WIE3008

PRACTICAL ASSESSMENT

Name: Amir Firdaus Bin Abdul Hadi

Matric Number: 17204620

Submission Date: 16 January 2023 6pm

Instructions: Answer the following questions by Business Analytic Tool tool such as SAS. Explain how you perform each step (include an appropriate print screen for each of the questions).

1. **Muat turun "final.csv" dari Spectrum. Muatkan ke dalam alatan.**
   Download "final.csv" from the Spectrum. Load to the tool.

(1 markah/*mark*)

Title: Insurance Claim Analysis: Demographic and Health

Dataset: https://www.kaggle.com/datasets/thedevastator/insurance-claim-analysis-demographic-and-health?select=insurance_data.csv

2. **Laksanakan TIGA (3) teknik prapemprosesan. Justifikasikan teknik yang dipilih.**
   Apply THREE (3) preprocessing techniques. Justify the selected techniques.

(12 markah/*marks*)

3 preprocessing techniques that I applied for the dataset are *drop*, *impute* and *replacement* by using built-in nodes in SAS Enterprise Miner. Overall processing diagram can be referred to figure 7. Before doing preprocessing, I need to explore the dataset first. I use "StatExplore" node to find out if there are any missing values in each variable. From Figure 1 and 2, there are 3 missing values from class variable "region" and 5 missing values from interval variables "age". And then uses "Multiplot" node to see the distribution of data and also to look into nominal values for nominal variables. From figure 3, we can see that variable "gender" and "smoker" have more than 2 labels due to different in spelling using uppercase and lowercase where for example "male" is different than "Male".

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	diabetic	INPUT	3	0	No	52.09	Yes	47.24
TRAIN	gender	INPUT	4	0	male	50.37	female	49.03
TRAIN	region	INPUT	5	3	southeast	33.06	northwest	26.04
TRAIN	smoker	INPUT	3	0	No	72.31	Ye	20.45

Figure 1: Exploring missing values with StatExplore (Class Variable)

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
age	INPUT	38.07865	11.10292	1335	5	18	38	60	0.113611	-0.94702
bloodpressure	INPUT	94.15746	11.43471	1340	0	80	92	140	1.483534	2.890032
bmi	INPUT	30.66896	6.106735	1340	0	16	30.4	53.1	0.285972	-0.0602
children	INPUT	1.093284	1.205334	1340	0	0	1	5	0.940299	0.205463
claim	TARGET	13252.75	12109.61	1340	0	1121.87	9361.33	63770.43	1.516747	1.610246

Figure 2: Exploring missing values with StatExplore (Interval Variable)

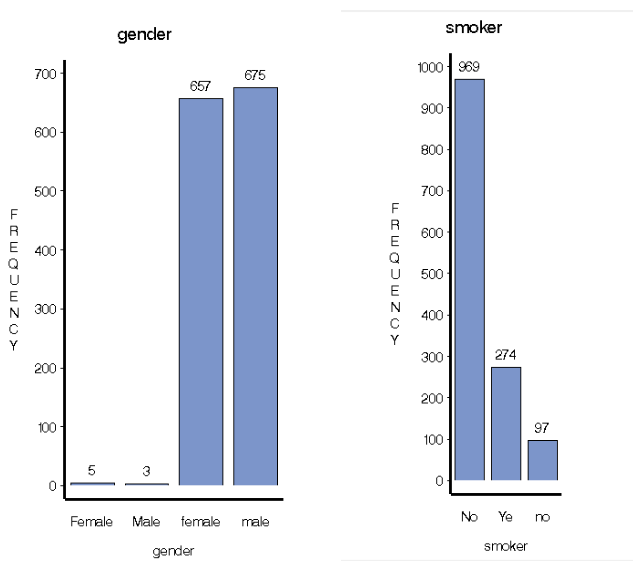


Figure 3: Finding Labelling Issues in Nominal Variable with Multiplot

Drop:

From figure 4, I found out there are 2 similar variables, ”PatientID” and ”Index” that are currently have the roles of Input. Hence, I chose ”PatientID” to be the ID and need to reject and drop the ”Index” column as it will be redundant and interfere with analysis.

