Mushroom Toxicity Classification: A Data-Driven Guide for Vegetarians

Aisha Anamika

Marina Soares de Souza

Nathania Mbeshi

BA 706 – Applied Analytic Modeling

December, 2024

Table of Contents

Mushroom Classification Analysis

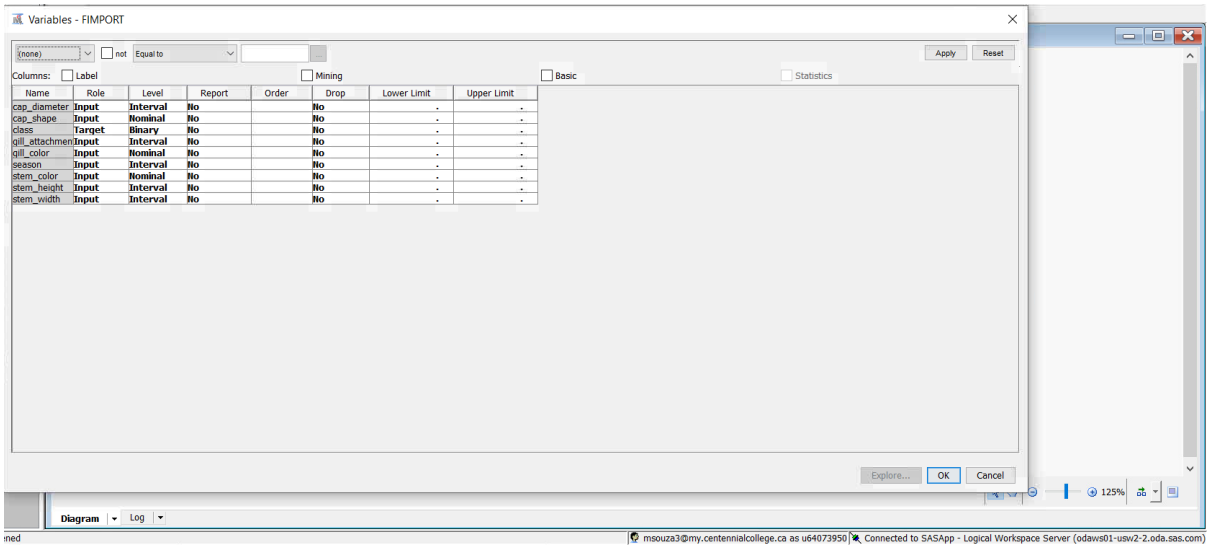
According to Forage Hyperfoods (Staicu, 2023), mushroom foraging is popular among vegetarians in Canada, offering a sustainable source of nutritious food. However, identifying whether a mushroom is poisonous, or edible can be a life-or-death decision. To support vegetarian foragers, we aim to create a reliable guide that uses data-driven insights to classify mushrooms based on key features.

This analysis utilizes the mushroom_cleaned dataset from Kaggle to identify the critical variables that determine mushroom edibility. By utilizing predictive models such as decision trees, logistic regression, and neural networks built in SAS Enterprise Miner, we will pinpoint the significant features that distinguish poisonous mushrooms from edible ones.

The results of this study will equip vegetarians with a scientifically grounded tool to make safer foraging decisions.

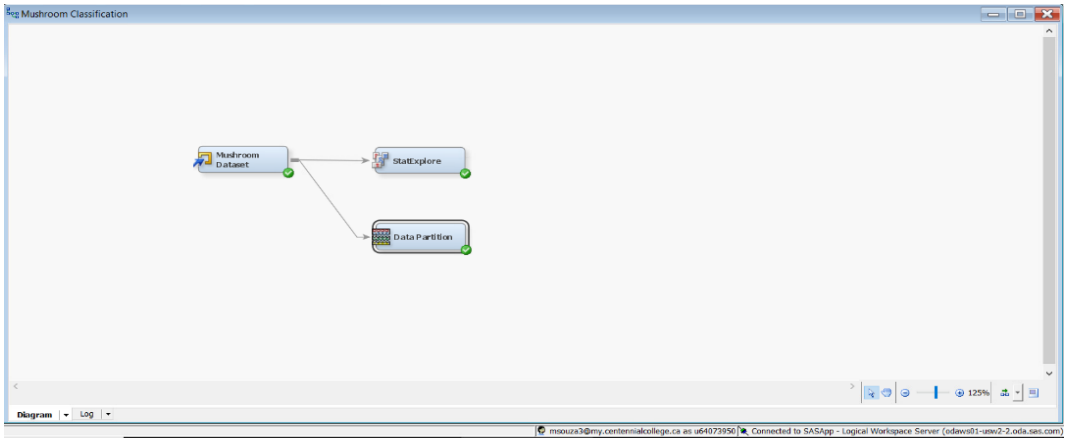
Data Selection and Target Variable

- **Dataset:** The dataset was sourced from Kaggle and contains 9 variables that explain mushroom features such as cap diameter, cap shape, class, gill attachment, gill color, season, stem height, stem width, and stem color.
- **Target Variable:** We identified our target variable as the mushroom class which was made up of a binary set of data that denotes whether a mushroom is poisonous (0) or edible (1).
- **Rejected Variables:** All the variables in our dataset were significant and therefore we did not need to reject any variable at the beginning of the analysis.



Statistical Analysis

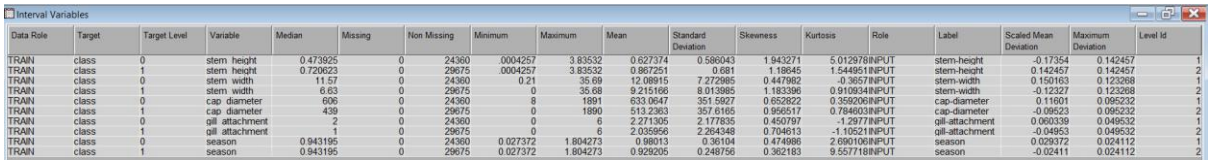
To initiate the analysis, we assessed the statistical properties of the dataset by incorporating the StatExplore node into the diagram and linking it to the Mushroom Dataset node.



Missing Values: The results revealed that the dataset contained no missing values, eliminating the need for imputation as we proceeded with the analysis.

Skewness: However, the variables stem width (for both poisonous and edible mushrooms) and stem height exhibited significant skewness, with values greater than 1. This indicates that

their distributions are not symmetrical, necessitating replacements or transformations to ensure accurate and reliable modeling outcomes.

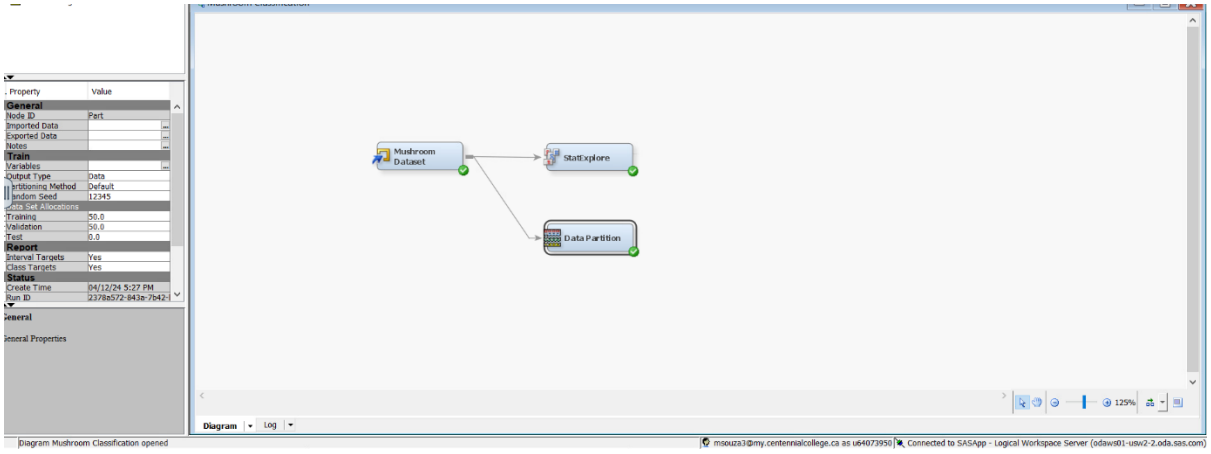

Interval Variables

Data Role	Target	Target Level	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label	Scaled Mean Deviation	Maximum Deviation	Level Id
TRAIN	class	0	stem_height	0.473925	0	24360	.0004257	3.83532	0.627374	0.586043	1.943271	5.012978	INPUT	stem-height	-0.17354	0.142457	1
TRAIN	class	1	stem_height	0.720623	0	29675	.0004257	3.83532	0.867251	0.681	1.18645	1.544951	INPUT	stem-height	0.142457	0.142457	2
TRAIN	class	0	stem_width	11.57	0	24360	0.21	35.69	12.08915	7.272985	0.447982	-0.36571	INPUT	stem-width	0.150163	0.123268	1
TRAIN	class	1	stem_width	6.63	0	29675	0	35.68	9.215166	8.013985	1.183396	0.910934	INPUT	stem-width	-0.12327	0.123268	2
TRAIN	class	0	cap_diameter	606	0	24360	8	1891	633.0647	351.5927	0.652822	0.359206	INPUT	cap-diameter	0.11601	0.095232	1
TRAIN	class	1	cap_diameter	439	0	29675	0	1890	513.2363	357.6165	0.956517	0.784603	INPUT	cap-diameter	-0.09623	0.095232	2
TRAIN	class	0	gill_attachment	2	0	24360	0	6	2.271305	2.177835	0.450797	-1.29778	INPUT	gill-attachment	0.060339	0.049532	1
TRAIN	class	1	gill_attachment	1	0	29675	0	6	2.035956	2.264348	0.704613	-1.10521	INPUT	gill-attachment	-0.04953	0.049532	2
TRAIN	class	0	season	0.943195	0	24360	0.027372	1.804273	0.98013	0.36104	0.474986	2.690106	INPUT	season	0.029372	0.024112	1
TRAIN	class	1	season	0.943195	0	29675	0.027372	1.804273	0.929205	0.248756	0.362183	9.557718	INPUT	season	-0.02411	0.024112	2

Data Partitioning

To evaluate model performance and mitigate overfitting, the dataset was partitioned into training and validation subsets with a 50:50 split. This was achieved by incorporating the *Data Partition* node into the diagram and linking it to the mushroom dataset node.

The data allocation ratios were adjusted to 50% for training, 50% for validation, and 0% for testing, as illustrated below.

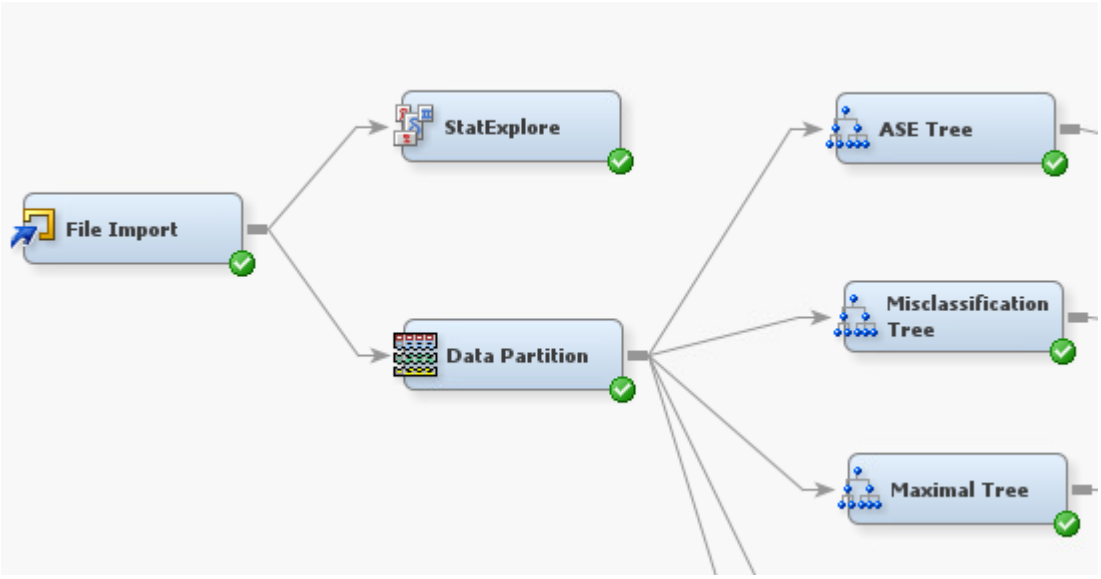


Modelling

We built three predictive models to classify mushroom edibility: a Decision Tree, Regression, and Neural Network.

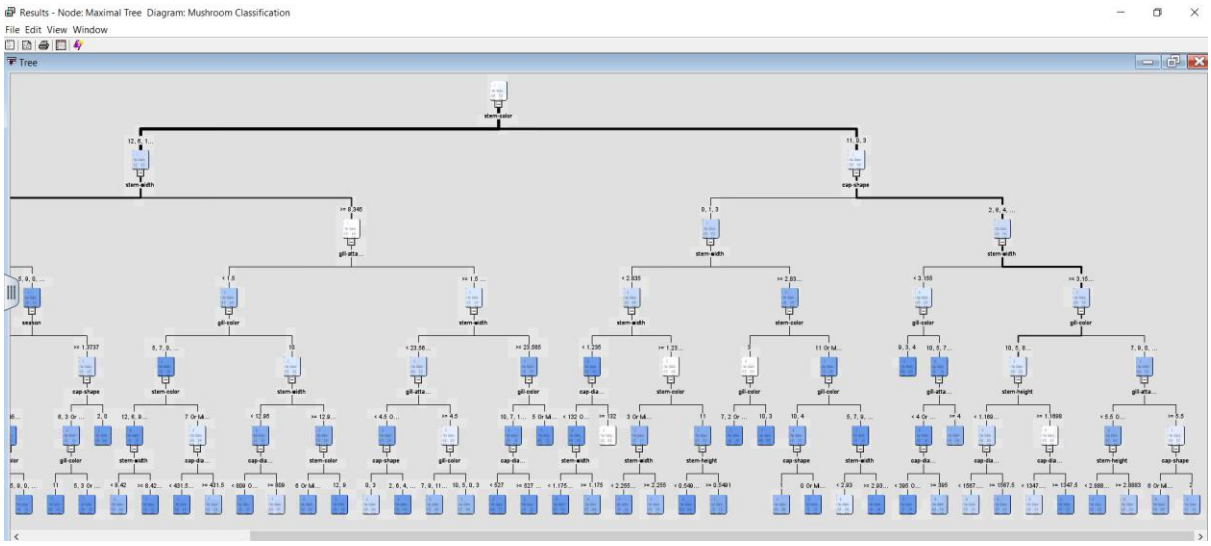
1. Decision Tree Model

After running the data partition, we dragged in decision tree nodes and connected it to the data partition node.

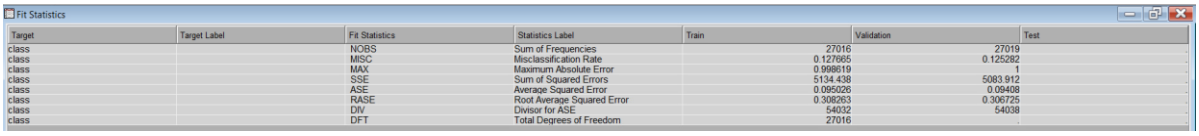


a. Maximal Tree

We modelled a maximal tree by making no changes to the node and ran it. This was to consider all possible splits of the independent variables. We found that the tree had 107 nodes and the three most significant predictors in order of importance were: Stem Color, Stem Width and Cap Shape.



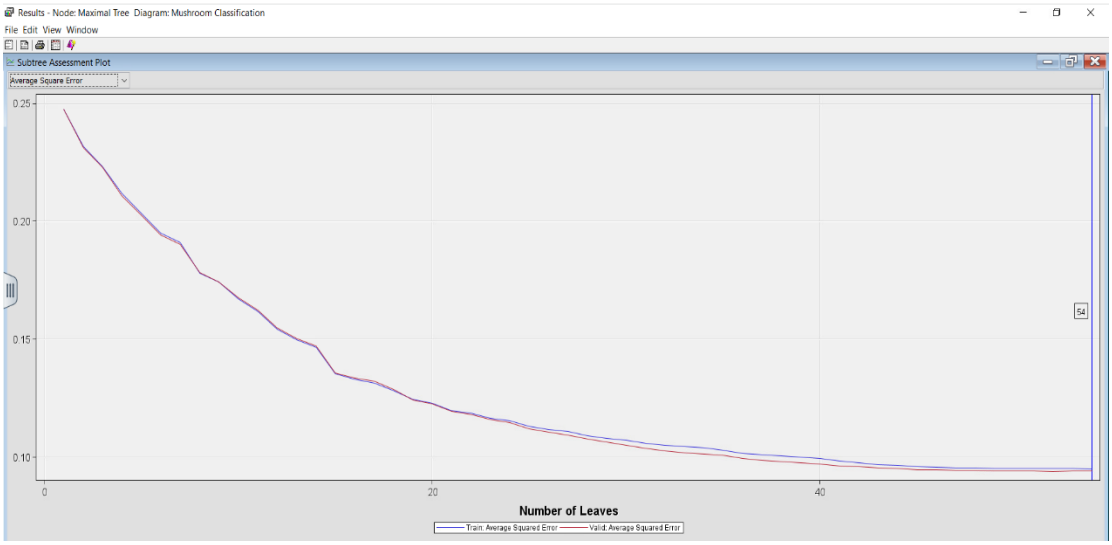
We expanded the results to examine the fit statistics and noted that the ASE (Average Squared Error) of the maximal tree was 0.09408.



Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
class		NOBS	Sum of Frequencies	27016	27019	
class		MISC	Misclassification Rate	0.127665	0.125282	
class		MAX	Maximum Absolute Error	0.998619	1	
class		SSE	Sum of Squared Errors	5134.438	5083.912	
class		ASE	Average Squared Error	0.095026	0.09408	
class		RASE	Root Average Squared Error	0.308263	0.306725	
class		DIV	Divisor for ASE	54032	54038	
class		DFT	Total Degrees of Freedom	27016		

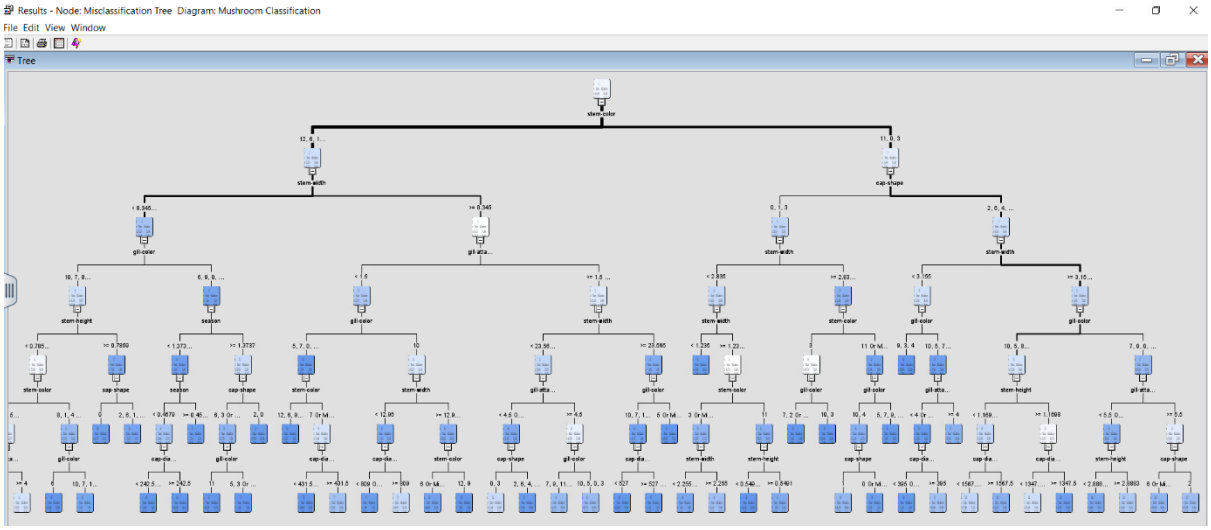
Based on the subtree assessment plot shown below, we observed that the maximal tree contained 54 leaves.



