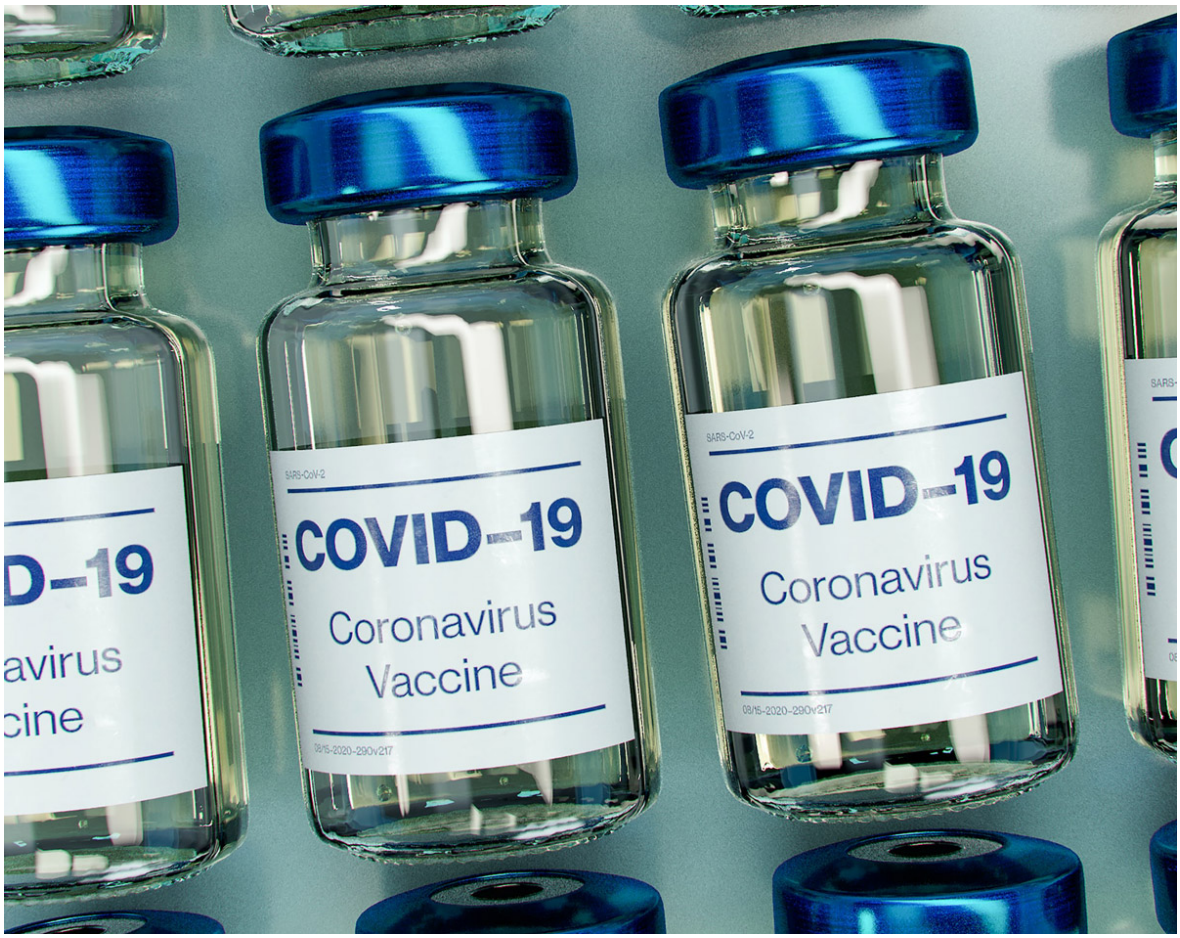


Predicting COVID-19 Vaccine Hesitancy in Canada

BA706 Project Revision / Fortification Assignment

Professor David Parent

Group 5:

Bautista, Glen George (301181547)

Manzano, Wilma Ellaine (301181591)

Rimando, Karon Mae (301168240)

TABLE OF CONTENTS

1. INTRODUCTION 3

2. DATA GATHERING 3

3. EXPLORATORY DATA ANALYSIS 4

4. ANALYSIS: PREDICTIVE MODELLING 17

4.1 DECISION TREE 18

4.2 LOGISTIC REGRESSION 21

4.3 NEURAL NETWORK 28

5. CONCLUSION 33

6. RECOMMENDATION 34

REFERENCES 36

APPENDIX 38

1. INTRODUCTION

According to the World Health Organization (WHO), as of 14 December 2021, there have been more than 270 million people infected globally, including more than 5.3 million deaths due to the COVID-19 virus (WHO, 2021). While vaccination remains to be the most effective protection against the virus, the success depends on the acceptance and uptake of these vaccines. In fact, the World Health Organization listed vaccine hesitancy as one of the top 10 threats to global health in 2019 because it "threatens to reverse progress made in tackling vaccine-preventable diseases" (WHO, 2019).

Although several vaccines have been developed and made available since 2020, the challenge of public vaccination rates prevails with a significant proportion of the population being hesitant or skeptical about it. One way to control the spread of the virus is through herd immunity which can be achieved through vaccination or prior infection. Experts estimated that at least 90% of the population needs to be immune to achieve herd immunity (Herhalt, 2021). In Canada, only about 82% of the population has received atleast one dose of the available vaccines.

The aim of our project is to understand and predict the likelihood of an individual living in Canada to become hesitant of the vaccine. The hesitation is associated with several factors such as: trust ratings (government, news, professionals, and social media), responders demographics, and perceptions relating to the COVID-19 virus.

2. DATA GATHERING

Our study was based on data derived from kaggle.com on "COVID-19 Vaccine Hesitancy Canada COSMO Survey". The study was conducted by the Privy Council Office (PCO) through a web-based survey over a stretch of 8 waves from 10 April 2020 to 16 September 2020. A sample of 2,000 survey respondents were randomly

selected each wave, and consisted of Canadian population aged 18 years and above, who speaks either English or French. All respondents were invited through email and were provided with information regarding the nature of the research, the consent information, and the rights and obligations of the respondents.

For the purpose of this analysis, we have taken the sample from the last survey wave to remove the effects of repeated measures. All observations in this analysis are independent to meet the assumptions of logistic regression.

3. EXPLORATORY DATA ANALYSIS

Our first objective was understanding the variables and responses in the dataset, in order to identify potential variables that would help us predict vaccine hesitancy. To ensure that all variables are usable and reliable, out-of-range values such as 98 for *'Don't know'* need to be recoded to missing. The dataset was also reduced to exclude unnecessary variables that will not be used in the analysis. Next, we proceeded to find the variable that corresponds to COVID-19 vaccine hesitancy which will be the target variable in our model. The data source suggested "*E1r1: If an effective COVID-19 vaccine becomes available and is recommended for me, I would get it*". Therefore, this is the first variable that we have considered and explored. Below are the frequency counts of E1r1.

E1r1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	144	7.24	144	7.24
2	76	3.82	220	11.07
3	68	3.42	288	14.49
4	140	7.04	428	21.53
5	197	9.91	625	31.44
6	319	16.05	944	47.48
7	1044	52.52	1988	100.00

Frequency Missing = 129

Based on this table, 68% have shown strong agreement (top 2 box) that they will be taking the COVID-19 vaccine. This percentage seems to deviate from the Canada's current achievement rate of 75% fully vaccinated individuals. We then continued to explore additional options which would give us more accurate estimate for population hesitancy. We found "*E1r2: If a safe COVID-19 vaccine becomes available and is recommended for me, I would get it"* and *"E1r16: I would be willing to get vaccinated in order to return to work, travel, or attend large gatherings"* as viable options; hence, we have generated their frequencies too.

E1r2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	138	6.89	138	6.89
2	76	3.79	214	10.68
3	59	2.95	273	13.63
4	137	6.84	410	20.47
5	173	8.64	583	29.11
6	332	16.58	915	45.68
7	1088	54.32	2003	100.00

Frequency Missing = 114

E1r16	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	211	10.70	211	10.70
2	96	4.87	307	15.57
3	96	4.87	403	20.44
4	212	10.75	615	31.19
5	252	12.78	867	43.97
6	302	15.31	1169	59.28
7	803	40.72	1972	100.00

Frequency Missing = 145

It seems that using E1r2 and E1r16 on their own would also give us inaccurate estimate. Upon further exploration, we decided to use all the three variables E1r1, E1r2, and E1r16 to determine vaccine hesitancy. Those who have answered top 2 box for all three questions, clearly do not have any reservations in getting the vaccine, while those who have answered bottom 5 scale in any of the questions show mild, moderate, or extreme hesitancy with the vaccine. 27% have expressed hesitancy with the vaccine through any of the three questions. Since this gives us a close estimate that's aligned with the current vaccination rate in the country, we shall use this as our target variable. The frequency counts for the binary variable on hesitancy is provided below.

vax_hesitant	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1532	72.37	1532	72.37
1	585	27.63	2117	100.00

All initial data cleaning and manipulation mentioned are documented in *Appendix A*. The next step was identifying factors that would possibly explain vaccine hesitancy, the independent variables in the model. We have reviewed the list of questions and selected demographic and psychographic questions that could be considered as predictors.

The variables we used for this project is described in the following data dictionary.

Question ID	New Variable Name	Description	Values After Recode	Level
E1r1, E1r2, E1r16	vax_hesitant	Vaccine hesitancy determination *(Based on effective vaccine availability, safe vaccine availability and getting vaccinated inorder to return to work, travel and attend large gatherings)*	0 - Non-hesitant (Willing) 1 - Hesitant (Not Willing)	Binary
QS1	Age	Age of respondent as of year of survey	18 and above	Interval
QS2a	Gender	Respondent's Gender	1 - Male 2 - Female 3 - Gender Diverse	Nominal
QS3	Edu	Highest Level of Education	1 - Grade 8 or Less 2 - Partial High School 3 - High School or equivalent 4 - Apprenticeship / Trade Certificate / Diploma 5 - Partial College / University Education 6 - College, CEGEP / Non-university certificate or diploma 7 - University Certificate or diploma below Bachelor's Degree	Nominal

			8 - Bachelor's Degree 9 - Post Graduate	
QS5	illness	If any serious, long-term illness, like diabetes, emphysema, or high blood pressure	0 - No 1 - Yes	Binary
QS6	community_size	Size of the community lived in	1 - Major Metropolitan *(1M pop and above)* 2 - Large Urban area *(100K pop and above)* *3* - Medium Population area *(pop between 30K and under 100K)* 4 - Small Population area *(pop between 1K and under 30K)* 5 - Small rural area *(under 1K pop)*	Nominal
QS7	province	Province lived in	1 - Alberta 2 - British Columbia 3 - Manitoba 4 - New Brunswick 5 - Newfoundland and Labrador 6 - Northwest Territories 7 - Nova Scotia 8 - Nunavut 9 - Ontario	Nominal

			10 - Prince Edward Island 11 - Quebec 12 - Saskatchewan 13 - Yukon	
Q16	race	Population group the respondent belongs to	1 - White 2 - South Asia 3 - East/Southeast Asia 4 - Black 5 - Latin America 6 - West Asia 7 - Indigenous 8 - Canadian 9 - Other	Nominal
QS8	kids	Whether respondent have any children 18 years old and below	0 - No 1 - Yes	Binary
QS10	employment_impact	Whether COVID-19 had any impact on respondent's employment status	0 - Lost job temporarily / permanently / reduced hours 1 - No impact/ changes	Binary
QS17	income	Total Household Income	1 - Under 20K 2 - 20K to under 40K 3 - 40K to under 60K 4 - 60K to under 80K 5 - 80K to under 100K 6 - 100K to under 150K 7 - 150K to under 200K	Nominal

			8 - 200K and above	
QS18	HH_size	Size of Household	1 and above	Interval
A3	infected	Whether the respondent have been infected by COVID-19	0 - No 1 - Yes / Maybe	Binary
A4	knowinfected	Whether the respondent know anyone else having infected	0 - No 1 - Yes / Maybe	Binary
A5	C19knowledge	Self-rating on level of knowledge of COVID-19	1 - Very Poor 2 - 2 3 - 3 4 - 4 5 - 5 6 - 6 7 - Excellent	Interval
B7r1 to B7r21	measures_freq	Mean frequency of following measures *(hand washing, sanitizing, social distancing, etc)* to keep from getting sick with COVID-19	1 - Never 2 - 2 3 - 3 4 - 4 5 - 5 6 - 6 7 - Always	Interval
B8r1 to B8r8	sentiment	Mean sentiments regarding COVID-19 *(I want to protect others by avoiding public areas, etc)*	1 - Strongly Disagree 2 - 2 3 - 3 4 - 4 5 - 5	Interval

			6 - 6 7 - Strongly Agree	
C1r1 to C1r3, C1r8, C1r11 to C1r14	news_trust	Mean trust rating on news as source of information about COVID-19	1 - Very little Trust 2 - 2 3 - 3 4 - 4 5 - 5 6 - 6 7 - Great Deal of Trust	Interval
C1r9, C1r10, C1r15, C1r16, C1r21	gov_trust	Mean trust rating on government as source of information about COVID-19	1 - Very little Trust 2 - 2 3 - 3 4 - 4 5 - 5 6 - 6 7 - Great Deal of Trust	Interval
C1r4 to C1r6	circle_trust	Mean trust rating on family/ friends/ colleagues as source of information about COVID-19	1 - Very little Trust 2 - 2 3 - 3 4 - 4 5 - 5 6 - 6 7 - Great Deal of Trust	Interval
C1r7	pro_trust	Mean trust rating on health professionals as source of information about COVID-19	1 - Very little Trust 2 - 2 3 - 3 4 - 4 5 - 5	Interval

			6 - 6 7 - Great Deal of Trust	
C1r17 to C1r20, C1r22	socmed_trust	Mean trust rating on social media *(Facebook, Twitter, YouTube, etc)* as source of information about COVID-19	1 - Very little Trust 2 - 2 3 - 3 4 - 4 5 - 5 6 - 6 7 - Great Deal of Trust	Interval

We further examined the patterns and characteristics of these variables through frequency tables and descriptive statistics.