Name	Role	Level
age	Input	Interval
bloodpressure	Input	Interval
bmi	Input	Interval
children	Input	Interval
claim	Target	Interval
diabetic	Input	Nominal
gender	Input	Nominal
index	Input	Interval
PatientID	Input	Interval
region	Input	Nominal
smoker	Input	Nominal

Name	Role	Level
PatientID	ID	Interval
age	Input	Interval
bloodpressure	Input	Interval
bmi	Input	Interval
children	Input	Interval
claim	Target	Interval
diabetic	Input	Nominal
gender	Input	Nominal
index	Rejected	Interval
region	Input	Nominal
smoker	Input	Nominal

Figure 4: Variables and Roles (Before and After)

Impute:

As per figure 1 and 2, I need to impute missing values for ”age” and ”region”. For ”age”, the imputation will be using mean as it is numerical while for ”region”, the imputation will be using count(mode) as it is categorical. From figure 5, the imputed value for age is 38.08 and for region is southeast.

Imputation Summary
Number Of Observations

Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
age	MEAN	IMP_age	38.078651685	INPUT	INTERVAL		5
region	COUNT	IMP_region	southeast	INPUT	NOMINAL		3

Figure 5: Imputation summary

Replacement:

For replacement (figure 6), I standardize to change labels for class variables to uppercase first word, for example ”male” to ”Male” and ”no” to ”No”. This will solves the issues with extra labelling issues in figure 3. I also changes labels for ”region” into code for ease of use. For example, ”southwest” into ”SW” and ”northeast” into ”NE”.

Variable	Formatted Value	Type	Character Unformatted Value	Numeric Value	Replacement Value
IMP_region	southeast	C	southeast	.	SE
IMP_region	northwest	C	northwest	.	NW
IMP_region	southwest	C	southwest	.	SW
IMP_region	northeast	C	northeast	.	NE
diabetic	yes	C	yes	.	Yes
gender	male	C	male	.	Male
gender	female	C	female	.	Female
smoker	no	C	no	.	No

Figure 6: Replacement summary

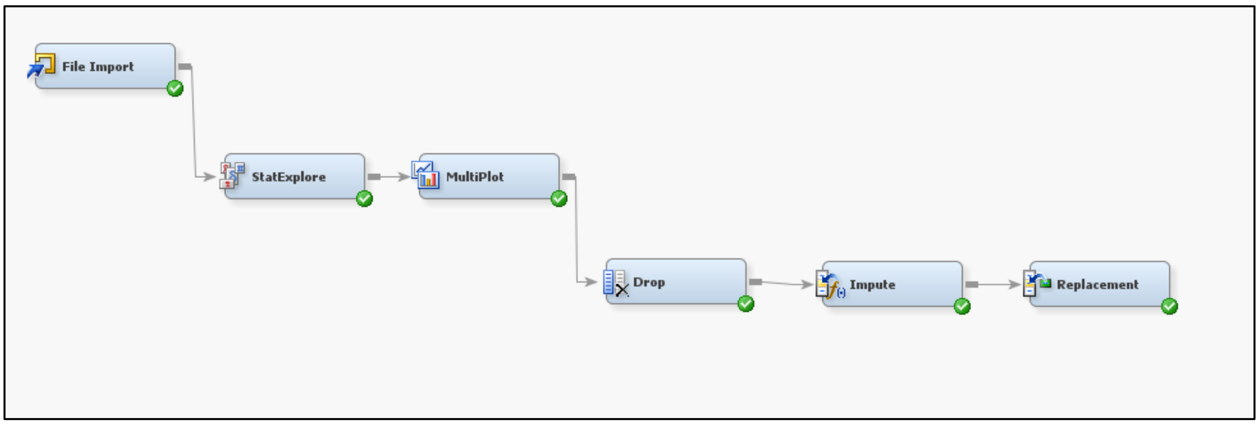
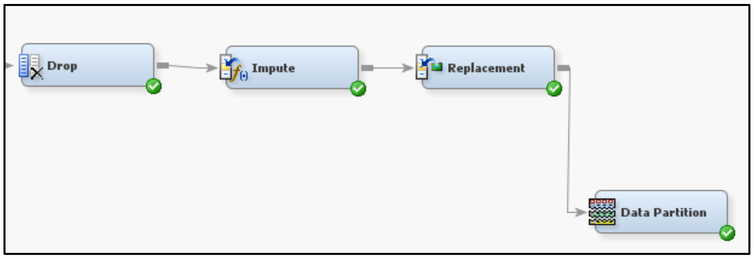


Figure 7: Overall diagram for pre-processing steps

3. **Pisahkan set data kepada set latihan dan set ujian.**
   Split the dataset into a training set and a testing set.