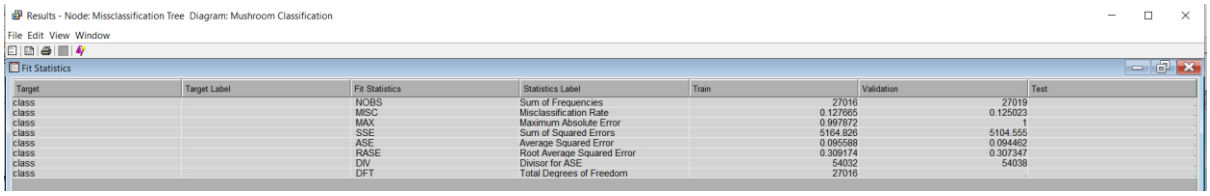
b. Optimal Tree

To evaluate potential overfitting, additional decision trees were incorporated into the analysis. The subtree assessment measures were adjusted to include ASE and misclassification.

Misclassification Tree: Similar to the maximal tree, the misclassification tree consisted of 107 nodes. Additionally, we identified stem color, cap shape, and stem width as its three significant predictors.

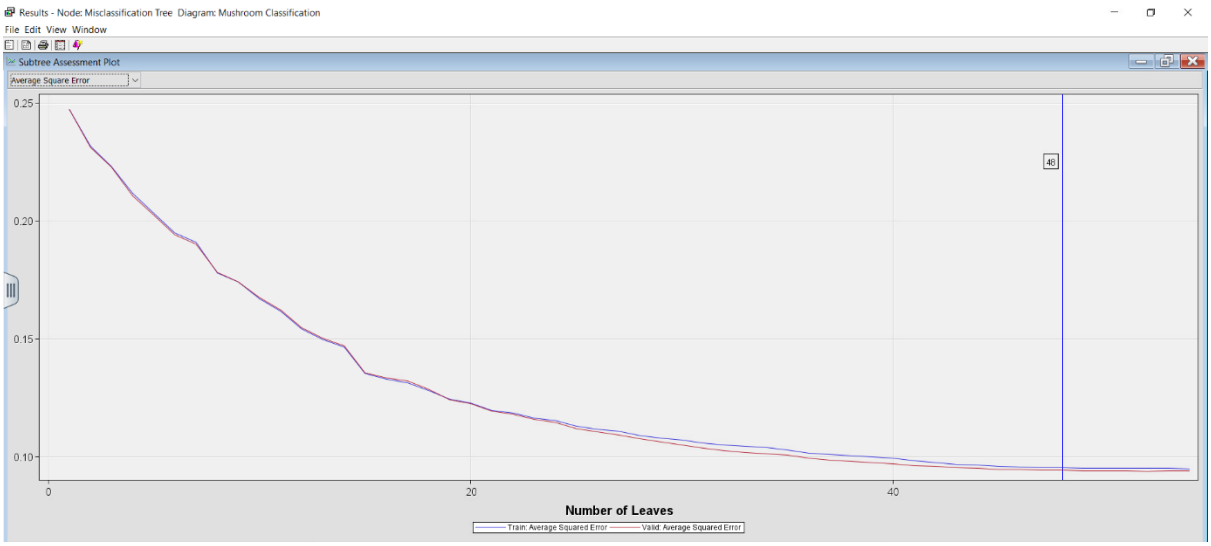


We expanded the results to examine the fit statistics and noted that the ASE of the misclassification tree was 0.094462.

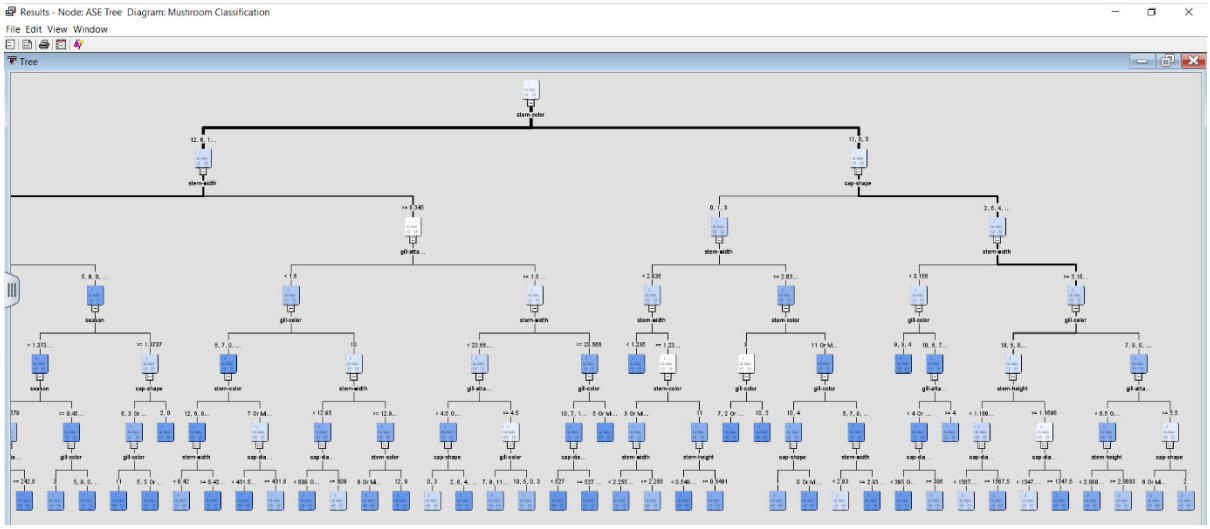


Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
class		NOBS	Sum of Frequencies	27016	27019	
class		MISC	Misclassification Rate	0.127665	0.125023	
class		MAX	Maximum Absolute Error	0.997872	1	
class		SSE	Sum of Squared Errors	5164.826	5104.555	
class		ASE	Average Squared Error	0.095588	0.094462	
class		RASE	Root Average Squared Error	0.309174	0.307347	
class		DIV	Divisor for ASE	54032	54038	
class		DFT	Total Degrees of Freedom	27016		

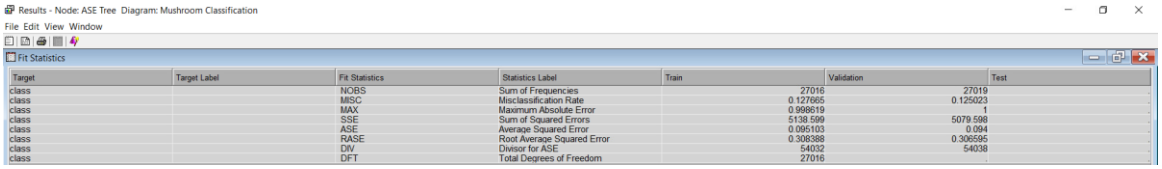
Based on the subtree assessment plot shown below, we observed that the misclassification tree contained 48 leaves.



The Average Squared Error (ASE) Tree: The ASE tree had 107 leaves as well and, consistent with the other decision tree models, identified Stem Color, Stem Width, and Cap Shape as its significant predictors.

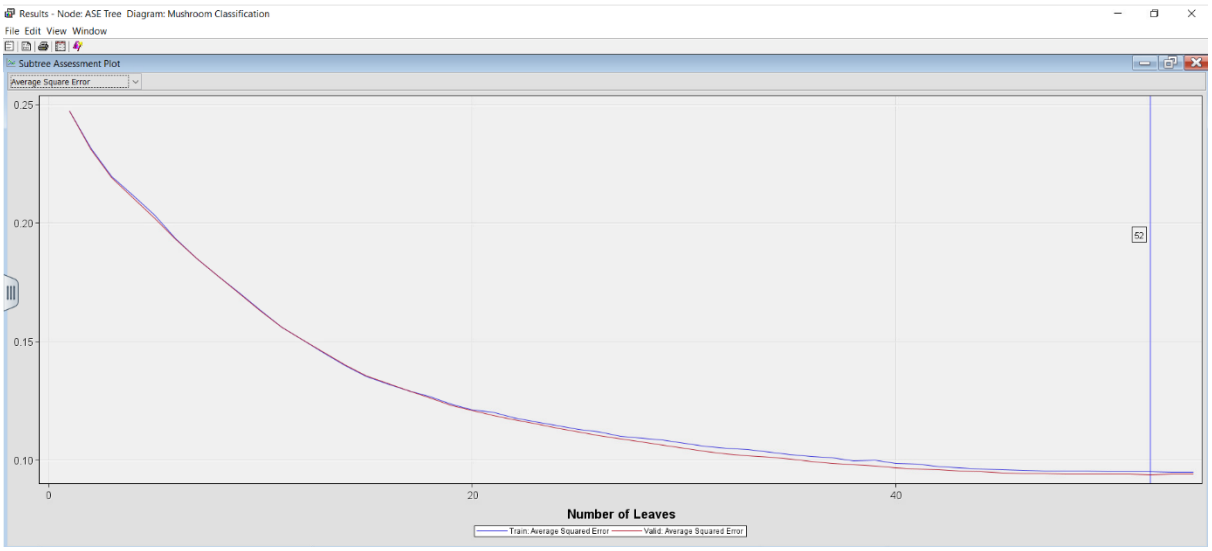


We expanded the results to examine the fit statistics and noted that the ASE of the ASE tree was 0.094.



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
class		NOBS	Sum of Frequencies	27016	27019	
class		MISC	Misclassification Rate	0.127665	0.125023	
class		MAX	Maximum Absolute Error	0.998619	1	
class		SSE	Sum of Squared Errors	5138.599	5079.598	
class		ASE	Average Squared Error	0.095103	0.094	
class		RASE	Root Average Squared Error	0.308388	0.306595	
class		DIV	Divisor for ASE	54032	54038	
class		DFT	Total Degrees of Freedom	27016		

Based on the subtree assessment plot shown below, we observed that the ASE tree contained 52 leaves.



Below is the comparison table for the decision tree statistics to select the better predictor of the three models:

Decision Tree	**ASE Fit Statistic**	**Number of Leaves**	**Number of Nodes**
Maximal Tree	0.094080	54	107
Misclassification Tree	0.094462	48	107
ASE Tree	0.094000	52	107

To determine the better decision tree, the ASE measure was compared across the three models. Compared to the rest, the ASE Tree achieved the lowest ASE of 0.094.
These results indicate that the ASE Tree was the best predictor among the three models with the lowest ASE fit statistic.

2. Regression Models

Since Regressions are less intolerant to skewness, the variables exhibiting skewness underwent data transformation to fix this. This included doing a cap and floor adjustment and finally a log transformation.

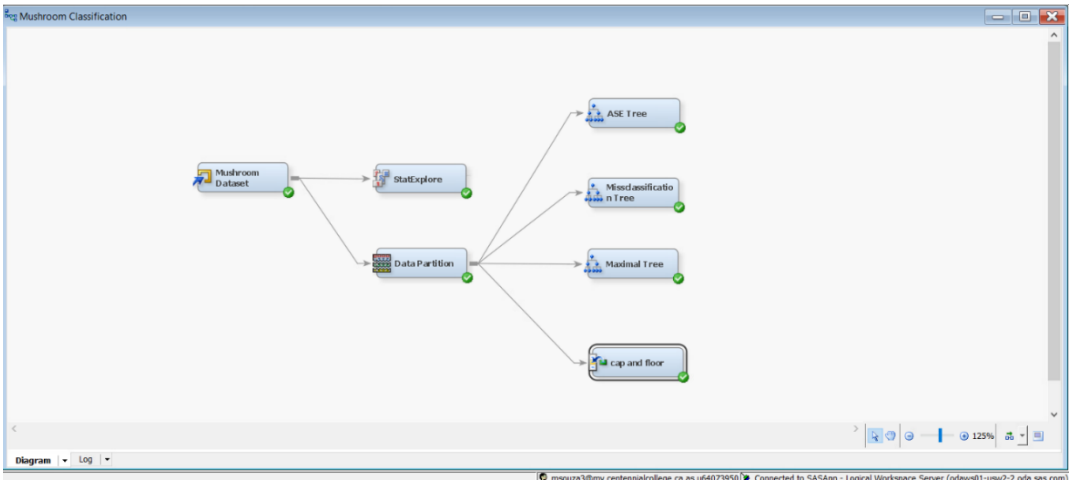
a. Data Transformation

Cap and Floor

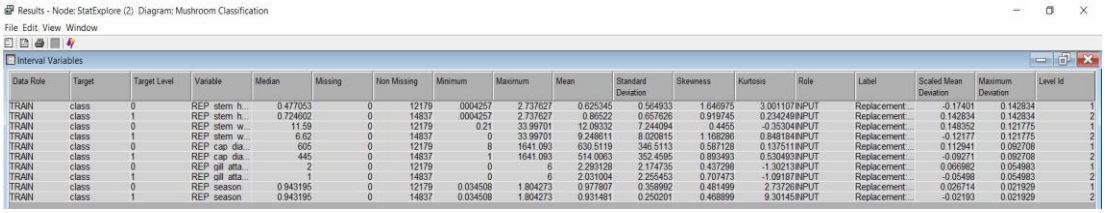
To address the skewness of the variables, we introduced a replacement node from the "Modify" tab and renamed it "Cap and Floor." The properties of the cap and floor node were left unchanged. This node was then connected to the data partition, and the process was executed to assess its impact on skewness.

Following this step, we introduced another Stat Explore(2) node to compare the skewness before and after the adjustment. A significant improvement was observed in the variables' skewness.

However, among the three variables examined, two still exhibited skewness greater than 1. These were REP Stem Height - 1.64 and REP Stem Width - 1.168. Therefore, we decided to implement further adjustments to reduce their skewness. Nonetheless, the cap and floor technique proved to be effective overall.



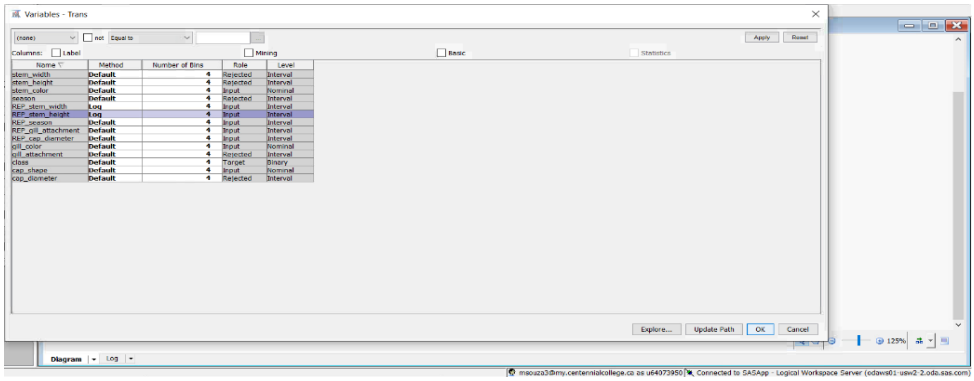
The screenshot attached below shows the reduction in skewness for the second StatExplore(2) after running cap and floor.



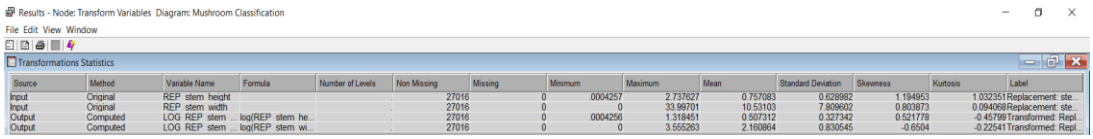
Data Role	Target	Target Level	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label	Scaled Mean Deviation	Maximum Deviation	Level Id
TRAIN	class	0	REP stem h...	0.477053	0	12179	.0004257	2.737627	0.625345	0.564933	1.646975	3.001107	INPUT	Replacement:...	-0.17401	0.142834	1
TRAIN	class	1	REP stem h...	0.724602	0	14837	.0004257	2.737627	0.86522	0.657626	0.919745	0.234249	INPUT	Replacement:...	0.142834	0.142834	2
TRAIN	class	0	REP stem w...	11.59	0	12179	0.21	33.99701	12.08332	7.244094	0.4455	-0.35304	INPUT	Replacement:...	0.148352	0.121775	1
TRAIN	class	1	REP stem w...	6.62	0	14837	0	33.99701	9.248611	8.020815	1.168286	0.848184	INPUT	Replacement:...	-0.12177	0.121775	2
TRAIN	class	0	REP cap dia...	605	0	12179	8	1641.093	630.5119	346.5113	0.587128	0.137511	INPUT	Replacement:...	0.112941	0.092708	1
TRAIN	class	1	REP cap dia...	445	0	14837	1	1641.093	514.0063	352.4595	0.893493	0.530493	INPUT	Replacement:...	-0.09271	0.092708	2
TRAIN	class	0	REP gill atta...	2	0	12179	0	6	2.293128	2.174735	0.437298	-1.30213	INPUT	Replacement:...	0.069862	0.054983	1
TRAIN	class	1	REP gill atta...	1	0	14837	0	6	2.031004	2.255453	0.707473	-1.09187	INPUT	Replacement:...	-0.05498	0.054983	2
TRAIN	class	0	REP season	0.943195	0	12179	0.034508	1.804273	0.977807	0.358992	0.481499	2.737261	INPUT	Replacement:...	0.026714	0.021929	1
TRAIN	class	1	REP season	0.943195	0	14837	0.034508	1.804273	0.931481	0.250201	0.468899	9.301451	INPUT	Replacement:...	-0.02193	0.021929	2

Transform Variables

Although the cap and floor method was effective in reducing the skewness of our variables, further adjustments were necessary as some variables still exhibited skewness values above 1, REP Stem Height and REP Stem Width. To address this, we introduced the "Transform Variables" node from the "Modify" tab and connected it to the cap and floor node.

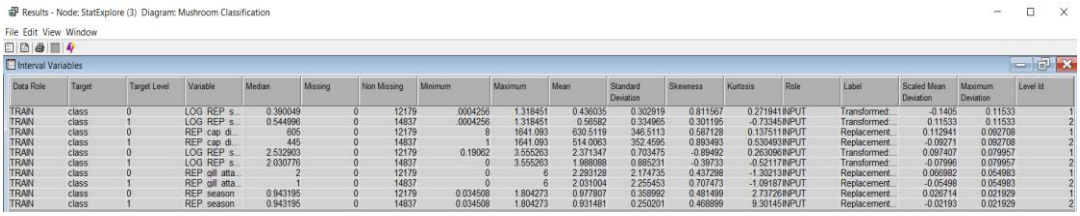


In the properties section, the variables with the highest skewness were transformed using a logarithmic function to further reduce their skewness. After running the transform variables process, we observed a significant reduction in skewness.



Source	Method	Variable Name	Formula	Number of Levels	Non Missing	Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Label
Input	Original	REP_stem_height			27016	0	.0004257	2.737627	0.757083	0.628982	1.194953	1.032351	Replacement: ste...
Input	Original	REP_stem_width			27016	0	0	33.99701	10.53103	7.809602	0.803873	0.094068	Replacement: ste...
Output	Computed	LOG_REP_stem_...	log(REP_stem_he...		27016	0	.0004256	1.318451	0.507312	0.327342	0.521778	-0.45799	Transformed: Repl...
Output	Computed	LOG_REP_stem_...	log(REP_stem_wi...		27016	0	0	3.555263	2.160864	0.830545	-0.6504	-0.22541	Transformed: Repl...

