A grey rectangular sign with white text

Description automatically generated

**BA-706 (001)**

**Applied Analytic Modelling**

**SAS Project Fortification**

**Predictive Modelling on student performance**

**July 15, 2024**

**Course Professor: David Parent**

**Campus: Progress**

|  |  |
| --- | --- |
| Done by: | |
| * Alester Joshua D Costa | 301321745 |
| * Devansh Soni | 301239463 |
| * Khawja Nayeem Ather | 301187396 |
| * Miguel Polanco Sanchez | 301299027 |

**Table of Contents**

[I. Introduction 4](#_Toc171967828)

[II. Data Exploration 5](#_Toc171967829)

[III. Decision Tree 10](#_Toc171967830)

[i. Maximal Decision Tree 10](#_Toc171967831)

[ii. Misclassification Decision Tree 11](#_Toc171967832)

[iii. Average Square Error (ASE) Decision Tree 12](#_Toc171967833)

[iv. Analysis and Interpretation of Decision Tree 13](#_Toc171967834)

[IV. Impute 15](#_Toc171967835)

[V. Handling Extreme Values 16](#_Toc171967836)

[VI. Handling Skewed Values 17](#_Toc171967837)

[VII. Accounting for Curse of Dimensionality 18](#_Toc171967838)

[VIII. Regression 19](#_Toc171967839)

[i. Full Regression 19](#_Toc171967840)

[ii. Forward Regression 20](#_Toc171967841)

[iii. Backward Regression 22](#_Toc171967842)

[iv. Stepwise Regression 24](#_Toc171967843)

[v. Analysis and Interpretation of Regression 25](#_Toc171967844)

[IX. Neural Network 27](#_Toc171967845)

[i. Impute Neural Network 27](#_Toc171967846)

[ii. Cap and Floor Neural Network 28](#_Toc171967847)

[iii. Transform Neural Network 29](#_Toc171967848)

[iv. Stepwise Regression Neural Network 30](#_Toc171967849)

[v. Forward Regression Neural Network 31](#_Toc171967850)

[vi. Backward Regression Neural Network 32](#_Toc171967851)

[vii. Full Regression Neural Network 33](#_Toc171967852)

[viii. 1H Forward Regression Neural Network 35](#_Toc171967853)

[ix. 2H Forward Regression Neural Network 36](#_Toc171967854)

[x. 4H Forward Regression Neural Network 37](#_Toc171967855)

[xi. 5H Forward Regression Neural Network 38](#_Toc171967856)

[xii. 6H Forward Regression Neural Network 39](#_Toc171967857)

[xiii. 7H Forward Regression Neural Network 40](#_Toc171967858)

[xiv. 8H Forward Regression Network 41](#_Toc171967859)

[xv. Analysis of Neural Network 42](#_Toc171967860)

[X. Model Comparison 43](#_Toc171967861)

[i. Analysis of Model Comparison 44](#_Toc171967862)

[a. Interpretation of Backward Regression Model 44](#_Toc171967863)

[ii. Changing Certain Data Types to Compare Model Performance 45](#_Toc171967864)

[XI. Conclusion and Recommendations 46](#_Toc171967865)

[XII. References 47](#_Toc171967866)

# Introduction

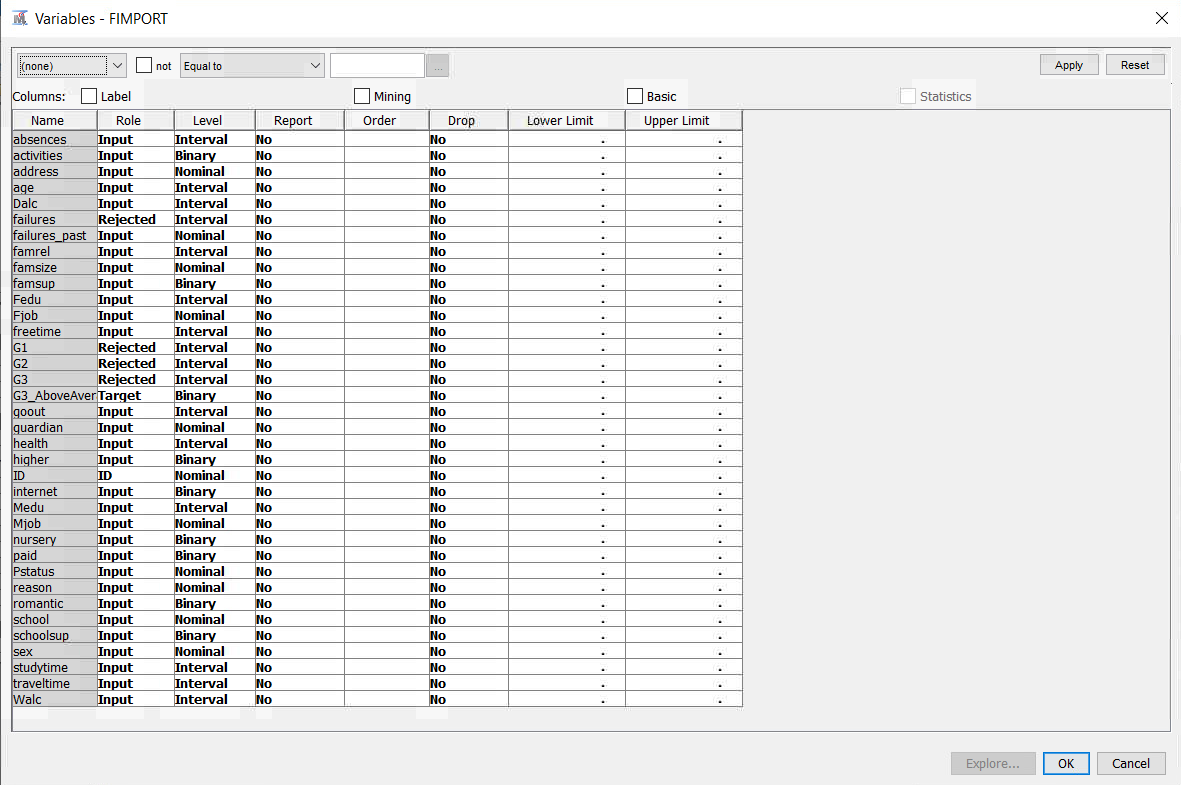
In any educational setting, academic performance refers to the scale on which students have obtained their short-term or long-term educational objectives (Tadese, 2022, p. 2). There may be numerous determinants that could affect the academic performance of students; these components comprise of learning capabilities, financial circumstances, attitudes, family and more (Abu Bakar, 2023, p. 1498). According to Regier (2011), students are more likely to make the transition into adulthood and attain professional and lucrative success if they excel in their academics at school.

The main purpose of this report is to understand the student’s academic performance at two secondary schools in Portugal, namely Gabriel Pereira and Mousinho Da Silveria, through building predictive models on SAS Enterprise Miner. By utilizing predictive models such as decision trees, regressions, and neural networks, an analysis is carried out to predict if these students achieve an above-average grade in their final period of exams in the Portuguese subject. This report's main assessment measure utilized to compare the best predictive model is the **Receiver Operating Characteristic (ROC) index**. Additionally, the Average Squared Error (ASE) is also used to aid in the assessment. The data present within the chosen dataset of student performance in the Portuguese subject has been collected through school reports and questionnaires, and the numerous variables present in the student performance dataset consist of student grades marked with the Portuguese grading system, demographics, social and school-affiliated characteristics (Chauhan, 2022).

# Data Exploration

To carry out the process of predicting whether secondary school students would score an above-average grade in their final period grade of the Portuguese subject, a total of 35 variables are present in the student performance dataset. Out of these 35 variables, the *G3\_AboveAverage* variable represents a target variable whose binary values are modeled and predicted by the other 29 variables. Furthermore, there are four variables that are excluded from the predictive modeling process because they have a very strong correlation with the target variable. In essence, the variables *G1*, *G2* and *G3* have been excluded since each variable would not have been helpful to develop a predictive model with these variables present. Additionally, the *failures* variable is excluded because it only consists of a range of numbers from 0 to 4, representing the count of prior student failures in their academics. To mitigate this concern, the *failures\_past* variable has been created with a binary measurement level, i.e., values of 0 and 1, indicating whether students have had previous failures. While the following table describes each of the variables present within the dataset of student performance in the Portuguese subject, the later screenshot highlights the necessary amendments made to the dataset variables of student performance in the Portuguese subject on SAS Enterprise Miner.

| **Variable** | **Model Role** | **Measurement Level** | **Description** |
| --- | --- | --- | --- |
| *absences* | Input | Interval | Number of student’s absences from school  (From 0 to 93) |
| *activities* | Input | Binary | Participation of student in extra-curricular activities  (Either yes or no) |
| *address* | Input | Nominal | Student’s home address type  (Either ‘U’: Urban or ‘R’: Rural) |
| *age* | Input | Interval | Student’s age  (From 15 to 22) |
| *Dalc* | Input | Interval | Workday consumption of alcohol  (From 1: very low to 5: very high) |
| *failures* | Rejected | Interval | Number of previous failures by students at class  (From 0 to 4) |
| *failures\_past* | Input | Nominal | Previous failures encountered by students based on the *failures* variable  (Either yes or no) |
| *famrel* | Input | Interval | Quality of student’s relationship with their family  (From 1: very bad to 5: excellent |
| *famsize* | Input | Nominal | Size of student’s family  (Either ‘LE3’: Less than or equal to 3 or ‘GT3’: Greater than 3) |
| *famsup* | Input | Binary | Family educational support  (Either yes or no) |
| *Fedu* | Input | Interval | Educational level of student’s father  (Either 0 : None or  1 : Primary education,  2 : 5th to 9th,  3 : Secondary education or  4 : Higher education) |
| *Fjob* | Input | Nominal | Job of student’s father  (Either 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other') |
| *freetime* | Input | Interval | Free time after school  (From 1 - very low to 5 - very high) |
| *G1* | Rejected | Interval | First period grade  (From 0 to 20) |
| *G2* | Rejected | Interval | Second period grade  (From 0 to 20) |
| *G3* | Rejected | Interval | Third period grade  (From 0 to 20) |
| *G3\_AboveAverage* | Target | Binary | Above average grade attained by students based on the *G3* variable  (Either yes or no) |
| *goout* | Input | Interval | Going out with friends  (From 1: very low to 5: very high) |
| *guardian* | Input | Nominal | Student’s guardian  (Either ‘mother’, ‘father’ or ‘other’) |
| *health* | Input | Interval | Current health status  (From 1: very bad to 5: very good) |
| *ID* | ID | Nominal | Unique identifier for each student |
| *internet* | Input | Binary | Internet access available for students at home  (Either yes or no) |
| *Medu* | Input | Interval | Educational level of student’s mother  (Either 0: none,  1: primary education (4th Grade),  2: (5th to 9th Grade),  3: Secondary Education or  4: higher Education) |
| *Mjob* | Input | Nominal | Job of student’s mother  (Either ‘teacher’, ‘health’ care related, civil ‘services’, ‘at\_home’ or ‘other’) |
| *nurserv* | Input | Binary | Nursery attended by student  (Either yes or no) |
| *paid* | Input | Binary | Extra paid classes taken for the Portuguese subject  (Either yes or no) |
| *Pstatus* | Input | Nominal | Cohabitation status of student’s parents  (Either ‘T’: living Together or  ‘A’: Apart) |
| *reason* | Input | Nominal | Reason for selecting the school  (Either close to ‘home’, school ‘reputation’, ‘course’ preference or ‘other’) |
| *romantic* | Input | Binary | Student in a romantic relationship (Either yes or no) |
| *school* | Input | Nominal | Student’s school  (Either ‘GP’: Gabriel Pereira or  ‘MS’: Mousinho da Silveira) |
| *schoolsup* | Input | Binary | Extra educational support  (Either yes or no) |
| *sex* | Input | Nominal | Student’s gender  (Either ‘F’: Female, or ‘M’: Male) |
| *studytime* | Input | Interval | Student’s weekly study time  (1: Less than 2 hours,  2: Between 2 to 5 hours,  3: Between 5 to 10 hours, or  4: More than 10 hours) |
| *traveltime* | Input | Interval | Home to school travel time  (1: Less than 15 minutes,  2: Between 15 to 30 minutes,  3: Between 30 minutes to 1 hour, or  4: More than 1 hour) |
| *Walc* | Input | Interval | Weekend consumption of alcohol  (1: very low to 5: very high) |



# Decision Tree

After carrying out the process of choosing the Portuguese schools’ dataset and configuring the relevant roles and levels of each variable, three decision tree models have been developed on SAS Enterprise Miner: the maximal Tree, the misclassification Tree, and the average square error (ASE) tree. In each developed decision tree, the thickest black line represents the optimal path and the dark blue leaves depict the most optimal leaves.