Class Variables

Data Role	Target	Target Level ▲	Variable Name	Frequency Count	Percent Within	CODE	Level
TRAIN	vax_hesitant	0	Edu	1	0.065274	81	
TRAIN	vax_hesitant	0	Edu	24	1.56658	22	
TRAIN	vax_hesitant	0	Edu	202	13.18538	13	
TRAIN	vax_hesitant	0	Edu	77	5.02611	74	
TRAIN	vax_hesitant	0	Edu	132	8.616188	45	
TRAIN	vax_hesitant	0	Edu	297	19.38642	06	
TRAIN	vax_hesitant	0	Edu	86	5.613577	57	
TRAIN	vax_hesitant	0	Edu	487	31.78851	38	
TRAIN	vax_hesitant	0	Edu	226	14.75196	69	
TRAIN	vax_hesitant	0	Gender	766	50	01	
TRAIN	vax_hesitant	0	Gender	765	49.93473	12	
TRAIN	vax_hesitant	0	Gender	1	0.065274	23	
TRAIN	vax_hesitant	0	community...	491	32.04961	41	
TRAIN	vax_hesitant	0	community...	483	31.52742	32	
TRAIN	vax_hesitant	0	community...	231	15.07833	23	
TRAIN	vax_hesitant	0	community...	228	14.88251	04	
TRAIN	vax_hesitant	0	community...	99	6.462141	15	
TRAIN	vax_hesitant	0	employmen...	604	39.42559	1.	
TRAIN	vax_hesitant	0	employmen...	562	36.68407	00	
TRAIN	vax_hesitant	0	employmen...	366	23.89034	21	
TRAIN	vax_hesitant	0	illness	17	1.109661	2.	
TRAIN	vax_hesitant	0	illness	1080	70.49608	00	
TRAIN	vax_hesitant	0	illness	435	28.39426	11	
TRAIN	vax_hesitant	0	income	123	8.028721	8.	
TRAIN	vax_hesitant	0	income	59	3.851175	21	
TRAIN	vax_hesitant	0	income	218	14.22977	12	
TRAIN	vax_hesitant	0	income	224	14.62141	53	
TRAIN	vax_hesitant	0	income	235	15.33943	04	
TRAIN	vax_hesitant	0	income	228	14.88251	35	
TRAIN	vax_hesitant	0	income	281	18.34204	46	
TRAIN	vax_hesitant	0	income	101	6.592689	77	
TRAIN	vax_hesitant	0	income	63	4.112272	68	
TRAIN	vax_hesitant	0	infected	34	2.219321	2.	
TRAIN	vax_hesitant	0	infected	1405	91.71018	00	
TRAIN	vax_hesitant	0	infected	93	6.070496	11	
TRAIN	vax_hesitant	0	kids	1149	75	00	
TRAIN	vax_hesitant	0	kids	383	25	11	
TRAIN	vax_hesitant	0	knowinfected	15	0.979112	2.	
TRAIN	vax_hesitant	0	knowinfected	1207	78.7859	00	
TRAIN	vax_hesitant	0	knowinfected	310	20.23499	11	

Class Variables

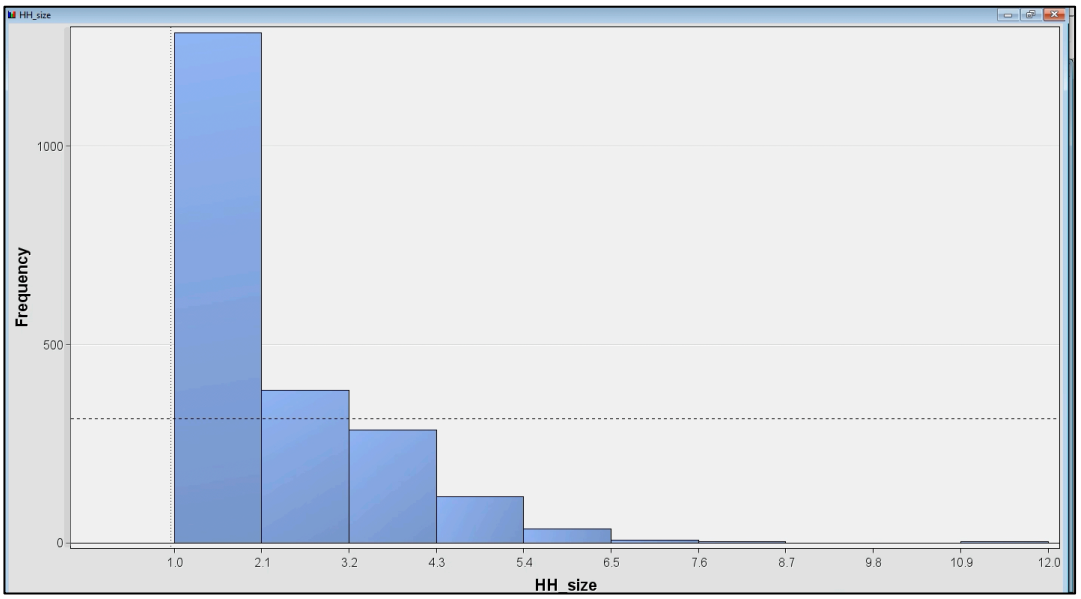
Data Role	Target	Target Level ▲	Variable Name	Frequency Count	Percent Within	CODE	Level	
TRAIN	vax_hesitant	0	province	192	12.53264	31		N
TRAIN	vax_hesitant	0	province	219	14.29504	22		N
TRAIN	vax_hesitant	0	province	98	6.396867	63		N
TRAIN	vax_hesitant	0	province	53	3.45953	84		N
TRAIN	vax_hesitant	0	province	44	2.872063	15		N
TRAIN	vax_hesitant	0	province	1	0.065274	116		N
TRAIN	vax_hesitant	0	province	69	4.503916	57		N
TRAIN	vax_hesitant	0	province	459	29.96084	79		N
TRAIN	vax_hesitant	0	province	22	1.436031	910		N
TRAIN	vax_hesitant	0	province	295	19.25587	011		N
TRAIN	vax_hesitant	0	province	78	5.091384	412		N
TRAIN	vax_hesitant	0	province	2	0.130548	1013		N
TRAIN	vax_hesitant	0	race	152	9.921671	1.		N
TRAIN	vax_hesitant	0	race	1242	81.0705	01		N
TRAIN	vax_hesitant	0	race	61	3.981723	42		N
TRAIN	vax_hesitant	0	race	16	1.044386	24		N
TRAIN	vax_hesitant	0	race	21	1.370757	65		N
TRAIN	vax_hesitant	0	race	13	0.848564	56		N
TRAIN	vax_hesitant	0	race	3	0.195822	87		N
TRAIN	vax_hesitant	0	race	2	0.130548	78		N
TRAIN	vax_hesitant	0	race	22	1.436031	396		N

Interval Variables

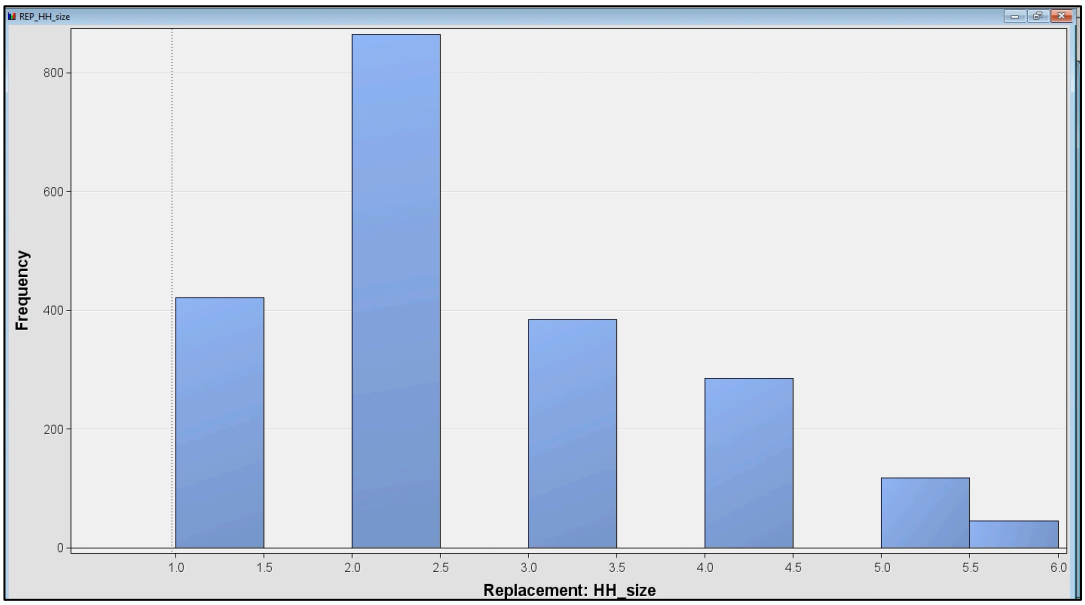
Data Role	Target	Target Level	Variable ▲	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
TRAIN	vax_hesitant	0	Age	51	2	1530	18	95	50.46797	17.31418	-0.01792	-1.12165
TRAIN	vax_hesitant	1	Age	44	3	582	18	78	44.67698	15.01538	0.315951	-0.82781
TRAIN	vax_hesitant	0	circle_trust	4	15	1517	1	7	3.975939	1.258927	-0.23934	-0.23409
TRAIN	vax_hesitant	1	circle_trust	4	11	574	1	7	3.916957	1.301643	-0.29018	0.021972
TRAIN	vax_hesitant	0	gov_trust	5.8	4	1528	1	7	5.515085	1.192306	-1.15234	1.703344
TRAIN	vax_hesitant	1	gov_trust	4.6	5	580	1	7	4.420546	1.586683	-0.54223	-0.37073
TRAIN	vax_hesitant	0	HH_size	2	0	1532	1	8	2.456266	1.214992	0.957526	0.712481
TRAIN	vax_hesitant	1	HH_size	2	0	585	1	12	2.663248	1.404896	1.583699	6.105238
TRAIN	vax_hesitant	0	measures_freq	5.2	0	1532	1.923077	7	5.161673	0.809653	-0.4363	0.3056
TRAIN	vax_hesitant	1	measures_freq	4.714286	0	585	1	7	4.650339	1.007935	-0.55018	0.595637
TRAIN	vax_hesitant	0	news_trust	4.833333	4	1528	1	7	4.66332	1.180113	-0.69292	0.500831
TRAIN	vax_hesitant	1	news_trust	4	3	582	1	7	3.86718	1.387704	-0.30295	-0.57211
TRAIN	vax_hesitant	0	pro_trust	6	162	1370	1	7	5.608029	1.182571	-1.02055	1.431429
TRAIN	vax_hesitant	1	pro_trust	5	66	519	1	7	4.786127	1.3903	-0.68616	0.398565
TRAIN	vax_hesitant	0	socmed_trust	2	33	1499	1	7	2.474983	1.502515	0.795448	-0.19846
TRAIN	vax_hesitant	1	socmed_trust	3	15	570	1	7	2.784211	1.543999	0.442188	-0.73968

The variables '*infected'* and *'knowinfected'* have very high 'NO' percentage response, as respondents might be too reluctant to answer whether they got infected or even know anyone else who did. Despite of the importance of these factors to the model, it would still be better to reject them instead since the results are inaccurate. Meanwhile, *'Employment_impact'*, *'C19Knowledge', and 'Sentiment'* have high missing rate (more than 30%) and should also be rejected. As for the internal

variables, skewness and kurtosis values are within acceptable range except for '*HH_size'*. This is verified with the histogram below.



We can observe from the histogram that there are extreme values at the right side of the plot. To tighten the distribution, we impose upper limit and capping the household size to 6. The histogram and descriptive summary post-transformation is provided below. We shall use this new household size variable in the succeeding analysis.