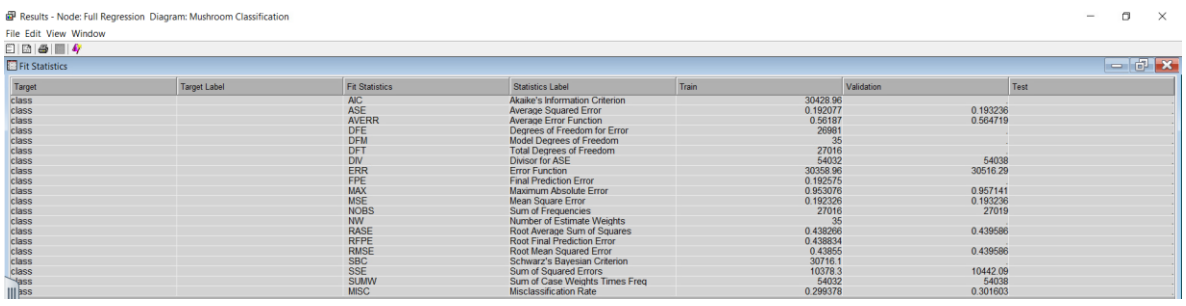
Finally, we added a third "Stat Explore" node and connected it to the transform variables node to compare the updated skewness values. As shown in the attached screenshot, the skewness for all variables was successfully reduced to values below 1**.**



Data Role	Target	Target Level	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label	Scaled Mean Deviation	Maximum Deviation	Level Id
TRAIN	class	0	LOG_REP_s...	0.390049	0	12179	.0004256	1.318451	0.436035	0.302919	0.811567	0.271941	INPUT	Transformed...	-0.1405	0.11533	1
TRAIN	class	1	LOG_REP_s...	0.544996	0	14837	.0004256	1.318451	0.56582	0.334965	0.301195	-0.73345	INPUT	Transformed...	0.11533	0.11533	2
TRAIN	class	0	REP_cap_di...	605	0	12179	8	1641.093	630.5119	346.5113	0.587128	0.137511	INPUT	Replacement...	0.112941	0.092708	1
TRAIN	class	1	REP_cap_di...	445	0	14837	1	1641.093	514.0063	352.4595	0.893493	0.530493	INPUT	Replacement...	-0.09271	0.092708	2
TRAIN	class	0	LOG_REP_s...	2.532903	0	12179	0.19062	3.555263	2.371347	0.703475	-0.89492	0.263096	INPUT	Transformed...	0.097407	0.079957	1
TRAIN	class	1	LOG_REP_s...	2.030778	0	14837	0	3.555263	1.988088	0.885231	-0.39733	-0.52117	INPUT	Transformed...	-0.07998	0.079957	2
TRAIN	class	0	REP_gill_atta...	2	0	12179	0	6	2.293128	2.174735	0.437298	-1.30213	INPUT	Replacement...	0.066982	0.054983	1
TRAIN	class	1	REP_gill_atta...	1	0	14837	0	6	2.031004	2.255453	0.707473	-1.09187	INPUT	Replacement...	-0.05498	0.054983	2
TRAIN	class	0	REP_season	0.943195	0	12179	0.034508	1.804273	0.977807	0.358992	0.481499	2.73726	INPUT	Replacement...	0.026714	0.021929	1
TRAIN	class	1	REP_season	0.943195	0	14837	0.034508	1.804273	0.931481	0.250201	0.468899	9.30145	INPUT	Replacement...	-0.02193	0.021929	2

b. Full Regression

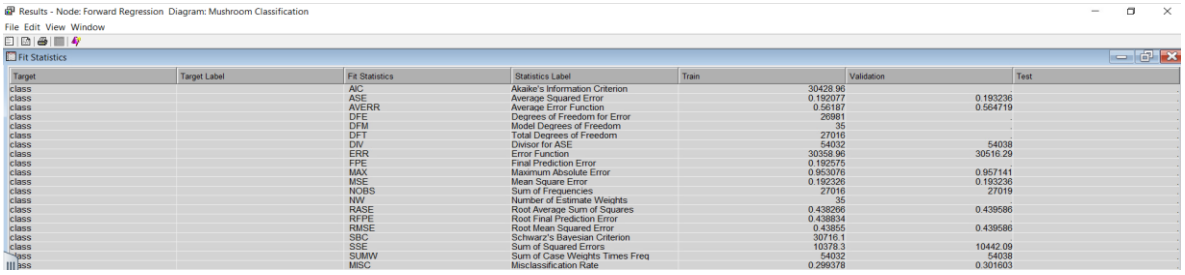
The first regression node introduced from the Model tab was the full regression model. This node was connected to the transform variables node to execute the analysis and evaluate the results. Upon reviewing the output from the full regression model, we observed that the ASE was 0.193236, which is higher than the previously achieved lowest ASE from the ASE Tree.



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
class		AIC	Akaike's Information Criterion	30428.96		
class		ASE	Average Squared Error	0.192077	0.193236	
class		AVERR	Average Error Function	0.56187	0.564719	
class		DFE	Degrees of Freedom for Error	26981		
class		DFM	Model Degrees of Freedom	35		
class		DFT	Total Degrees of Freedom	27016		
class		DIV	Divisor for ASE	54032	54038	
class		ERR	Error Function	30358.96	30516.29	
class		FPE	Final Prediction Error	0.192575		
class		MAX	Maximum Absolute Error	0.953076	0.957141	
class		MSE	Mean Square Error	0.192326	0.193236	
class		NOBS	Sum of Frequencies	27016	27019	
class		NW	Number of Estimate Weights	35		
class		RASE	Root Average Sum of Squares	0.438266	0.439586	
class		RFPE	Root Final Prediction Error	0.438834		
class		RMSE	Root Mean Squared Error	0.43855	0.439586	
class		SBC	Schwarz's Bayesian Criterion	30716.1		
class		SSE	Sum of Squared Errors	10378.3	10442.09	
class		SUMW	Sum of Case Weights Times Freq	54032	54038	
class		MISC	Misclassification Rate	0.299378	0.301603	

c. Forward Regression

The second regression node we brought was Forward Regression from the model tab and we connected it to transform variables. However, we did see that the ASE was the same as Full Regression. Our regression models were taking into account all the variables to determine the result.


Results - Node: Forward Regression Diagram: Mushroom Classification

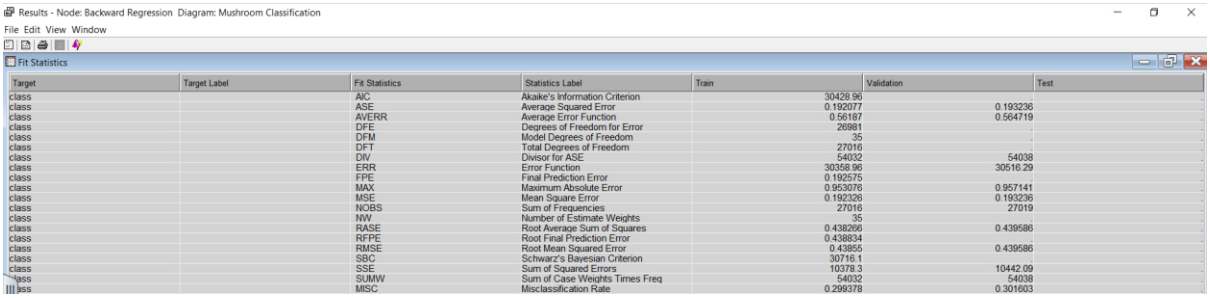
File Edit View Window

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
class		AIC	Akaike's Information Criterion	30428.96		
class		ASE	Average Squared Error	0.192077	0.193236	
class		AVERR	Average Error Function	0.56187	0.564719	
class		DFE	Degrees of Freedom for Error	26981		
class		DFM	Model Degrees of Freedom	35		
class		DFT	Total Degrees of Freedom	27016		
class		DIV	Divisor for ASE	54032	54038	
class		ERR	Error Function	30358.96	30516.29	
class		FPE	Final Prediction Error	0.192575		
class		MAX	Maximum Absolute Error	0.953076	0.957141	
class		MSE	Mean Square Error	0.192326	0.193236	
class		NOBS	Sum of Frequencies	27016	27019	
class		NW	Number of Estimate Weights	35		
class		RASE	Root Average Sum of Squares	0.438266	0.439586	
class		RFPE	Root Final Prediction Error	0.438834		
class		RMSE	Root Mean Squared Error	0.43855	0.439586	
class		SBC	Schwarz's Bayesian Criterion	30716.1		
class		SSE	Sum of Squared Errors	10378.3	10442.09	
class		SUMW	Sum of Case Weights Times Freq	54032	54038	
class		MISC	Misclassification Rate	0.299378	0.301603	

d. Backward Regression

Our backward regression also ended up having the same ASE as other models.


Results - Node: Backward Regression Diagram: Mushroom Classification

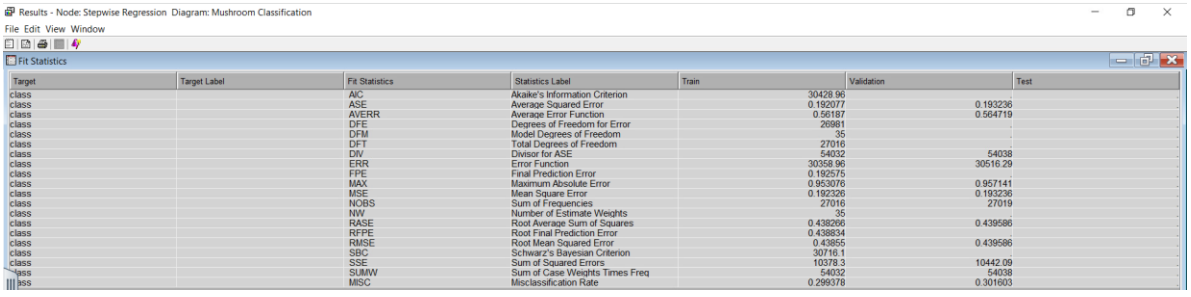
File Edit View Window

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
class		AIC	Akaike's Information Criterion	30428.96		
class		ASE	Average Squared Error	0.192077	0.193236	
class		AVERR	Average Error Function	0.56187	0.564719	
class		DFE	Degrees of Freedom for Error	26981		
class		DFM	Model Degrees of Freedom	35		
class		DFT	Total Degrees of Freedom	27016		
class		DIV	Divisor for ASE	54032	54038	
class		ERR	Error Function	30358.96	30516.29	
class		FPE	Final Prediction Error	0.192575		
class		MAX	Maximum Absolute Error	0.953076	0.957141	
class		MSE	Mean Square Error	0.192326	0.193236	
class		NOBS	Sum of Frequencies	27016	27019	
class		NW	Number of Estimate Weights	35		
class		RASE	Root Average Sum of Squares	0.438266	0.439586	
class		RFPE	Root Final Prediction Error	0.438834		
class		RMSE	Root Mean Squared Error	0.43855	0.439586	
class		SBC	Schwarz's Bayesian Criterion	30716.1		
class		SSE	Sum of Squared Errors	10378.3	10442.09	
class		SUMW	Sum of Case Weights Times Freq	54032	54038	
class		MISC	Misclassification Rate	0.299378	0.301603	

e. Stepwise Regression

Finally, our stepwise regression also showed the same ASE as the other models**.**


Results - Node: Stepwise Regression Diagram: Mushroom Classification

File Edit View Window

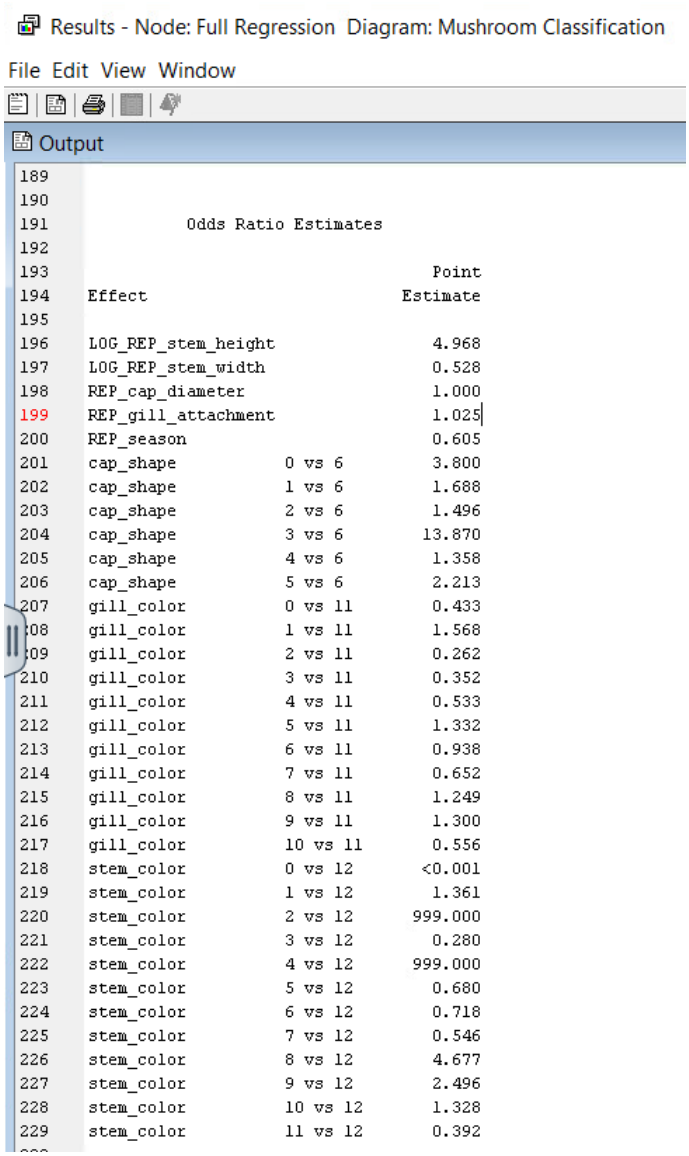
Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
class		AIC	Akaike's Information Criterion	30428.96		
class		ASE	Average Squared Error	0.192077	0.193236	
class		AVERR	Average Error Function	0.56187	0.564719	
class		DFE	Degrees of Freedom for Error	26981		
class		DFM	Model Degrees of Freedom	35		
class		DFT	Total Degrees of Freedom	27016		
class		DIV	Divisor for ASE	54032	54038	
class		ERR	Error Function	30358.96	30516.29	
class		FPE	Final Prediction Error	0.192575		
class		MAX	Maximum Absolute Error	0.953076	0.957141	
class		MSE	Mean Square Error	0.192326	0.193236	
class		NOBS	Sum of Frequencies	27016	27019	
class		NW	Number of Estimate Weights	35		
class		RASE	Root Average Sum of Squares	0.438266	0.439586	
class		RFPE	Root Final Prediction Error	0.438834		
class		RMSE	Root Mean Squared Error	0.43855	0.439586	
class		SBC	Schwarz's Bayesian Criterion	30716.1		
class		SSE	Sum of Squared Errors	10378.3	10442.09	
class		SUMW	Sum of Case Weights Times Freq	54032	54038	
class		MISC	Misclassification Rate	0.299378	0.301603	

Odds Ratio Interpretation

Since our regression models produced consistent results across all analyses, the odds ratio values were also similar, indicating a clean dataset. These odds ratios allow us to identify the relative importance and impact of each predictor variable in determining whether a mushroom is classified as poisonous or edible.

Below, we present the key insights derived from the node results, highlighting the specific variables that significantly influence the classification and their corresponding effects on the outcome.



Odds Ratio Estimates

Effect		Point Estimate
LOG_REP_stem_height		4.968
LOG_REP_stem_width		0.528
REP_cap_diameter		1.000
REP_gill_attachment		1.025
REP_season		0.605
cap_shape	0 vs 6	3.800
cap_shape	1 vs 6	1.688
cap_shape	2 vs 6	1.496
cap_shape	3 vs 6	13.870
cap_shape	4 vs 6	1.358
cap_shape	5 vs 6	2.213
gill_color	0 vs 11	0.433
gill_color	1 vs 11	1.568
gill_color	2 vs 11	0.262
gill_color	3 vs 11	0.352
gill_color	4 vs 11	0.533
gill_color	5 vs 11	1.332
gill_color	6 vs 11	0.938
gill_color	7 vs 11	0.652
gill_color	8 vs 11	1.249
gill_color	9 vs 11	1.300
gill_color	10 vs 11	0.556
stem_color	0 vs 12	<0.001
stem_color	1 vs 12	1.361
stem_color	2 vs 12	999.000
stem_color	3 vs 12	0.280
stem_color	4 vs 12	999.000
stem_color	5 vs 12	0.680
stem_color	6 vs 12	0.718
stem_color	7 vs 12	0.546
stem_color	8 vs 12	4.677
stem_color	9 vs 12	2.496
stem_color	10 vs 12	1.328
stem_color	11 vs 12	0.392

- **Stem Height**

  A 2.718-inch increase in the stem height increases the probability of the mushroom being poisonous by **396%**

- **Stem Width**

  A 2.178-inch increase in the stem width decreases the probability of the mushroom being poisonous by **47.2%**.

- **Cap Diameter**

  A unit change in cap diameter has no significant effect on the probability of being classified as poisonous.

- **Gill Attachment**

  A unit increase in gill attachment increases the probability of the mushroom being poisonous by 2.5%.

- **Season**

  A unit increase in season variable decreases the probability of the mushroom being poisonous by 39.5%.

  *It should be noted however; we did not have enough information to deduce the season variable from the decimals that were provided in the dataset.

- **Cap Shape**

Comparison	**Interpretation**
0 vs 6	Mushrooms with cap shape 0 are 2.8 times more likely to be poisonous compared to cap shape 6
1 vs 6	Mushrooms with cap shape 1 are 68% more likely to be poisonous compared to cap shape 6.
2 vs 6	Mushrooms with cap shape 2 are 50% more likely to be poisonous compared to cap shape 6.

3 vs 6	Mushrooms with cap shape 3 are 12.87 times (1287%) more likely to be poisonous compared to cap shape 6 (a very strong effect).
4 vs 6	Mushrooms with cap shape 4 are 36% more likely to be poisonous compared to cap shape 6.
5 vs 6	Mushrooms with cap shape 5 are 121% more likely to be poisonous compared to cap shape 6.

*It should be noted that we did not have enough information on the specific metrics for the cap shapes 1-6 from the dataset

- **Gill Colour**

Comparison	**Interpretation**
0 vs 11	Mushrooms with gill color 0 are 56.7% less likely to be poisonous compared to gill color 11.
1 vs 11	Mushrooms with gill color 1 are 56.8% more likely to be poisonous compared to gill color 11.
2 vs 11	Mushrooms with gill color 2 are 73.8% less likely to be poisonous compared to gill color 11.
3 vs 11	Mushrooms with gill color 3 are 64.8% less likely to be poisonous compared to gill color 11.
4 vs 11	Mushrooms with gill color 4 are 46.7% less likely to be poisonous compared to gill color 11.
5 vs 11	Mushrooms with gill color 5 are 33% more likely to be poisonous compared to gill color 11.

6 vs 11	Mushrooms with gill color 6 are 6.2% less likely to be poisonous compared to gill color 11.
7 vs 11	Mushrooms with gill color 7 are 34.8% less likely to be poisonous compared to gill color 11.
8 vs 11	Mushrooms with gill color 8 are 25% more likely to be poisonous compared to gill color 11.
8 vs 11	Mushrooms with gill color 9 are 30% more likely to be poisonous compared to gill color 11.
10 vs 11	Mushrooms with gill color 10 are 44.4% less likely to be poisonous compared to gill color 11.