(1 markah/*mark*)

The dataset is splitted into 70:30 training and testing set using ”Data Partition” node (figure 8).



Partition Summary

Type	Data Set	Number of Observations
DATA	EMWS1.Repl_TRAIN	1340
TRAIN	EMWS1.Part_TRAIN	938
TEST	EMWS1.Part_TEST	402

Figure 8: Data Partition Summary

4. **Laksanakan DUA (2) algoritma pembelajaran mesin yang bersesuaian (contoh: untuk klasifikasi). Justifikasikan pemilihan tersebut.**
   Apply TWO (2) suitable machine learning algorithms (e.g., for classification). Justify the selection.

(6 markah/*marks*)

2 machine learning algorithms that I used are Regression and Neural Network (figure 9). This is because the target variable is ”claim” is a numerical value that represents the amount of the insurance claim. The purpose of the machine learning is to analyse key factors across geographical areas and across different demographics such as age or gender so we can gain a greater understanding of who is most likely to receive an insurance claim. Hence, linear regression and neural network able to predict amount of claims based on available variables.

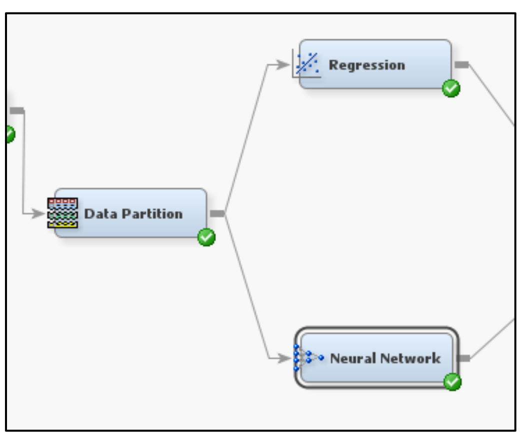


Figure 9: Machine learning models

Regression:

I am using linear regression with stepwise method. Linear Regression is a supervised learning technique that involves learning the relationship between the features and the target. The target values are continuous, which means that the values can take any values between an interval. Stepwise method is where training begins as in the forward model but may remove effects already in the model. This continues until the stay significance level or the stop criterion is met. Figure 10 below shows the curve line for mean predicted(blue) and mean target(red).