## Maximal Decision Tree

A diagram of a company

Description automatically generatedThe first decision tree that has been built for predicting whether students score an above-average grade in their final period exam (*G3*) in the Portuguese subject is a maximal tree. Upon running the maximal tree, it has been found that the ASE value of the model for predicting the average grade of students stands at **0.194757**.

A screenshot of a computer

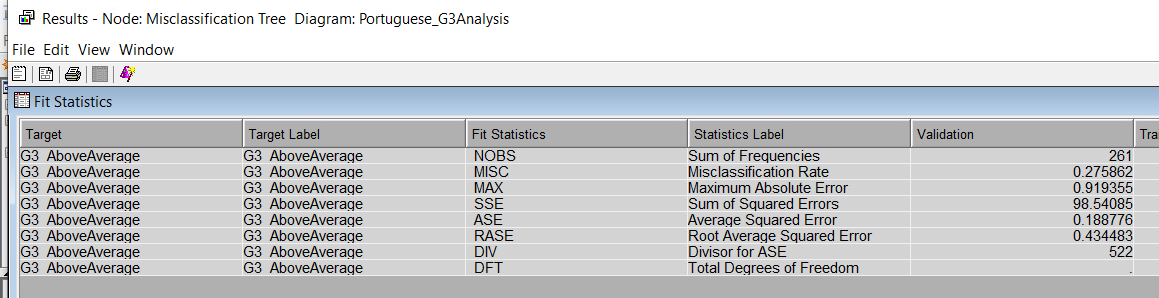
Description automatically generated

## Misclassification Decision Tree

A misclassification tree is a type of decision tree specifically designed for categorical target variables. It uses the misclassification rate as the splitting criterion to build the tree, aiming to minimize the overall misclassification of observation.

A diagram of a flowchart

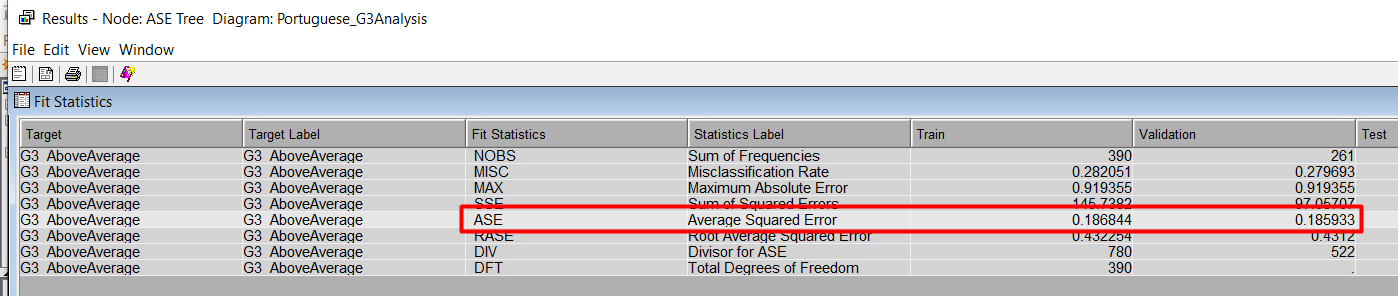
Description automatically generatedUpon running the built misclassification tree, it has been discovered that the ASE value stands at **0.188776**.



## Average Square Error (ASE) Decision Tree

A diagram of a missing person

Description automatically generated with medium confidenceThe third and final decision tree of the ASE tree has been developed. After running the created ASE tree model, the ASE value of the model is **0.185933.**



## Analysis and Interpretation of Decision Tree

Before carrying out the analysis, there are certain terminologies that need to be addressed, considering the interpretation of decision trees. Firstly, **root nodes** are the starting point of the decision tree and represent the very first question asked when trying to make a decision. Secondly, **leaf nodes** are the endpoints of decision trees and depict the final decision or outcome after all the questions have been asked. Lastly, **interior nodes** are the points along the way where more questions are asked after the initial question at the root node. Each interior node represents a decision point where the tree splits into branches based on the answer to a question.

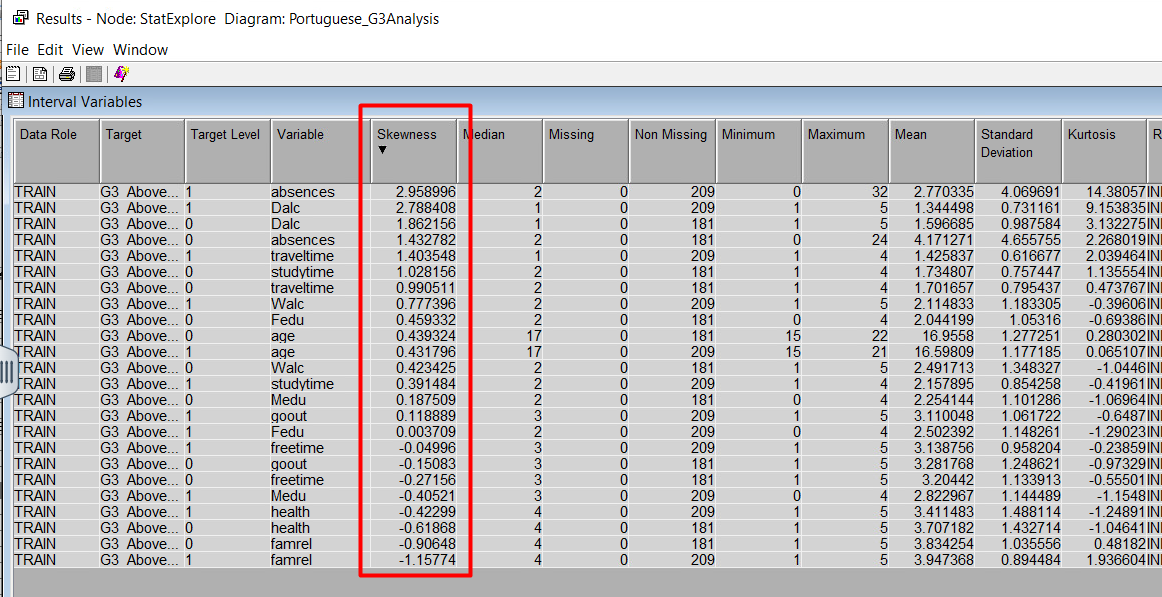
Based on the generated ASE figures that have been generated for each created decision tree model, consisting of the maximal tree, misclassification tree and ASE tree, the most optimal decision tree is the ASE tree. By having the lowest ASE figure of **0.185933** among all the decision tree models, the ASE tree represents the best model in comparison to the misclassification tree and maximal tree, with ASE figures of **0.188776** and **0.194757**, respectively.

As part of the developed ASE decision tree model, the root node that has been generated for identifying the likelihood of students scoring or not scoring an above average grade, i.e., *G3*, in the Portuguese subject is the previous failures encountered by students in their previous academic years, i.e., the *failures\_past* variable. Out of the possible aspects, the most optimal path of the ASE tree highlights that students are more likely to achieve an above-average grade without having any past failures. If students have not had prior failures in their previous academic periods, there is a 61.88% probability that they will score an above-average grade in the final period of the Portuguese subject. In contrast, students who have encountered failures in their previous academic years have a 94.74% chance of not obtaining an above-average grade in their final grade, which represents the highest probability value among all leaves present in the ASE tree model.

The interior nodes of the ASE tree model indicate that if a student is pursuing their current education at the school of Gabriel Pereira (GP) and has an inclination to seek higher education, i.e., *higher*, upon the completion of their secondary education, there is a 69.48% and 72.11% probability, respectively, of obtaining an above-average grade in their final period, i.e., *G3* of the Portuguese subject. On the contrary, another leaf node indicates that students pursuing their secondary education at the school of Mousinho da Silveira (MS) have a lesser probability of attaining an above-average grade in their final period grade at 44.93%. Finally, another significant leaf of the model highlights that students who take fewer than 3.5 hours to travel to campus, i.e., *traveltime*, have a 73.43% probability of achieving an above-average grade in their final period of the Portuguese subject.

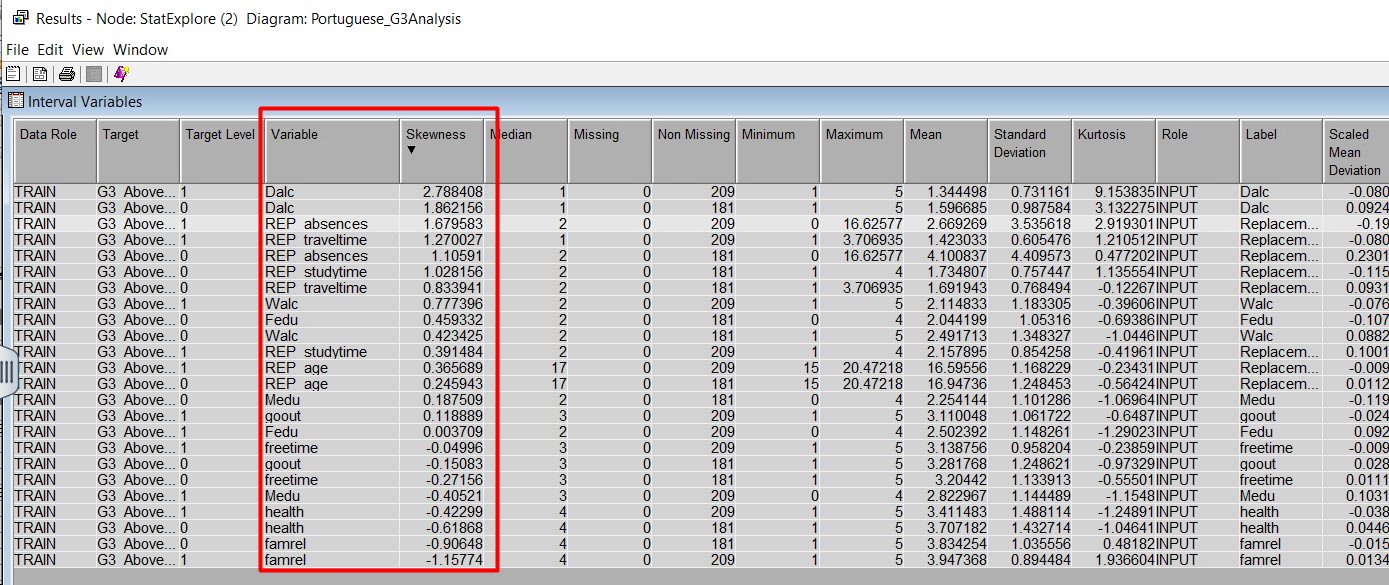
# Impute

In business data analysis, ensuring robust handling of missing information is essential for reliable decision-making. An Impute Node serves as a tool that fills in missing data points within a dataset, preparing for scenarios where data might be incomplete in the future, even if the current dataset is complete. This proactive approach ensures the analysis remains dependable over time.