*It should be noted that we did not have enough information to deduce the gill colors 1-11 from the dataset

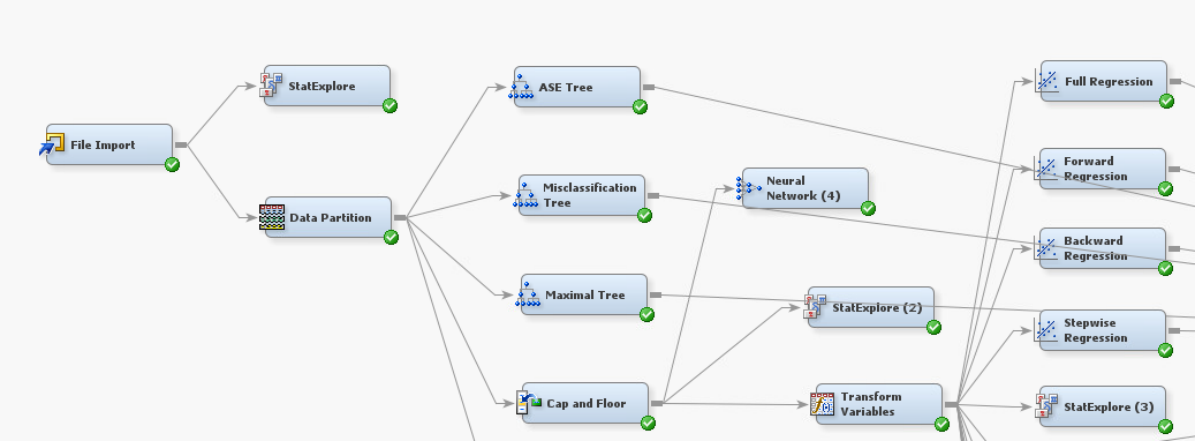
- **Stem Color**

Comparison	**Interpretation**
0 vs 12	Mushrooms with stem color 0 have an extremely low probability of being poisonous compared to stem color 12.
1 vs 12	Mushrooms with stem color 1 are 36% more likely to be poisonous compared to stem color 12.
2 vs 12	Mushrooms with stem color 2 have an extremely high probability of being poisonous compared to stem color 12.
3 vs 12	Mushrooms with stem color 3 are 72% less likely to be classified as poisonous compared to stem color 12.

4 vs 12	Mushrooms with stem color 4 have an extremely high probability of being poisonous compared to stem color 12.
5 vs 12	Mushrooms with stem color 5 are 32% less likely to be poisonous compared to stem color 12.
6 vs 12	Mushrooms with stem color 6 are 28.2% less likely to be poisonous compared to stem color 12.
7 vs 12	Mushrooms with stem color 7 are 45.4% less likely to be poisonous compared to stem color 12.
8 vs 12	Mushrooms with stem color 8 are 367% (3.677x) more likely to be poisonous compared to stem color 12.
8 vs 12	Mushrooms with stem color 9 are 150% (1.496x) more likely to be poisonous compared to stem color 12.
10 vs 12	Mushrooms with stem color 10 are 33% times more likely to be poisonous compared to stem color 12.
11 vs 12	Mushrooms with stem color 11 are 60.8% less likely to be poisonous compared to stem color 12.

Regression Analysis

Based on the results of the four regression analyses, we conclude that no single model outperformed the others in terms of achieving the lowest ASE, as the ASE values were identical across all models. All variables included in the analyses were statistically significant. Consequently, we retained the full regression model as the primary framework for interpreting the odds ratios, which serve to identify whether mushrooms are poisonous or safe for consumption.



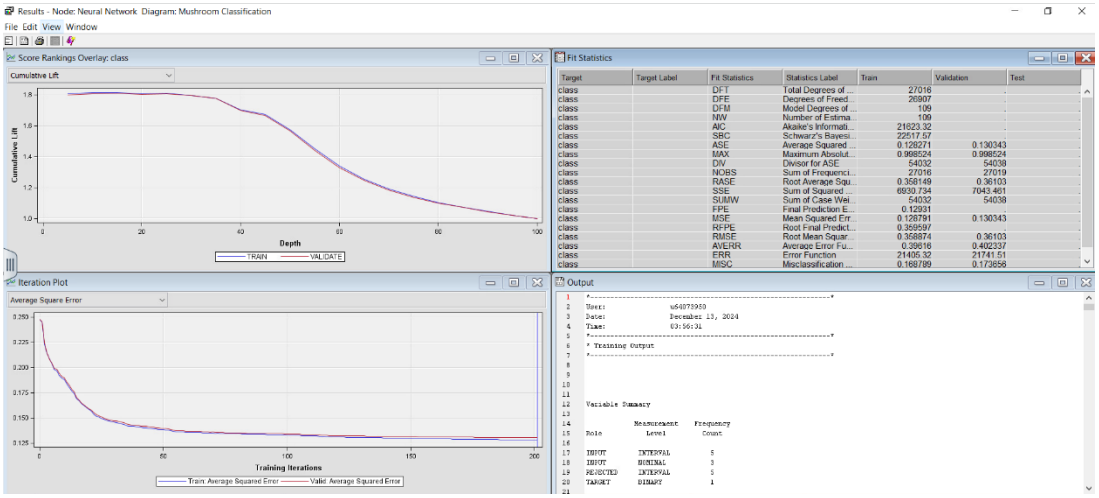
3. Neural Network

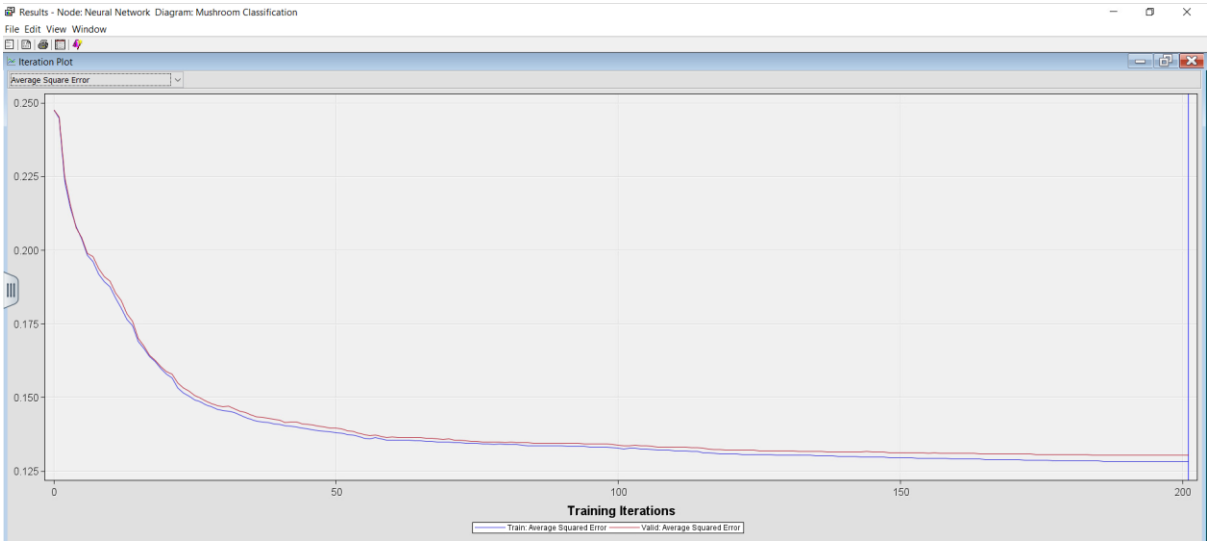
a. **Transform Variables**

The next step in our project would be to run NEURAL NETWORKS. We brought in our first neural network node from the model tab, and we connected it to transform variables. From the results we found the ASE for this to be 0.130343.

The maximum iterations were set as 1000 and as we can see from the graph below that after about 200 iterations the graph starts to converge. There is also no overfitting of data. However, since all our regression models had the same ASE and so far, our ASE tree had the best output, we observed that in the best output the variable ‘season’ was not accounted for.

So, moving forward we conduct neural network models without the season variable as well to test which gives us the better result.





Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
class		DFT	Total Degrees of Freedom	27016		
class		DFE	Degrees of Freedom for Error	26907		
class		DFM	Model Degrees of Freedom	109		
class		NW	Number of Estimated Weights	109		
class		AIC	Akaike's Information Criterion	21623.32		
class		SBC	Schwarz's Bayesian Criterion	22517.57		
class		ASE	Average Squared Error	0.128271	0.130343	0.998524
class		MAX	Maximum Absolute Error	0.998524	0.998524	54038
class		DIV	Divisor for ASE	54032	54038	27019
class		NOBS	Sum of Frequencies	27016	27019	0.36103
class		RASE	Root Average Squared Error	0.358149	0.36103	7043.461
class		SSE	Sum of Squared Errors	6930.734	7043.461	54038
class		SUMW	Sum of Case Weights Times Freq	54032	54038	
class		FPE	Final Prediction Error	0.12931		
class		MSE	Mean Squared Error	0.128791	0.130343	
class		RFPE	Root Final Prediction Error	0.359597		
class		RMSE	Root Mean Squared Error	0.358874	0.36103	
class		AVERR	Average Error Function	0.39616	0.402337	
class		ERR	Error Function	21405.32	21741.51	
class		MISC	Misclassification Rate	0.168789	0.173656	
class		WRONG	Number of Wrong Classifications	4560	4692	

Results - Node: ASE Tree Diagram: Mushroom Classification

File Edit View Window

Output

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
gill_color	gill-color	11	1.0000	1.0000	1.0000
stem_width	stem-width	9	0.9909	0.9935	1.0026
stem_color	stem-color	7	0.7968	0.8093	1.0157
gill_attachment	gill-attachment	5	0.7601	0.7578	0.9970
cap_shape	cap-shape	6	0.6650	0.6782	1.0199
stem_height	stem-height	4	0.5796	0.5621	0.9698
cap_diameter	cap-diameter	7	0.4594	0.4615	1.0044
season		2	0.1552	0.1253	0.8074

Variables - Neural11

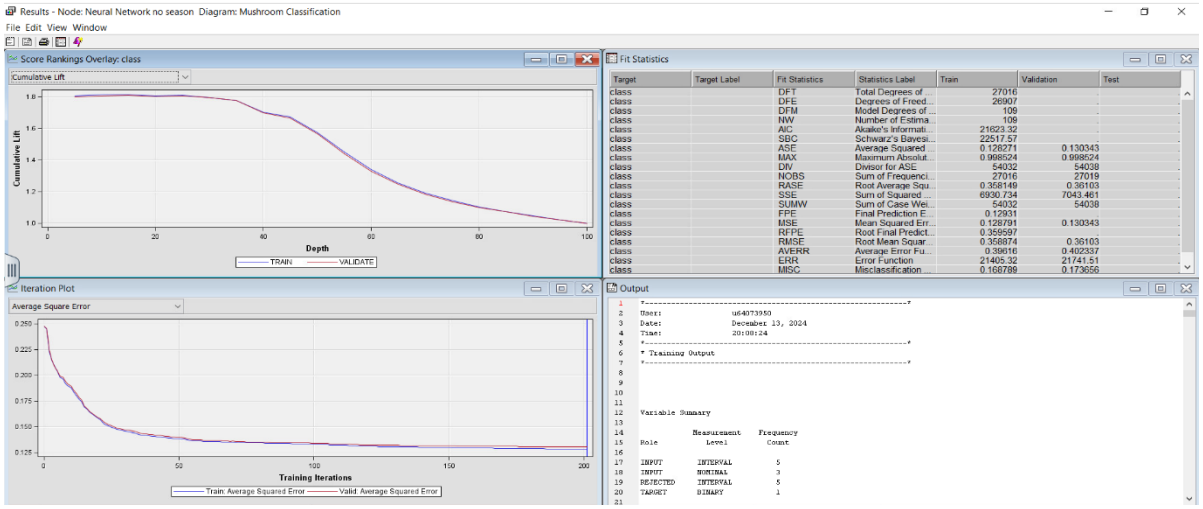
(none) not Equal to

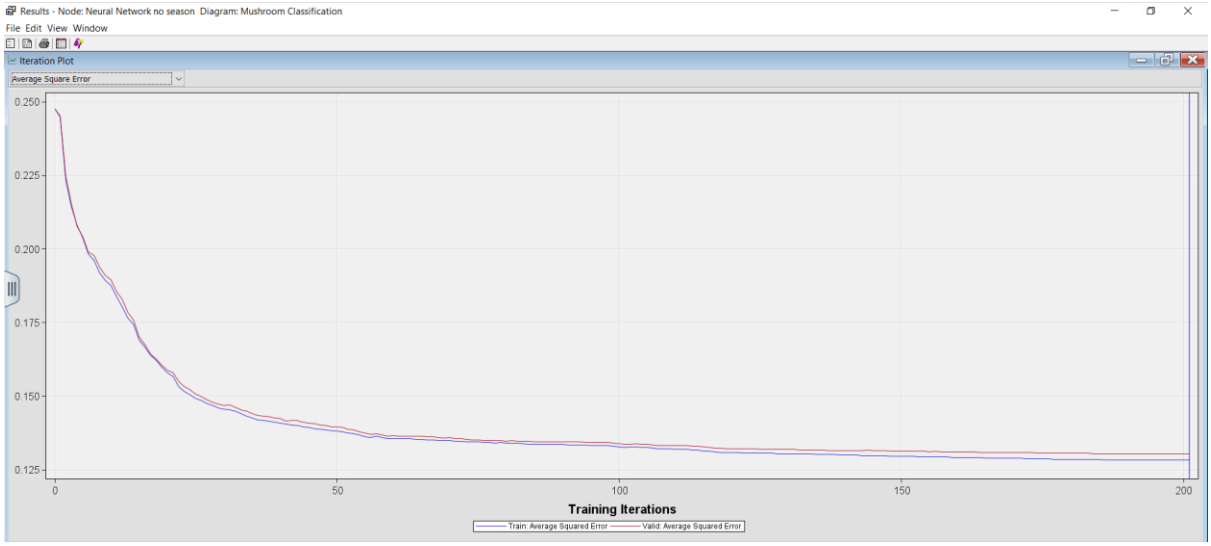
Columns: Label

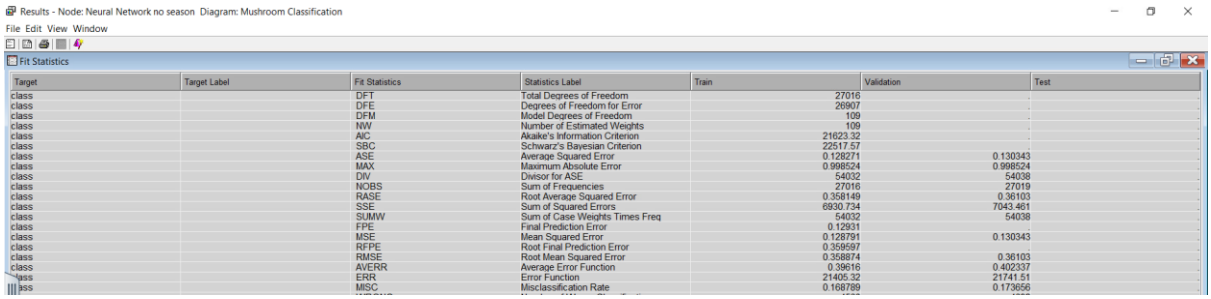
Name	Use	Report	Role	Level
LOG_REP_stem	Default	No	Input	Interval
LOG_REP_stem	Default	No	Input	Interval
REP_cap_diam	Default	No	Input	Interval
REP_gill_attach	Default	No	Input	Interval
REP_season	Default	No	Input	Interval
cap_diameter	Default	No	Rejected	Interval
cap_shape	Default	No	Input	Nominal
class	Yes	No	Target	Binary
gill_attachmen	Default	No	Rejected	Interval
gill_color	Default	No	Input	Nominal
season	No	No	Rejected	Interval
stem_color	Default	No	Input	Nominal
stem_height	Default	No	Rejected	Interval
stem_width	Default	No	Rejected	Interval

Neural Network No Season – Transform Variables

In this neural network, we ran the model after removing the season variable to observe how the model gets affected. However, after running the node we see that the ASE remained 0.130343.



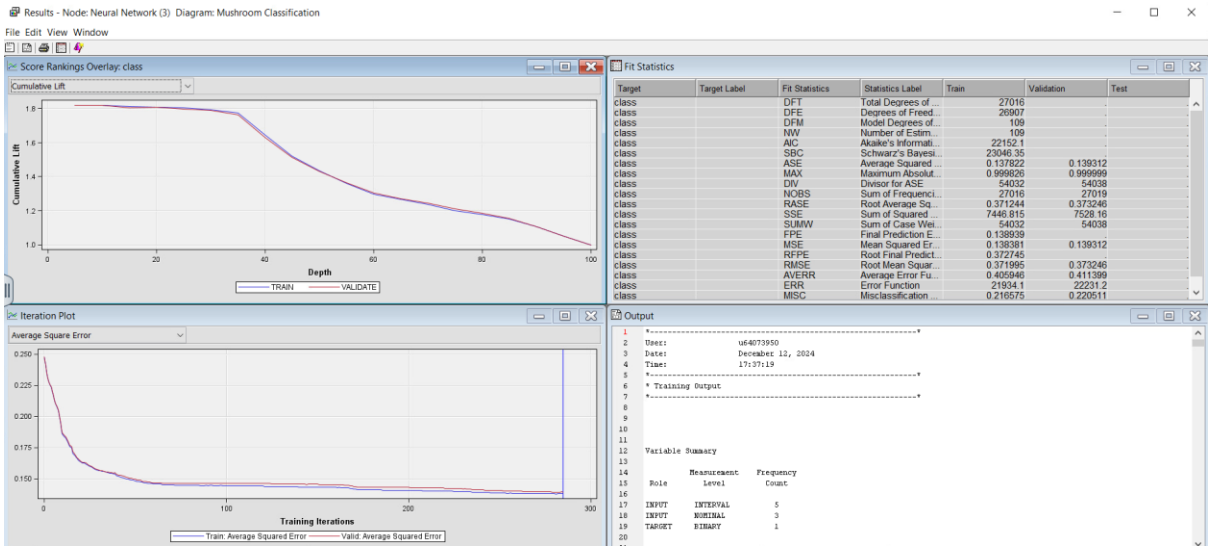


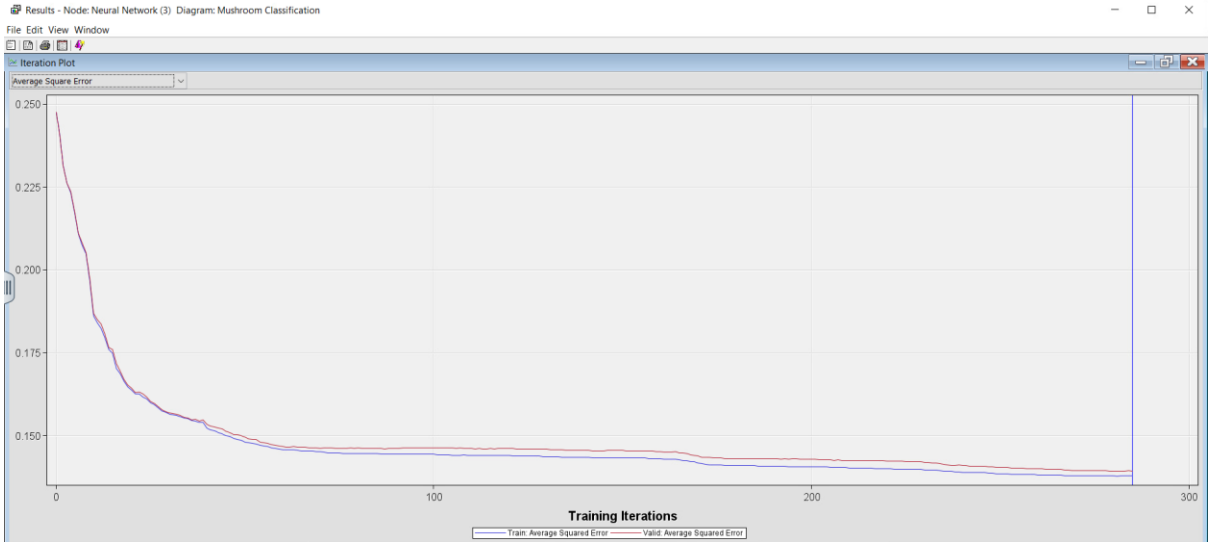


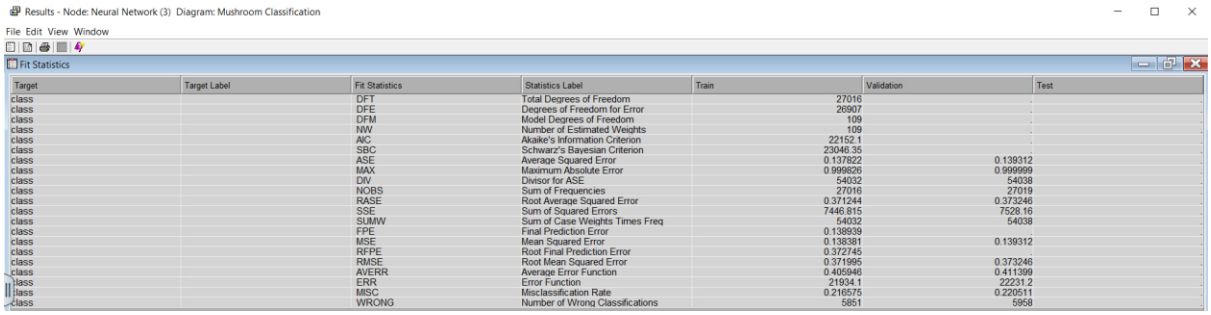
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
class		DFT	Total Degrees of Freedom	27016		
class		DFE	Degrees of Freedom for Error	26907		
class		DFM	Model Degrees of Freedom	109		
class		NW	Number of Estimated Weights	109		
class		AIC	Akaike's Information Criterion	21623.32		
class		SBC	Schwarz's Bayesian Criterion	22517.57		
class		ASE	Average Squared Error	0.128271	0.130343	
class		MAX	Maximum Absolute Error	0.998524	0.998524	
class		DIV	Divisor for ASE	54032	54038	
class		NOBS	Sum of Frequencies	27016	27019	
class		RASE	Root Average Squared Error	0.358149	0.36103	
class		SSE	Sum of Squared Errors	6930.734	7043.461	
class		SUMW	Sum of Case Weights Times Freq	54032	54038	
class		FPE	Final Prediction Error	0.12931		
class		MSE	Mean Squared Error	0.128791	0.130343	
class		RFPE	Root Final Prediction Error	0.359597		
class		RMSE	Root Mean Squared Error	0.358874	0.36103	
class		AVERR	Average Error Function	0.39616	0.402337	
class		ERR	Error Function	21405.32	21741.51	
class		MISC	Misclassification Rate	0.166789	0.173656	