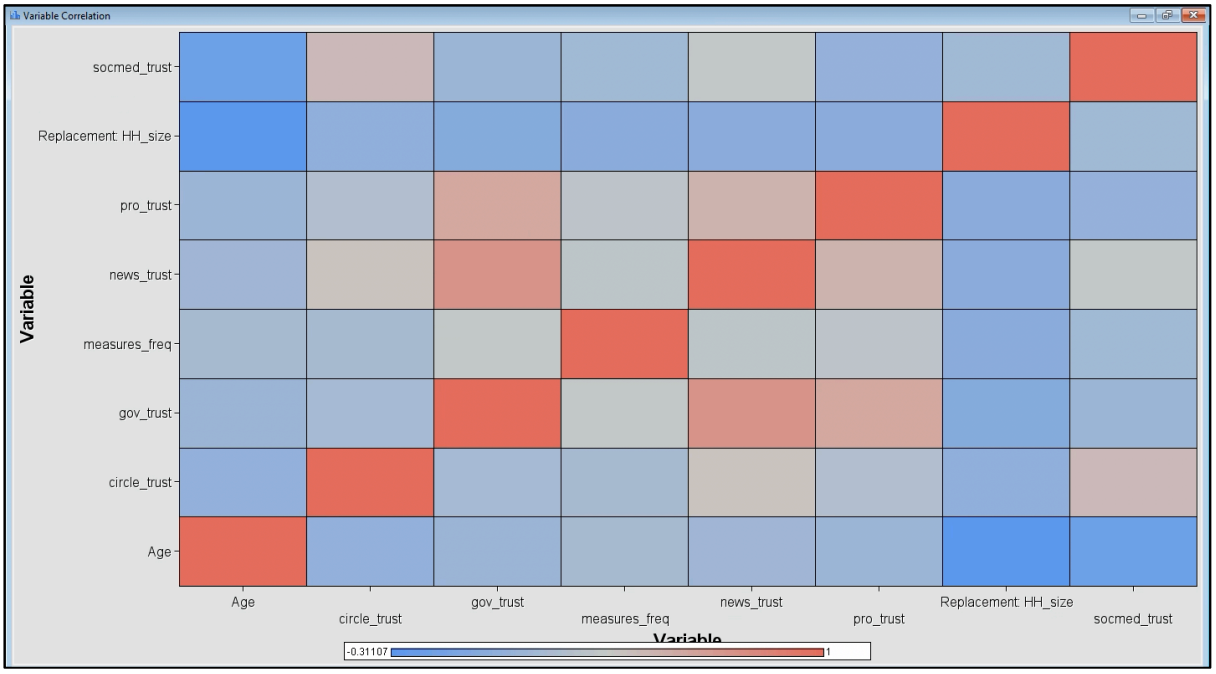


Interval Variables

Data Role	Target	Target Level	Variable ▲	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
TRAIN	vax_hesitant	0	Age	51	2	1530	18	95	50.46797	17.31418	-0.01792	-1.12165
TRAIN	vax_hesitant	1	Age	44	3	582	18	78	44.67698	15.01538	0.315951	-0.82781
TRAIN	vax_hesitant	0	circle_trust	4	15	1517	1	7	3.975939	1.258927	-0.23934	-0.23409
TRAIN	vax_hesitant	1	circle_trust	4	11	574	1	7	3.916957	1.301643	-0.29018	0.021972
TRAIN	vax_hesitant	0	gov_trust	5.8	4	1528	1	7	5.515085	1.192306	-1.15234	1.703344
TRAIN	vax_hesitant	1	gov_trust	4.6	5	580	1	7	4.420546	1.586683	-0.54223	-0.37073
TRAIN	vax_hesitant	0	measures_freq	5.2	0	1532	1.923077	7	5.161673	0.809653	-0.4363	0.3056
TRAIN	vax_hesitant	1	measures_freq	4.714286	0	585	1	7	4.650339	1.007935	-0.55018	0.595637
TRAIN	vax_hesitant	0	news_trust	4.833333	4	1528	1	7	4.66332	1.180113	-0.69292	0.500831
TRAIN	vax_hesitant	1	news_trust	4	3	582	1	7	3.86718	1.387704	-0.30295	-0.57211
TRAIN	vax_hesitant	0	pro_trust	6	162	1370	1	7	5.608029	1.182571	-1.02055	1.431429
TRAIN	vax_hesitant	1	pro_trust	5	66	519	1	7	4.786127	1.3903	-0.68616	0.398565
TRAIN	vax_hesitant	0	REP_HH_size	2	0	1532	1	6	2.451697	1.199125	0.866231	0.251009
TRAIN	vax_hesitant	1	REP_HH_size	2	0	585	1	6	2.635897	1.287084	0.707524	-0.15853
TRAIN	vax_hesitant	0	socmed_trust	2	33	1499	1	7	2.474983	1.502515	0.795448	-0.19846
TRAIN	vax_hesitant	1	socmed_trust	3	15	570	1	7	2.784211	1.543999	0.442188	-0.73968

After completing the data exploration, we were left with 16 predictor variables: ***age, education, gender, household size (capped), circle trust, community size, illness, income, kids, measures frequency, news trust, pro trust, gov_trust, province, race, and social media trust.***

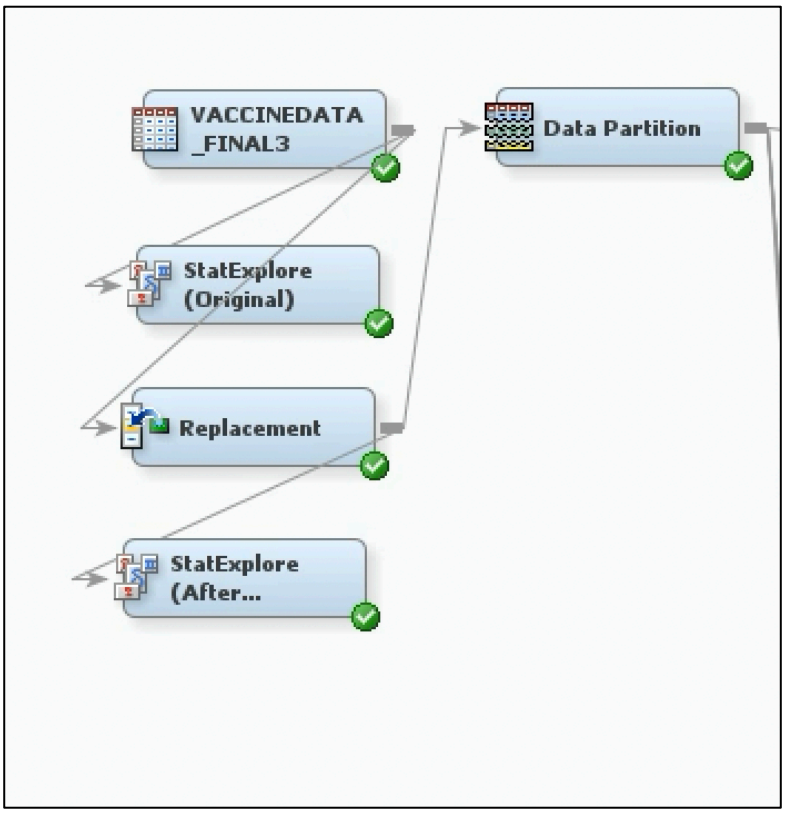
Lastly, we checked for presence of multicollinearity among our variables through correlation matrix and heatmap. A moderately strong correlation is evident between trust rating on government and medical professionals, but still passed our threshold of less than 0.8; hence, both variables were retained.



4. ANALYSIS: PREDICTIVE MODELLING

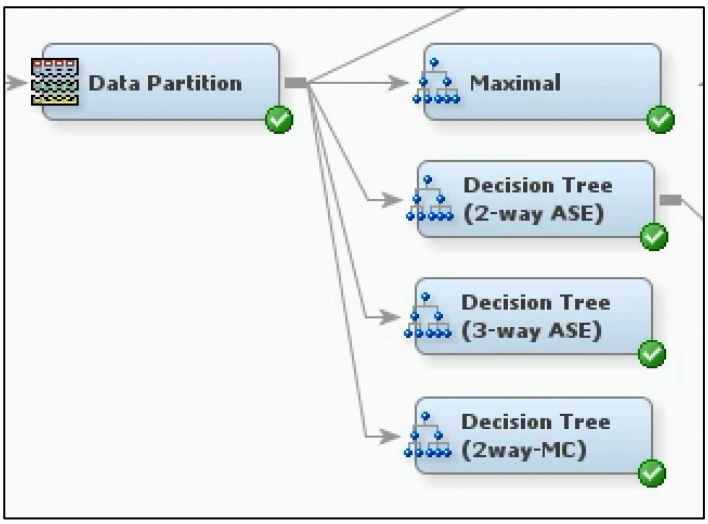
To understand factors driving vaccine hesitancy among Canadian residents, we have applied three predictive modelling techniques: Decision Tree, Regression, and Neural Network. After running several possible models, we will decide on the best model having the best fit statistics (Average Square Error (ASE), Misclassification Rate, and ROC Index).

Prior to running any of the models, we have partitioned our sample equally between training and validation. The training data would be used during the training process to generate several models which will then be evaluated using the validation data to confirm which is the most optimal.

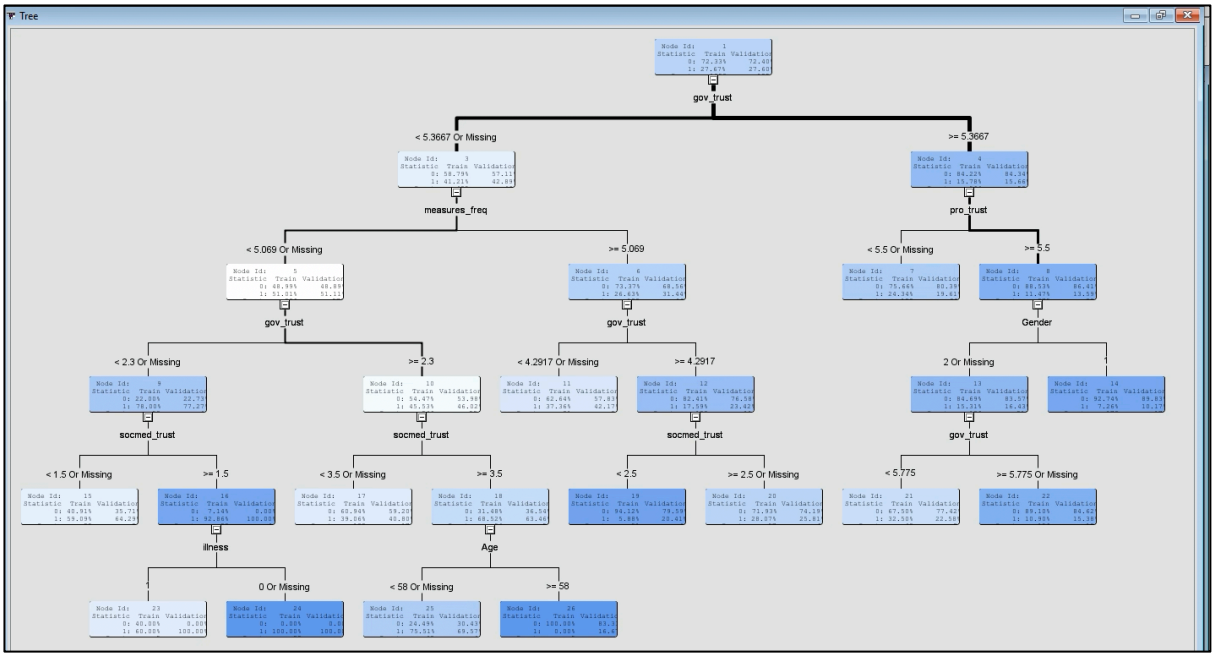


4.1 DECISION TREE

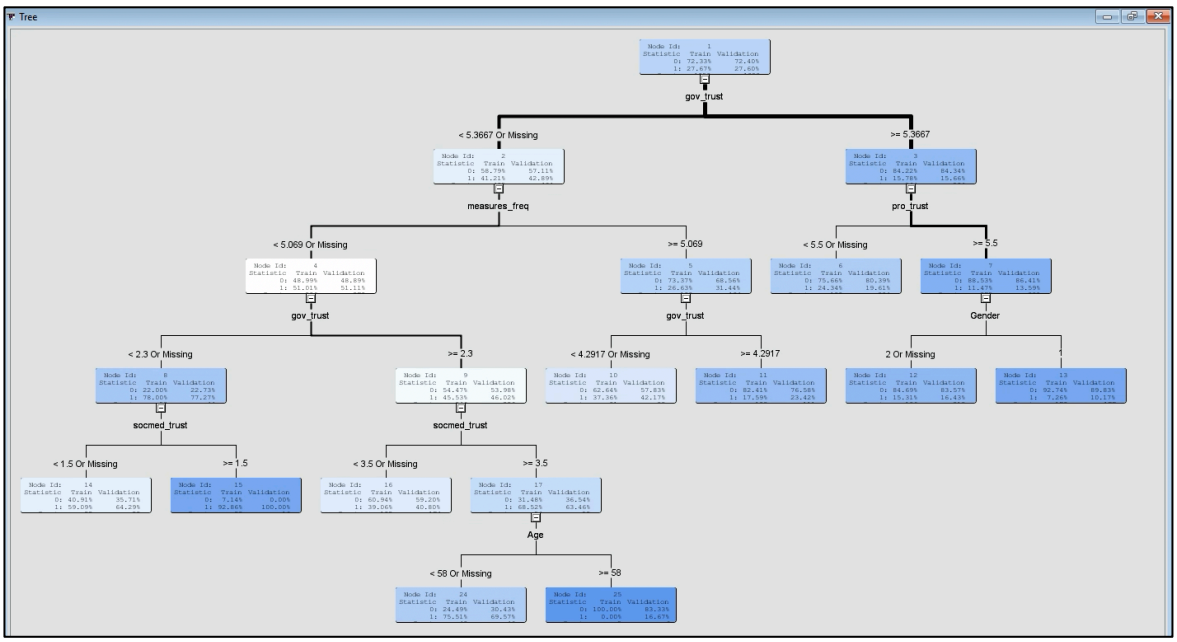
Decision tree allows us to identify worthy variables and compare likelihood of hesitancy across segment splits. We generated maximal tree and several optimal trees using ASE and misclassification rate model assessment.