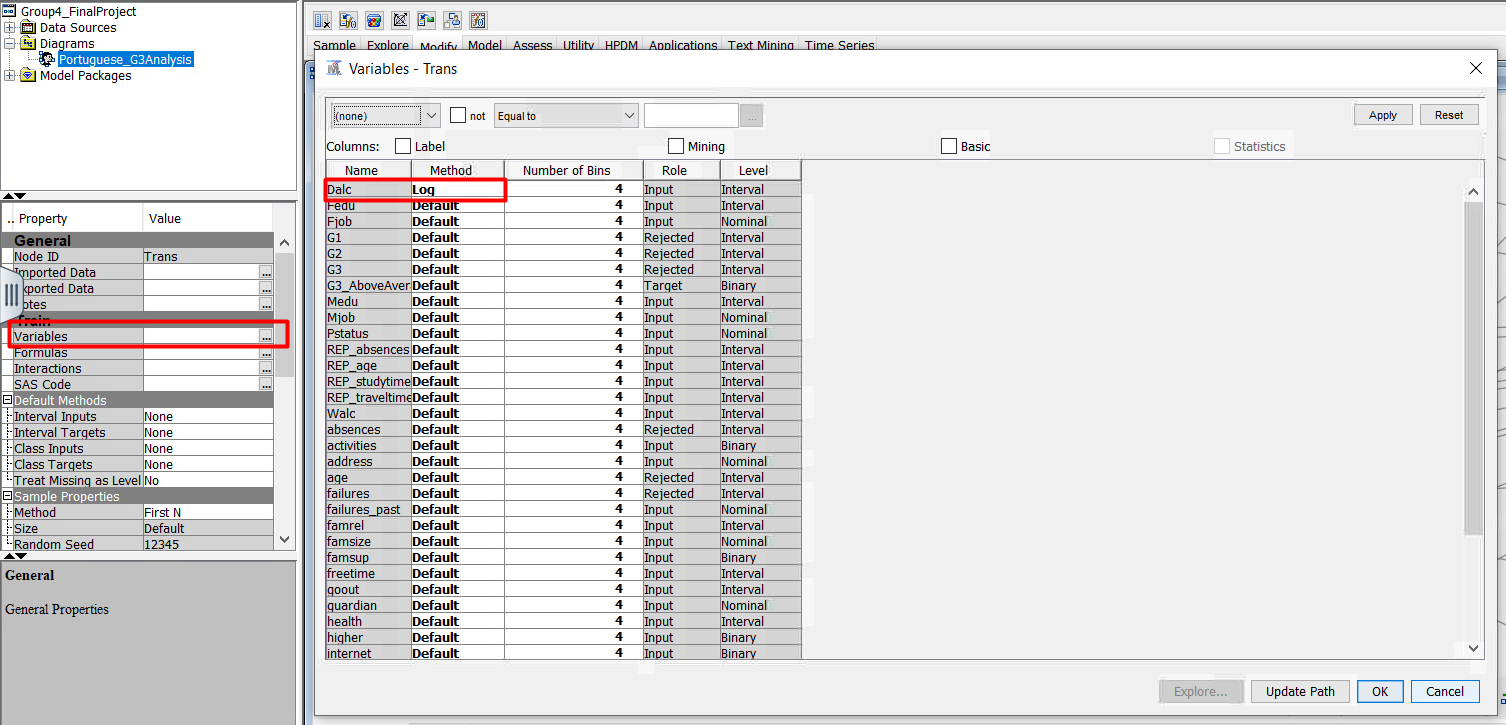
The Impute Node is integrated into the data analysis process, ensuring missing values do not disrupt decision-making. As part of this process, data characteristics such as skewness need to be identified using the StatExplorer Node tool. Skewness shows if the data is evenly distributed or if it leans heavily to one side. For example, variables like *absences* and *Dalc* may exhibit skewness levels above 2, indicating that most data points cluster around specific values, with outliers affecting the distribution. This approach enhances the reliability and effectiveness of the results obtained.

# Handling Extreme Values

After handling missing data, a method to limit the range of certain variables (*absences*, *traveltime*, *studytime*, and *age*) by setting maximum and minimum values was developed. This process helped to reduce the skewness of the *absence* variable. Initially, the *absences* variable had a high skewness of 2.96; however, after applying this range-limiting method, the skewness decreased to a more acceptable level of 1.68, improving the suitability of the data for the model.

On the other hand, the *Dalc* variable has not been affected by the Cap and Floor Node, which enables to account for extreme values, and the variable continues to represent a skewness above 2. For the next step, it is important to create a Transform Node to address the level of skewness for the *Dalc* variable.

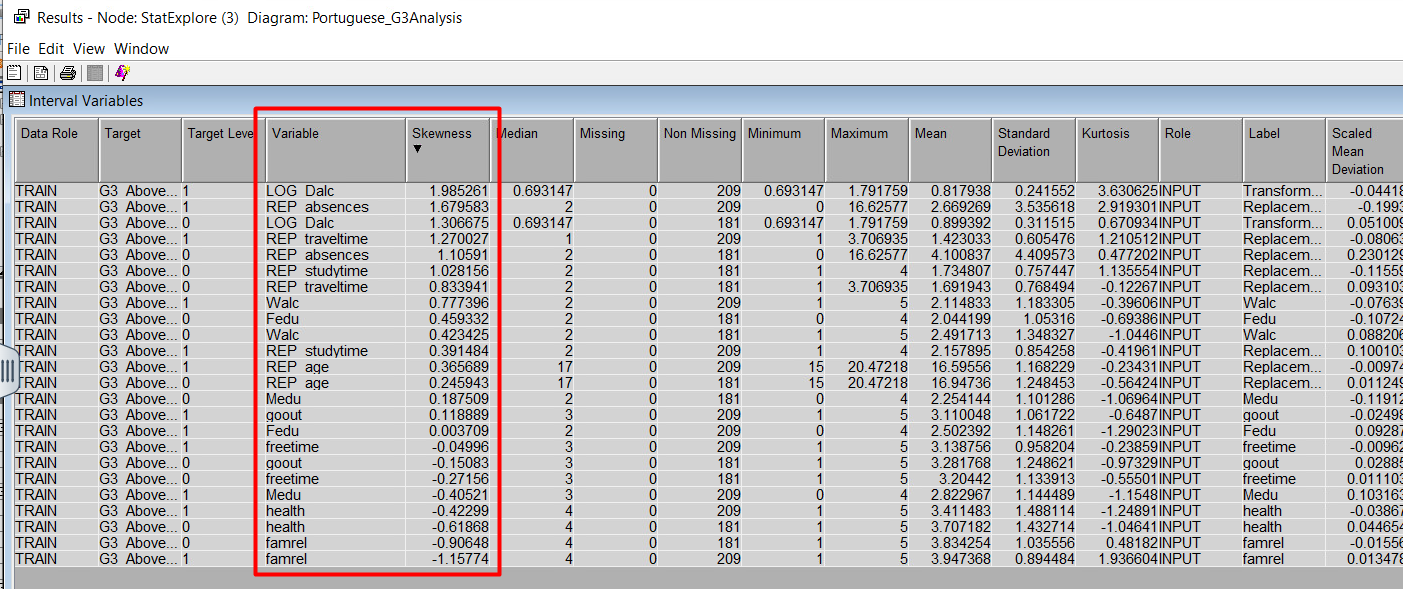
# Handling Skewed Values

Following statistical guidelines, the skewness of the *Dalc* (daily alcohol consumption) variable needed to be reduced. To do this, a logarithmic transformation was applied. By changing the transformation method of *Dalc* to a logarithmic scale, it successfully reduced its skewness from above 2 to 1.98. With all variables now having a skewness below 2, the next stage of developing regression models has been carried out.

# Accounting for Curse of Dimensionality

In managing data analysis, it is crucial to address the curse of dimensionality, which refers to the challenges posed by datasets with numerous variables that can complicate model interpretation and effectiveness. To manage this, a Replacement Node is generally employed. This tool combines different categories within variables to simplify models while maintaining interpretability.

However, for the current report, each variable in the dataset holds unique business significance that could lead to different strategic actions. Using a Replacement Node is unnecessary in this context because simplifying variables could overlook critical nuances. By maintaining the distinctness of each variable, the analysis captures comprehensive details, ensuring clearer insights and supporting strategic decision-making effectively.

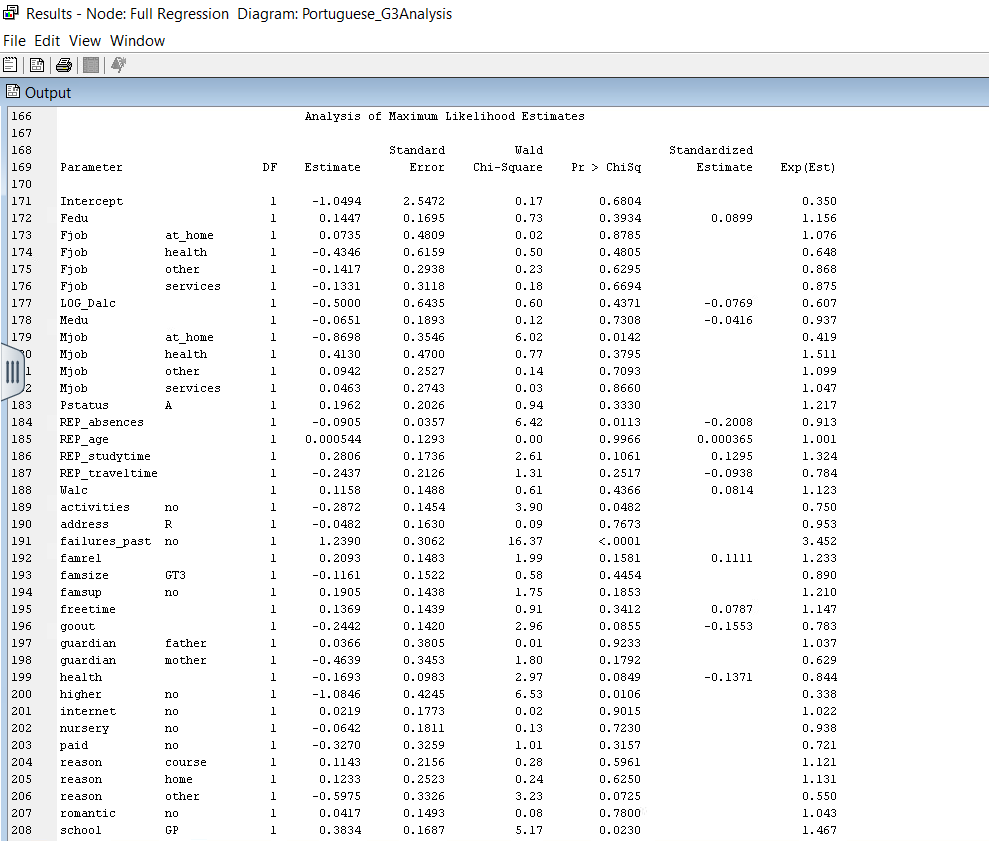
Please refer to the model comparison section (section X) of the report titled **“Changing Certain Data Types to Compare Model Performance”** on [page 45](#_Changing_Certain_Data), which outlines the specific adjustments made to data types and values to enable comparison with the original model outcome.