Neural Network – Data Partition

In this neural network, we connected the node to data partition and observed the change. The ASE seems to have gone up to 0.139312 so has the number of maximum iterations. The **Iteration Plot** shows the model converges around **150-200 iterations** but continues up until ~ 270.



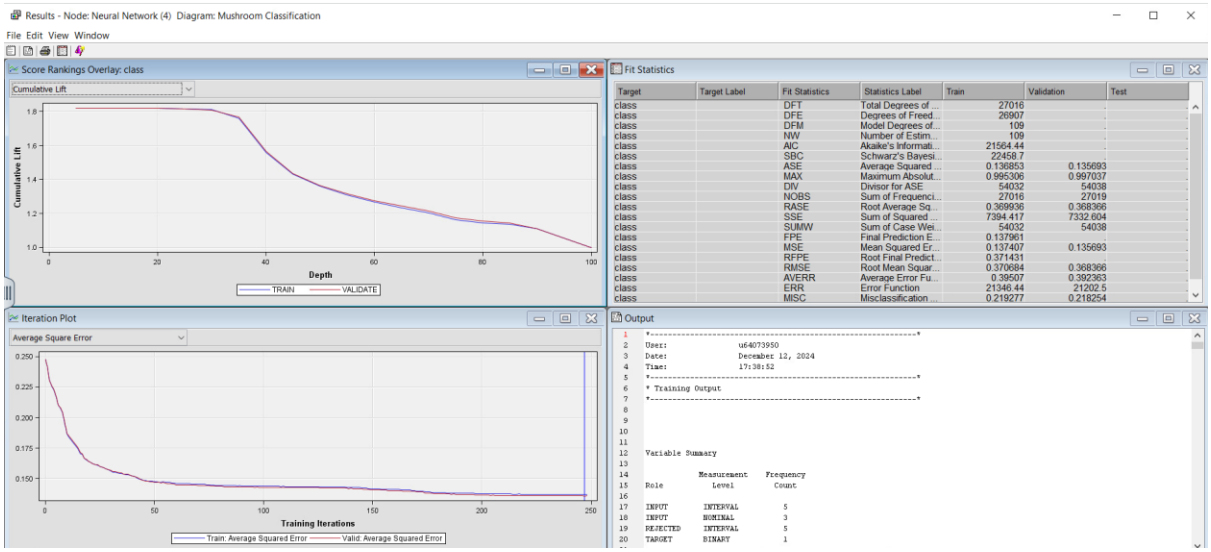


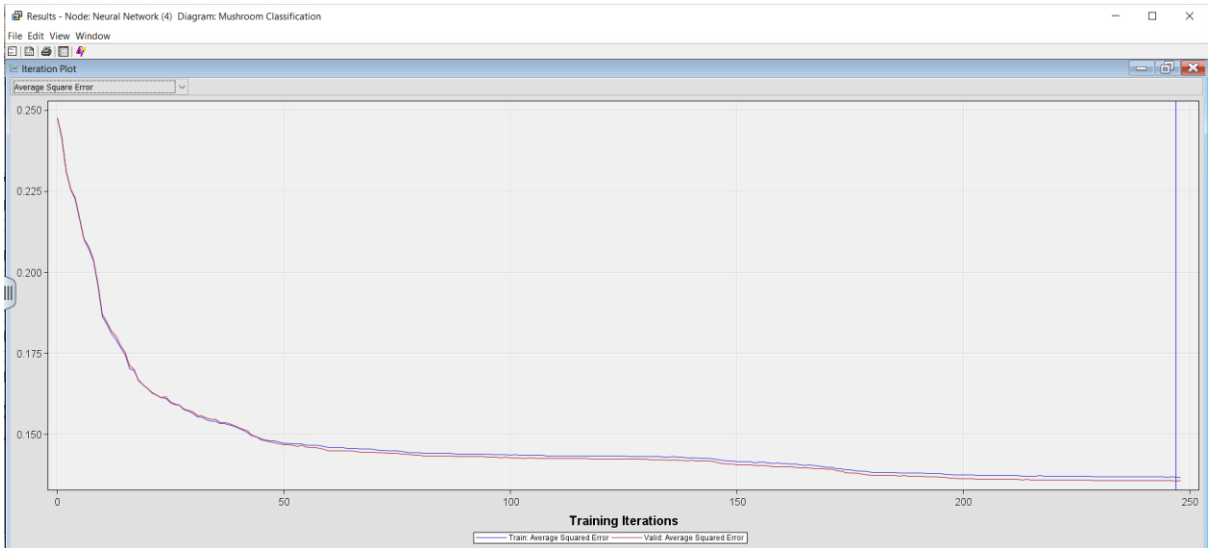


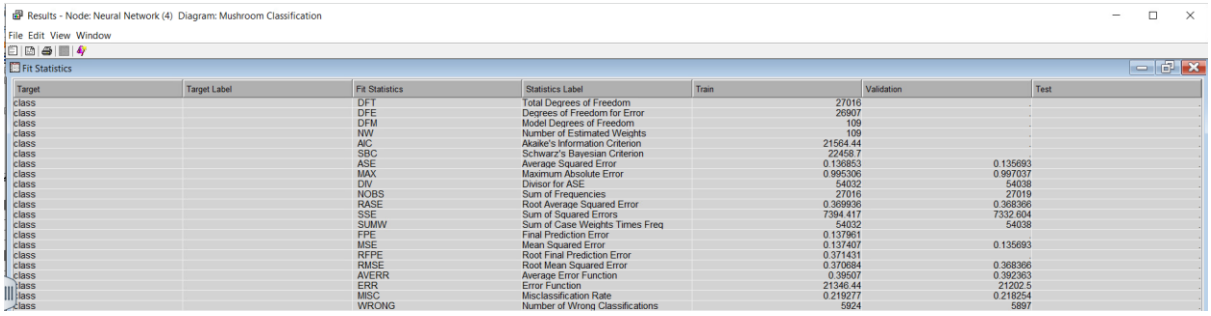
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
class		DFT	Total Degrees of Freedom	27016		
class		DFE	Degrees of Freedom for Error	26907		
class		DFM	Model Degrees of Freedom	109		
class		NW	Number of Estimated Weights	109		
class		AIC	Akaike's Information Criterion	22152.1		
class		SBC	Schwarz's Bayesian Criterion	23046.35		
class		ASE	Average Squared Error	0.137822	0.139312	
class		MAX	Maximum Absolute Error	0.999826	0.999999	
class		DIV	Divisor for ASE	54032	54038	
class		NOBS	Sum of Frequencies	27016	27019	
class		RASE	Root Average Squared Error	0.371244	0.373246	
class		SSE	Sum of Squared Errors	7446.815	7528.16	
class		SUMW	Sum of Case Weights Times Freq	54032	54038	
class		FPE	Final Prediction Error	0.138939		
class		MSE	Mean Squared Error	0.138381	0.139312	
class		RFPE	Root Final Prediction Error	0.372745		
class		RMSE	Root Mean Squared Error	0.371995	0.373246	
class		AVERR	Average Error Function	0.405946	0.411399	
class		ERR	Error Function	21934.1	22231.2	
class		MISC	Misclassification Rate	0.216575	0.220511	
class		WRONG	Number of Wrong Classifications	5851	5958	

Neural Network – Cap and Floor

In our last neural network, we see the ASE stands at 0.135693 and the number of iterations has reduced to ~240. This shows an improvement in our ASE and accuracy in predicting the classification of the mushrooms, but it is still not our best model yet.

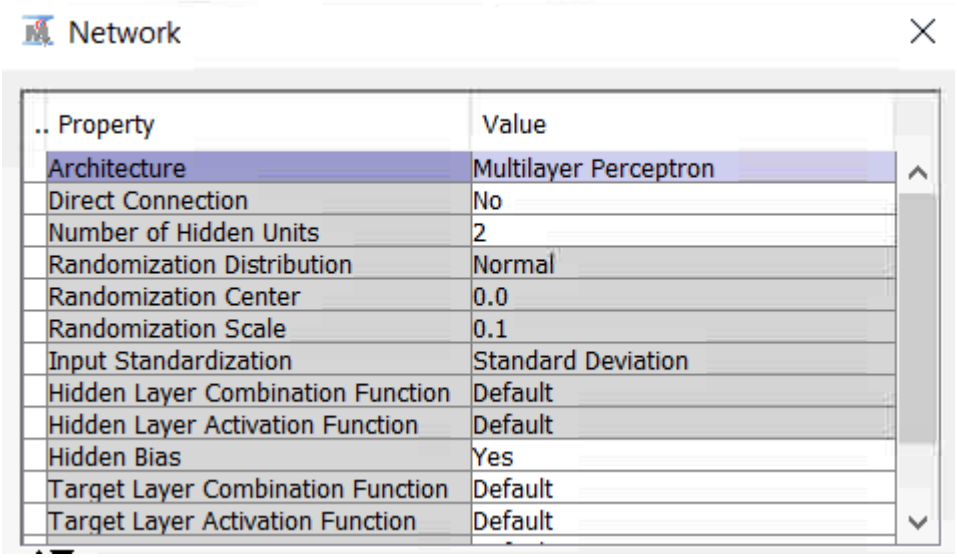

Results - Node: Neural Network (4) Diagram: Mushroom Classification
File Edit View Window
Score Rankings Overlay: class
Cumulative Lift
Cumulative Lift
Depth
TRAIN
VALIDATE
Fit Statistics
Target
Target Label
Fit Statistics
Statistics Label
Train
Validation
Test
class DFT Total Degrees of ... 27016
class DFE Degrees of Freed... 26907
class DFM Model Degrees of... 109
class NW Number of Estim... 109
class AIC Akaike's Informati... 21564.44
class SBC Schwarz's Bayesi... 22458.7
class ASE Average Squared ... 0.136853 0.135693
class MAX Maximum Absolut... 0.995306 0.997037
class DIV Divisor for ASE 54032 54038
class NOBS Sum of Frequenci... 27016 27019
class RASE Root Average Sq... 0.369936 0.368366
class SSE Sum of Squared ... 7394.417 7332.604
class SUMW Sum of Case Wei... 54032 54038
class FPE Final Prediction E... 0.137961
class MSE Mean Squared Er... 0.137407 0.135693
class RFPE Root Final Predict... 0.371431
class RMSE Root Mean Squar... 0.370684 0.368366
class AVERR Average Error Fu... 0.39507 0.392363
class ERR Error Function 21346.44 21202.5
class MISC Misclassification ... 0.219277 0.218254
Iteration Plot
Average Square Error
Training Iterations
Train: Average Squared Error
Valid: Average Squared Error
Output
User: u64073950
Date: December 12, 2024
Time: 17:38:52
* Training Output
Variable Summary
Role Measurement Level Frequency Count
INPUT INTERVAL 5
INPUT NOMINAL 3
REJECTED INTERVAL 5
TARGET BINARY 1

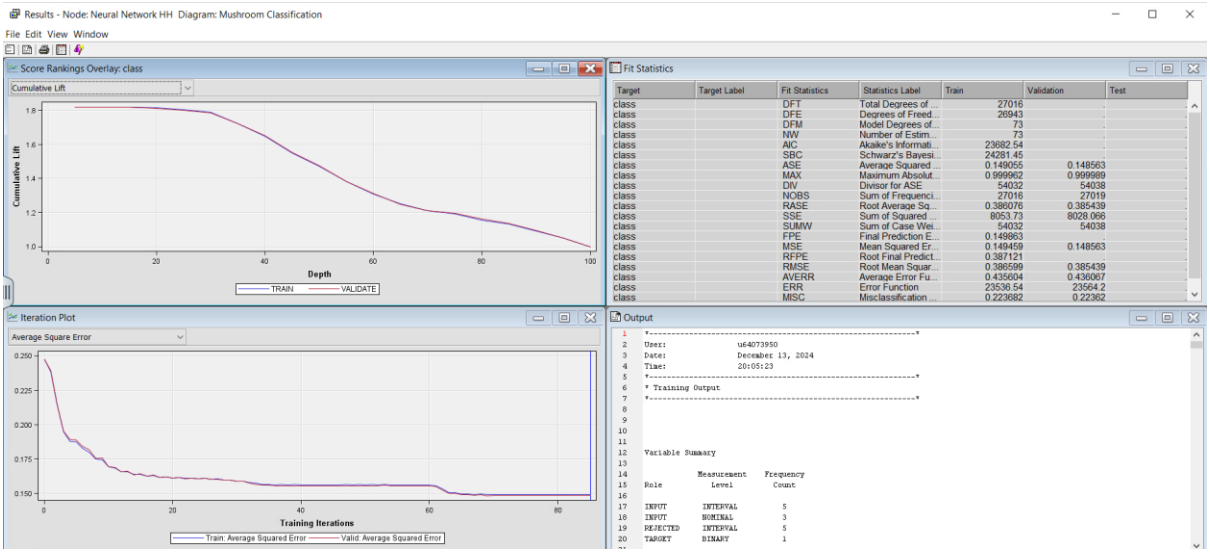

Results - Node: Neural Network (4) Diagram: Mushroom Classification
File Edit View Window
Iteration Plot
Average Square Error
0.250
0.225
0.200
0.175
0.150
0
50
100
150
200
250
Training Iterations
Train: Average Squared Error
Valid: Average Squared Error

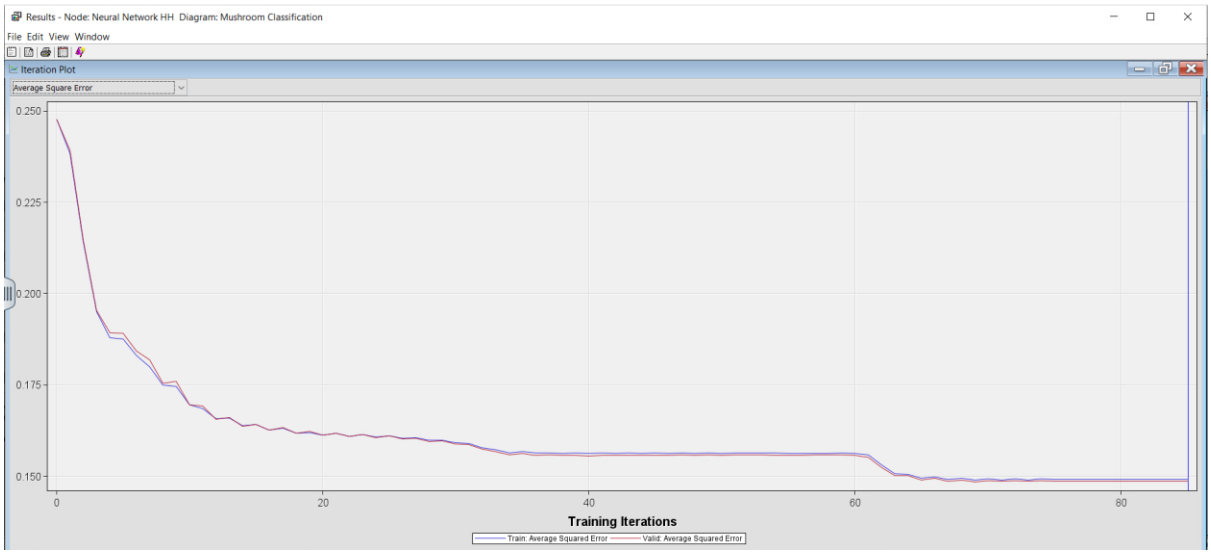

Results - Node: Neural Network (4) Diagram: Mushroom Classification
File Edit View Window
Fit Statistics
Target Target Label Fit Statistics Statistics Label Train Validation Test
class DFT Total Degrees of Freedom 27016
class DFE Degrees of Freedom for Error 26907
class DFM Model Degrees of Freedom 109
class NW Number of Estimated Weights 109
class AIC Akaike's Information Criterion 21564.44
class SBC Schwarz's Bayesian Criterion 22458.7
class ASE Average Squared Error 0.136853 0.135693
class MAX Maximum Absolute Error 0.995306 0.997037
class DIV Divisor for ASE 54032 54038
class NOBS Sum of Frequencies 27016 27019
class RASE Root Average Squared Error 0.369936 0.368366
class SSE Sum of Squared Errors 7394.417 7332.604
class SUMW Sum of Case Weights Times Freq 54032 54038
class FPE Final Prediction Error 0.137961
class MSE Mean Squared Error 0.137407 0.135693
class RFPE Root Final Prediction Error 0.371431
class RMSE Root Mean Squared Error 0.370684 0.368366
class AVERR Average Error Function 0.39507 0.392363
class ERR Error Function 21346.44 21202.5
class MISC Misclassification Rate 0.219277 0.218254
class WRONG Number of Wrong Classifications 5924 5897

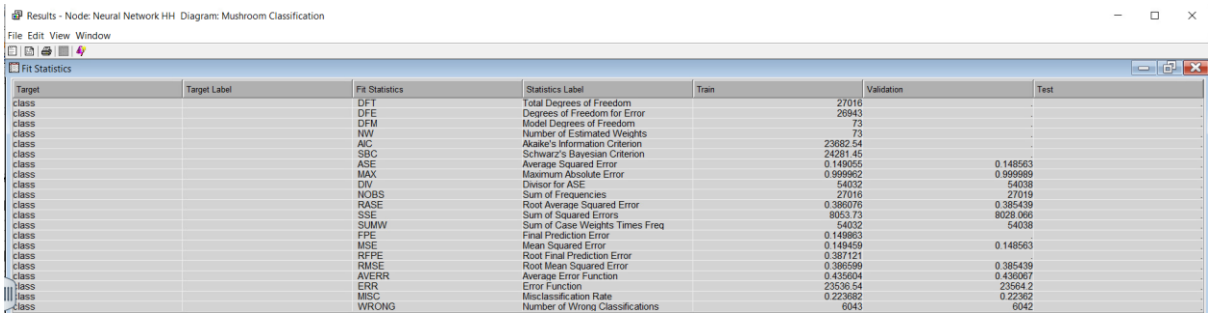
Neural Network No Season - HH

From this point onward, we run 6 neural networks with different hidden units. The default hidden unit is 3 which have been conducted and with attached results above. However, we also must run with hidden units of – 2,4,5,6,7,8 to determine the best output. We are going to monitor the ASE for each of them and see if it increases or decreases, giving us an indication of which model gives us the best output. These neural networks are also run without the 'season' variable because so far, we have had our best output without this variable.





