# Predicting Drug (Cannabis) Use

In this project we try to use **Logistic regression** to perform **supervised classification** to identify if a subject is a **drug (Cannabis) user** or not.

In Logistic Regression, we model the probability of a case being drug user based on different characteristic or demographic features of that case.

The predicted posterior probabilities from the Logistic regression model is a nonlinear function of the input variables. To linearize the model, we use logit transformation that measures the log odds of the posterior probability which is a linear function of the input function just as in linear regression.

# Project Goal

* Use **Logistic regression** to perform **supervised classification** to identify if a subject is a **drug (Cannabis) user** or not Defining the Workflow

# Defining the Workflow

* Gathering the data
* Data processing
* Data visualization and Test of Association between Predictor Variables and Response Variable
* Handling categorical variables and redundancy to resolve high dimensionality and overfitting.
* Training and testing the model
* Evaluating the model

# Data Gathering

* Data for training/testing obtained from:
  + **Data Source**: "Drug consumption (quantified) Data Set" from the UCI Machine Learning Repository: <https://archive.ics.uci.edu/ml/machine-learning-databases/00373/>
  + **Data**: <https://archive.ics.uci.edu/ml/machine-learning-databases/00373/drug_consumption.data>

# Data Processing

* **Libraries: SAS**
* **Features for Machine Learning Analysis**
* **Demographics**: age, gender, education
* **Personality Traits:** neuroticism, extraversion, openness, agreeableness, conscientiousness, impulsiveness, sensation
* **Drug**: Cannabis

The country and ethnicity categories were not chosen because for cannabis users, the country data was heavily skewed to "USA" and "UK", and the ethnicity data was heavily skewed to "White".

# Project Description with Tasks

In this project we use Logistic regression to perform supervised classification to identify if a subject is a drug (Cannabis) user or not. In Logistic Regression, we model the probability of a case being drug user based on different characteristic or demographic features of that case. The predicted posterior probabilities from the Logistic regression model is a nonlinear function of the input variables. To linearize the model, we use logit transformation that measures the log odds of the posterior probability which is a linear function of the input function just as in linear regression.

* Data processing starts with first **importing “drug\_consumption.data" data into a csv file**. 1885 records were read from the data file.

/\*Importing Data into csv file\*/  
data drug.drug\_usage;  
 infile "/folders/myfolders/project-drug\_usage/drug\_consumption.data" dlm=',' dsd;  
 input id Age $ Gender $ Education $ Country $ Ethnicity $  
 Neuroticism Extraversion Openness Agreeableness Conscientiousness  
 Impulsiveness Sensation Alcohol $ Amphet $ Amyl $ Benzos $  
 Caff $ Cannabis $ Choc $ Coke $ Crack $ Ecstasy $ Heroin $   
 Ketamine $ Legalh $ LSD $ Meth $ Mushrooms $ Nicotine $  
 Semer $ VSA $;  
run;

* The next step was to **filter only the columns that are needed** in the model. Since we are going to predict the usage of Cannabis using Logistic Regression, we will filter the  
  required columns from dataset.

data drug.drug\_usage;

set drug.drug\_usage;

keep id Age Gender Education Country Ethnicity Neuroticism Extraversion

Openness Agreeableness Conscientiousness Impulsiveness Sensation Cannabis;

run;

* During **Data Wrangling** the next step was to **clean the data,** rename the values in each column according to the data description provided on the website: [https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29#](https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29)

Values in following columns were renamed:

Age, Gender, Education, Country, Ethnicity, also values in the following columns were reassigned: Neuroticism, Extraversion, Agreeableness, Conscientiousness, Impulsiveness, Sensation,

* Now the **Proc Freq** was run and **Distribution** of Age, Gender, Education, Country, and Ethnicity by **Cannabis** was plotted.

proc freq data=drug.drug\_usage;

tables (Age Gender Education Country Ethnicity)\*Cannabis/

plots=freqplot(scale=percent);

run;