We examined the maximal tree first which has resulted to 13 leaves. The first split was the *gov_trust*, where lower government trust rating (5.3 and below) has higher likelihood for vaccine hesitancy. Variables *measure_freq, pro_trust, socmed_trust, gender, age, and illness* were the competing splits. The variables were found to be appearing and splitting several times within the same branch, which implies that the tree can be further improved through pruning or running several trees autonomously.



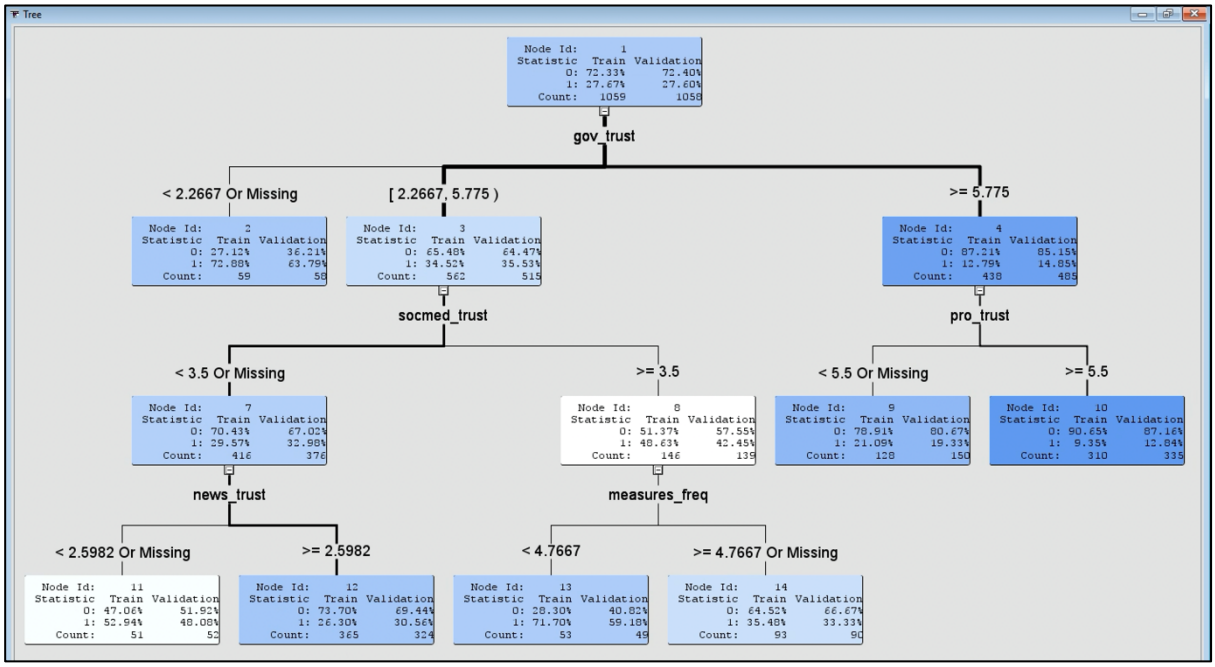
We have generated 2- and 3-way splits using ASE model assessment. In the 2-way splits, we were able to reduce the number of leaves to 10. The validation ASE statistics is 0.1687. We noticed that the *gov_trust* continues to split repetitively halfway through the tree. Increasing the number of splits could potentially resolve this issue and provide us with a finer model.



Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
vax_hesitant		_NOBS_	Sum of Frequencies	1059	1058
vax_hesitant		_MISC_	Misclassification Rate	0.226629	0.236295
vax_hesitant		_MAX_	Maximum Absolute Error	0.928571	1
vax_hesitant		_SSE_	Sum of Squared Errors	342.3241	356.9946
vax_hesitant		_ASE_	Average Squared Error	0.161626	0.168712
vax_hesitant		_RASE_	Root Average Squared Error	0.402027	0.410746
vax_hesitant		_DIV_	Divisor for ASE	2118	2116
vax_hesitant		_DFT_	Total Degrees of Freedom	1059	.

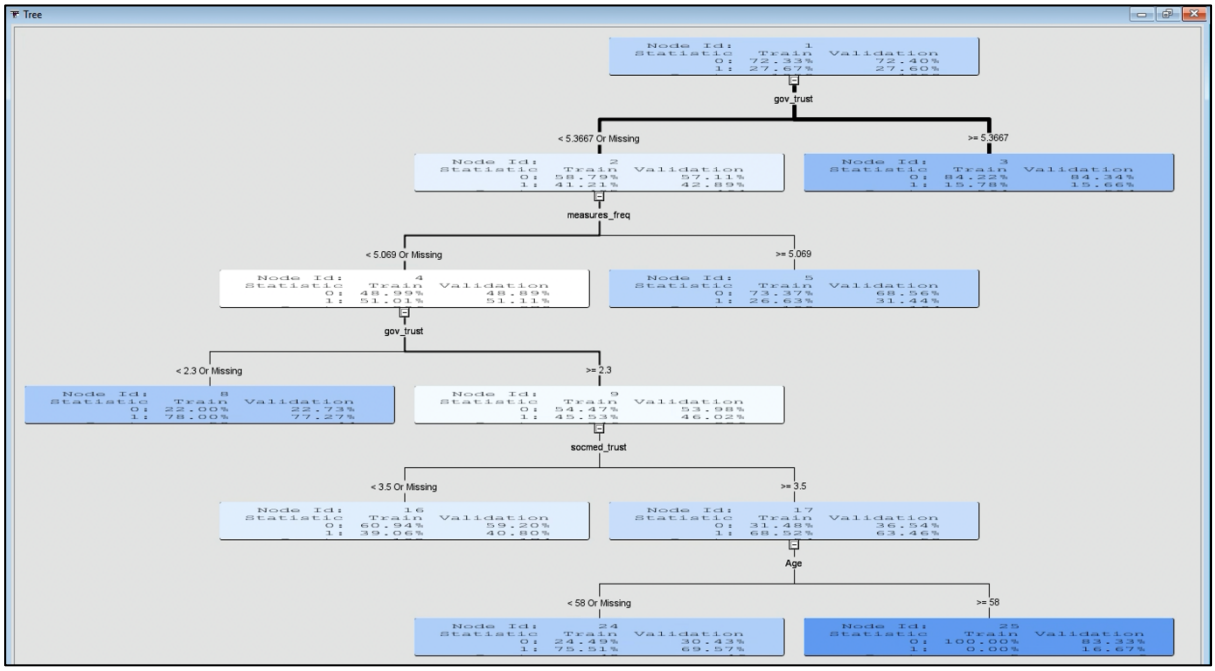
The number of leaves were further reduced after creating 3-way splits. However, validation ASE increased to 0.179. Since the error term was getting large, and the number of leaves is already too low, we did not run for additional number of branches.



Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
vax_hesitant		_NOBS_	Sum of Frequencies	1059	1058
vax_hesitant		_MISC_	Misclassification Rate	0.226629	0.254253
vax_hesitant		_MAX_	Maximum Absolute Error	0.906452	0.906452
vax_hesitant		_SSE_	Sum of Squared Errors	349.5088	380.5758
vax_hesitant		_ASE_	Average Squared Error	0.165018	0.179856
vax_hesitant		_RASE_	Root Average Squared Error	0.406224	0.424095
vax_hesitant		_DIV_	Divisor for ASE	2118	2116
vax_hesitant		_DFT_	Total Degrees of Freedom	1059	.

Using misclassification rate for the model assessment, the validation ASE remained higher at 0.17, and did not produce a better tree model.

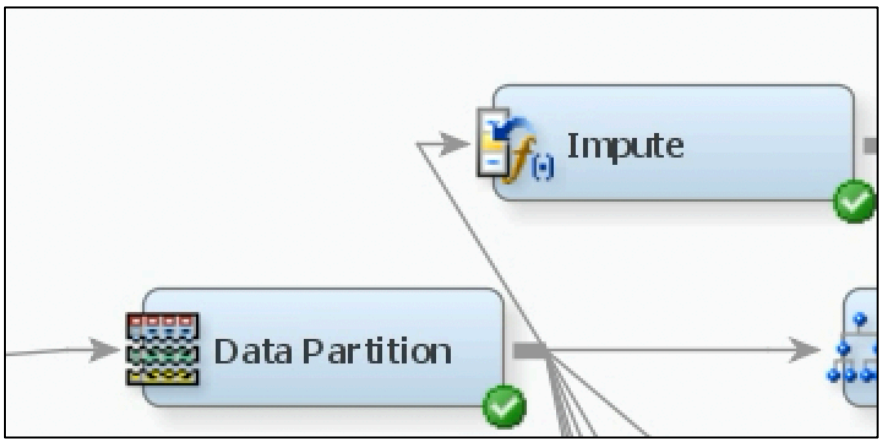


Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
vax_hesitant		_NOBS_	Sum of Frequencies	1059	1058
vax_hesitant		_MISC_	Misclassification Rate	0.226629	0.236295
vax_hesitant		_MAX_	Maximum Absolute Error	0.842199	1
vax_hesitant		_SSE_	Sum of Squared Errors	354.3689	362.8432
vax_hesitant		_ASE_	Average Squared Error	0.167313	0.171476
vax_hesitant		_RASE_	Root Average Squared Error	0.409039	0.414097
vax_hesitant		_DIV_	Divisor for ASE	2118	2116
vax_hesitant		_DFT_	Total Degrees of Freedom	1059	.

4.2 LOGISTIC REGRESSION

Our second predictive modelling technique is logistic regression, which would allow us to identify which among our independent variables contributes to the changes in vaccine hesitancy likelihood and remove those that have no impact. In preparation for regression, we had to impute those with missing responses so that no observations will be discarded from the model.



In the imputation, missing cases for categorical variables were replaced by modal class while interval variables were replaced by the mean. We have also created indicator variables for the missing responses so that we can capture any confounding effects from the missing data.

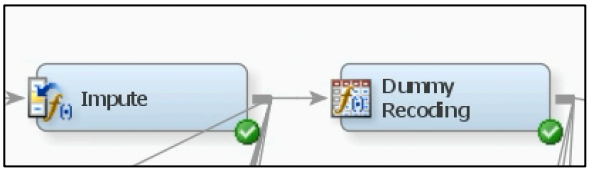
Property	Value
General	
Node ID	Impt
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Nonmissing Variables	No
Missing Cutoff	50.0
Class Variables	
Default Input Method	Count
Default Target Method	None
Normalize Values	Yes
Interval Variables	
Default Input Method	Mean
Default Target Method	None
Default Constant Value	
Default Character Value	
Default Number Value	.

Skewness and kurtosis values remains stable after imputation of missing cases which implies that further data transformation is not required prior to regression.