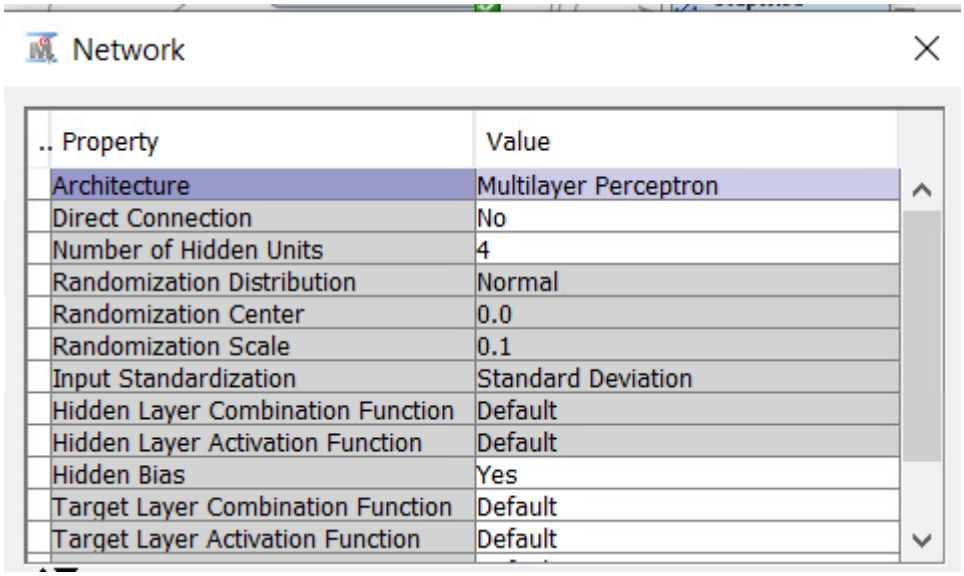




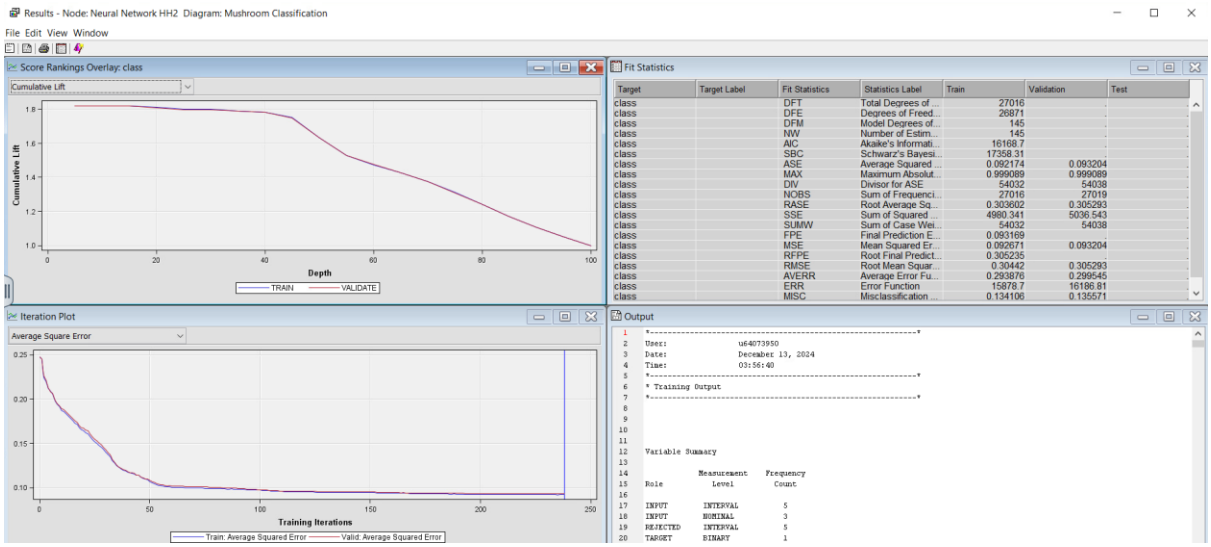
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
class		DFT	Total Degrees of Freedom	27016		
class		DFE	Degrees of Freedom for Error	26943		
class		DFM	Model Degrees of Freedom	73		
class		NW	Number of Estimated Weights	73		
class		AIC	Akaike's Information Criterion	23682.54		
class		SBC	Schwarz's Bayesian Criterion	24281.45		
class		ASE	Average Squared Error	0.149055		0.148563
class		MAX	Maximum Absolute Error	0.999962		0.999989
class		DIV	Divisor for ASE	54032		54038
class		NOBS	Sum of Frequencies	27016		27019
class		RASE	Root Average Squared Error	0.386076		0.385439
class		SSE	Sum of Squared Errors	8053.73		8028.066
class		SUMW	Sum of Case Weights Times Freq	54032		54038
class		FPE	Final Prediction Error	0.149863		
class		MSE	Mean Squared Error	0.149459		0.148563
class		RFPE	Root Final Prediction Error	0.387121		
class		RMSE	Root Mean Squared Error	0.386599		0.385439
class		AVERR	Average Error Function	0.435604		0.436067
class		ERR	Error Function	23536.54		23564.2
class		MISC	Misclassification Rate	0.223682		0.22362
class		WRONG	Number of Wrong Classifications	6043		6042

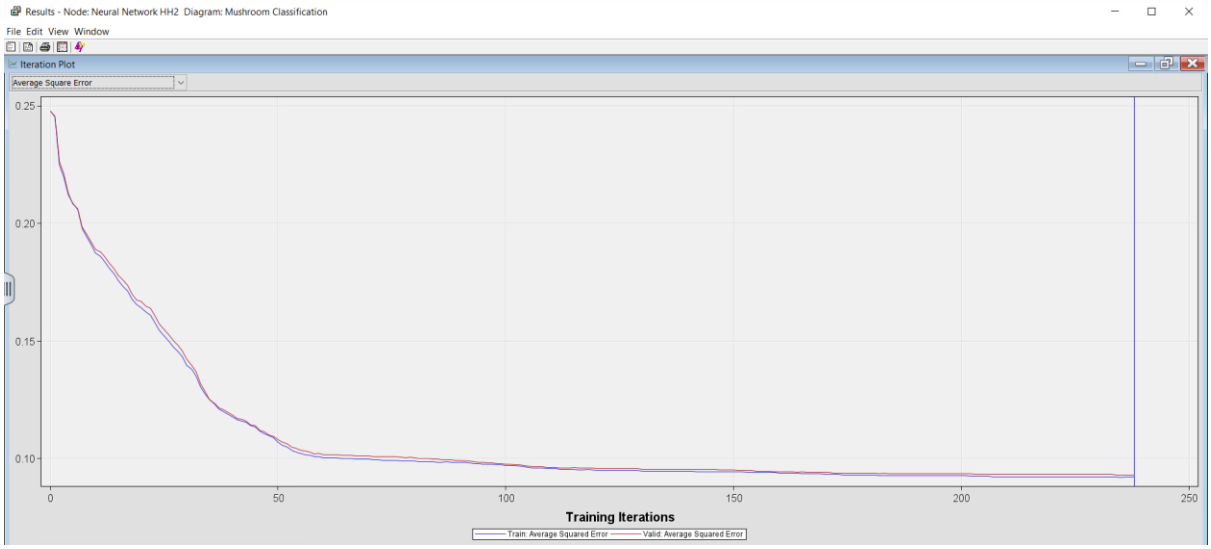
Neural Network No Season – HH2

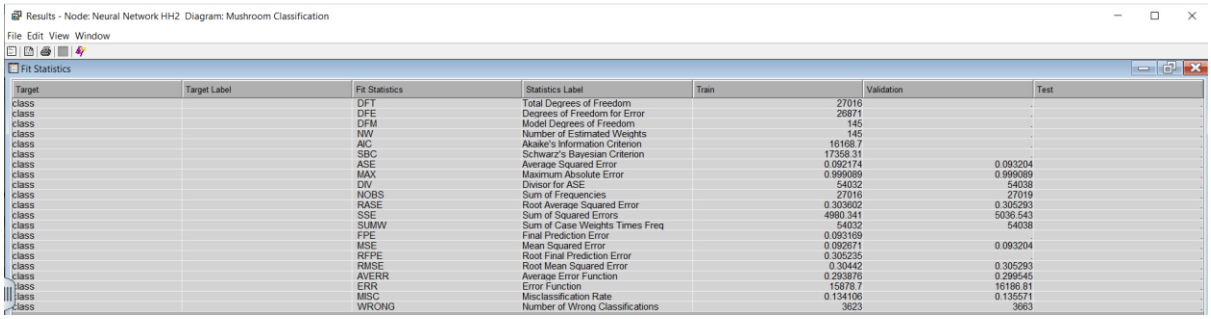
Neural Network with 4 hidden units and no season variable had **ASE – 0.093204**



.. Property	Value
Architecture	Multilayer Perceptron
Direct Connection	No
Number of Hidden Units	4
Randomization Distribution	Normal
Randomization Center	0.0
Randomization Scale	0.1
Input Standardization	Standard Deviation
Hidden Layer Combination Function	Default
Hidden Layer Activation Function	Default
Hidden Bias	Yes
Target Layer Combination Function	Default
Target Layer Activation Function	Default



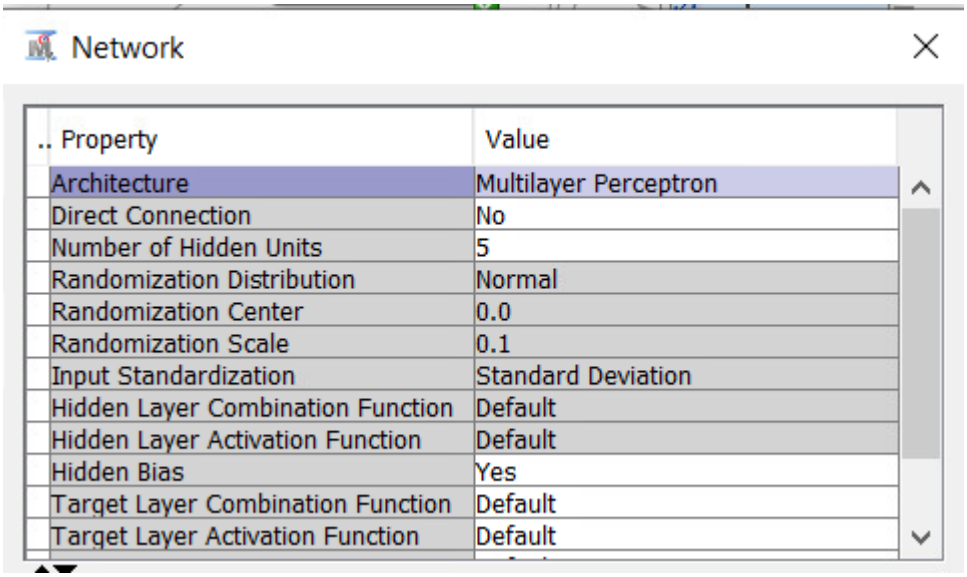


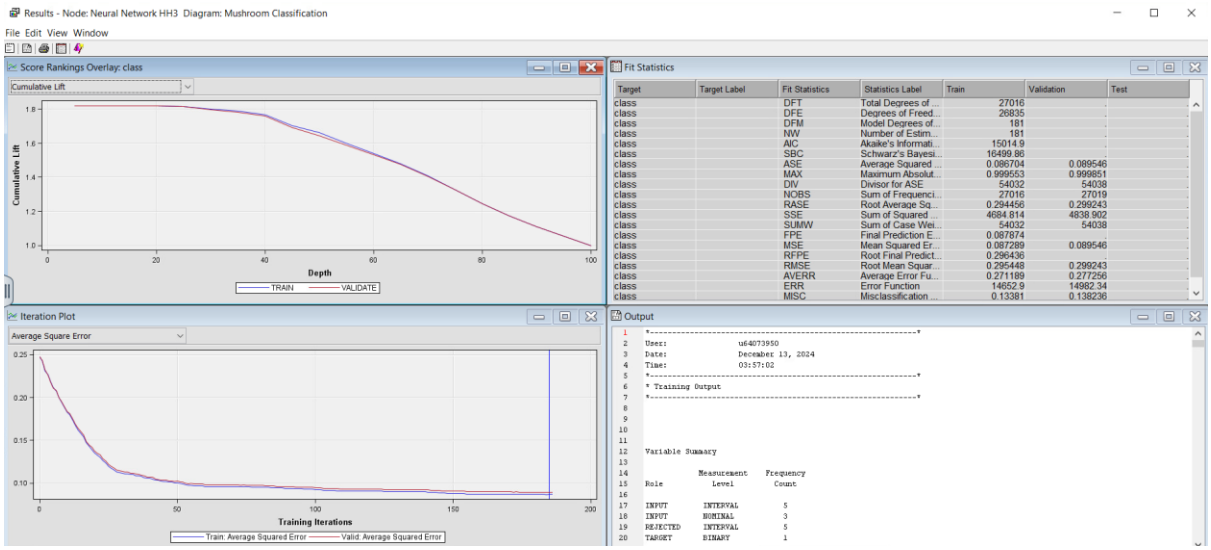


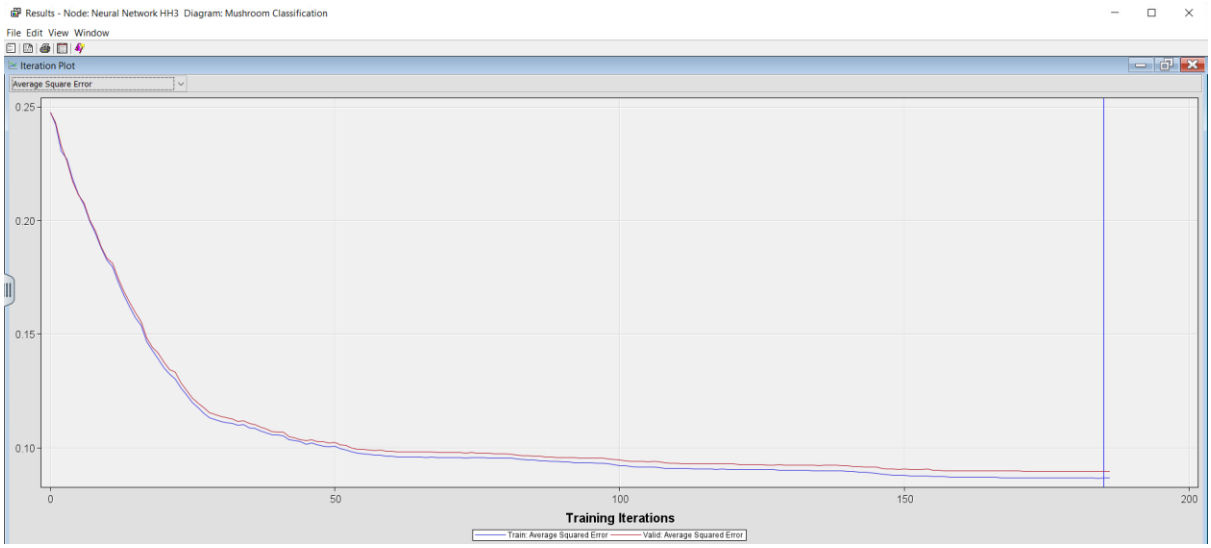
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
class		DFT	Total Degrees of Freedom	27016		
class		DFE	Degrees of Freedom for Error	26871		
class		DFM	Model Degrees of Freedom	145		
class		NW	Number of Estimated Weights	145		
class		AIC	Akaike's Information Criterion	16168.7		
class		SBC	Schwarz's Bayesian Criterion	17358.31		
class		ASE	Average Squared Error	0.092174	0.093204	
class		MAX	Maximum Absolute Error	0.999089	0.999089	
class		DIV	Divisor for ASE	54032	54038	
class		NOBS	Sum of Frequencies	27016	27019	
class		RASE	Root Average Squared Error	0.303602	0.305293	
class		SSE	Sum of Squared Errors	4980.341	5036.543	
class		SUMW	Sum of Case Weights Times Freq	54032	54038	
class		FPE	Final Prediction Error	0.093169		
class		MSE	Mean Squared Error	0.092671	0.093204	
class		RFPE	Root Final Prediction Error	0.305235		
class		RMSE	Root Mean Squared Error	0.30442	0.305293	
class		AVERR	Average Error Function	0.293876	0.299545	
class		ERR	Error Function	15878.7	16186.81	
class		MISC	Misclassification Rate	0.134106	0.135571	
class		WRONG	Number of Wrong Classifications	3623	3663	

