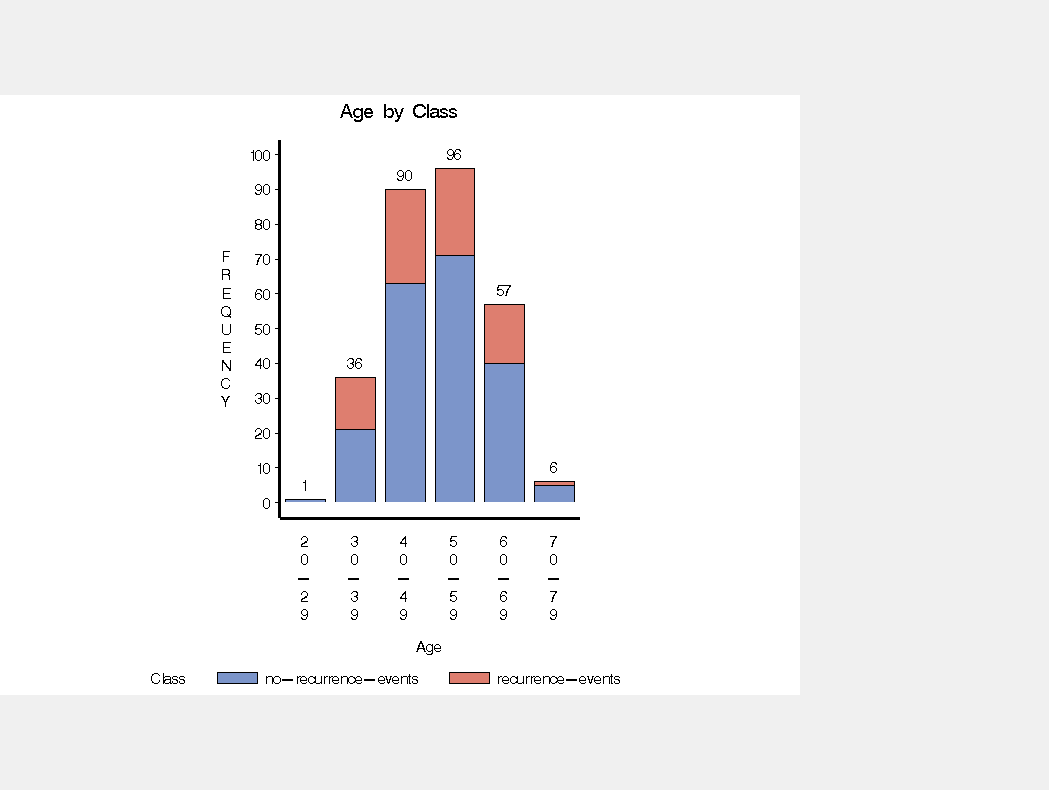


* The formula selected is log(deg\_malig +1)