**Observations from the plots**:

* Distribution of Cannabis seems to vary in different age groups. The highest amount of drug usage is found in early age groups of 18-24 and 25-34. It seems to reduce in later ages. Those above between 55 and 64 age groups shows highly reduced level of drug usage and those above 65+ hardly show any drug usage.
* Among the different gender groups, **male** shows much higher drug usage than females.
* When looking at the distribution of drug users and non-users at various levels of Education, it seems like **those went to some College or University but did not attain a degree** tend to show the largest amount of drug consumption. It might be the reason that these people get demotivated and start taking drugs and are not able to complete the degree. After that those who complete University tend to show high drug usage followed by those who attain a Professional diploma or certificate and those who are in Masters Program. This clearly shows drugs are mainly used among those **who are affiliated to higher level academic educational institutes**.
* It is also seen that drug consumption is very higher at higher percentages of **impulsiveness** such as 6.29% and 8.83%.
* The drug consumption is also highest for those who show **sensation** at 1.19% and 3.21%.
* Among the different Countries, USA and UK shows highest amount of drug usage.  
  Among different ethnic groups, those belonging to **White** Ethnicity shows the highest and extremely large amount of drug consumption.
* Also we see several cases of Quasi complete separation where we have 100% of drug assumptions in certain categorical levels of the categorical input variables such as **Country**, and Ethnicity.
* This will give problems when trying to fit a Logistic Regression model as the model will not converge as the maximum likelihood estimation function will lead to infinite probability for that level leading to an infinite parameter estimate for that level. Therefore, to avoid that problem, we will use Smoothed weight of evidence method to convert the categorical column to numerical columns.

The tables and plots are appended here under after this description.

After plotting of distribution, **chi-square test of association between the categorical variables and the target variable Cannabis** is performed.

proc freq data=drug.drug\_usage;  
 tables (Age Gender Education Country Ethnicity)\*Cannabis/  
 chisq expected cellchi2 nocol nopercent relrisk;  
run;

Chi-Square test of Association shows a significant association between the categorical variables:

Age, Gender, Impulsiveness, Sensation, Country, and Ethnicity with the target response variable Cannabis with a significant p-value for the chi-square statistics.

This shows that the difference in the distribution of Cannabis users and non-users among different levels of the categorical variables is significant at a significance level of 0.05.

Thereafter, we split the dataset in to training and validation dataset using **proc surveyselect**.

**Note**: When we do stratified sampling, we basically separate sample or oversample the events in the training and validation data set.

proc surveyselect data=drug.drug\_usage1 samprate=0.6667 seed=1256345

out=drug.drug\_usage\_split outall;

strata Cannabis;

run;

For numerical or ordinal variables, we would like to verify if the relationship with the target response variable is linear/ monotonic or nonlinear/ nonmonotonic. The variables that shows high **Spearman correlation** statistics will show a linear/ monotonic association with the target variable regardless of the Hoeffding statistics.

But those variables which shows low value for Spearman correlation statistic and high value of

Hoeffding statistics shows a nonlinear/ non-monotonic association with the target variable.

For this, we compute the Spearman and Hoeffding correlation statistics which shows association between the rank orders of two ordinal variables.

The Spearman and Hoeffding statistics are rank ordered in the descending order. A plot of Spearman and Hoeffding ranks will then plotted to determine those variables with high Spearman rank (corresponding to low Spearman correlation statistic) and low Hoeffding rank (corresponding to high Hoeffding correlation statistic). These variables are nonmonotonic in nature.

%let intervall= Neuroticism Agreeableness Openness Extraversion Conscientiousness Impulsiveness Sensation;  
ods select none;  
ods output spearmancorr=drug.spearmann  
 hoeffdingcorr=drug.hoeffding;  
proc corr data=drug.drug\_usage\_train spearman hoeffding;  
 var Cannabis;  
 with &intervall;  
run;

We can see that in correlations table, the spearman correlation statistic has the smallest p-value as 0.008. Therefore, we choose the minimum boundary line to separate the variables with higher Spearman ranks or lower spearman correlation statistics at 0.005 in the scatter plot of Spearman and Hoeffding ranks.