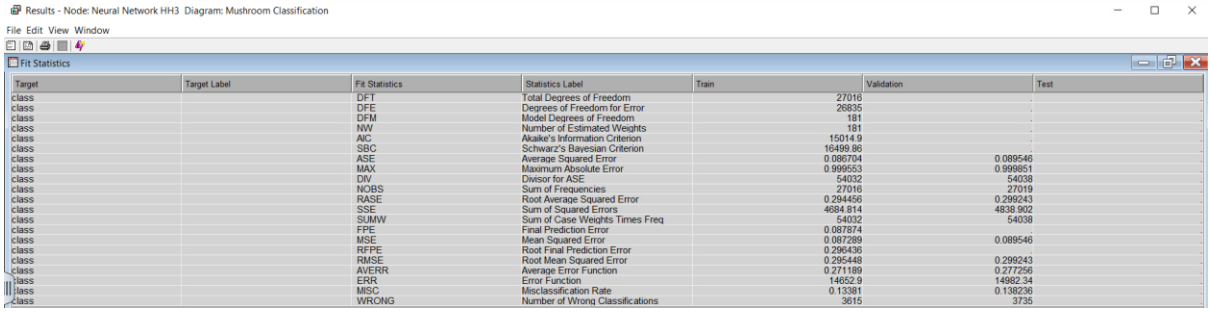
Neural Network No Season – HH3

Neural Network with 5 hidden units and no season variable had **ASE – 0.089546**









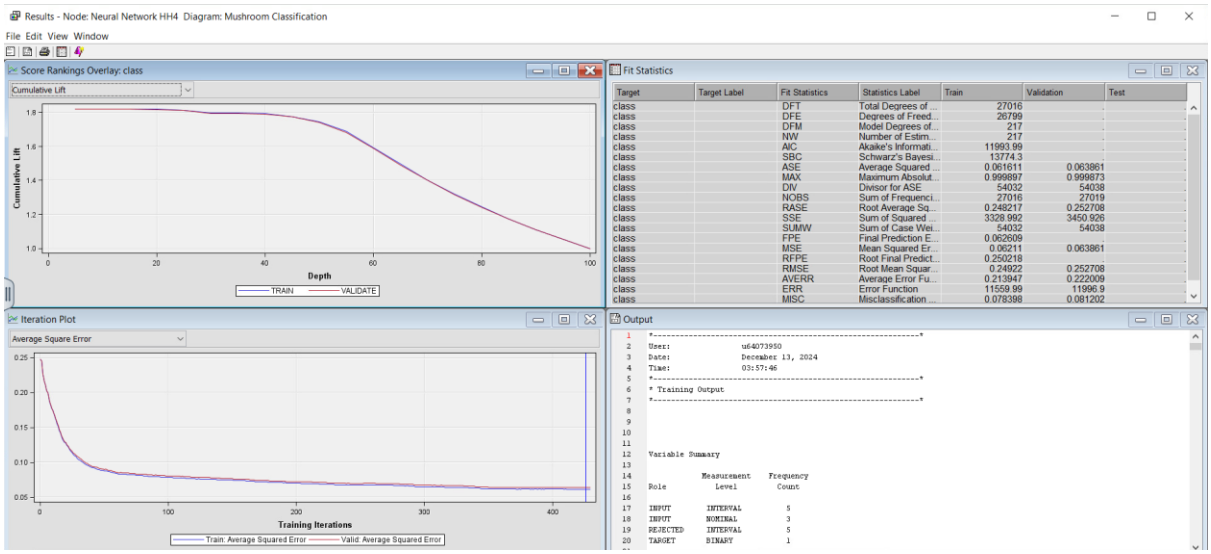
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
class		DFT	Total Degrees of Freedom	27016		
class		DFE	Degrees of Freedom for Error	26835		
class		DFM	Model Degrees of Freedom	181		
class		NW	Number of Estimated Weights	181		
class		AIC	Akaike's Information Criterion	15014.9		
class		SBC	Schwarz's Bayesian Criterion	16499.86		
class		ASE	Average Squared Error	0.086704	0.089546	
class		MAX	Maximum Absolute Error	0.999553	0.999851	
class		DIV	Divisor for ASE	54032	54038	
class		NOBS	Sum of Frequencies	27016	27019	
class		RASE	Root Average Squared Error	0.294456	0.299243	
class		SSE	Sum of Squared Errors	4684.814	4838.902	
class		SUMW	Sum of Case Weights Times Freq	54032	54038	
class		FPE	Final Prediction Error	0.087874		
class		MSE	Mean Squared Error	0.086722	0.089546	
class		RFPE	Root Final Prediction Error	0.296436		
class		RMSE	Root Mean Squared Error	0.295448	0.299243	
class		AVERR	Average Error Function	0.271189	0.277256	
class		ERR	Error Function	14652.9	14982.34	
class		MISC	Misclassification Rate	0.13381	0.138236	
class		WRONG	Number of Wrong Classifications	3615	3735	

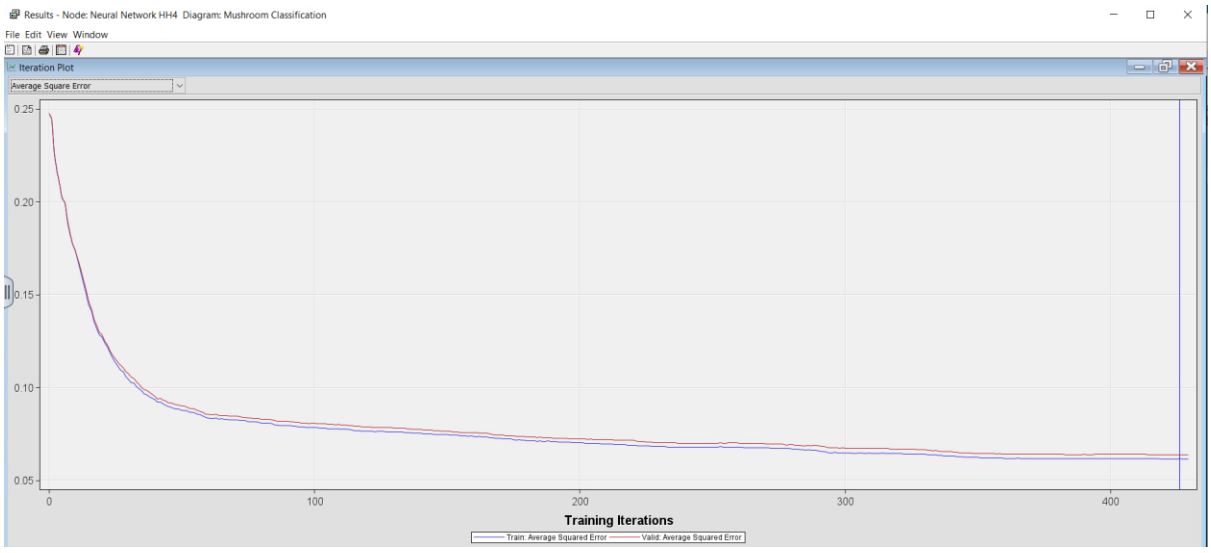
Neural Network No Season – HH4

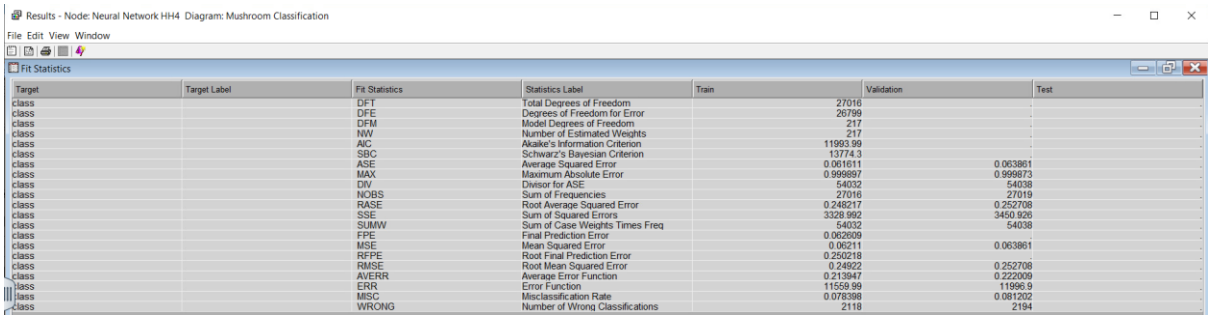
Neural Network with 6 hidden units and no season variable had **ASE – 0.063861.** We observed that the ASE keeps going down with increasing hidden units for our Neural Networks.

Network

.. Property	Value
Architecture	Multilayer Perceptron
Direct Connection	No
Number of Hidden Units	6
Randomization Distribution	Normal
Randomization Center	0.0
Randomization Scale	0.1
Input Standardization	Standard Deviation
Hidden Layer Combination Function	Default
Hidden Layer Activation Function	Default
Hidden Bias	Yes
Target Layer Combination Function	Default
Target Layer Activation Function	Default



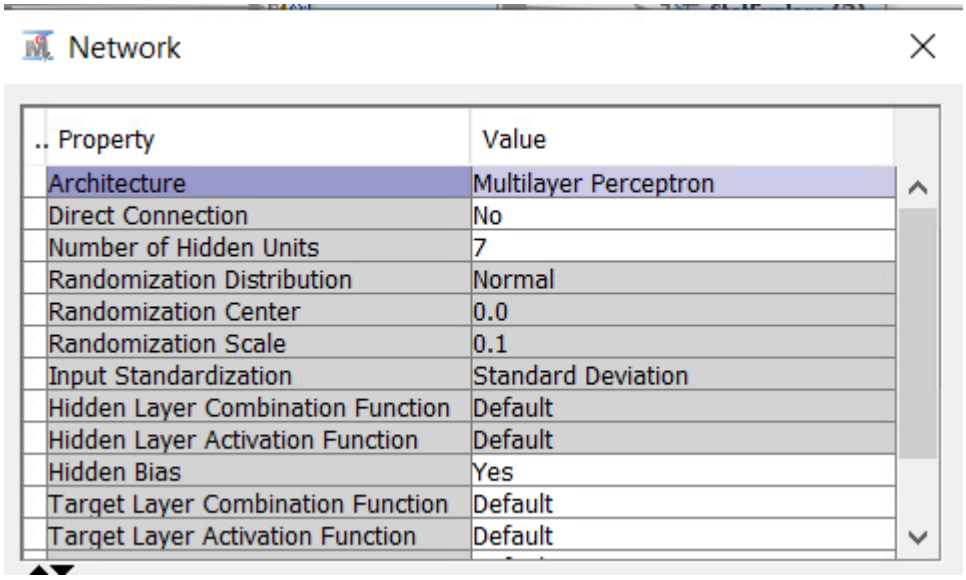




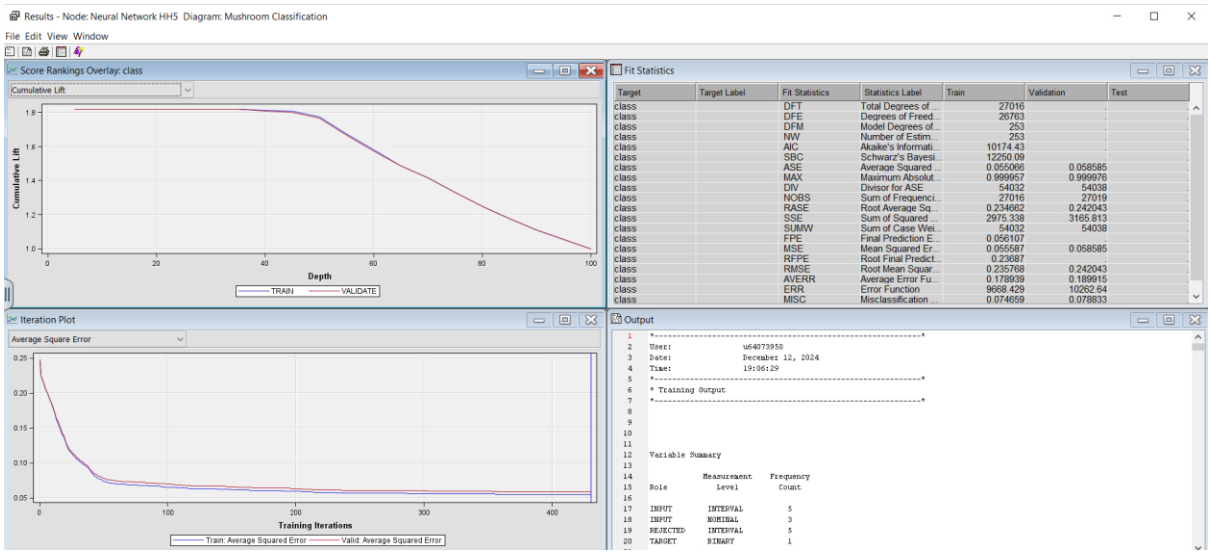
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
class		DFT	Total Degrees of Freedom	27016		
class		DFE	Degrees of Freedom for Error	26799		
class		DFM	Model Degrees of Freedom	217		
class		NW	Number of Estimated Weights	217		
class		AIC	Akaike's Information Criterion	11993.99		
class		SBC	Schwarz's Bayesian Criterion	13774.3		
class		ASE	Average Squared Error	0.061611	0.063861	
class		MAX	Maximum Absolute Error	0.999897	0.999873	
class		DIV	Divisor for ASE	54032	54038	
class		NOBS	Sum of Frequencies	27016	27019	
class		RASE	Root Average Squared Error	0.248217	0.252708	
class		SSE	Sum of Squared Errors	3328.992	3450.926	
class		SUMW	Sum of Case Weights Times Freq	54032	54038	
class		FPE	Final Prediction Error	0.062609		
class		MSE	Mean Squared Error	0.06211	0.063861	
class		RFPE	Root Final Prediction Error	0.250218		
class		RMSE	Root Mean Squared Error	0.24922	0.252708	
class		AVERR	Average Error Function	0.213947	0.222009	
class		ERR	Error Function	11559.99	11996.9	
class		MISC	Misclassification Rate	0.078398	0.081202	
class		WRONG	Number of Wrong Classifications	2118	2194	

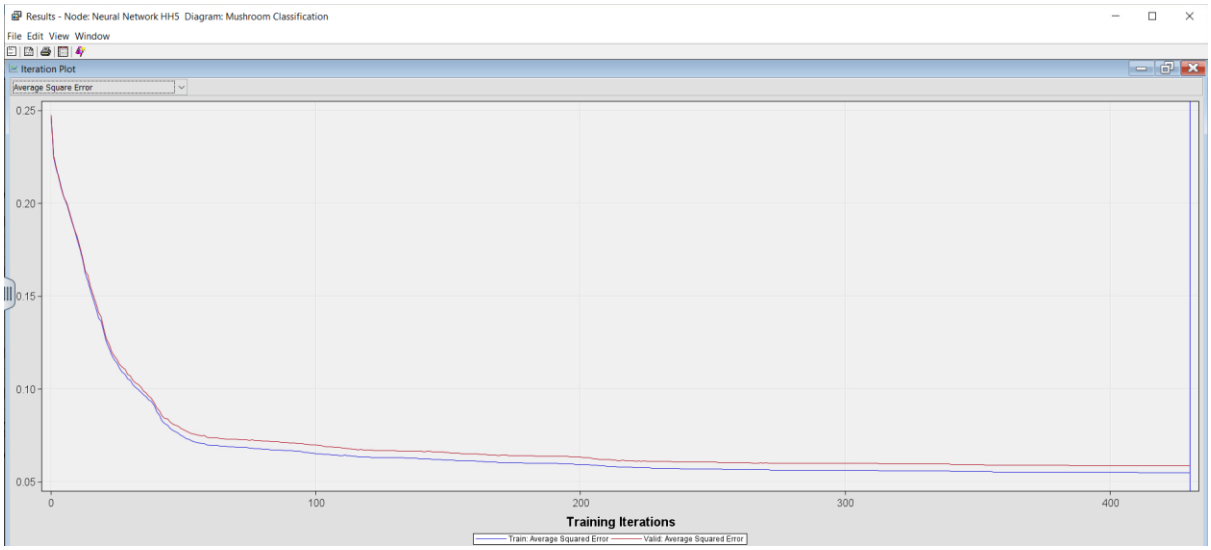
Neural Network No Season – HH5

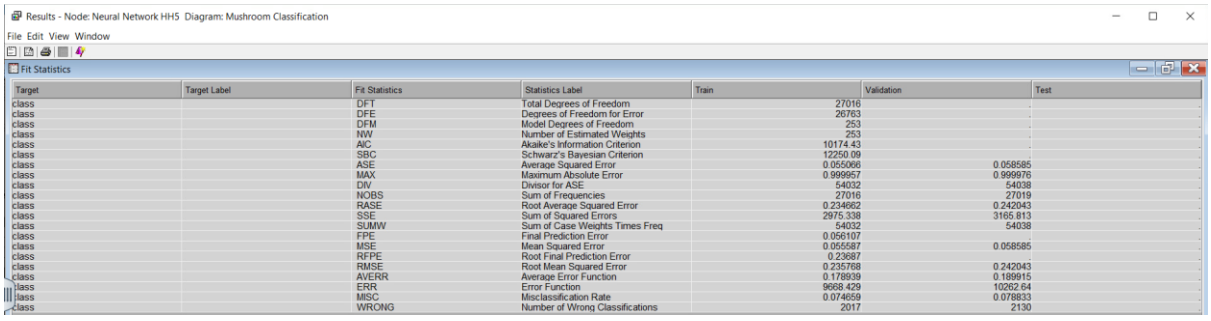
Neural Network with 7 hidden units and no season variable had **ASE – 0.058585**



.. Property	Value
Architecture	Multilayer Perceptron
Direct Connection	No
Number of Hidden Units	7
Randomization Distribution	Normal
Randomization Center	0.0
Randomization Scale	0.1
Input Standardization	Standard Deviation
Hidden Layer Combination Function	Default
Hidden Layer Activation Function	Default
Hidden Bias	Yes
Target Layer Combination Function	Default
Target Layer Activation Function	Default