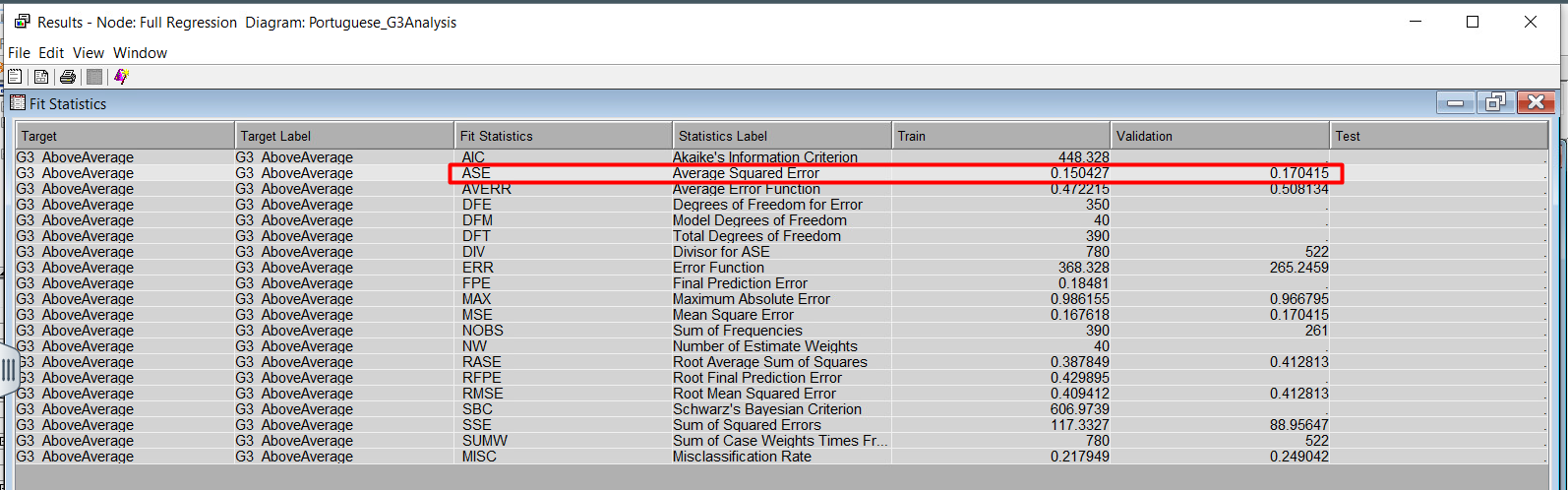
# Regression

The second type of predictive model that has been developed to comprehend if students score an above-average grade in their final period grade of the Portuguese subject is regression. From the Portuguese student performance dataset, the four regression models that have been constructed comprise of full regression, forward regression, backward regression and stepwise regression.

## Full Regression

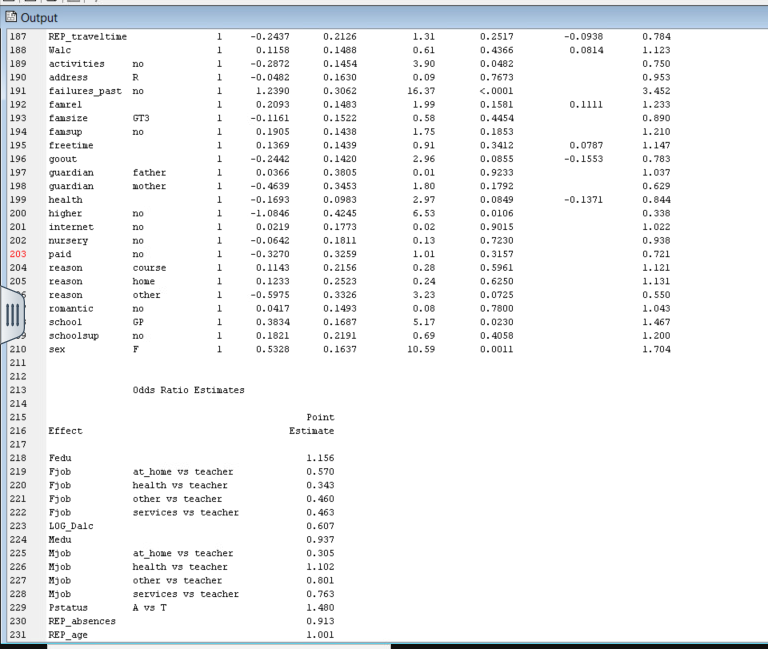
A full regression model helps in identifying which factors have a significant impact on the outcome and how they interact with each other, allowing for a comprehensive analysis of their relationships.

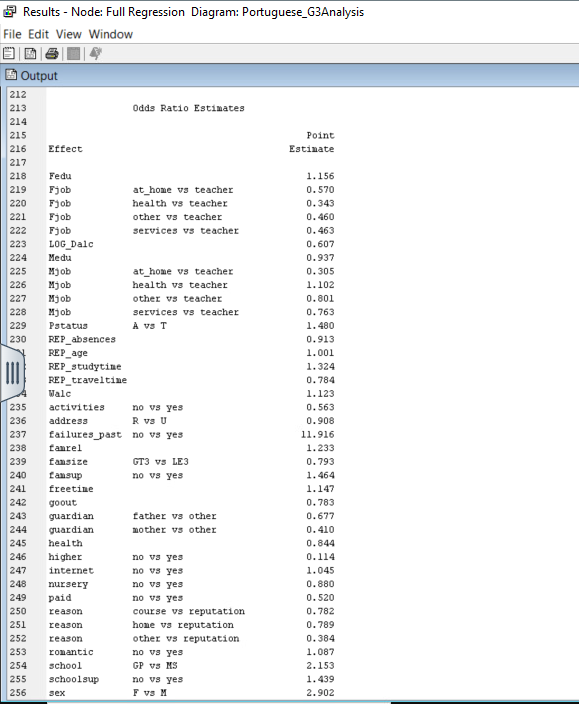
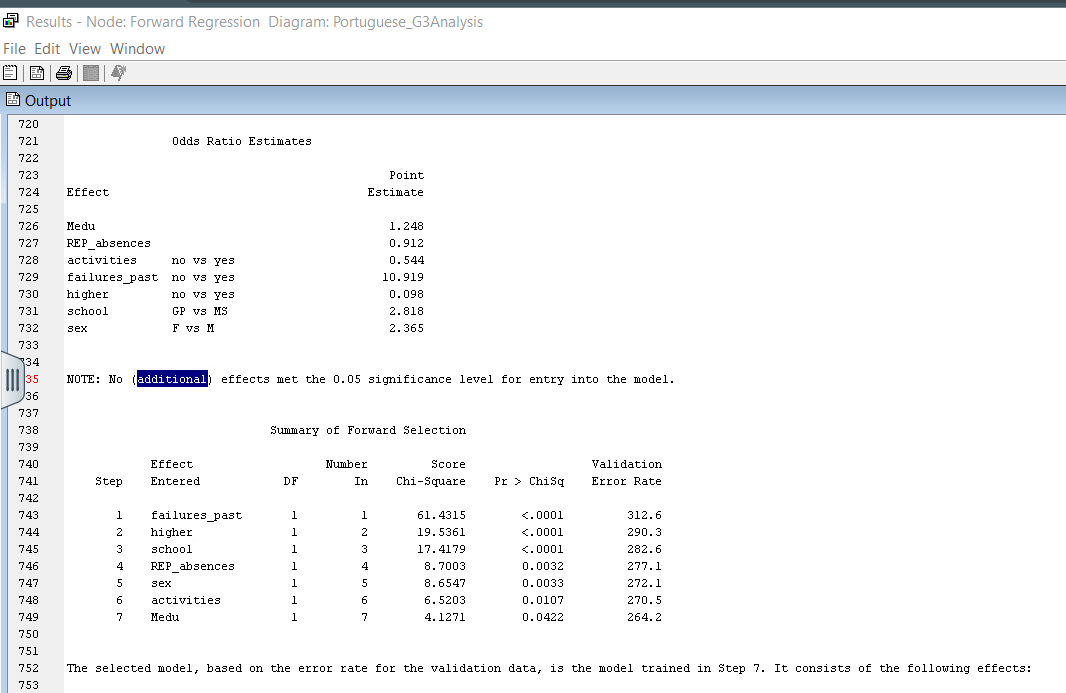
Upon running the full regression model, the ASE result was **0.170415.** The screenshots below highlight the ASE results and output.

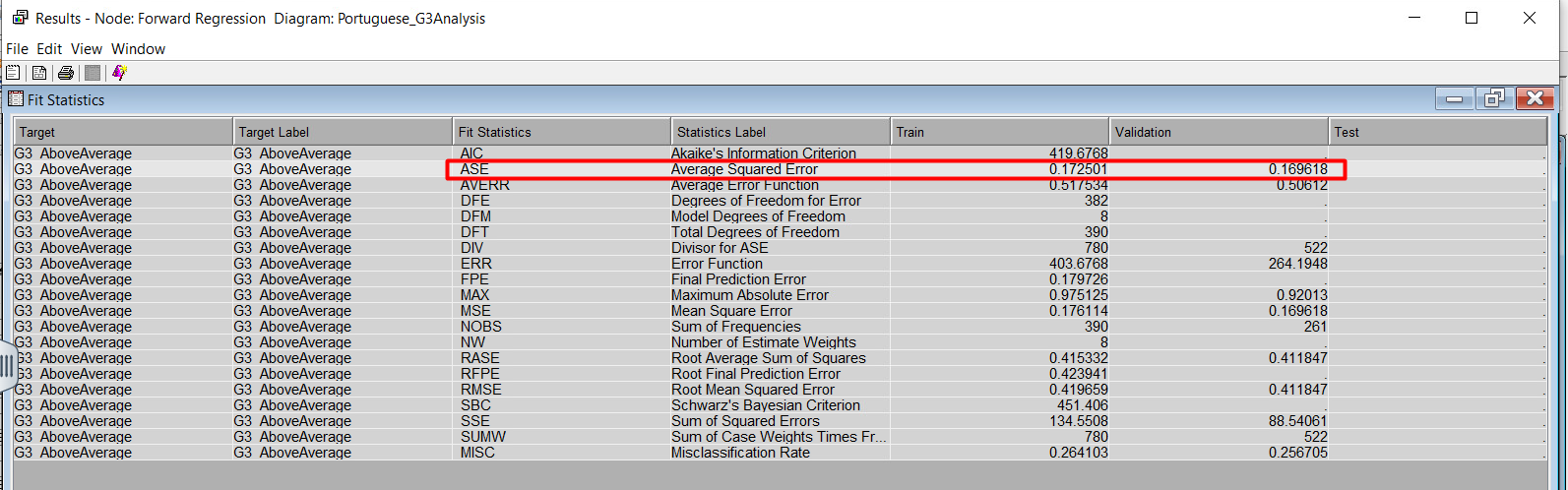


## Forward Regression

A forward regression gradually adds predictor variables to a model, selecting the most impactful ones that enhance the model's predictive power until no further variables significantly improve its performance.