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	vax_hesitant	0	IMP_gov_tr...	5.75	0	766	1	7	5.486557	1.190895	-1.03694	1.346239	INPUT	Imputed go...	0.054818	0.143313	1
TRAIN	vax_hesitant	1	IMP_gov_tr...	4.8	0	293	1	7	4.455992	1.57114	-0.62563	-0.20984	INPUT	Imputed go...	-0.14331	0.143313	2
TRAIN	vax_hesitant	0	IMP_news_...	4.75	0	766	1	7	4.654704	1.162262	-0.61483	0.319456	INPUT	Imputed ne...	0.04673	0.122168	1
TRAIN	vax_hesitant	1	IMP_news_...	4.125	0	293	1	7	3.903633	1.363247	-0.28127	-0.54635	INPUT	Imputed ne...	-0.12217	0.122168	2
TRAIN	vax_hesitant	0	IMP_socme...	2	0	766	1	7	2.425086	1.459126	0.888857	0.13835	INPUT	Imputed so...	-0.04307	0.112597	1
TRAIN	vax_hesitant	1	IMP_socme...	3	0	293	1	7	2.81958	1.517399	0.434474	-0.71051	INPUT	Imputed so...	0.112597	0.112597	2
TRAIN	vax_hesitant	0	IMP_pro_tr...	6	0	766	1	7	5.589157	1.105384	-0.92311	1.458344	INPUT	Imputed pr...	0.038206	0.099884	1
TRAIN	vax_hesitant	1	IMP_pro_tr...	5	0	293	1	7	4.845751	1.299793	-0.88815	1.045127	INPUT	Imputed pr...	-0.09988	0.099884	2
TRAIN	vax_hesitant	0	IMP_Age	50	0	766	18	95	49.78895	17.20521	0.016268	-1.08553	INPUT	Imputed Age	0.030083	0.078646	1
TRAIN	vax_hesitant	1	IMP_Age	42	0	293	18	77	44.53357	15.15416	0.376729	-0.79885	INPUT	Imputed Age	-0.07865	0.078646	2
TRAIN	vax_hesitant	0	measures_...	5.2	0	766	2.266667	7	5.129601	0.817341	-0.46652	0.129704	INPUT	measures_...	0.025253	0.06602	1
TRAIN	vax_hesitant	1	measures_...	4.692308	0	293	1	6.833333	4.672937	0.956264	-0.44779	0.317837	INPUT	measures_...	-0.06602	0.06602	2
TRAIN	vax_hesitant	0	REP_HH_s...	2	0	766	1	6	2.442559	1.197477	0.891625	0.307546	INPUT	Replaceme...	-0.01834	0.047951	1
TRAIN	vax_hesitant	1	REP_HH_s...	2	0	293	1	6	2.607509	1.257666	0.711639	-0.01619	INPUT	Replaceme...	0.047951	0.047951	2
TRAIN	vax_hesitant	0	IMP_circle_...	4	0	766	1	7	3.959124	1.236416	-0.26946	-0.21269	INPUT	Imputed cir...	0.008369	0.02188	1
TRAIN	vax_hesitant	1	IMP_circle_...	4	0	293	1	7	3.840356	1.323176	-0.26946	-0.01509	INPUT	Imputed cir...	-0.02188	0.02188	2

To make interpretation of the regression output and odds ratio viable, dummy indicator variables were created for the categorical variables using transformation node.



Property	Value
General	
Node ID	Trans
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Formulas	
Interactions	
SAS Code	
Default Methods	
Interval Inputs	None
Interval Targets	None
Class Inputs	Dummy Indicators
Class Targets	None
Treat Missing as Level	No

After the preparatory steps were done on the data, we ran full regression to check significance of each of our independent variables.

Based on the full regression, *gov_trust, news_trust, pro_trust, socmed_trust,* and *measures*_freq were found to be significant in predicting vaccine hesitancy.

Output

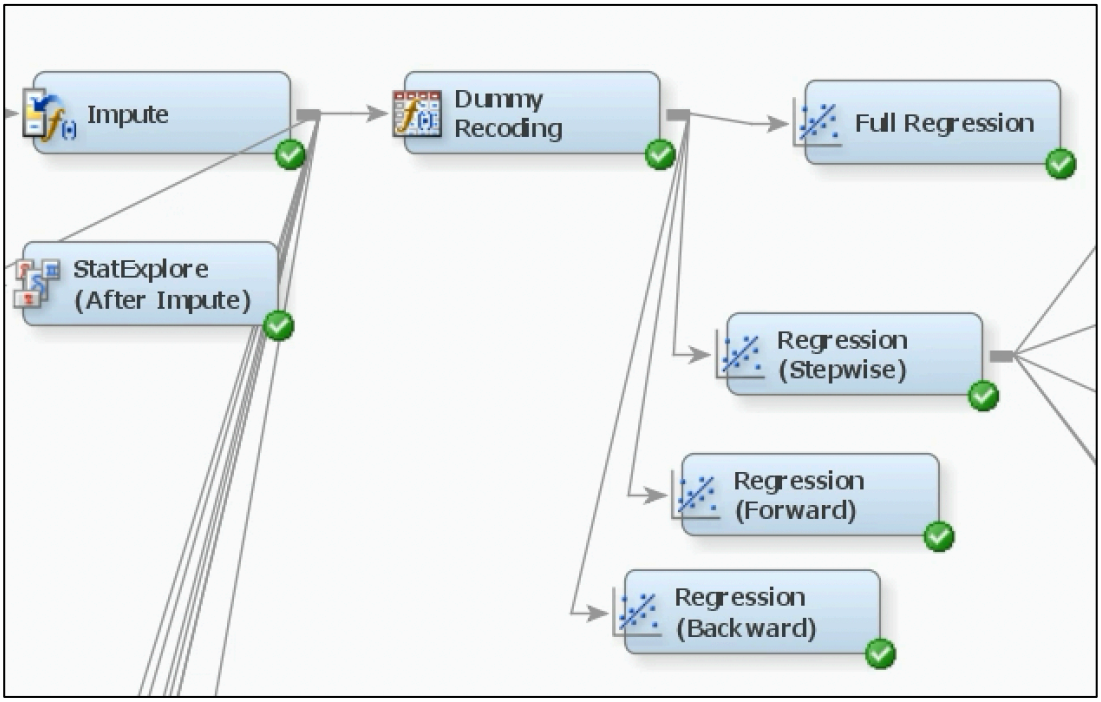
Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	0.1166	63.9876	0.00	0.9985		1.124
IMP_Age		1	-0.00701	0.00589	1.41	0.2343	-0.0650	0.993
IMP_circle_trust		1	-0.0248	0.0798	0.10	0.7558	-0.0173	0.975
IMP_gov_trust		1	-0.2932	0.0903	10.55	0.0012	-0.2240	0.746
IMP_news_trust		1	-0.2106	0.1055	3.99	0.0459	-0.1470	0.810
IMP_pro_trust		1	-0.2373	0.0778	9.30	0.0023	-0.1581	0.789
IMP_socmed_trust		1	0.2927	0.0677	18.68	<.0001	0.2397	1.340
REP_HH_size		1	0.0607	0.0895	0.46	0.4974	0.0407	1.063
TI_Edu1	0	1	-0.2964	0.2848	1.08	0.2980		0.744
TI_Edu2	0	1	-0.0924	0.1677	0.30	0.5816		0.912
TI_Edu3	0	1	-0.2157	0.2112	1.04	0.3073		0.806
TI_Edu4	0	1	-0.2844	0.1728	2.71	0.0998		0.752
TI_Edu5	0	1	-0.1570	0.1506	1.09	0.2970		0.855
TI_Edu6	0	1	-0.2731	0.1911	2.04	0.1530		0.761
TI_Edu7	0	1	0.0363	0.1412	0.07	0.7970		1.037
TI_Edu8	0	0	0	.	.	.	.	.
TI_Gender1	0	1	0.1005	0.0839	1.43	0.2312		1.106
TI_Gender2	0	0	0	.	.	.	.	.
TI_IMP_illness1	0	1	-0.0388	0.0984	0.16	0.6931		0.962
TI_IMP_illness2	0	0	0	.	.	.	.	.
TI_IMP_income1	0	1	-0.5833	0.3146	3.44	0.0637		0.558
TI_IMP_income2	0	1	-0.4251	0.2869	2.20	0.1383		0.654
TI_IMP_income3	0	1	-0.2856	0.2840	1.01	0.3146		0.752
TI_IMP_income4	0	1	-0.2772	0.2856	0.94	0.3319		0.758
TI_IMP_income5	0	1	-0.4967	0.2804	3.14	0.0765		0.609
TI_IMP_income6	0	1	-0.2254	0.2761	0.67	0.4143		0.798
TI_IMP_income7	0	1	-0.1082	0.3082	0.12	0.7256		0.897
TI_IMP_income8	0	0	0	.	.	.	.	.
TI_IMP_race1	0	1	0.2636	0.2631	1.00	0.3164		1.302
TI_IMP_race2	0	1	0.3324	0.3195	1.08	0.2983		1.394
TI_IMP_race3	0	1	-0.2682	0.3876	0.48	0.4891		0.765
TI_IMP_race4	0	1	0.2546	0.4494	0.32	0.5710		1.290
TI_IMP_race5	0	1	-0.0354	0.4136	0.01	0.9318		0.965
TI_IMP_race6	0	1	-0.6837	0.9280	0.54	0.4613		0.505
TI_IMP_race7	0	1	3.9623	45.2091	0.01	0.9302		52.576
TI_IMP_race8	0	0	0	.	.	.	.	.
TI_M_Age1	0	1	0.6410	0.7470	0.74	0.3908		1.898
TI_M_Age2	0	0	0	.	.	.	.	.
TI_M_circle_trust1	0	1	0.3127	0.3084	1.03	0.3106		1.367

Output

255	TI_M_circle_trust2	0	0	0	.	.	.	.	.
256	TI_M_gov_trust1	0	1	0.6102	0.6600	0.85	0.3552		1.841
257	TI_M_gov_trust2	0	0	0	.	.	.	.	.
258	TI_M_illness1	0	1	0.0752	0.3705	0.04	0.8392		1.078
259	TI_M_illness2	0	0	0	.	.	.	.	.
260	TI_M_income1	0	1	0.1834	0.1723	1.13	0.2871		1.201
261	TI_M_income2	0	0	0	.	.	.	.	.
262	TI_M_news_trust1	0	1	0.2412	0.7699	0.10	0.7541		1.273
263	TI_M_news_trust2	0	0	0	.	.	.	.	.
264	TI_M_pro_trust1	0	1	0.1479	0.1326	1.24	0.2646		1.159
265	TI_M_pro_trust2	0	0	0	.	.	.	.	.
266	TI_M_race1	0	1	-0.00373	0.1405	0.00	0.9788		0.996
267	TI_M_race2	0	0	0	.	.	.	.	.
268	TI_M_socmed_trust1	0	1	-0.5690	0.3874	2.16	0.1419		0.566
269	TI_M_socmed_trust2	0	0	0	.	.	.	.	.
270	TI_community_size1	0	1	0.0987	0.1900	0.27	0.6033		1.104
271	TI_community_size2	0	1	0.0656	0.1843	0.13	0.7219		1.068
272	TI_community_size3	0	1	0.0808	0.1978	0.17	0.6828		1.084
273	TI_community_size4	0	1	0.0895	0.1933	0.21	0.6433		1.094
274	TI_community_size5	0	0	0	.	.	.	.	.
275	TI_kids1	0	1	0.0383	0.1168	0.11	0.7429		1.039
276	TI_kids2	0	0	0	.	.	.	.	.
277	TI_province1	0	1	-0.00732	0.1444	0.00	0.9596		0.993
278	TI_province10	0	1	0.0198	0.1976	0.01	0.9201		1.020
279	TI_province11	0	1	4.1148	45.2050	0.01	0.9275		61.239
280	TI_province2	0	1	-0.1548	0.1396	1.23	0.2675		0.857
281	TI_province3	0	1	-0.0185	0.1894	0.01	0.9224		0.982
282	TI_province4	0	1	0.1576	0.2429	0.42	0.5165		1.171
283	TI_province5	0	1	0.3450	0.2855	1.46	0.2269		1.412
284	TI_province6	0	1	0.3047	0.2336	1.70	0.1920		1.356
285	TI_province7	0	1	-0.1674	0.1191	1.98	0.1597		0.846
286	TI_province8	0	1	0.4597	0.5714	0.65	0.4211		1.584
287	TI_province9	0	0	0	.	.	.	.	.
288	measures_freq		1	-0.4458	0.0984	20.53	<.0001	-0.2167	0.640
289									

To eliminate insignificant variables, we have run three regression models using Stepwise, Forward, and Backward Elimination.



Due to the high number of independent variables, backward elimination performed poorly and failed to remove insignificant variables from the model.