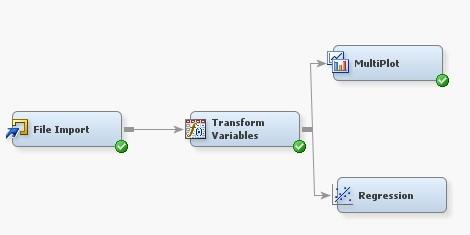




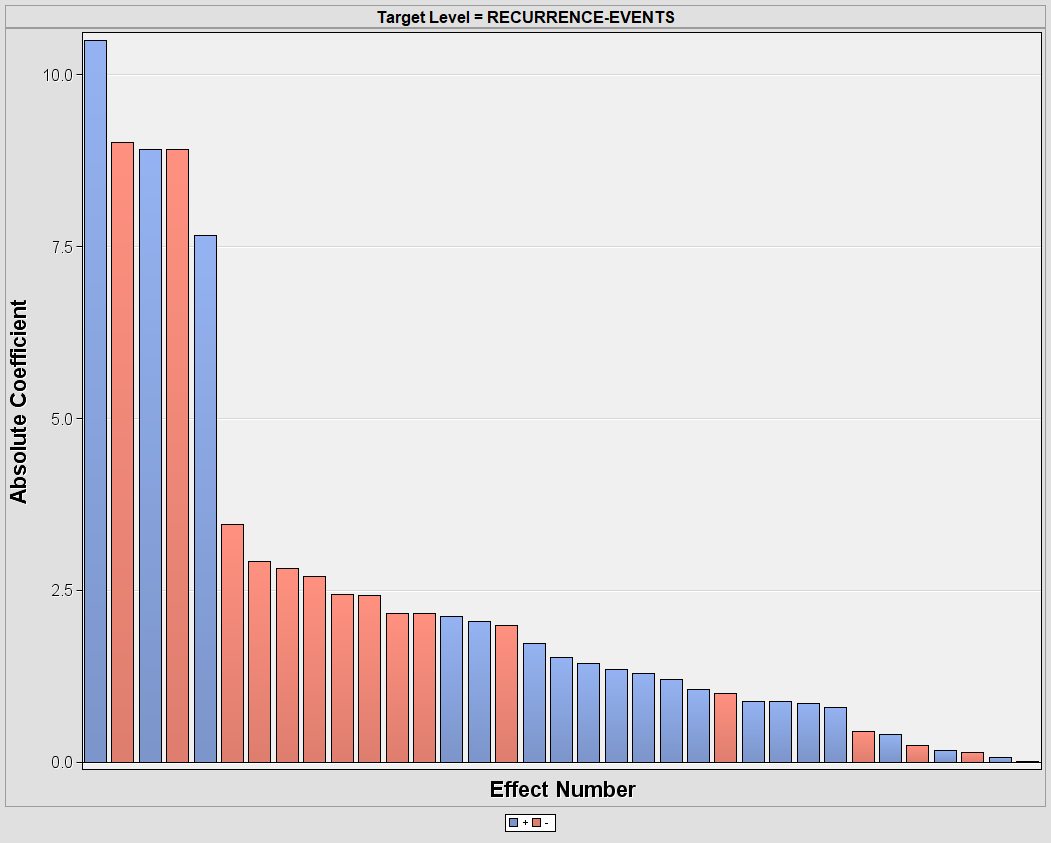
In the **MultiPlot** node, choose **BAR CHART** plot as the **Type of Charts**.