From the correlations table, we definitely see that all the variables shows a significant association with target response variable at a significance of 0.05. However, the value of Spearman correlation statistic is not very large. Therefore, it is a relatively weaker association. Therefore, we choose 0.005 as significance level to draw the boundary line for including variables in the model. The variables with high Spearman and Hoeffding ranks (that is with high spearman and correlation statistics) are extremely weak predictors and can be excluded from the model development process.

From the **scatter plot**, it is evident that **Extraversion** is a relatively weaker predictor of target variable  
Cannabis and will therefore, be **excluded from the model**.

We can see from above plots of adjusted log odds of the events (i.e. drug usage) vs different ordinal features such as Agreeableness, Openness, Conscientiousness and Neuroticism after binning these variables into groups or bins is approximately linear with some deviations possibly due to sampling variability. Therefore, instead of using the actual variables, we will use the binned variables in the final model. For that we create a dataset that will contain the binned variables for Agreeableness, Openness, Conscientiousness and Neuroticism. We first write set of rules to a SAS file and then include that file in a dataset to create binned variables.

However, for variables such as Impulsiveness and Sensation, there seems to be a non-monotonic relation.

Even though Logistic regression assumes a linear relation between input variables and the predicted logits (or log odds), it has been seen that Logistic Regression is a fairly robust model towards such non-linearity occurrences. **Therefore, we will continue to fit a logistic regression model to the data.**

**Output Data: Table: DRUG.DRUG\_USAGE** Total Rows: 1885 Total Columns: 14

**Top 20 rows**:

| **Obs** | **id** | **Age** | **Gender** | **Neuroticism** | **Extraversion** | **Openness** | **Agreeableness** | **Conscientiousness** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | 1 | 35-44 | Female | 39 | 36 | 42 | 37 | 42 |
| **2** | 2 | 25-34 | Male | 29 | 52 | 55 | 48 | 41 |
| **3** | 3 | 35-44 | Male | 31 | 45 | 40 | 32 | 34 |
| **4** | 4 | 18-24 | Female | 34 | 34 | 46 | 47 | 46 |
| **5** | 5 | 35-44 | Female | 43 | 28 | 43 | 41 | 50 |
| **6** | 6 | 65+ | Female | 29 | 106 | 35 | 55 | 52 |
| **7** | 7 | 45-54 | Male | 31 | 32 | 43 | 41 | 48 |
| **8** | 8 | 35-44 | Male | 24 | 52 | 40 | 41 | 52 |
| **9** | 9 | 35-44 | Female | 42 | 55 | 39 | 48 | 49 |
| **10** | 10 | 55-64 | Male | 33 | 40 | 36 | 47 | 43 |
| **11** | 11 | 25-34 | Female | 26 | 45 | 38 | 38 | 53 |
| **12** | 12 | 45-54 | Male | 24 | 40 | 47 | 30 | 38 |
| **13** | 13 | 55-64 | Female | 56 | 41 | 49 | 32 | 36 |
| **14** | 14 | 55-64 | Female | 28 | 45 | 46 | 49 | 59 |
| **15** | 15 | 55-64 | Female | 27 | 49 | 49 | 39 | 52 |
| **16** | 16 | 55-64 | Male | 19 | 29 | 35 | 36 | 49 |
| **17** | 17 | 35-44 | Female | 22 | 34 | 34 | 45 | 47 |
| **18** | 18 | 45-54 | Male | 41 | 31 | 44 | 40 | 31 |
| **19** | 19 | 55-64 | Male | 49 | 39 | 45 | 30 | 30 |
| **20** | 20 | 35-44 | Male | 32 | 27 | 29 | 30 | 47 |