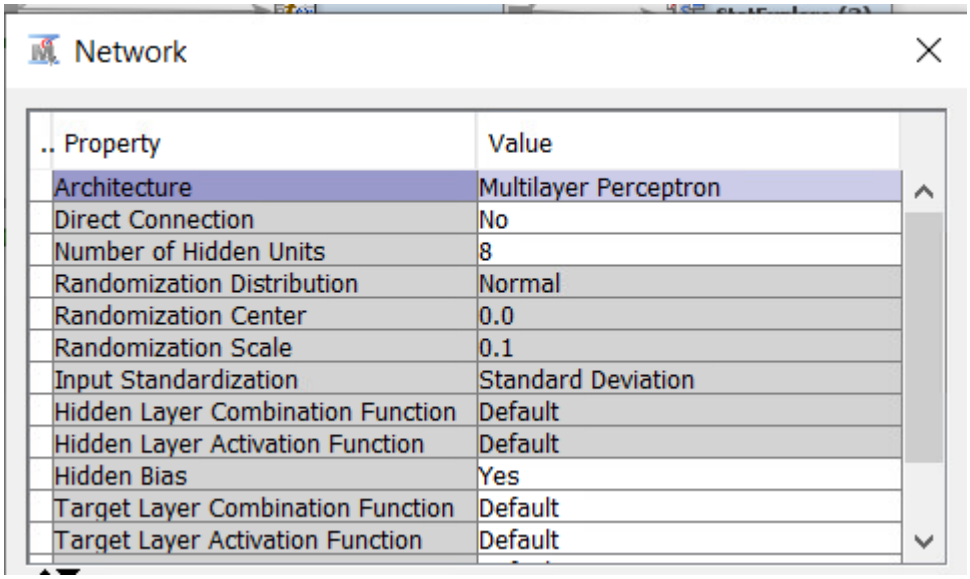


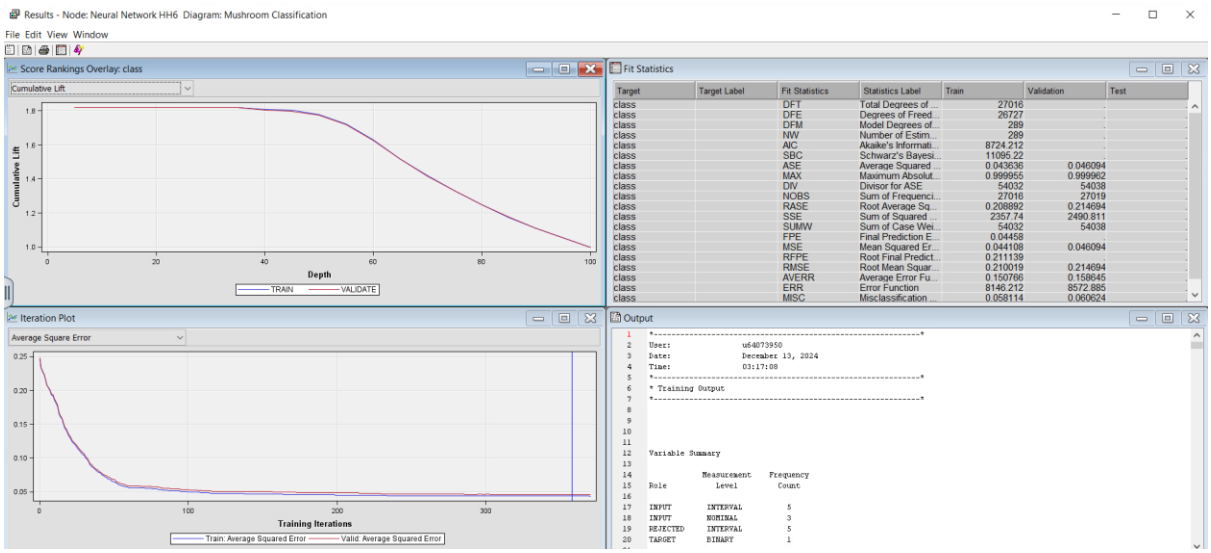


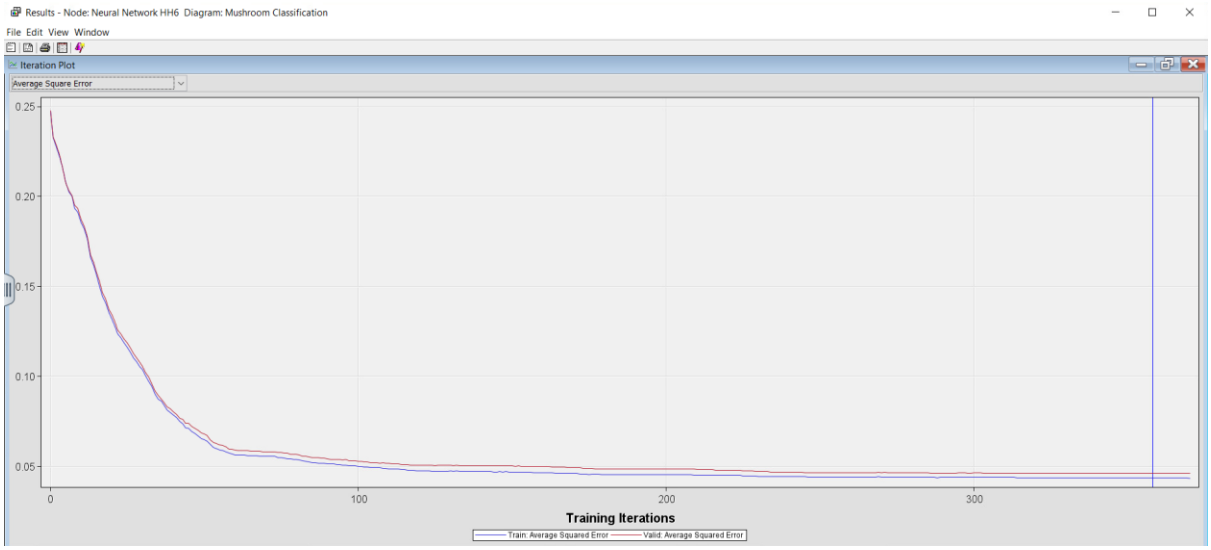
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
class		DFT	Total Degrees of Freedom	27016		
class		DFE	Degrees of Freedom for Error	26763		
class		DFM	Model Degrees of Freedom	253		
class		NW	Number of Estimated Weights	253		
class		AIC	Akaike's Information Criterion	10174.43		
class		SBC	Schwarz's Bayesian Criterion	12250.09		
class		ASE	Average Squared Error	0.055066	0.058585	
class		MAX	Maximum Absolute Error	0.999957	0.999976	
class		DIV	Divisor for ASE	54032	54038	
class		NOBS	Sum of Frequencies	27016	27019	
class		RASE	Root Average Squared Error	0.234682	0.242043	
class		SSE	Sum of Squared Errors	2975.338	3165.813	
class		SUMW	Sum of Case Weights Times Freq	54032	54038	
class		FPE	Final Prediction Error	0.056107		
class		MSE	Mean Squared Error	0.055587	0.058585	
class		RFPE	Root Final Prediction Error	0.23687		
class		RMSE	Root Mean Squared Error	0.235768	0.242043	
class		AVERR	Average Error Function	0.178939	0.189915	
class		ERR	Error Function	9668.429	10262.64	
class		MISC	Misclassification Rate	0.074659	0.078833	
class		WRONG	Number of Wrong Classifications	2017	2130	

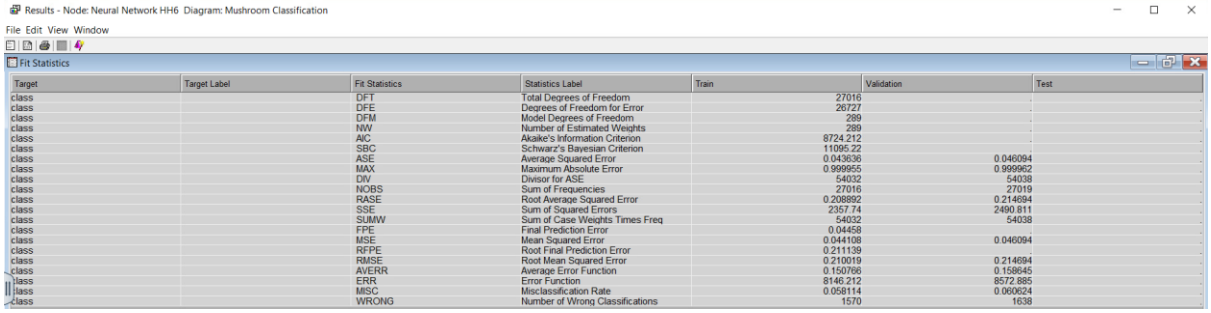
Neural Network No Season – HH6

Neural Network with 8 hidden units and no season variable had **ASE – 0.046094.** So far, this has been the lowest ASE we got from any of the models.





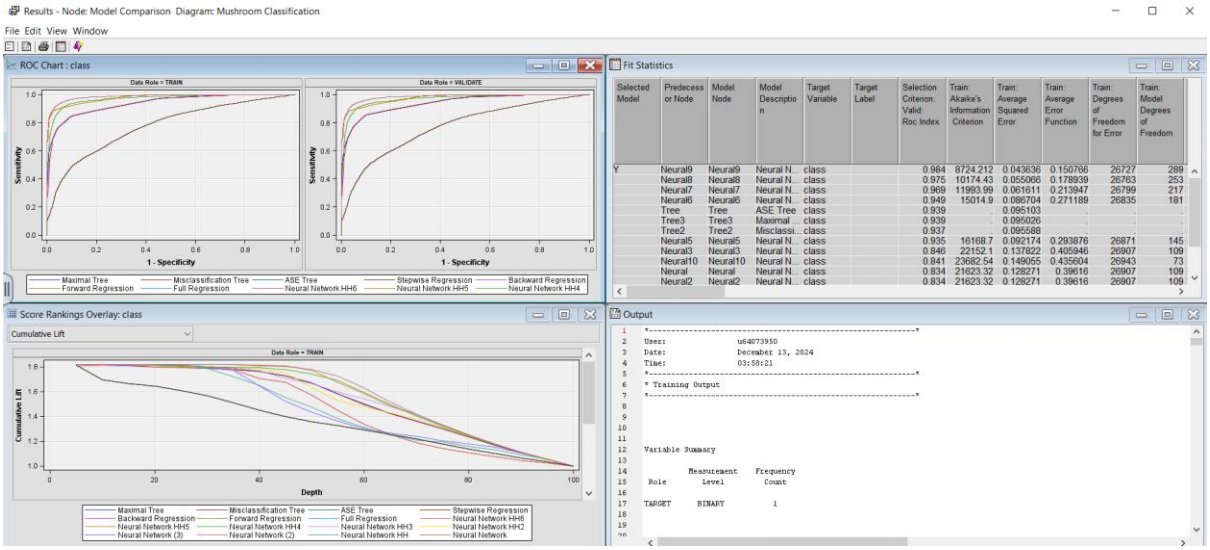


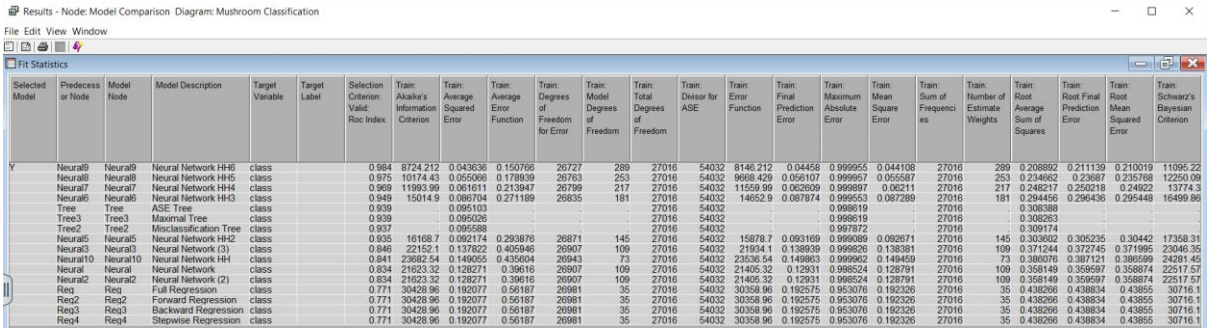


Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
class		DFT	Total Degrees of Freedom	27016		
class		DFE	Degrees of Freedom for Error	26727		
class		DFM	Model Degrees of Freedom	289		
class		NW	Number of Estimated Weights	289		
class		AIC	Akaike's Information Criterion	8724.212		
class		SBC	Schwarz's Bayesian Criterion	11095.22		
class		ASE	Average Squared Error	0.043636	0.046094	
class		MAX	Maximum Absolute Error	0.999955	0.999962	
class		DIV	Divisor for ASE	54032	54038	
class		NOBS	Sum of Frequencies	27016	27019	
class		RASE	Root Average Squared Error	0.208892	0.214694	
class		SSE	Sum of Squared Errors	2357.74	2490.811	
class		SUMW	Sum of Case Weights Times Freq	54032	54038	
class		FPE	Final Prediction Error	0.04458		
class		MSE	Mean Squared Error	0.044108	0.046094	
class		RFPE	Root Final Prediction Error	0.211139		
class		RMSE	Root Mean Squared Error	0.210019	0.214694	
class		AVERR	Average Error Function	0.150768	0.158645	
class		ERR	Error Function	8146.212	8572.885	
class		MISC	Misclassification Rate	0.058114	0.060624	
class		WRONG	Number of Wrong Classifications	1570	1638	

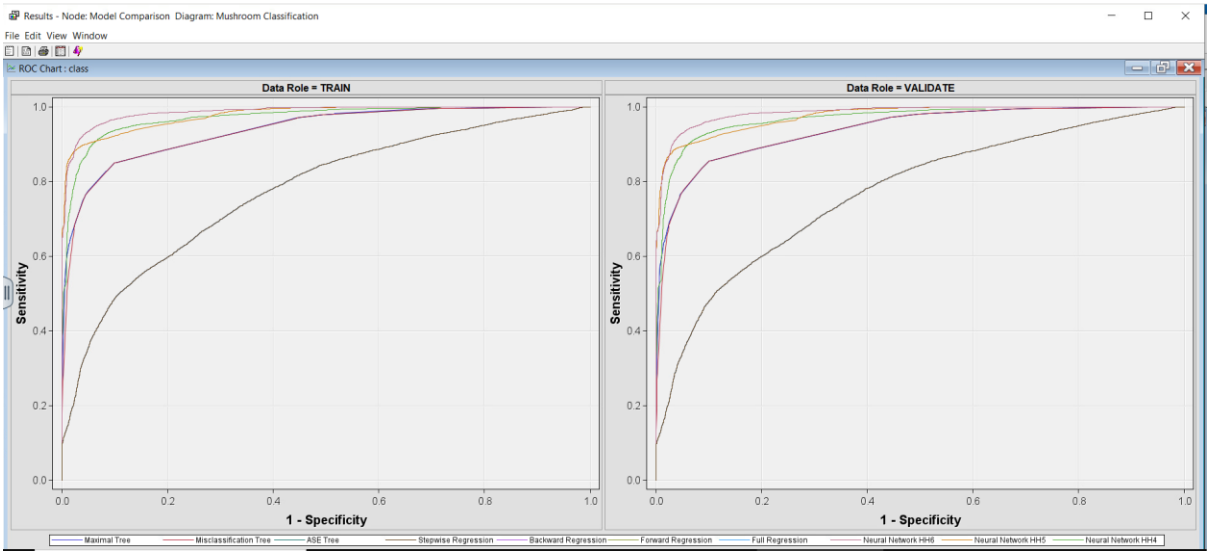
4. Model Comparison

- As we can see from the observations above, Neural9 has the **lowest ASE** (0.043036), confirming its accuracy. This neural network also has the highest ROC index of 0.984, meaning it performs best in differentiating between poisonous and safe mushrooms.
- **Decision Trees** offer decent but slightly lower performance, due to their inability to fully model complex interactions between features.
- **Logistic Regression** performs the worst, showing that the relationships are **non-linear**, and linear models are not suitable for this classification problem.

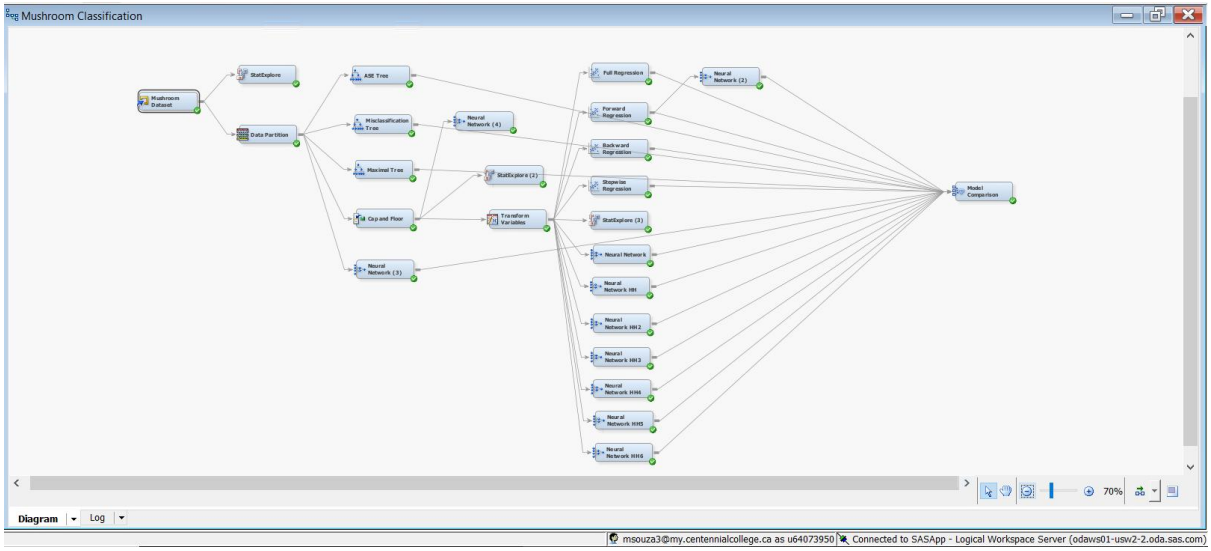




Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Roc Index	Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Average Error Function	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Train: Total Degrees of Freedom	Train: Divisor for ASE	Train: Error Function	Train: Final Prediction Error	Train: Maximum Absolute Error	Train: Mean Square Error	Train: Sum of Frequencies	Train: Number of Estimate Weights	Train: Root Average Sum of Squares	Train: Root Final Prediction Error	Train: Root Mean Squared Error	Train: Schwarz's Bayesian Criterion
Y	Neural9	Neural9	Neural Network HH6	class		0.984	8724.212	0.043636	0.150766	26727	289	27016	54032	8146.212	0.04459	0.999955	0.044108	27016	289	0.208892	0.211139	0.210019	11095.22
	Neural8	Neural8	Neural Network HH5	class		0.975	10174.43	0.055066	0.178939	26763	253	27016	54032	9668.429	0.056107	0.999957	0.055587	27016	253	0.234662	0.23687	0.235768	12250.09
	Neural7	Neural7	Neural Network HH4	class		0.969	11993.99	0.061611	0.213947	26799	217	27016	54032	11559.99	0.062609	0.999897	0.06211	27016	217	0.248217	0.250218	0.24922	13774.3
	Neural6	Neural6	Neural Network HH3	class		0.949	15014.9	0.086704	0.271189	26835	181	27016	54032	14652.9	0.087874	0.999553	0.087289	27016	181	0.294456	0.296436	0.295448	16499.86
	Tree	Tree	ASE Tree	class		0.939		0.095103				27016	54032			0.998619		27016		0.308388			
	Tree3	Tree3	Maximal Tree	class		0.939		0.095026				27016	54032			0.998619		27016		0.308263			
	Tree2	Tree2	Misclassification Tree	class		0.937		0.095588				27016	54032			0.997872		27016		0.309174			
	Neural5	Neural5	Neural Network HH2	class		0.935	16168.7	0.092174	0.293876	26871	145	27016	54032	15878.7	0.093169	0.999089	0.092671	27016	145	0.303602	0.305235	0.30442	17358.31
	Neural3	Neural3	Neural Network (3)	class		0.846	22152.1	0.137822	0.405946	26907	109	27016	54032	21934.1	0.138939	0.999826	0.138381	27016	109	0.371244	0.372745	0.371995	23046.35
	Neural10	Neural10	Neural Network HH	class		0.841	23682.54	0.149055	0.435604	26943	73	27016	54032	23536.54	0.149863	0.999962	0.149459	27016	73	0.386076	0.387121	0.386599	24281.45
	Neural	Neural	Neural Network	class		0.834	21623.32	0.128271	0.39616	26907	109	27016	54032	21405.32	0.12931	0.998524	0.128791	27016	109	0.358149	0.359597	0.358874	22517.57
	Neural2	Neural2	Neural Network (2)	class		0.834	21623.32	0.128271	0.39616	26907	109	27016	54032	21405.32	0.12931	0.998524	0.128791	27016	109	0.358149	0.359597	0.358874	22517.57
	Reg	Reg	Full Regression	class		0.771	30428.96	0.192077	0.56187	26981	35	27016	54032	30358.96	0.192575	0.953076	0.192326	27016	35	0.438266	0.438834	0.43855	30716.1
	Reg2	Reg2	Forward Regression	class		0.771	30428.96	0.192077	0.56187	26981	35	27016	54032	30358.96	0.192575	0.953076	0.192326	27016	35	0.438266	0.438834	0.43855	30716.1
	Reg3	Reg3	Backward Regression	class		0.771	30428.96	0.192077	0.56187	26981	35	27016	54032	30358.96	0.192575	0.953076	0.192326	27016	35	0.438266	0.438834	0.43855	30716.1
	Reg4	Reg4	Stepwise Regression	class		0.771	30428.96	0.192077	0.56187	26981	35	27016	54032	30358.96	0.192575	0.953076	0.192326	27016	35	0.438266	0.438834	0.43855	30716.1



5. Diagram



6. Conclusion

The final analysis concluded that critical aspects of the stem – including stem color, stem width and cap shape – were the most reliable predictors of poison or edible mushrooms. With sophisticated predictive models, especially Neural Networks (that we used 8-hidden units), the classification got the best results averaging ASE 0.046. This proves that the brain is dependable in distinguishing poisonous from non-poisonous mushrooms.

It is especially important for vegetarian foragers to be aware of the following relationships:

Stem Color: Different combinations of colors have significant effects on the risk of toxication. Mushrooms of stem color 2, for instance, are highly poisonous.

Stem Length: Wider stems are less prone to pointing out poisonous mushrooms.

Shape and Gill Color: Certain shapes and colors are strongly associated with toxicity.

These insights yield a robust, data-driven basis for managing safe foraging practices. Vegans can use these observations to stay away from poisonous mushrooms and live a sustainable

lifestyle. Yet field validation and expertise continue to play a key role in making sense of these traits, because classification errors can be very costly.

REFERENCES

Sawhney, P. (n.d.). *Mushroom dataset*. Kaggle. Retrieved August 4, 2024, from https://www.kaggle.com/datasets/prishasawhney/mushroom-dataset

Forage Hyperfoods Reference

Staicu, C. (2023). *Forage hyperfoods*.