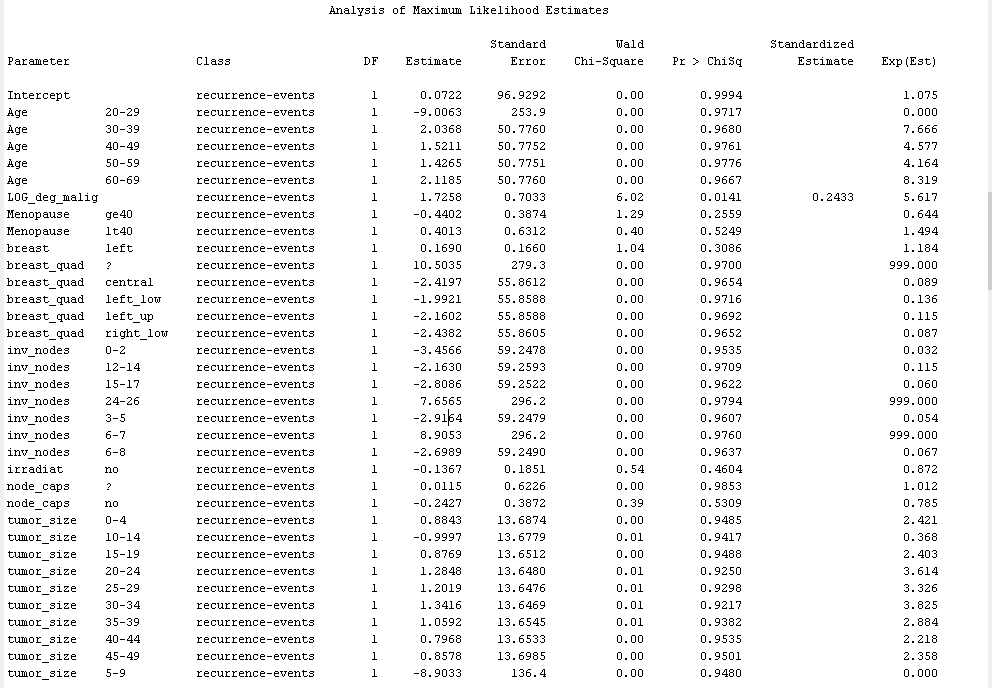
**REGRESSION MODEL**



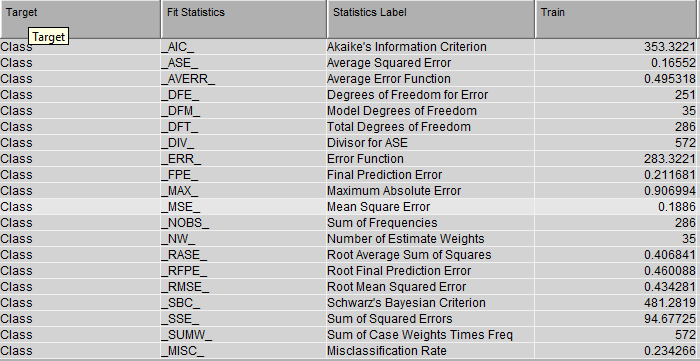
* no-recurrence-events have the highest number of coefficients



* Based on this model, the predicted standardized for \_deg\_malig is 0.2433 most of the Class target has recurrence-events.



Mean Square Error (MSE) for the target *Class* is 0.1886



## Deliverables

1. **Final report**. MS Word document for your report. Add this declaration to your file:

I, Krystel Venice Lanzaderas, 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 web sites, unless specified as references. I have not distributed my work to other students.

1. SAS Enterprise Miner **project** files. (Github link preferred)
2. Prepare a presentation and explain your findings in **less than 12 slides.** The talk should not take longer than **10 minutes**.

Note: Upload report file (Word or PDF), SAS EM project files (Zipped) and presentation to Blackboard. **Do not zip all three deliverables together**.

Evaluation Rubric

# Exploratory Data Analysis [30%]

Apply Exploratory Data Analysis (EDA) to discover patterns, spot anomalies, test hypothesis and to check assumptions with the help of summary statistics and graphical representations. Explain your understanding of the data present insights using appropriate graphs. Elaborate on the requirement of any transformations and feature engineering needed for your project.

# Predictive models [30%]

* Train **at least two models** using appropriate algorithms (i.e. decision tree, logistic regression, linear regression, neural network, knn).
* Apply **k-Fold cross validation** to ensure generalizability of your models.
* Assess, analyze, and compare the performance of your models

# Report [20%]

Your report (less than 10 pages) should include the following:

1. The goal and summary of your work on the data set and your choice of the algorithms.
2. Explain results of EDA. What business decision should be made based on your data analysis.
3. Explain the training and validation methodology along with evaluation results. Use appropriate metrics for evaluation.

# Presentation [20%]

* Powerpoint slide deck (less than 12 slides and 10 minutes talk),
* Record your presentation in concise and clear manner. Explain key finding from EDA, predictive modeling and evaluation methodology and key results.
* The outlines and conclusions in the ppt must be clear. Use your findings in graph and short note format and the speech should lead audience to conclusion.

I will be available to answer questions in Microsoft Teams and via email. If you have technical issues, please contact the ITS service desk at [servicedesk@senecacollege.ca](mailto:servicedesk@senecacollege.ca) , or login to our Technician’s Microsoft Teams site (available 08:00 AM EST – 05:00 PM EST).

[**Academic Integrity Policy**](http://www.senecacollege.ca/about/policies/academic-integrity-policy.html). Seneca upholds a learning community that values academic integrity, honesty, fairness, trust, respect, responsibility and courage. These values enhance Seneca’s commitment to students by delivering high-quality education and teaching excellence, while supporting a positive learning environment. The AI policy is always in effect. Note **Sections** **2.3 and 2.4:**

***“…2.3*** *Should there be a suspected violation of this policy (e.g.…cheating, falsification, impersonation or plagiarism), the academic integrity sanctions will be applied according to the severity of the offence committed. Refer to* [*Appendix B*](http://www.senecacollege.ca/about/policies/academic-integrity-policy.html) *for the academic integrity sanctions.* ***2.4*** *Should a suspected violation of this policy be a result of, or in combination with, a suspected violation of Seneca’s Student Code of Conduct and/or another non-academic-related Seneca policy, the matter will be investigated and adjudicated through the process found in the Student Code of Conduct.”*

|  |
| --- |
| ***TO BE COMPLETED BY STUDENT*** |
| *SUBJECT SECTION NUMBER (e.g. QNM223 AA): BAN ZZB 2022* |
| *STUDENT NAME: Krystel Venice Lanzaderas* |
| *STUDENT NUMBER: 106889215* |
| *STUDENT SIGNATURE:* |