| **Obs** | **Impulsiveness** | **Sensation** | **Education** | **Country** | **Ethnicity** | **Cannabis** |
| --- | --- | --- | --- | --- | --- | --- |
| **1** | 8.83% | 7.00% | Professional certificate/ diploma | UK | Mixed-White/Asian | 0 |
| **2** | 6.29% | 1.83% | Doctorate degree | UK | White | 1 |
| **3** | 4.64% | 3.21% | Professional certificate/ diploma | UK | White | 1 |
| **4** | 4.64% | 7.00% | Masters degree | UK | White | 1 |
| **5** | 8.83% | 1.83% | Doctorate degree | UK | White | 1 |
| **6** | 4.64% | 4.62% | Left school at 18 years | Canada | White | 0 |
| **7** | 8.83% | 1.62% | Masters degree | USA | White | 0 |
| **8** | 3.63% | 1.19% | Left school at 16 years | UK | White | 0 |
| **9** | 4.64% | 4.62% | Professional certificate/ diploma | Canada | White | 0 |
| **10** | 4.64% | 8.97% | Masters degree | UK | White | 0 |
| **11** | 3.63% | 1.62% | University degree | UK | White | 1 |
| **12** | 1.46% | 1.14% | Some college or university, no certificate or degree | Other | White | 1 |
| **13** | 7.85% | 1.62% | University degree | UK | White | 1 |
| **14** | 6.29% | 8.97% | Professional certificate/ diploma | Canada | White | 0 |
| **15** | 7.85% | 1.19% | Professional certificate/ diploma | UK | White | 0 |
| **16** | 6.29% | 1.19% | University degree | UK | White | 0 |
| **17** | 8.83% | 3.77% | Some college or university, no certificate or degree | UK | White | 1 |
| **18** | 4.64% | 8.97% | Left school at 16 years | UK | White | 1 |
| **19** | 6.29% | 1.83% | University degree | Australia | White | 1 |
| **20** | 4.64% | 3.77% | Professional certificate/ diploma | UK | White | 0 |

| **Table of Age by Cannabis** | | | |
| --- | --- | --- | --- |
| **Age** | **Cannabis** | | |
| **Frequency Percent Row Pct Col Pct** | **0** | **1** | **Total** |
| **18-24** | 61 3.24 9.49 9.84 | 582 30.88 90.51 46.01 | 643 34.11 |
| **25-34** | 138 7.32 28.69 22.26 | 343 18.20 71.31 27.11 | 481 25.52 |
| **35-44** | 163 8.65 45.79 26.29 | 193 10.24 54.21 15.26 | 356 18.89 |
| **45-54** | 185 9.81 62.93 29.84 | 109 5.78 37.07 8.62 | 294 15.60 |
| **55-64** | 56 2.97 60.22 9.03 | 37 1.96 39.78 2.92 | 93 4.93 |
| **65+** | 17 0.90 94.44 2.74 | 1 0.05 5.56 0.08 | 18 0.95 |
| **Total** | 620 32.89 | 1265 67.11 | 1885 100.00 |



| **Table of Gender by Cannabis** | | | |
| --- | --- | --- | --- |
| **Gender** | **Cannabis** | | |
| **Frequency Percent Row Pct Col Pct** | **0** | **1** | **Total** |
| **Female** | 415 22.02 44.06 66.94 | 527 27.96 55.94 41.66 | 942 49.97 |
| **Male** | 205 10.88 21.74 33.06 | 738 39.15 78.26 58.34 | 943 50.03 |
| **Total** | 620 32.89 | 1265 67.11 | 1885 100.00 |



| **Table of Education by Cannabis** | | | |
| --- | --- | --- | --- |
| **Education** | **Cannabis** | | |
| **Frequency Percent Row Pct Col Pct** | **0** | **1** | **Total** |
| **Doctorate degree** | 41 2.18 46.07 6.61 | 48 2.55 53.93 3.79 | 89 4.72 |
| **Left school at 16 years** | 44 2.33 44.44 7.10 | 55 2.92 55.56 4.35 | 99 5.25 |
| **Left school at 17 years** | 7 0.37 23.33 1.13 | 23 1.22 76.67 1.82 | 30 1.59 |
| **Left school at 18 years** | 23 1.22 23.00 3.71 | 77 4.08 77.00 6.09 | 100 5.31 |
| **Left school before 16 years** | 8 0.42 28.57 1.29 | 20 1.06 71.43 1.58 | 28 1.49 |
| **Masters degree** | 141 7.48 49.82 22.74 | 142 7.53 50.18 11.23 | 283 15.01 |
| **Professional certificate/ diploma** | 115 6.10 42.59 18.55 | 155 8.22 57.41 12.25 | 270 14.32 |
| **Some college or university, no certificate or degree** | 53 2.81 10.47 8.55 | 453 24.03 89.53 35.81 | 506 26.84 |
| **University degree** | 188 9.97 39.17 30.32 | 292 15.49 60.83 23.08 | 480 25.46 |
| **Total** | 620 32.89 | 1265 67.11 | 1885 100.00 |