Summary of Backward Elimination

Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq	Validation Error Rate
1	TI_M_race1	1	54	0.0007	0.9788	1098.2
2	TI_province1	1	53	0.0024	0.9608	1098.3
3	TI_IMP_race5	1	52	0.0069	0.9336	1098.4
4	TI_IMP_race7	1	51	0.0074	0.9316	1084.6
5	TI_province3	1	50	0.0073	0.9320	1084.6
6	TI_province11	1	49	0.0193	0.8894	1084.2
7	TI_province10	1	48	0.0163	0.8983	1084.2
8	TI_M_illness1	1	47	0.0404	0.8407	1084.3
9	TI_Edu7	1	46	0.0422	0.8373	1084.4
10	IMP_circle_trust	1	45	0.0675	0.7951	1083.1
11	TI_kids1	1	44	0.1038	0.7474	1083.0
12	TI_IMP_income7	1	43	0.1038	0.7473	1083.3
13	TI_M_news_trust1	1	42	0.1160	0.7334	1081.8
14	TI_community_size2	1	41	0.1457	0.7027	1083.2
15	TI_community_size3	1	40	0.0399	0.8418	1083.5
16	TI_community_size4	1	39	0.0865	0.7687	1082.9
17	TI_community_size1	1	38	0.0580	0.8097	1083.3
18	TI_IMP_illness1	1	37	0.2080	0.6483	1082.0
19	TI_IMP_race4	1	36	0.3946	0.5299	1083.5
20	TI_province4	1	35	0.4097	0.5221	1081.9
21	TI_Edu2	1	34	0.6811	0.4092	1082.4
22	TI_M_Age1	1	33	0.7505	0.3863	1083.4
23	TI_IMP_race6	1	32	0.8005	0.3709	1082.2
24	TI_IMP_race3	1	31	0.7855	0.3755	1082.1
25	TI_IMP_income6	1	30	0.7197	0.3963	1080.4
26	TI_IMP_income4	1	29	0.5408	0.4621	1078.4
27	TI_IMP_income3	1	28	0.6985	0.4033	1082.3
28	REP_HH_size	1	27	0.8514	0.3561	1081.9
29	TI_province8	1	26	1.0977	0.2948	1080.7
30	TI_M_income1	1	25	1.0144	0.3139	1079.0
31	TI_province5	1	24	1.2622	0.2612	1076.5
32	TI_M_circle_trust1	1	23	1.3612	0.2433	1078.4
33	TI_province6	1	22	1.2425	0.2650	1076.6
34	TI_Edu1	1	21	1.1649	0.2804	1075.2
35	TI_IMP_income2	1	20	1.3257	0.2496	1069.5
36	TI_M_socmed_trust1	1	19	1.2717	0.2595	1060.6
37	TI_Edu3	1	18	1.6330	0.2013	1065.0
38	TI_Edu5	1	17	1.3404	0.2470	1067.5
39	TI_M_pro_trust1	1	16	1.4533	0.2280	1067.4

The stepwise and forward elimination has produced similar output, and their validation ASE is both at 0.1625. Therefore, we can use either of the two as our best regression model.

NOTE: No (additional) effects met the 0.05 significance level for entry into the model.

Summary of Stepwise Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Validation Error Rate
1	IMP_gov_trust	1	1	117.3744		<.0001	1112.5
2	IMP_socmed_trust	1	2	22.9116		<.0001	1098.2
3	measures_freq	1	3	26.1522		<.0001	1062.5
4	IMP_pro_trust	1	4	12.5496		0.0004	1064.3
5	IMP_news_trust	1	5	7.7091		0.0055	1062.6
6	TI_province7	1	6	6.2757		0.0122	1070.0
7	TI_IMP_incomel	1	7	4.2633		0.0389	1065.0
8	TI_IMP_race3	1	8	4.5873		0.0322	1059.7
9	TI_IMP_income5	1	9	4.5805		0.0323	1072.8
10	TI_M_gov_trustl	1	10	4.8978		0.0269	1075.0

NOTE: No (additional) effects met the 0.05 significance level for entry into the model.

Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq	Validation Error Rate
1	IMP_gov_trust	1	1	117.3744	<.0001	1112.5
2	IMP_socmed_trust	1	2	22.9116	<.0001	1098.2
3	measures_freq	1	3	26.1522	<.0001	1062.5
4	IMP_pro_trust	1	4	12.5496	0.0004	1064.3
5	IMP_news_trust	1	5	7.7091	0.0055	1062.6
6	TI_province7	1	6	6.2757	0.0122	1070.0
7	TI_IMP_incomel	1	7	4.2633	0.0389	1065.0
8	TI_IMP_race3	1	8	4.5873	0.0322	1059.7
9	TI_IMP_income5	1	9	4.5805	0.0323	1072.8
10	TI_M_gov_trustl	1	10	4.8978	0.0269	1075.0

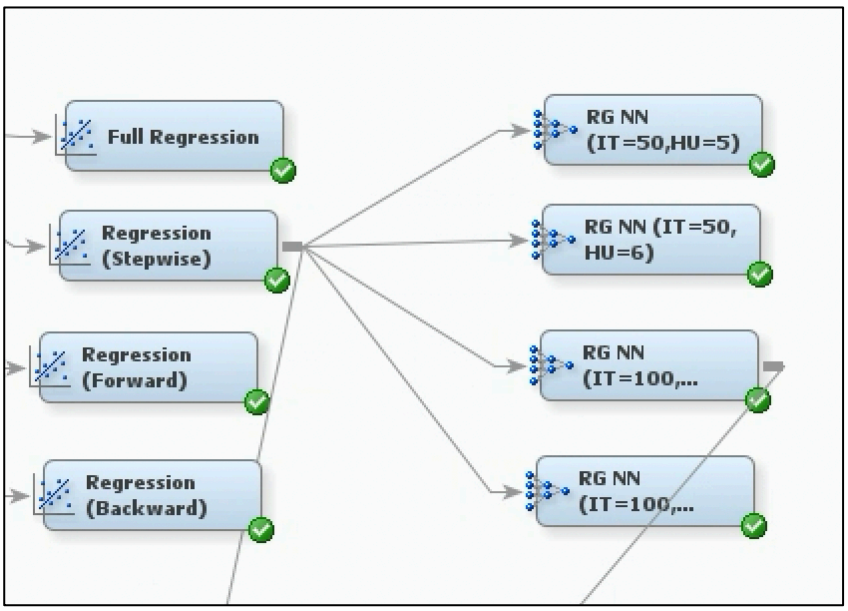
We have also examined the odds ratio estimates. It is quite notable that increase in trust rating on government, news, and professionals result to decrease in likelihood in vaccine hesitancy by at least 20%. On the other hand, reliance on social media information result to 37% increase in likelihood. This output concurred with the probabilities from the decision tree model. In relation to demographic characteristics, least income bracket, blacks, and those living in Ontario are found to be more likely to be hesitant. Same trend is evident for those who are less frequently practicing measures to prevent getting the virus.

Odds Ratio Estimates

Effect		Point Estimate
IMP_gov_trust		0.759
IMP_news_trust		0.777
IMP_pro_trust		0.802
IMP_socmed_trust		1.370
TI_IMP_incomel	0 vs 1	0.526
TI_IMP_race3	0 vs 1	0.325
TI_province7	0 vs 1	0.660
measures_freq		0.649

4.3 NEURAL NETWORK

The third predictive modelling technique is Neural Network. This technique recognizes patterns in the data to yield probabilities. Due to its complexity, it is quite difficult to apply and interpret in business decision-making. However, it would still be beneficial to see if adding neural network to our regression output have better model. We have added neural network nodes to the stepwise regression output, running several iterations and hidden units to search for the most optimal model.



The model failed to converge at the initial run of 50 iterations and 5 as well as 6 hidden units, which indicates that there were too few iterations to produce accurate results.

Optimization Results

Iterations	50	Function Calls	60
Jacobian Calls	52	Active Constraints	0
Objective Function	0.4545803026	Max Abs Gradient Element	0.0014620119
Lambda	0.0087028253	Actual Over Pred Change	0.2869349895
Radius	0.1434318026		

LEVMAR needs more than 50 iterations or 2147483647 function calls.

WARNING: LEVMAR Optimization cannot be completed.

Convergence criterion was satisfied upon increasing the number of iterations to 100 and 5 hidden units, with validation ASE of 0.1642. We see no improvement even after attempting to increase hidden units to 6. Hence, it would be sufficient to settle with the neural network having 5 hidden units.

Optimization Results

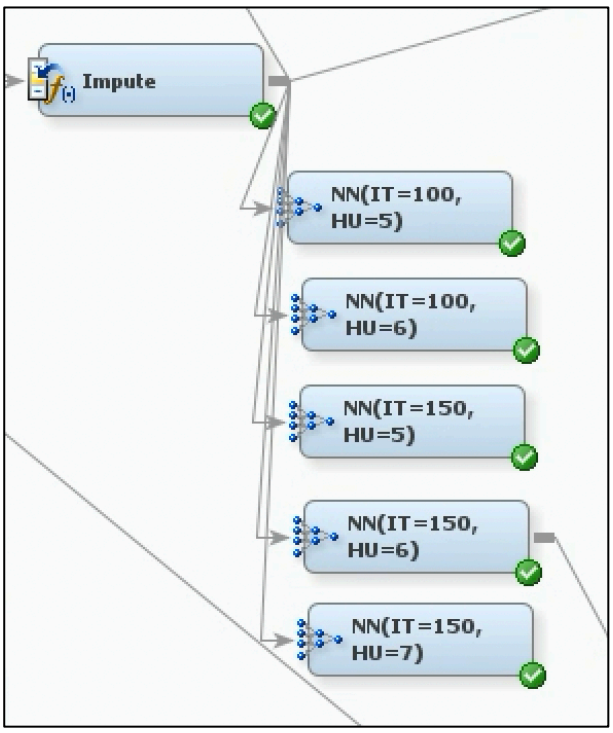
Iterations	82	Function Calls	100
Jacobian Calls	85	Active Constraints	0
Objective Function	0.4463351952	Max Abs Gradient Element	0.0007517526
Lambda	1.4840650068	Actual Over Pred Change	0.9654595834
Radius	0.0070471764		

Convergence criterion (FCONV=0.0001) satisfied.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
vax_hesitant		_DFT_	Total Degrees of Freedom	1059	.
vax_hesitant		_DFE_	Degrees of Freedom for Error	1008	.
vax_hesitant		_DFM_	Model Degrees of Freedom	51	.
vax_hesitant		_NW_	Number of Estimated Weights	51	.
vax_hesitant		_AIC_	Akaike's Information Criterion	1128.151	.
vax_hesitant		_SBC_	Schwarz's Bayesian Criterion	1381.37	.
vax_hesitant		_ASE_	Average Squared Error	0.15817	0.164235
vax_hesitant		_MAX_	Maximum Absolute Error	0.951215	0.96349
vax_hesitant		_DIV_	Divisor for ASE	2118	2116
vax_hesitant		_NOBS_	Sum of Frequencies	1059	1058
vax_hesitant		_RASE_	Root Average Squared Error	0.397706	0.405259
vax_hesitant		_SSE_	Sum of Squared Errors	335.0048	347.521
vax_hesitant		_SUMW_	Sum of Case Weights Times Freq	2118	2116
vax_hesitant		_FPE_	Final Prediction Error	0.174176	.
vax_hesitant		_MSE_	Mean Squared Error	0.166173	0.164235
vax_hesitant		_RFPE_	Root Final Prediction Error	0.417344	.
vax_hesitant		_RMSE_	Root Mean Squared Error	0.407643	0.405259
vax_hesitant		_AVERR_	Average Error Function	0.484491	0.506222
vax_hesitant		_ERR_	Error Function	1026.151	1071.165
vax_hesitant		_MISC_	Misclassification Rate	0.228517	0.226843
vax_hesitant		_WRONG_	Number of Wrong Classifications	242	240

Additional neural network nodes were connected to the impute mode to test effects of altering the model. Convergence was achieved using 150 iterations and 6 hidden units, with validation ASE of 0.1644 which is close to the neural network output even after regression. To check if this model can still be further improved, hidden units were increased to 7; however, it resulted to higher validation ASE.