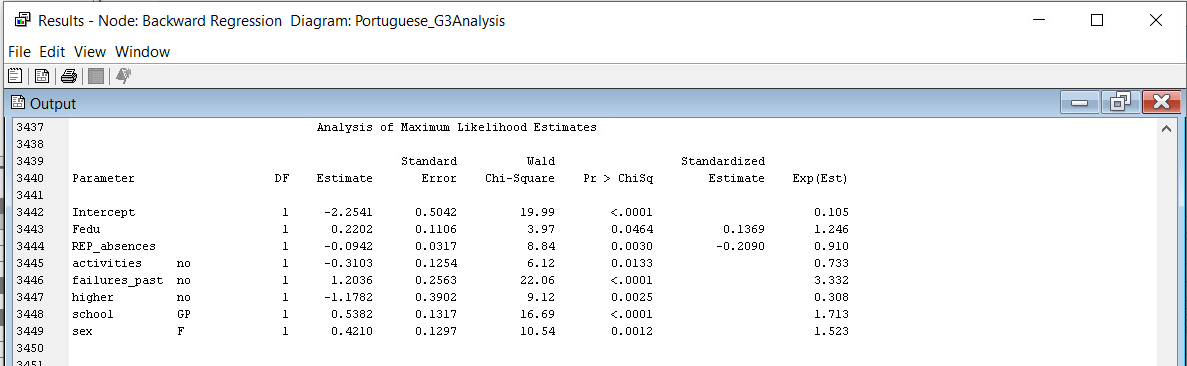
The ASE result of the forward regression model was **0.169618.** The illustrations below show the ASE results and output.

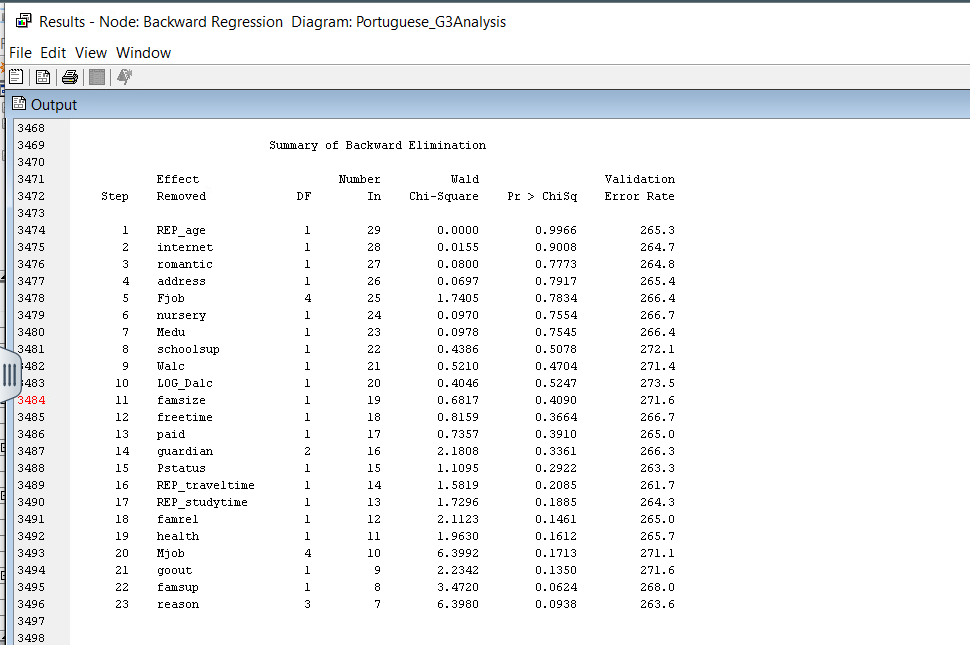




## Backward Regression

A backward regression begins with a model containing all predictors and eliminates the least impactful variables one by one until the model's predictive performance no longer significantly improves.

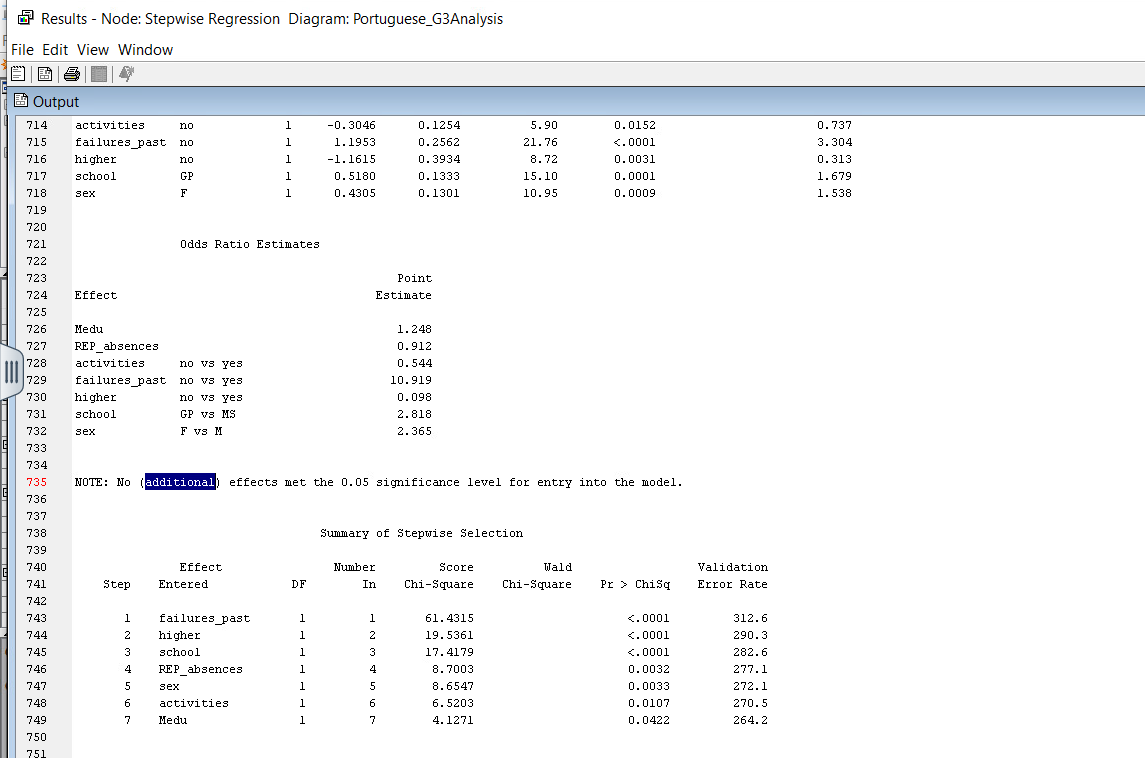
Upon running the backward regression model, the ASE result was **0.166314.** The figures below highlight the results of the model.





## Stepwise Regression

The final regression model that has been created is stepwise regression, which is a method that automatically selects the most important variables for a model by adding or removing them based on their impact on improving the model's accuracy.

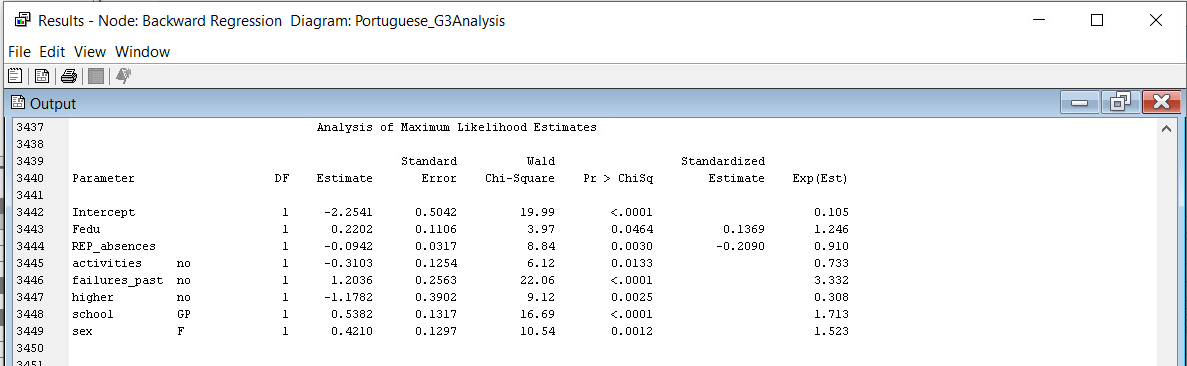
The ASE value of the stepwise regression model resulted in **0.169618.** The illustrations below display the results of the model.



## Analysis and Interpretation of Regression

After analyzing the four regression models, the backward regression is selected as the most optimal model because it has the lowest ASE score. While analyzing the output of the model, a notable observation is that the backward regression model did 23 steps while eliminating the variables with the higher Pr > ChiSq. After the completion of this process, the variables with the maximum likelihood of success were the ones with a Pr > ChiSq below 0.05. The variables with the lowest Pr > ChiSq and the highest Wald Chi-Square were, in order of importance, the following:

1. *failures\_past* with a No result*,* a Wald Chi-Square of 22.06 and a Pr > ChiSq of <0.0001.
2. *school* with a GP result, a Wald Chi-Square of 16.69 and a Pr > ChiSq of <0.0001.
3. *sex* with a F result, a Wald Chi-Square of 10.54 and a Pr > ChiSq of 0.0012.
4. *higher* with a No result, a Wald Chi-Square of 9.12 and a Pr > ChiSq of 0.0025.
5. *REP\_absences* with a Wald Chi-Square of 8.84 and a Pr > ChiSq of 0.0030.
6. *activities* with a No result, a Wald Chi-Square of 6.12 and a Pr > ChiSq of 0.0133.
7. *Fedu* with a Wald Chi-Square of 3.97 and a Pr > ChiSq of 0.0464.