| **Table of Country by Cannabis** | | | |
| --- | --- | --- | --- |
| **Country** | **Cannabis** | | |
| **Frequency Percent Row Pct Col Pct** | **0** | **1** | **Total** |
| **Australia** | 5 0.27 9.26 0.81 | 49 2.60 90.74 3.87 | 54 2.86 |
| **Canada** | 20 1.06 22.99 3.23 | 67 3.55 77.01 5.30 | 87 4.62 |
| **New Zealand** | 0 0.00 0.00 0.00 | 5 0.27 100.00 0.40 | 5 0.27 |
| **Other** | 18 0.95 15.25 2.90 | 100 5.31 84.75 7.91 | 118 6.26 |
| **Republic of Ireland** | 4 0.21 20.00 0.65 | 16 0.85 80.00 1.26 | 20 1.06 |
| **UK** | 543 28.81 52.01 87.58 | 501 26.58 47.99 39.60 | 1044 55.38 |
| **USA** | 30 1.59 5.39 4.84 | 527 27.96 94.61 41.66 | 557 29.55 |
| **Total** | 620 32.89 | 1265 67.11 | 1885 100.00 |



| **Table of Ethnicity by Cannabis** | | | |
| --- | --- | --- | --- |
| **Ethnicity** | **Cannabis** | | |
| **Frequency Percent Row Pct Col Pct** | **0** | **1** | **Total** |
| **Asian** | 19 1.01 73.08 3.06 | 7 0.37 26.92 0.55 | 26 1.38 |
| **Black** | 23 1.22 69.70 3.71 | 10 0.53 30.30 0.79 | 33 1.75 |
| **Mixed-Black/Asian** | 0 0.00 0.00 0.00 | 3 0.16 100.00 0.24 | 3 0.16 |
| **Mixed-White/Asian** | 4 0.21 20.00 0.65 | 16 0.85 80.00 1.26 | 20 1.06 |
| **Mixed-White/Black** | 6 0.32 30.00 0.97 | 14 0.74 70.00 1.11 | 20 1.06 |
| **Other** | 13 0.69 20.63 2.10 | 50 2.65 79.37 3.95 | 63 3.34 |
| **White** | 555 29.44 32.27 89.52 | 1165 61.80 67.73 92.09 | 1720 91.25 |
| **Total** | 620 32.89 | 1265 67.11 | 1885 100.00 |



| **Variable** | **N** | **Mean** | **Std Dev** | **Minimum** | **Maximum** |
| --- | --- | --- | --- | --- | --- |
| id Neuroticism Extraversion Openness Agreeableness Conscientiousness Impulsiveness Sensation Cannabis | 1885 1885 1885 1885 1885 1885 1885 1885 1885 | 945.2949602 35.9214854 41.8420689 45.7623342 42.8663130 41.4371353 13.5175385 10.1886790 0.6710875 | 545.1676411 9.1358693 18.3630559 6.5796414 6.4381061 6.9666250 4.0780971 2.6723837 0.4699428 | 1.0000000 12.0000000 1.1140600 24.0000000 12.0000000 17.0000000 0.3700000 3.7700000 0 | 1888.00 60.0000000 106.0000000 60.0000000 60.0000000 59.0000000 18.8300000 13.2100000 1.0000000 |