Fit Statistics

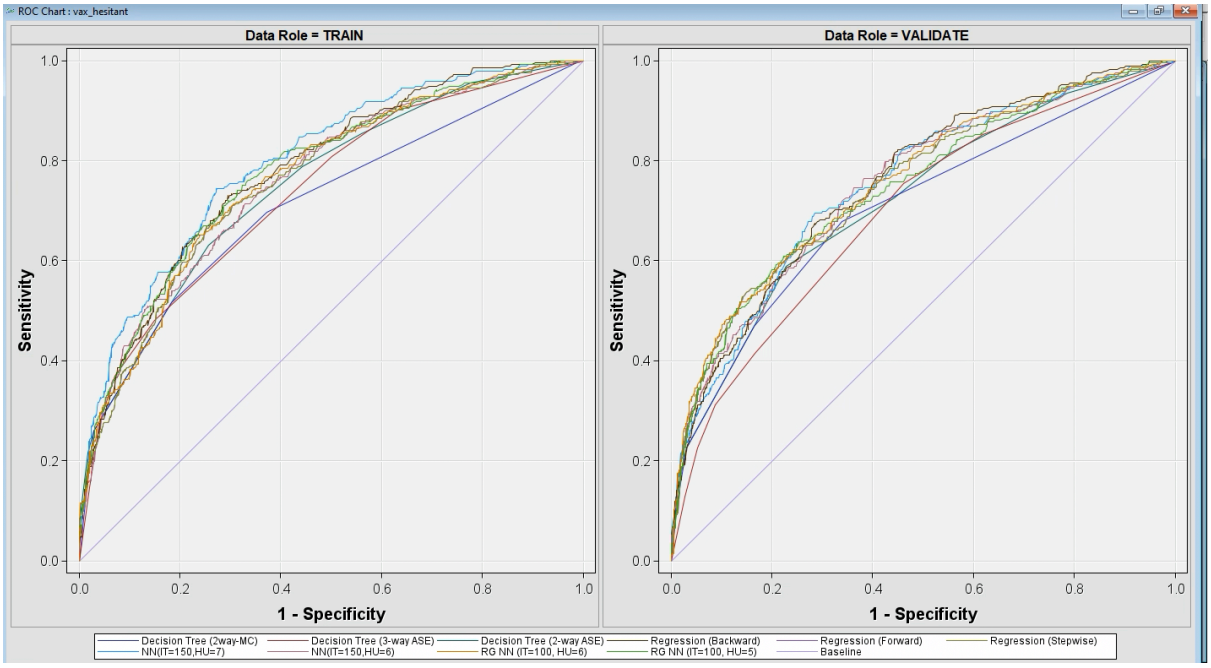
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
vax_hesitant		_DFT_	Total Degrees of Freedom	1059	.
vax_hesitant		_DFE_	Degrees of Freedom for Error	716	.
vax_hesitant		_DFM_	Model Degrees of Freedom	343	.
vax_hesitant		_NW_	Number of Estimated Weights	343	.
vax_hesitant		_AIC_	Akaike's Information Criterion	1741.663	.
vax_hesitant		_SBC_	Schwarz's Bayesian Criterion	3444.686	.
vax_hesitant		_ASE_	Average Squared Error	0.162871	0.164408
vax_hesitant		_MAX_	Maximum Absolute Error	0.967211	0.970102
vax_hesitant		_DIV_	Divisor for ASE	2118	2116
vax_hesitant		_NOBS_	Sum of Frequencies	1059	1058
vax_hesitant		_RASE_	Root Average Squared Error	0.403573	0.405473
vax_hesitant		_SSE_	Sum of Squared Errors	344.9607	347.8876
vax_hesitant		_SUMW_	Sum of Case Weights Times Freq	2118	2116
vax_hesitant		_FPE_	Final Prediction Error	0.318918	.
vax_hesitant		_MSE_	Mean Squared Error	0.240894	0.164408
vax_hesitant		_RFPE_	Root Final Prediction Error	0.564728	.
vax_hesitant		_RMSE_	Root Mean Squared Error	0.49081	0.405473
vax_hesitant		_AVERR_	Average Error Function	0.498425	0.505233
vax_hesitant		_ERR_	Error Function	1055.663	1069.072
vax_hesitant		_MISC_	Misclassification Rate	0.225685	0.227788
vax_hesitant		_WRONG_	Number of Wrong Classifications	239	241

4.4 MODEL COMPARISON

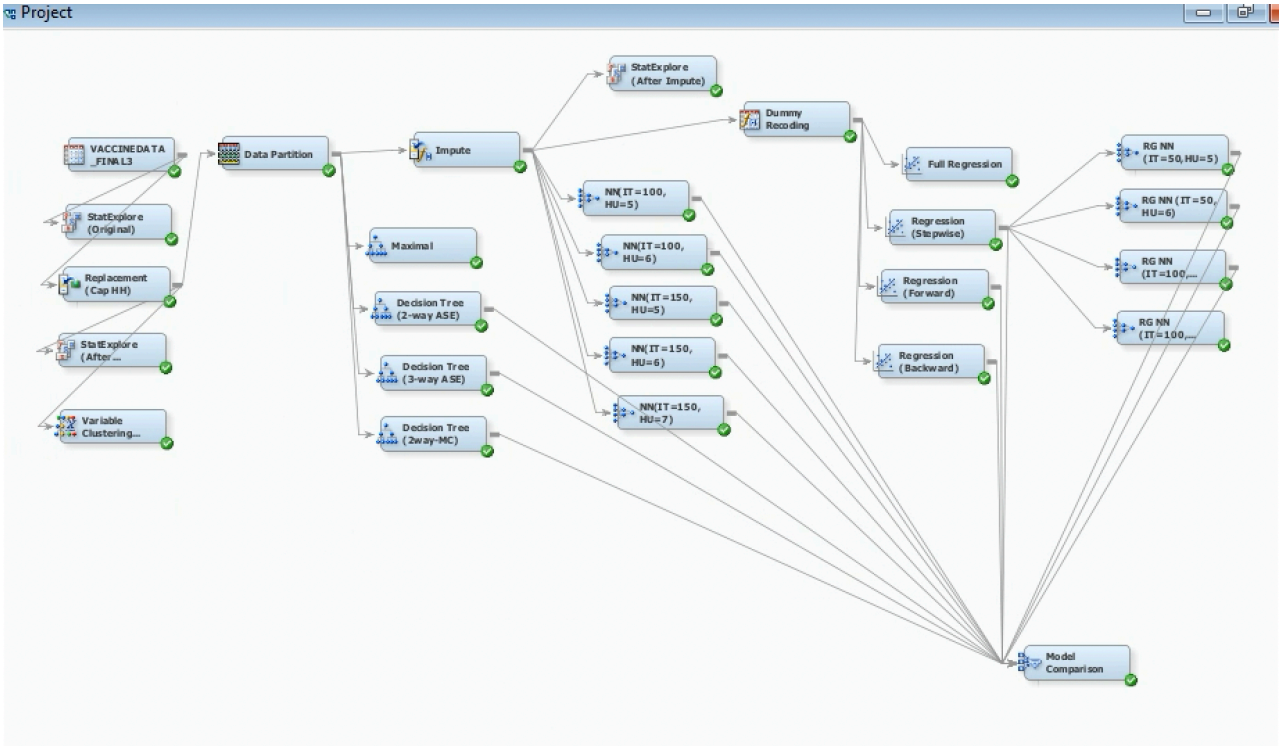
Below are the results for the comparison of all the generated models (Decision Tree, Logistic Regression, and Neural Network) in terms of average squared error, misclassification rate, and ROC Index.

Model Description	Valid: Average Squared Error ▲	Valid: Misclassification Rate	Valid: Roc Index
RG NN (IT=100, HU=6)	0.160352	0.217391	0.756v
Regression (Stepwise)	0.162488	0.224008	0.751v
Regression (Forward)	0.162488	0.224008	0.751v
RG NN (IT=100, HU=5)	0.164235	0.226843	0.745v
Regression (Backward)	0.164304	0.229679	0.755v
NN(IT=150,HU=6)	0.164408	0.227788	0.75v
NN(IT=150,HU=7)	0.16523	0.232514	0.751v
Decision Tree (2-way ASE)	0.168712	0.236295	0.721v
Decision Tree (2way-MC)	0.171476	0.236295	0.704v
Decision Tree (3-way ASE)	0.179856	0.254253	0.696v

Neural network model after regression produced the least average squared error and misclassification rate, closely followed by regression (using either stepwise or forward), whereas models from the decision tree were the highest. Neural network also has the best ROC index among all the models. The ROC chart is provided below.



Below is a screenshot of the final project diagram:



We can infer that adding neural network has slightly improved the regression fit. However, since the difference is minute, it is still recommended to utilize regression due to interpretability.

5. CONCLUSION

In this case, the best model for determining the vaccine hesitancy of people in Canada is the forward/stepwise regression model.

Based on our final model, we conclude that the major predictors for vaccine hesitancy are the people's trust and reliance on various sources of information. As trust ratings on government, health news sources, and health professionals increase, the likelihood of hesitancy decreases by 20 to 25%. This is supported by a study conducted by Douglas et. al. (2021), wherein “Those with confidence in the vaccine conveyed trust in science and their healthcare professional, expressed concerns about potential long-term COVID-19 effects and felt that the vaccine was necessary to return

to normal." On the other hand, higher reliance on social media corresponds to 37% increase in likelihood.

In terms of demographics, individuals from low-income bracket are found to be more likely to refuse taking the vaccine. This is backed up by an article of CTV where Yun (2021) explained that "those living in lower-income and rural communities are more likely to say no to the vaccine." Blacks and people living in Ontario also tend to be more hesitant. This finding was also supported in the study about "Understanding national trends in COVID-19 vaccine hesitancy in Canada: results from five sequential cross- sectional representative surveys spanning April 2020–March 2021" where the authors mentioned that aside from other factors, vaccine hesitancy was also high in "those living in Western provinces and Ontario compared with Quebec and the Atlantic provinces" (Lavoie, et al., 2021) .

We also found significance in their behavior in practicing measures to prevent getting the virus, such as hand washing, sanitizing, and social distancing. Those who follow measures less frequently would tend to hesitate more.

6. RECOMMENDATION

We each have a vital part in preventing the spread of the virus. While we continue to follow safety precautions such as wearing masks, washing our hands correctly, sanitising, and maintaining social distance, vaccination is essential to protect ourselves and others around us; it is a societal obligation to establish herd immunity. Community vaccination reluctance exists, but it may be handled with the correct method.

Our team proposes conducting focus group discussions to better comprehend why individuals do not trust government leaders and medical professionals. In addition, being proactive in social media by forming partnerships with brands,

influencers, and social media companies will aid in the dissemination of information about COVID-19 and the COVID-19 vaccine. To further boost public confidence and trust in the government, a study from OECD (2021) suggested that government acts should be available to public examination, and public institutions should interact with the community by:

- Proactively releasing timely information about immunisation strategies, methodologies, and successes in disaggregated, user-friendly, and open source formats;
- Improving public communication openness and coherence to combat deception; and
- Involving the public in the formulation of vaccination strategies, as well as the style and content of crucial messages.

Furthermore, inorder to address COVID-19 vaccination misinformation in a community, the government and public health authorities should disseminate accurate, clear, and easy-to-find information that answers frequently asked concerns. This can be accomplished through their website, social media, and other locations where the general public searches for health information.

Combating COVID-19 disinformation is crucial to vaccine uptake. Boosting confidence in the safety and efficiency of the COVID-19 vaccinations, countering apathy about the pandemic, and increasing the convenience of getting vaccinated are ways to increase vaccine acceptance. Community-tailored engagement, outreach, and interventions are necessary to address reluctance factors and promote vaccine uptake.

References

Douglas, S., Laila, A., & Tang, L. (December, 2021). *Among sheeples and antivaxxers: Social media responses to COVID-19 vaccine news posted by Canadian news organizations, and recommendations to counter vaccine hesitancy.* Retrieved from https://www.canada.ca/en/public-health/services/reports-publications/canada-communicable-disease-report-ccdr/monthly-issue/2021-47/issue-12-december-2021/social-media-responses-covid-19-vaccine-posted-canadian-news-recommendations-counter-vaccine-hesitancy.html

Herhalt, C. (25 August, 2021). *'Herd immunity' no longer possible without vaccinating young children, Public Health Ontario says*. Retrieved from CTV News: https://toronto.ctvnews.ca/herd-immunity-no-longer-possible-without-vaccinating-young-children-public-health-ontario-says-1.5560431

Lavoie, K., Gosselin-Boucher, V., Stojanovic, J., Gupta, S., Gagné, M., Joyal-Desmarais, K., . . . Presseau, J. (22 November , 2021). *Understanding national trends in COVID-19 vaccine hesitancy in Canada: results from five sequential cross- sectional representative surveys spanning April 2020–March 2021*. Retrieved from BMJ Open: https://bmjopen.bmj.com/content/bmjopen/12/4/e059411.full.pdf

OECD. (10 May, 2021). *Enhancing public trust in COVID-19 vaccination: The role of governments.* Retrieved from Organization fro Economic Co-operation and Development: https://www.oecd.org/coronavirus/policy-responses/enhancing-public-trust-in-covid-19-vaccination-the-role-of-governments-eae0ec5a/

WHO. (15 August, 2021). *Canada Situation*. Retrieved from World Health Organization: https://covid19.who.int/region/amro/country/ca

WHO. (2019). *Ten threats to global health in 2019*. Retrieved from World Health Organization: https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019

Yun, T. (30 March, 2021). *Vaccine hesitancy higher among rural, low-income communities: study.* CTV News. Retrieved from https://www.ctvnews.ca/health/coronavirus/vaccine-hesitancy-higher-among-rural-low-income-communities-study-1.5368415

Appendix

Appendix A: SAS Code for Data Import and Data Cleaning

```
*Importing original CSV data from Kaggle*/


 proc import datafile="H:\BA706\Vaccine
Heistancy\COSMO_in_Canada_Waves_1-8_FINAL.csv"
        out=wm.vaccinedata
        dbms=csv
        replace;
 run;


data wm.vaccinedata_wave8;
set wm.vaccinedata;


/*getting final wave of dataset only*/


if wave=8;


run;


/*Checking frequency of potential target and independent variables*/


proc freq data=wm.vaccinedata_wave8;
tables
wave
```

E1r1 /*E1r1: If an effective COVID-19 vaccine becomes available and is recommended for me, I would get it. - Please give your opinion on the following statements.*/

E1r2 /*E1r2: If a safe COVID-19 vaccine becomes available and is recommended for me, I would get it. - Please give your opinion on the following statements*/