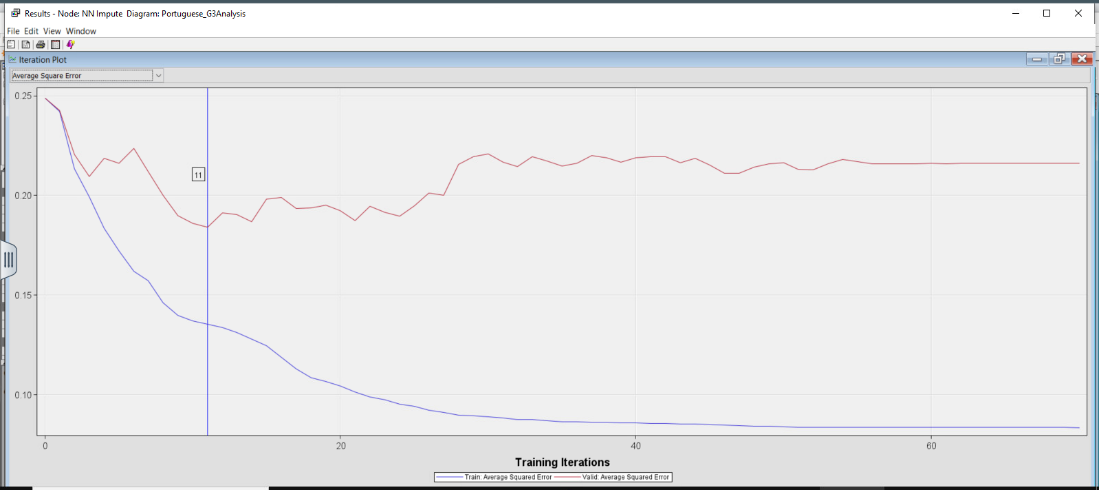
From the perspective of the odds ratio, the best variables for this model can be interpreted as follows in order of their impact on the target variable:

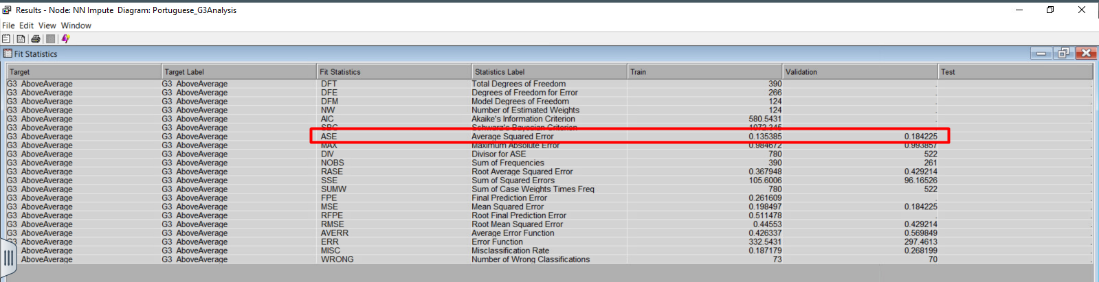
1. ***failures\_past:***Students who have not failed in the past are 11.103 times more likely to score above-average grades in *G3* than students who have failed in the past.
2. ***school*:** Students who go to GP school are 2.934 times more likely to score above-average grades in *G3* than students who go to MS school.
3. ***sex*:** Students who are Female are 2.321 times more likely to score above-average grades in *G3* than students who are Male.
4. ***higher:*** Students who do not want to pursue a higher education are 90.5% less likely to score above-average grades in *G3* than students who want to pursue a higher education.
5. ***REP\_absences*:** For every unit increase in *REP\_absences*, there is a 9% decrease in the probability of scoring an above-average grade in *G3*. This indicates that as students' absences increase, the probability of scoring an above-average grade in *G3* decreases.
6. ***activities*:** Students who do not perform any extracurricular activities are 46.2% less likely to score above-average grades in *G3* than students who perform any extracurricular activities.
7. ***Fedu*:** For every unit increase in *Fedu*, there is a 24.6% increase in the probability of scoring an above-average grade in *G3*. This indicates that as the father’s education level increases, the probability of scoring an above-average grade in G3 for the student increases as well.

# Neural Network

The third and final type of predictive model that has been created for predicting whether students attain an above average grade in their final period of the Portuguese subject is neural network. For this report, 13 neural networks have been constructed to discover the optimal neural network model.

## Impute Neural Network

The impute neural network displays the number of training iterations made by the model and as the properties of the neural network are set to a maximum of 100, the impute neural network converges and gets the optimal iterations at 11. This model has an ASE value of **0.184225**. The illustrations below depict the iteration plot and fit statistics of the model.

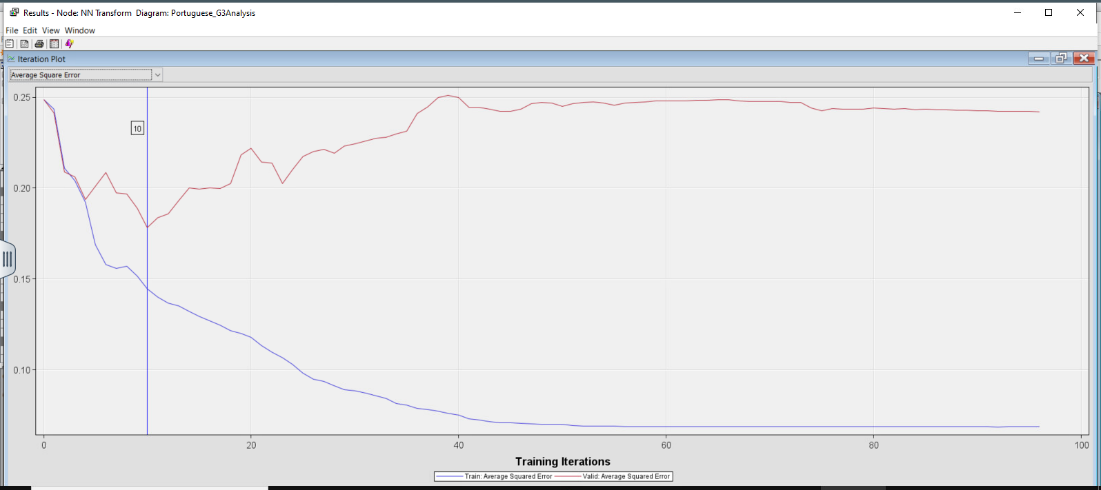


## Cap and Floor Neural Network

The cap and floor neural network show the number of training iterations the model made and as the properties of the neural network are set to a maximum of 100, the cap and floor neural network converges and gets the optimal iterations at 9. This model has an ASE value of **0.200161**. The screenshots below highlight the iteration plot and fit statistics of the model.

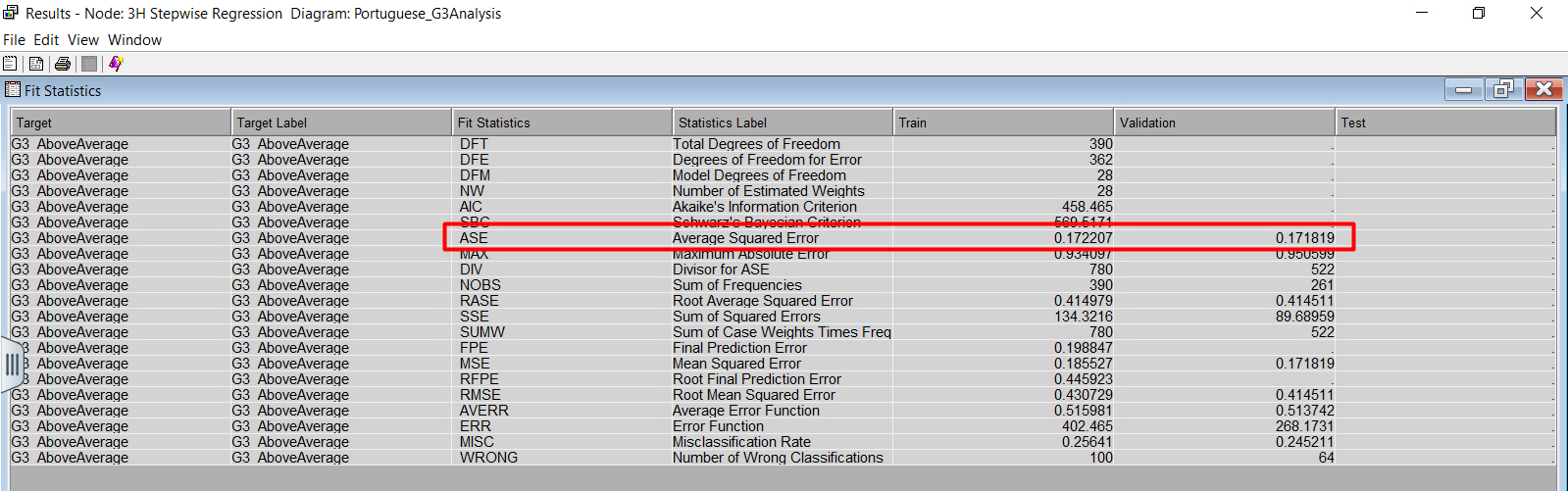


## Transform Neural Network

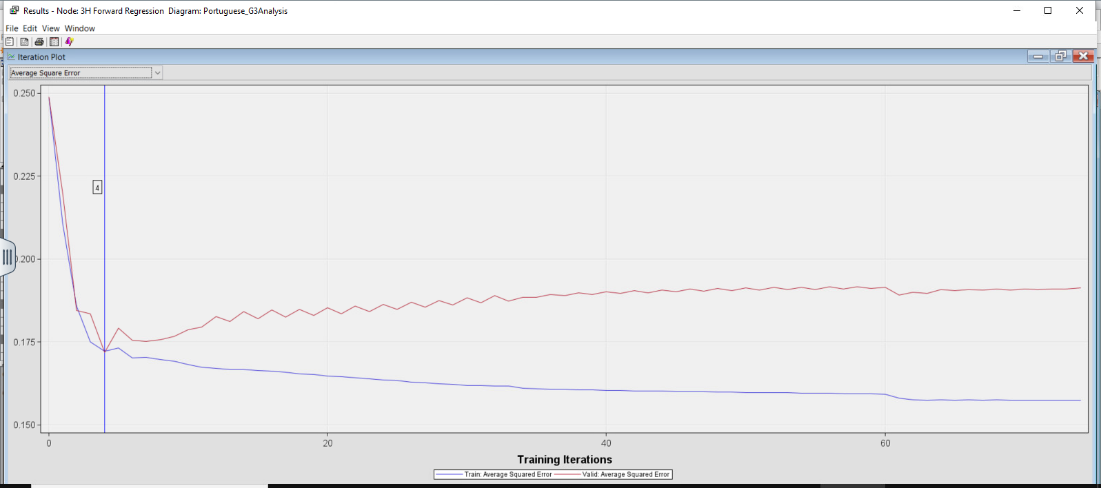
The transform neural network shows the number of training iterations the model made and as the properties of the Neural Network are set to a maximum of 100, the transform neural network converges and gets the optimal iterations at 10. This model has an ASE result of **0.178348**. The diagrams below illustrate the iteration plot and fit statistics of the model.