| **Neuroticism** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **12** | 1 | 0.05 | 1 | 0.05 |
| **13** | 1 | 0.05 | 2 | 0.11 |
| **14** | 7 | 0.37 | 9 | 0.48 |
| **15** | 4 | 0.21 | 13 | 0.69 |
| **16** | 3 | 0.16 | 16 | 0.85 |
| **17** | 4 | 0.21 | 20 | 1.06 |
| **18** | 10 | 0.53 | 30 | 1.59 |
| **19** | 16 | 0.85 | 46 | 2.44 |
| **20** | 24 | 1.27 | 70 | 3.71 |
| **21** | 31 | 1.64 | 101 | 5.36 |
| **22** | 26 | 1.38 | 127 | 6.74 |
| **23** | 29 | 1.54 | 156 | 8.28 |
| **24** | 35 | 1.86 | 191 | 10.13 |
| **25** | 56 | 2.97 | 247 | 13.10 |
| **26** | 57 | 3.02 | 304 | 16.13 |
| **27** | 65 | 3.45 | 369 | 19.58 |
| **28** | 70 | 3.71 | 439 | 23.29 |
| **29** | 60 | 3.18 | 499 | 26.47 |
| **30** | 61 | 3.24 | 560 | 29.71 |
| **31** | 87 | 4.62 | 647 | 34.32 |
| **32** | 78 | 4.14 | 725 | 38.46 |
| **33** | 68 | 3.61 | 793 | 42.07 |
| **34** | 76 | 4.03 | 869 | 46.10 |
| **35** | 69 | 3.66 | 938 | 49.76 |
| **36** | 73 | 3.87 | 1011 | 53.63 |
| **37** | 67 | 3.55 | 1078 | 57.19 |
| **38** | 63 | 3.34 | 1141 | 60.53 |
| **39** | 66 | 3.50 | 1207 | 64.03 |
| **40** | 80 | 4.24 | 1287 | 68.28 |
| **41** | 61 | 3.24 | 1348 | 71.51 |
| **42** | 77 | 4.08 | 1425 | 75.60 |
| **43** | 49 | 2.60 | 1474 | 78.20 |
| **44** | 51 | 2.71 | 1525 | 80.90 |
| **45** | 37 | 1.96 | 1562 | 82.86 |
| **46** | 67 | 3.55 | 1629 | 86.42 |
| **47** | 27 | 1.43 | 1656 | 87.85 |
| **48** | 49 | 2.60 | 1705 | 90.45 |
| **49** | 40 | 2.12 | 1745 | 92.57 |
| **50** | 24 | 1.27 | 1769 | 93.85 |
| **51** | 27 | 1.43 | 1796 | 95.28 |
| **52** | 17 | 0.90 | 1813 | 96.18 |
| **53** | 20 | 1.06 | 1833 | 97.24 |
| **54** | 15 | 0.80 | 1848 | 98.04 |
| **55** | 11 | 0.58 | 1859 | 98.62 |
| **56** | 10 | 0.53 | 1869 | 99.15 |
| **57** | 6 | 0.32 | 1875 | 99.47 |
| **58** | 3 | 0.16 | 1878 | 99.63 |
| **59** | 5 | 0.27 | 1883 | 99.89 |
| **60** | 2 | 0.11 | 1885 | 100.00 |

| **Table of Age by Cannabis** | | | |
| --- | --- | --- | --- |
| **Age** | **Cannabis** | | |
| **Frequency Expected Cell Chi-Square Row Pct** | **0** | **1** | **Total** |
| **18-24** | 61 211.49 107.08 9.49 | 582 431.51 52.484 90.51 | 643 |
| **25-34** | 138 158.21 2.5809 28.69 | 343 322.79 1.265 71.31 | 481 |
| **35-44** | 163 117.09 17.998 45.79 | 193 238.91 8.8213 54.21 | 356 |
| **45-54** | 185 96.7 80.629 62.93 | 109 197.3 39.518 37.07 | 294 |
| **55-64** | 56 30.589 21.11 60.22 | 37 62.411 10.346 39.78 | 93 |
| **65+** | 17 5.9204 20.734 94.44 | 1 12.08 10.162 5.56 | 18 |
| **Total** | 620 | 1265 | 1885 |