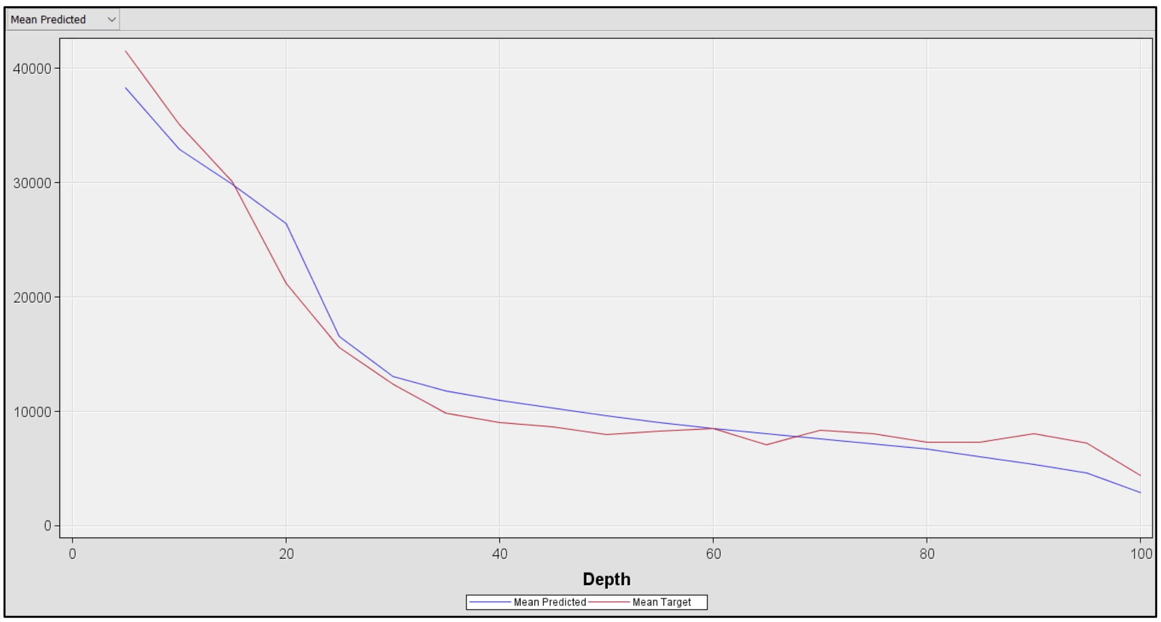


Figure 10: Means predicted vs Means Target (Linear Regression)

Neural Network:

The purpose of using Artificial Neural Networks for Regression over Linear Regression is that the linear regression can only learn the linear relationship between the features and target and therefore cannot learn the complex non-linear relationship. In order to learn the complex non-linear relationship between the features and target, we are in need of other techniques. One of those techniques is to use Neural Networks. Artificial Neural Networks have the ability to learn the complex relationship between the features and target due to the presence of activation function in each layer. Figure 11 below shows the curve line for mean predicted(blue) and mean target(red) for neural network while figure 12 shows the learning iterations against root mean square error.The Parameters used for this neural network are as follows:

- Model Selection Criteria = profit/loss
- Learning rate = 0.1
- Accelerate = 1.2
- Decelerate = 0.5
- Number of runs = 5

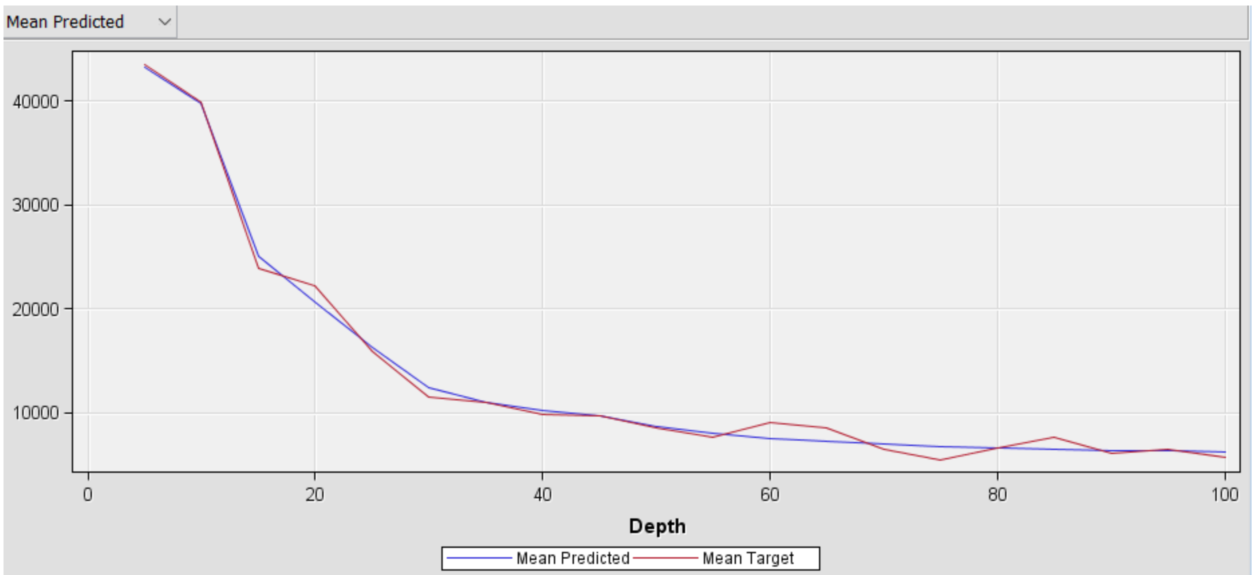


Figure 11: Means predicted vs Means Target (Neural Network)