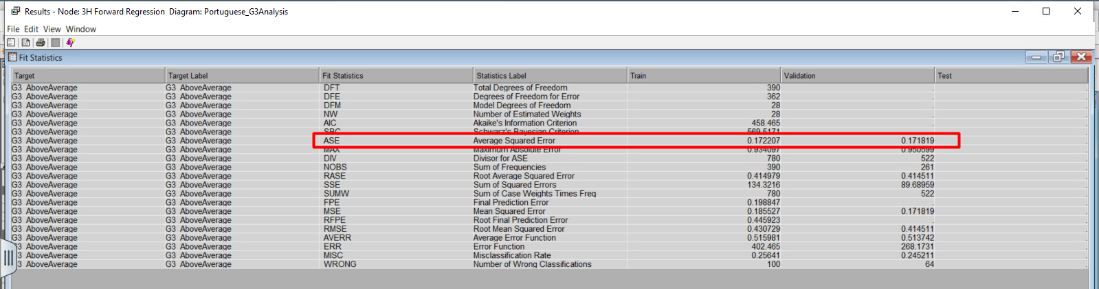


## Stepwise Regression Neural Network

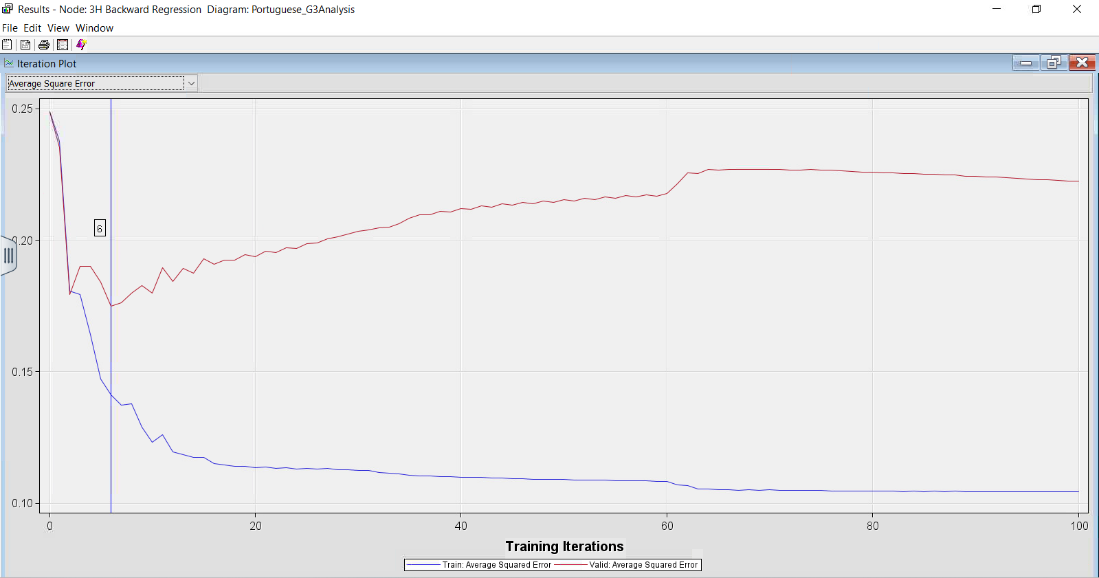
The stepwise regression neural network shows the number of training iterations the model made and as the properties of the neural network are set to a maximum of 100, the stepwise regression neural network converges and gets the optimal iterations at 4. This model has an ASE value of **0.171819**. The figures below explain the iteration plot and fit statistics of the model.

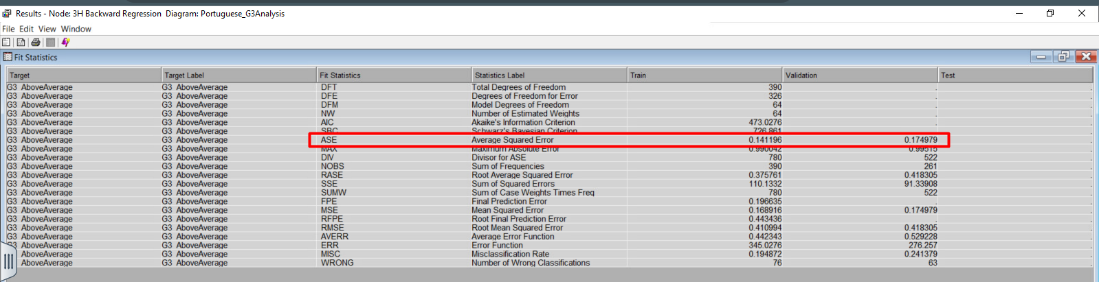
## Forward Regression Neural Network

The forward regression neural network shows the number of training iterations the model made and as the properties of the neural network are set to a maximum of 100, the forward regression neural network converges and gets the optimal iterations at 4. This model has an ASE result of **0.171819**. The iteration plot and fit statistics of the model are depicted below.



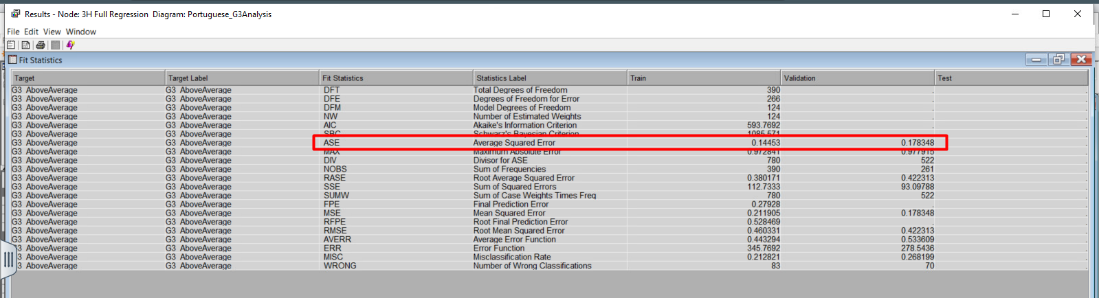
## Backward Regression Neural Network

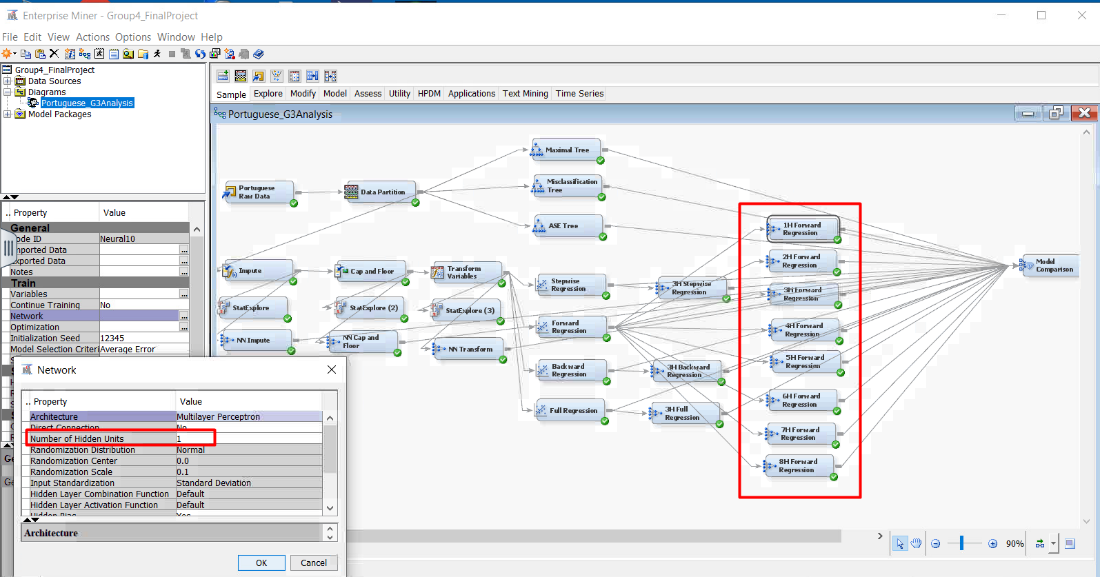
The backward regression neural network shows the number of training iterations the model made and as the properties of the neural network are set to a maximum of 100 iterations, the backward regression neural network gets the optimal iterations at 6. This model has an ASE figure of **0.174979**. The screenshots below depict the iteration plot and fit statistics of the model.



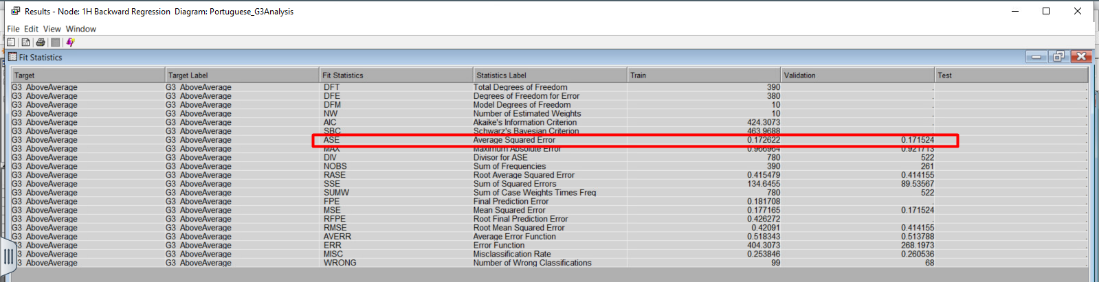
## Full Regression Neural Network

The full regression neural network shows the number of training iterations the model made and as the properties of the neural network are set to a maximum of 100, the full regression neural network converges and gets the optimal iterations at 10. This model has an ASE value of **0.178348**. The iteration plot and fit statistics of this model are provided in the illustrations below.

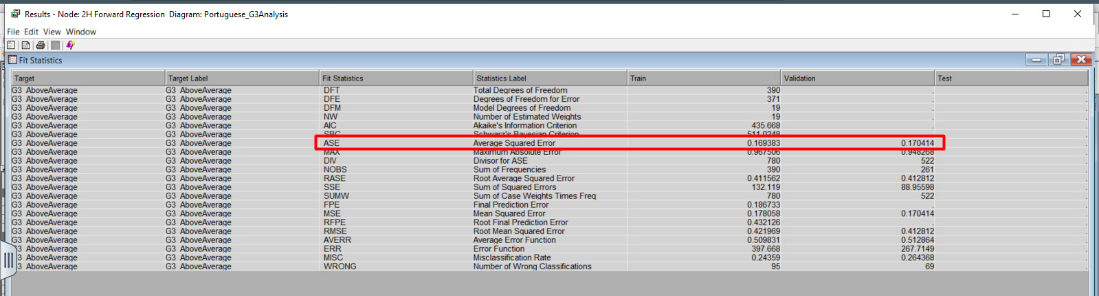


Based on all the neural networks that have been explained above, the 3H Forward Regression neural network and the 3H Stepwise Regression neural network have the lowest ASE (0.171819) out of all the neural network models. Between these two models, while one is not better than the other, the forward regression neural network has been selected for the next step. To find the optimal forward regression neural network, the number of hidden units has been changed within the range of 1 to 8.