E1r16 /*E1r16: I would be willing to get vaccinated in order to return to work, travel, or attend large gatherings - Please give your opinion on the following statements.*/

QS1 /*QS1: In what year were you born?*/

QS2 /*QS2a: What best describes your gender?*/

QS3 /*QS3: What is the highest level of formal education that you have completed?*/

QS4 /*QS4: Are you a health care provider (i.e., nurse, medical doctor, paramedic, first responder, nurse practitioner, pharmacist etc.)?*/

QS4a /*QS4a: Are you a frontline worker (i.e., gas station attendant, grocery store clerk, etc.)?*/

QS5 /*QS5: Do you have a serious, long-term illness, like diabetes, emphysema, or high blood pressure?*/

QS6 /*QS6: What is the size of the community you live in?*/

QS7 /*QS7: In which province or territory do you live?*/

QS8 /*QS8: Do you have or live with children under 18 years of age in your home?*/

QS10 /*QS10: Which of the following best describes the impact that COVID-19 has had on your employment?*/

QS14a /*QS14a: In what year did you first move to Canada?*/

QS16r1--QS16r96 /*Q16: race*/

QS17 /*QS17: Which of the following categories best describes your total household income? That is, the total income of all persons in your household combined, before taxes?*/

QS18 /*QS18: How many people live at your address, including yourself?*/

QS20Ar1--QS20Ar9 /*Do you currently own or operate a small (1-99 employees) or medium (100-499 employees) sized business?*/

QS23 /*QS23: Are you an Aboriginal person, that is, First Nations, Métis or Inuk (Inuit)? First Nations includes Status and Non–Status Indians.*/

QS22 /*QS22: How often did you use public transportation prior to the COVID-19 outbreak?*/

QS22a /*QS22a: How often did you use public transportation during the COVID-19 outbreak?*/

A1 /*A1: Do you think crisis situation is worse/same/improved?*/

A3 /*A3: Are you or have you been infected with COVID-19?*/

A4 /*A4: Do you know people in your immediate social network (i.e., friends or close family members) who are or have been infected with COVID-19?*/

A4A /*A4A: Do you know anyone who has died of COVID-19?*/

A5 /*A5: How would you rate your level of knowledge on COVID-19?*/

A6 /*A6: How would you rate your level of knowledge on how to prevent the spread of COVID-19?*/

A14 /*A14: I consider myself to be at high risk of contracting COVID-19.*/

A15 /*A15: I believe that if I get sick with COVID-19, I am at risk for poor health outcomes (e.g. requiring hospitalization)*/

A16 /*A16: I’m worried about the idea of transmitting COVID-19 to people around me.*/

B6r1--B6r20 /*Which of the following work to prevent the spread of COVID-19?Please evaluate all preventive measures listed below.*/

B7r1--B7r21 /*How often have you used the following measures to keep from getting sick with COVID-19?*/

B8r1--B8r8 /*Please indicate how much you disagree or agree with the following statements.*/

B9r1--B9r8 /*Please indicate your answer on the following scale. The numbers allow you to nuance your answer between the two statements.COVID-19 feels...*/

C1r1--C1r22 /*How much do you trust the following sources of information in their reporting about COVID-19?*/

C2r1-C2r22 /*How often do you use the following sources of information to stay informed about COVID-19?*/

;

run;

/*Reducing dataset to keep variables to be used only*/

data wm.vaccinedata2;

set wm.vaccinedata_wave8

(keep=

QS1

QS2

QS3

QS4

QS4a

QS5

QS6

QS7

QS8

QS10

QS14a

QS16r1-QS16r13 QS16r96

QS17

QS18

QS20Ar1-QS20Ar4 QS20Ar9

QS23

QS22

QS22a

A1

A3

A4

A4A

A5

A6

A14

A15

```
A16
B6r1-B6r20
B7r1-B7r21
B8r1-B8r8
B9r1-B9r8
C1r1-C1r22
C2r1-C2r22 C2r96
E1r1
E1r2
E1r16);
run;

data wm.vaccinedata3;
set wm.vaccinedata2;

/*data cleaning to recode values out of range to missing*/

if E1r1 > 7 then E1r1 = .;
if E1r2 > 7 then E1r2 = .;
if E1r16 > 7 then E1r16 =.;

if  B7r1>=95 then B7r1 = .;
if  B7r2>=95 then B7r2 = .;
if  B7r3>=95 then B7r3 = .;
if  B7r4>=95 then B7r4 = .;
```

```
if  B7r5>=95 then B7r5 = .;

if  B7r6>=95 then B7r6 = .;

if  B7r7>=95 then B7r7 = .;

if  B7r8>=95 then B7r8 = .;

if  B7r9>=95 then B7r9 = .;

if  B7r10>=95 then B7r10 = .;

if  B7r11>=95 then B7r11 = .;

if  B7r12>=95 then B7r12 = .;

if  B7r13>=95 then B7r13 = .;

if  B7r14>=95 then B7r14 = .;

if  B7r15>=95 then B7r15 = .;

if  B7r16>=95 then B7r16 = .;

if  B7r17>=95 then B7r17 = .;

if  B7r18>=95 then B7r18 = .;

if  B7r19>=95 then B7r19 = .;

if  B7r20>=95 then B7r20 = .;

if  B7r21>=95 then B7r21 = .;


if B8r1 >= 8 then B8r1 = .;

if B8r2 >= 8 then B8r2 = .;

if B8r3 >= 8 then B8r3 = .;

if B8r4 >= 8 then B8r4 = .;

if B8r5 >= 8 then B8r5 = .;

if B8r6 >= 8 then B8r6 = .;

if B8r7 >= 8 then B8r7 = .;
```

```
if B8r8 >= 8 then B8r8 = .;


if  C1r1>=8 then C1r1 = .;
if  C1r2>=8 then C1r2 = .;
if  C1r3>=8 then C1r3 = .;
if  C1r4>=8 then C1r4 = .;
if  C1r5>=8 then C1r5 = .;
if  C1r6>=8 then C1r6 = .;
if  C1r7>=8 then C1r7 = .;
if  C1r8>=8 then C1r8 = .;
if  C1r9>=8 then C1r9 = .;
if  C1r10>=8 then C1r10 = .;
if  C1r11>=8 then C1r11 = .;
if  C1r12>=8 then C1r12 = .;
if  C1r13>=8 then C1r13 = .;
if  C1r14>=8 then C1r14 = .;
if  C1r15>=8 then C1r15 = .;
if  C1r16>=8 then C1r16 = .;
if  C1r21>=8 then C1r21 = .;
if  C1r17>=8 then C1r17 = .;
if  C1r18>=8 then C1r18 = .;
if  C1r19>=8 then C1r19 = .;
if  C1r20>=8 then C1r20 = .;
if  C1r22>=8 then C1r22 = .;
```

```
if  C2r1>=8 then C2r1 = .;

if  C2r2>=8 then C2r2 = .;

if  C2r3>=8 then C2r3 = .;

if  C2r4>=8 then C2r4 = .;

if  C2r5>=8 then C2r5 = .;

if  C2r6>=8 then C2r6 = .;

if  C2r7>=8 then C2r7 = .;

if  C2r8>=8 then C2r8 = .;

if  C2r9>=8 then C2r9 = .;

if  C2r10>=8 then C2r10 = .;

if  C2r11>=8 then C2r11 = .;

if  C2r12>=8 then C2r12 = .;

if  C2r13>=8 then C2r13 = .;

if  C2r14>=8 then C2r14 = .;

if  C2r15>=8 then C2r15 = .;

if  C2r16>=8 then C2r16 = .;

if  C2r21>=8 then C2r21 = .;

if  C2r17>=8 then C2r17 = .;

if  C2r18>=8 then C2r18 = .;

if  C2r19>=8 then C2r19 = .;

if  C2r20>=8 then C2r20 = .;

if  C2r22>=8 then C2r22 = .;

if  C2r96>=8 then C2r96 = .;


if QS1=9999 then QS1=.;
```

```
QS1_Age = 2020 - QS1;

if QS5=2 then QS5=0;
if QS5 = 98 or QS5 = 9999 then QS5 = .;

if QS6 = 9999 then QS6 = .;

if QS10 <= 3 or QS10 = 5 then QS10 = 1;
if QS10 = 4 or  QS10 = 6 then QS10 = 0;
if QS10 = 98 or QS10=9999 then QS10 = .;

if QS16r1 = 1 then QS16 = 1; /*White*/
if QS16r2 = 1 then QS16 = 2; /*South Asia*/
if QS16r3 = 1 or QS16r5 = 1 or QS16r8 = 1 or QS16r10 = 1 or QS16r11 = 1
then SQ16=3; /*East/ SouthEast Asia*/
if QS16r4 = 1 then QS16 = 4; /*Black*/
if QS16r6 = 1 then  QS16 = 5; /*Latin America*/
if QS16r7 =1 or QS16r9 = 1 then QS16 = 6; /*West Asian*/
if QS16r12 =1 then QS16 = 7; /*Indigenous*/
if QS16r13 = 1 then QS16 = 8; /*Canadian*/
if QS16r96 =1 then QS16 = 96; /*Others*/

if QS8 = 2 then QS8 = 0; /*change No code from 2 to 0*/

if QS17 >= 98 then QS17 =.;
```

```
if QS18 = 9999 then QS18 = .;

if A1 = 98 then A1 = .;

if A3 <= 4 or A3 =6 then A3_recode = 1;
if A3 = 5 then A3_recode = 0;
if A3 = 98 then A3_recode = .;

if A4 <= 4 then A4_recode = 1;
if A4 = 5 then A4_recode = 0;
if A4 = 98 then A4_recode = .;

B7_mean = mean (B7r1,
B7r2,
B7r3,
B7r4,
B7r5,
B7r6,
B7r7,
B7r8,
B7r9,
B7r10,
B7r11,
B7r12,
```

```
B7r13,
B7r14,
B7r15,
B7r16,
B7r17,
B7r18,
B7r19,
B7r20,
B7r21);

B8_mean = mean (B8r1,
B8r2,
B8r3,
B8r4,
B8r5,
B8r6,
B8r7,
B8r8);

news_trust=mean(C1r1,
C1r2,
C1r3,
C1r8,
C1r11,
C1r12,
```

```
C1r13,
C1r14);

gov_trust=mean(C1r9,
C1r10,C1r15,
C1r16,
C1r21);

circle_trust=mean(C1r4,
C1r5,
C1r6);

pro_trust = C1r7;

socmed_trust=mean(C1r17,
C1r18,
C1r19,
C1r20,
C1r22);

run;

/*Creating target variable (vaccine hesitancy) based on T2B*/
```

```
data wm.vaccinedata4;
set wm.vaccinedata3;

T2B_E1r1 = 0;
T2B_E1r2 = 0;
T2B_E1r16 = 0;

if E1r1 = 6 or E1r1 = 7 then T2B_E1r1 = 1;
if E1r2 = 6 or E1r2 = 7 then T2B_E1r2 = 1;
if E1r16 = 6 or E1r16 = 7 then T2B_E1r16 = 1;

if T2B_E1r1 = 0 and T2B_E1r2 = 0 and T2B_E1r16 = 0 then vax_hesitant = 1;
else vax_hesitant = 0;

run;

/*Checking target variable*/

proc freq data=wm.vaccinedata4;
table E1r1 E1r2 E1r16 vax_hesitant;
run;

/*Data reduction to keep final variables*/
```

```
data wm.vaccinedata_final;
set wm.vaccinedata4 (keep=QS1_Age
QS2
QS3
QS5
QS6
QS7
QS8
QS10
QS16
QS17
QS18
QS22
QS22a
A3_recode
A4_recode
A5
A6
B7_mean
B8_mean
news_trust
gov_trust
circle_trust
pro_trust
socmed_trust
```

```
vax_hesitant);
run;


/*Renaming final variable for ease of reference*/

data wm.vaccinedata_final3;
set wm.vaccinedata_final;
rename
QS1_Age=Age
QS2=Gender
QS3=Edu
QS5=illness
QS6=community_size
QS7=province
QS8=kids
QS10=employment_impact
QS16=race
QS17=income
QS18=HH_size
A3_recode=infected
A4_recode=knowinfected
A5=C19knowledge
B7_mean=measures_freq
B8_mean=sentiment;
```

```
run;

/*Final variables checks*/

proc freq data=wm.vaccinedata_final3;
tables gender edu illness community_size province kids employment_impact
race income HH_size infected knowinfected C19knowledge measures_freq
sentiment / missing;
run;

proc means data=wm.vaccinedata_final3 missing;
var gender edu illness community_size province kids employment_impact
race income infected knowinfected
HH_size  C19knowledge measures_freq sentiment news_trust
gov_trust
circle_trust
pro_trust
socmed_trust;
run;

/*Checking for correlation among variables*/

proc corr data=wm.vaccinedata_final3;
var
age
```

HH_size

measures_freq

news_trust

gov_trust

circle_trust

pro_trust

socmed_trust;

run;