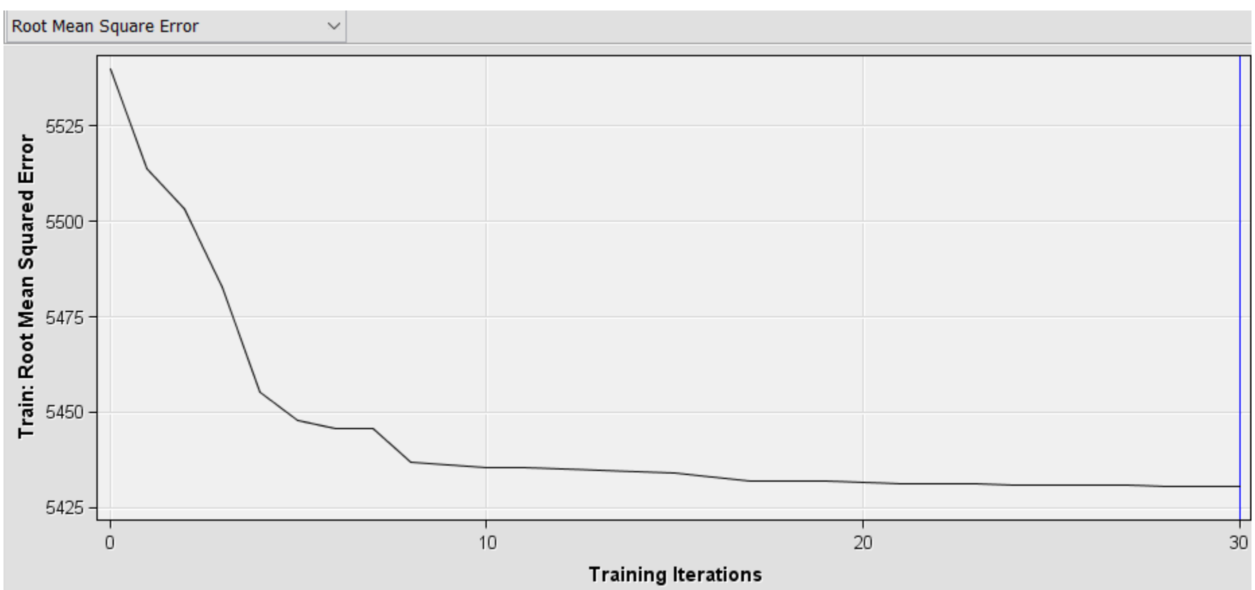


Figure 12: RMSE against training iterations

5. **Laksanakan DUA (2) kaedah penilaian menggunakan pengukuran yang bersesuaian. Justifikasikan pemilihan tersebut.**
Apply TWO (2) evaluation methods using suitable measurements. Justify the selection.

(4 markah/*marks*)

2 evaluation methods used are RMSE and AIC. “Model Comparison” node will be used (figure 13) to generate comparisons between regression and neural network models and better model will be chosen using the said metrics.

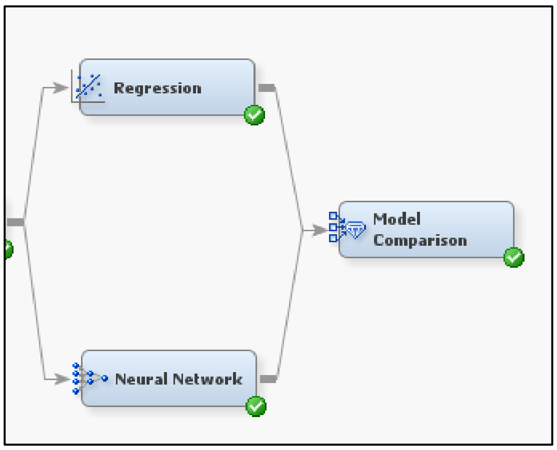


Figure 13: Model comparison node

Root Mean Square Error (RMSE):

RMSE measures the average difference between values predicted by a model and the actual values. It provides an estimation of how well the model is able to predict the target value (accuracy). The lower the value of the Root Mean Squared Error, the better the model is. The Root Mean Squared Error has the advantage of representing the amount of error in the same unit as the predicted column making it easy to interpret. From figure 14, RMSE for neural network and regression is 5430.44 and 6610.64 respectively. This shows that neural network has lower mean error than regression.