## 1H Forward Regression Neural Network



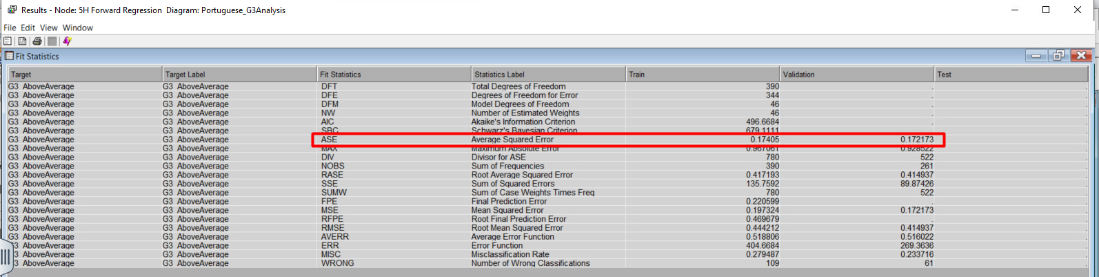
## 2H Forward Regression Neural Network



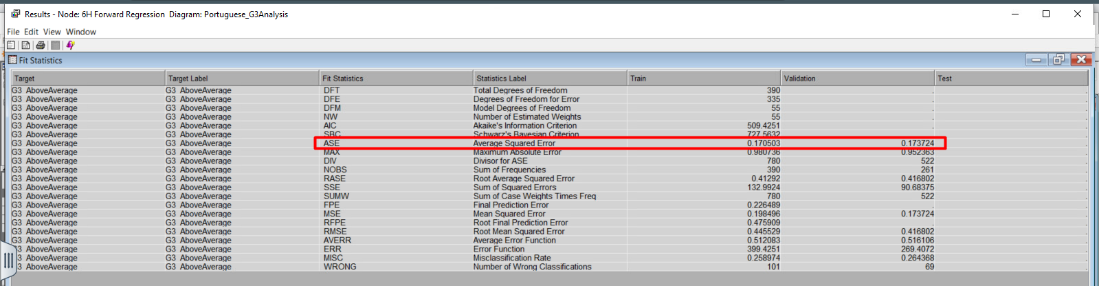
## 4H Forward Regression Neural Network



## Inserting image...5H Forward Regression Neural Network



## 6H Forward Regression Neural Network



## Inserting image...7H Forward Regression Neural Network



## 8H Forward Regression Network



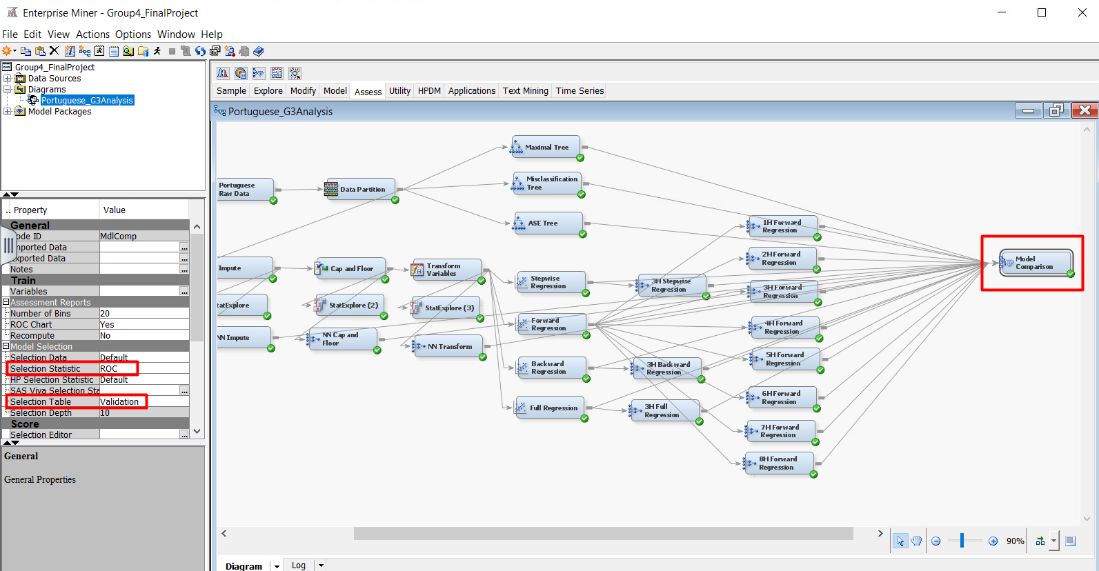
## Analysis of Neural Network

While Neural Network models are not subject to interpretation, the table below depicts which is the best model to use based on ASE values, the number of iterations, and convergence status of the created models. Based on the data, the best neural network model happens to be **7H Forward Regression NN** with ASE of **0.169274** and 2 iterations. It is followed up by **2H Forward Regression** **NN** model with an ASE score of **0.170414**. Although the 2H Forward Regression NN model has a greater number of iterations, it displayed convergence before reaching 100 iterations.

|  |  |  |  |
| --- | --- | --- | --- |
| **NN Model** | **ASE** | **Number of iterations** | **Convergence** |
| Neural Network Impute | 0.184225 | 11 | Yes |
| Neural Network Cap and Floor | 0.200161 | 9 | Yes |
| NN Transform | 0.178348 | 10 | Yes |
| 3H Stepwise NN | 0.171819 | 4 | Yes |
| 3H Forward NN | 0.171819 | 4 | Yes |
| 3H Backward NN | 0.174979 | 6 | No |
| 3H Full Regression NN | 0.178348 | 10 | Yes |
| 1H Forward Regression NN | 0.171524 | 7 | Yes |
| 2H Forward Regression NN | 0.170414 | 9 | Yes |
| 4H Forward Regression NN | 0.175181 | 4 | No |
| 5H Forward Regression NN | 0.172173 | 3 | No |
| 6H Forward Regression NN | 0.173724 | 3 | No |
| 7H Forward Regression NN | 0.169274 | 2 | No |
| 8H Forward Regression NN | 0.170721 | 3 | No |

# Model Comparison

In the final step, a Model Comparison node has been created and connected with every model node in the diagram. This step has been conducted to compare the ROC index of all the models and determine which one is the optimal model.



## Analysis of Model Comparison

Although the 7H Forward Regression NN is the best neural network model concerning the ASE value, the ROC index of **5H Forward Regression NN** is the best neural network model and the best model overall with a score of 0.83. While this is the optimal model at predicting success for students getting an above-average score for G3, this model cannot be interpreted.

The next best interpretable model is **backward regression** with a ROC index of 0.828, which is only 0.02 less than that 5H Forward Regression NN. Additionally, this model has the best ASE score (0.166314) among all the models.

Another notable result is the **ASE tree**. Although the ASE tree is the best decision tree in terms of ASE value, its ROC index is the third lowest out of all the predictive models (0.756).

### Interpretation of Backward Regression Model

During the model analysis, a notable finding was that the backward regression method involved 23 steps, eliminating variables with higher Pr > ChiSq values. Upon completion, it revealed that variables with a Pr > ChiSq below 0.05 had the highest probability of success. The variables with the lowest Pr > ChiSq and the highest Wald Chi-Square were, in order of importance, the following:

1. *failures\_past* with the No result*,* a Wald Chi-Square of 22.06 and a Pr > ChiSq of <0.0001.
2. *school* with a GP result, a Wald Chi-Square of 16.69 and a Pr > ChiSq of <0.0001.
3. *sex* with a F result, a Wald Chi-Square of 10.54 and a Pr > ChiSq of 0.0012.
4. *higher* with the No result, a Wald Chi-Square of 9.12 and a Pr > ChiSq of 0.0025.
5. *REP\_absences* with a Wald Chi-Square of 8.84 and a Pr > ChiSq of 0.0030.
6. *activities* with a No result, a Wald Chi-Square of 6.12 and a Pr > ChiSq of 0.0133.
7. *Fedu* with a Wald Chi-Square of 3.97 and a Pr > ChiSq of 0.0464.

As for the odds ratio, the best variables for this model can be interpreted as follows in order of their impact on the target variable:

1. ***failures\_past:***Students who have not failed in the past are 11.103 times more likely to score above-average grades in G3 than students who have failed in the past.
2. ***school*:** Students who go to GP school are 2.934 times more likely to score above-average grades in G3 than students who go to MS school.
3. ***sex*:** Students who are Female are 2.321 times more likely to score above-average grades in G3 than students who are Male.
4. ***higher:*** Students who do not want to pursue a higher education are 90.5% less likely to score above-average grades in G3 than students who want to pursue a higher education.
5. ***Rep\_absences*:** For every unit increase in *Rep\_absences*, there is a 9% decrease in the probability of scoring an above-average grade in G3. This indicates that as students' absences increase, the probability of scoring an above-average grade in G3 decreases.
6. ***activities*:** Students who do not perform any extracurricular activities are 46.2% less likely to score above-average grades in G3 than students who perform any extracurricular activities.

## Changing Certain Data Types to Compare Model Performance

An experiment was conducted to check if the model performance improved after changing certain data types and the corresponding categories. The new model was simplified by converting the data types of *Famrel*, *Fedu* and *Medu* from interval to nominal and consolidating their categories. After conducting a model comparison, it was observed that the initial results of the model, prior to the data type changes, yielded superior outcomes. This conclusion is drawn from the findings presented in the accompanying screenshot which represents the model with updated data types.

A screenshot of a computer

Description automatically generated

# Conclusion and Recommendations

Upon running each predictive model of decision trees, regressions and neural networks, along with the Model Comparison node, the optimal model has been identified to predict the students’ performance in the final period grade of the Portuguese subject. Additionally, the key variables that drive the final period grade above average have also been discovered.

Based on the findings, the following rank-ordered business recommendations can be made:

1. **Prioritize support for students with a history of academic failures:** Implement targeted interventions or support systems for students who have previously failed. These students are significantly less likely to score above average, indicating a need for tailored academic support or mentoring programs.
2. **Highlight the importance of school choice:** Emphasize the advantages of attending Gabriel Pereira over Mousinho da Silveira in terms of academic performance. Consider showcasing the positive outcomes associated with attending that school.
3. **Address gender disparities:** Develop initiatives to ensure equal opportunities for all genders in education. Provide additional support or programs specifically tailored to address any disparities that might affect male students' performance compared to female students.
4. **Promote higher education aspirations:** Encourage and support students' aspirations for higher education. Providing resources, guidance, and information about the benefits of pursuing higher education could positively impact their academic performance.
5. **Mitigate the impact of absences:** Implement strategies to reduce absenteeism among students. Addressing the issue of absenteeism could potentially improve students' likelihood of achieving above-average grades.
6. **Encourage extracurricular engagement:** Promote and facilitate extracurricular activities within the academic setting. Encouraging students to participate in extracurricular activities can potentially enhance their academic performance.

# References

Abu Bakar, N., Yusop, H., Ali, N.M., Abu Bakar, N.F. (2023). Determinants of Students’ Academic Performance in Higher Learning Institutions in Malaysia. International Journal of Academic Research in Business and Social Sciences, 13(2), 1496-1508. http://dx.doi.org/10.6007/IJARBSS/v13-i2/16250

Chauhan, A. (2022). *Student Performance: Predict student performance in secondary education (high school)*. Kaggle. https://www.kaggle.com/datasets/whenamancodes/student-performance

Regier, J. (2011). Why is Academic Success Important?. Saskatchewan School Boards Association. https://saskschoolboards.ca/wp-content/uploads/2015/08/2011SIAST.pdf

Tadese, M., Yeshaneh, A. & Mulu, G.B. (2022). Determinants of good academic performance among university students in Ethiopia: a cross-sectional study. BMC Medical Education, 22(395), 1-9. https://doi.org/10.1186/s12909-022-03461-0