Akaike information criterion (AIC):

AIC is a single number score that can be used to determine which of multiple models is most likely to be the best model for a given data set. It estimates models relatively, meaning that AIC scores are only useful in comparison with other AIC scores for the same data set. A lower AIC score is better and AIC penalizes models that use more parameters. So, if two models explain the same amount of variation, the one with fewer parameters will have a lower AIC score and will be the better-fit model. Hence, it is used to evaluate both models. From figure 14, the AIC value for neural network and regression is 16169.43 and 16510.08 respectively. Again, neural network model has lower AIC value than regression model

Thus, the better model for the dataset to predict amount of insurance claims is the neural network model with lower RMSE and AIC values than regression model.

Statistics	Neural	Reg
Train: Akaike's Information Criterion	16169.43	16510.08
Train: Average Squared Error	28326430.83	43327827.45
Train: Average Error Function	28326430.83	43327827.45
Selection Criterion: Train: Average Squared Error	28326430.83	43327827.45
Train: Degrees of Freedom for Error	901.00	930.00
Train: Model Degrees of Freedom	37.00	8.00
Train: Total Degrees of Freedom	938.00	938.00
Train: Divisor for ASE	938.00	938.00
Train: Error Function	26570192122.93	40641502144.30
Train: Final Prediction Error	30652907.95	44073252.43
Train: Maximum Absolute Error	26375.74	31772.30
Train: Misclassification Rate	.	.
Train: Mean Square Error	29489669.39	43700539.94
Train: Sum of Frequencies	938.00	938.00
Train: Number of Estimate Weights	37.00	8.00
Train: Root Average Sum of Squares	5322.26	6582.39
Train: Root Final Prediction Error	5536.51	6638.77
Train: Root Mean Squared Error	5430.44	6610.64
Train: Schwarz's Bayesian Criterion	16348.65	16548.83
Train: Sum of Squared Errors	26570192122.93	40641502144.30
Train: Sum of Case Weights Times Freq	938.00	938.00
Train: Number of Wrong Classifications	.	.

Figure 14: Model Comparison Evaluation Statistics

6. Lakukan DUA (2) teknik visualisasi yang sesuai. Justifikasikan pemilihan tersebut.
Apply TWO (2) suitable visualization techniques. Justify the selection.

(6 markah/*marks*)

2 visualization techniques are bar chart and pie chart. Bar chart is used to display distribution of data and histogram against the mean of the target variable "claim" while pie chart is used to show the data clustering segmentation profile (figure 15). These 2 visualizations will be done using "clustering" and "multiplot" node.

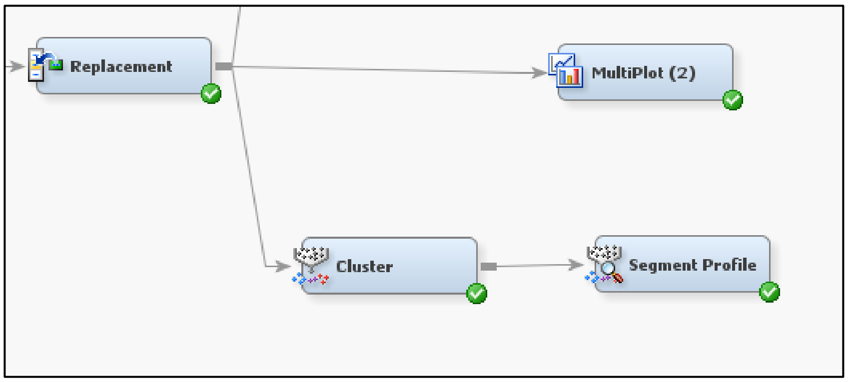


Figure 15: Cluster and Multiplot nodes

Clustering and Segment Profile:

From the figure 16, the data is segmented into 3 different clusters where cluster 1 consists of "smoker" and "blood pressure", cluster 2 and 3 consists of "gender", "age", "smoker" and "blood pressure". And from both the pie chart and bar chart, it shows the rank and worth of each variable's presence in the cluster. For cluster 1, "smoker" have higher worth, while "gender" has higher worth in cluster 2 and 3.

Figure 16: Cluster Segmentation

Multiplot:

When using the “multiplot” node, it will display charts about histogram of each variables distribution and also bar charts for each variable against the target variable. For example, figure 17 shows the distribution of “blood pressure” variable and figure 18 shows the bar chart for “blood pressure” against “claim”. From figure 17, we can see the data is skewed to the left meaning that most people that claims insurance have lower than 100mmHg. And from figure 18, we can see the higher the blood pressure, the higher the amount of claims they make from the insurance.

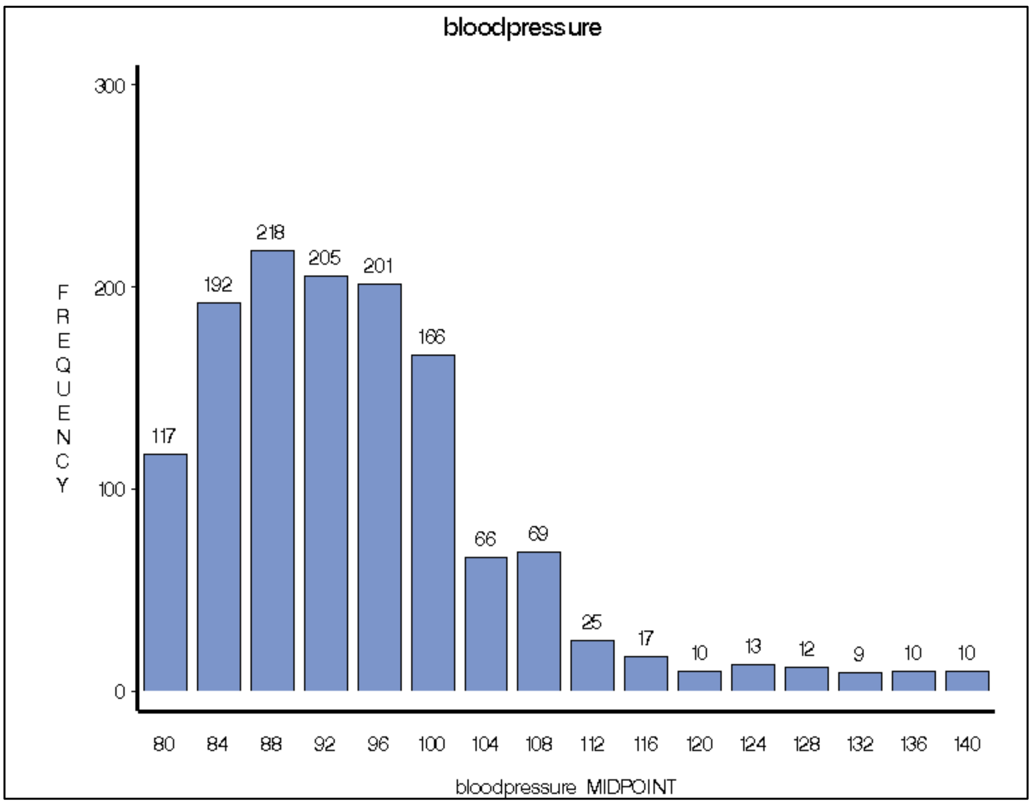


Figure 17: Histogram of blood pressure

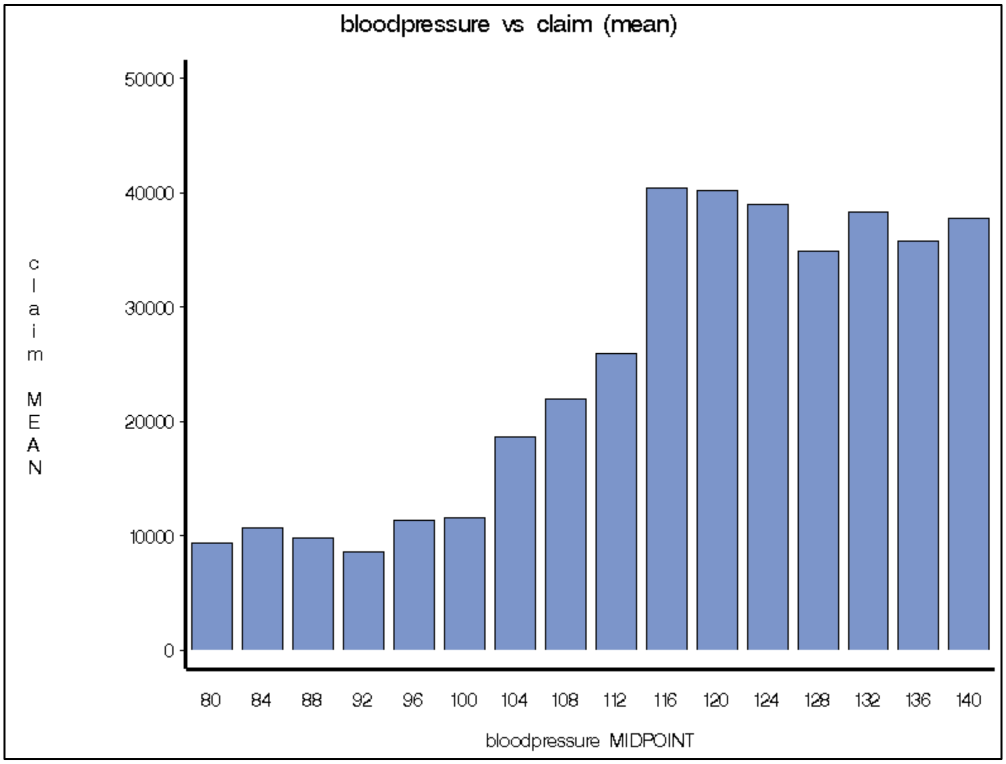


Figure 18: blood pressure vs claim (mean)

Overall Model Diagram:

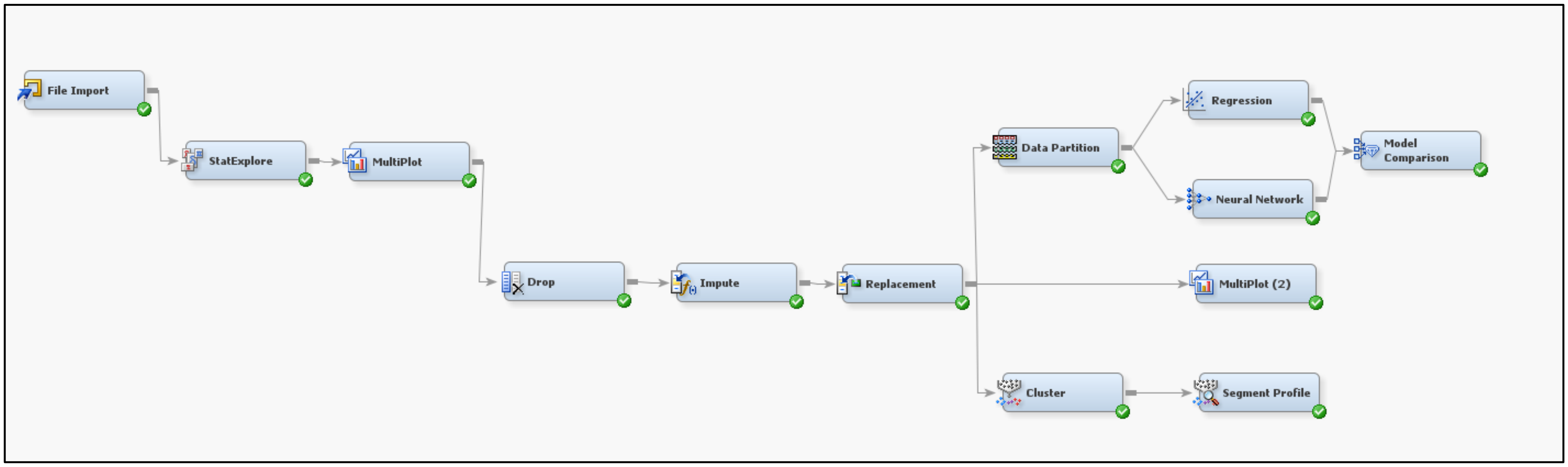


Figure 19: Overall model diagram covering loading, preprocessing, model analysis and visualization