|  |
| --- |
| ***Statistics for Table of Age by Cannabis*** |

| **Statistic** | **DF** | **Value** | **Prob** |
| --- | --- | --- | --- |
| **Chi-Square** | 5 | 372.7343 | <.0001 |
| **Likelihood Ratio Chi-Square** | 5 | 396.5673 | <.0001 |
| **Mantel-Haenszel Chi-Square** | 1 | 358.7558 | <.0001 |
| **Phi Coefficient** |  | 0.4447 |  |
| **Contingency Coefficient** |  | 0.4063 |  |
| **Cramer's V** |  | 0.4447 |  |

|  |
| --- |
| ***Sample Size = 1885*** |

| **Table of Gender by Cannabis** | | | |
| --- | --- | --- | --- |
| **Gender** | **Cannabis** | | |
| **Frequency Expected Cell Chi-Square Row Pct** | **0** | **1** | **Total** |
| **Female** | 415 309.84 35.695 44.06 | 527 632.16 17.495 55.94 | 942 |
| **Male** | 205 310.16 35.657 21.74 | 738 632.84 17.476 78.26 | 943 |
| **Total** | 620 | 1265 | 1885 |

|  |
| --- |
| ***Statistics for Table of Gender by Cannabis*** |

| **Statistic** | **DF** | **Value** | **Prob** |
| --- | --- | --- | --- |
| **Chi-Square** | 1 | 106.3230 | <.0001 |
| **Likelihood Ratio Chi-Square** | 1 | 107.9146 | <.0001 |
| **Continuity Adj. Chi-Square** | 1 | 105.3144 | <.0001 |
| **Mantel-Haenszel Chi-Square** | 1 | 106.2666 | <.0001 |
| **Phi Coefficient** |  | 0.2375 |  |
| **Contingency Coefficient** |  | 0.2311 |  |
| **Cramer's V** |  | 0.2375 |  |

| **Fisher's Exact Test** | |
| --- | --- |
| **Cell (1,1) Frequency (F)** | 415 |
| **Left-sided Pr <= F** | 1.0000 |
| **Right-sided Pr >= F** | <.0001 |
|  |  |
| **Table Probability (P)** | <.0001 |
| **Two-sided Pr <= P** | <.0001 |

| **Odds Ratio and Relative Risks** | | | |
| --- | --- | --- | --- |
| **Statistic** | **Value** | **95% Confidence Limits** | |
| **Odds Ratio** | 2.8349 | 2.3182 | 3.4668 |
| **Relative Risk (Column 1)** | 2.0265 | 1.7603 | 2.3331 |
| **Relative Risk (Column 2)** | 0.7149 | 0.6693 | 0.7635 |

|  |
| --- |
| ***Sample Size = 1885*** |

| **Table of Education by Cannabis** | | | |
| --- | --- | --- | --- |
| **Education** | **Cannabis** | | |
| **Frequency Expected Cell Chi-Square Row Pct** | **0** | **1** | **Total** |
| **Doctorate degree** | 41 29.273 4.6977 46.07 | 48 59.727 2.3024 53.93 | 89 |
| **Left school at 16 years** | 44 32.562 4.0175 44.44 | 55 66.438 1.9691 55.56 | 99 |
| **Left school at 17 years** | 7 9.8674 0.8332 23.33 | 23 20.133 0.4084 76.67 | 30 |
| **Left school at 18 years** | 23 32.891 2.9746 23.00 | 77 67.109 1.4579 77.00 | 100 |
| **Left school before 16 years** | 8 9.2095 0.1589 28.57 | 20 18.79 0.0779 71.43 | 28 |
| **Masters degree** | 141 93.082 24.668 49.82 | 142 189.92 12.09 50.18 | 283 |
| **Professional certificate/ diploma** | 115 88.806 7.7259 42.59 | 155 181.19 3.7866 57.41 | 270 |
| **Some college or university, no certificate or degree** | 53 166.43 77.308 10.47 | 453 339.57 37.89 89.53 | 506 |
| **University degree** | 188 157.88 5.7471 39.17 | 292 322.12 2.8167 60.83 | 480 |
| **Total** | 620 | 1265 | 1885 |

|  |
| --- |
| ***Statistics for Table of Education by Cannabis*** |

| **Statistic** | **DF** | **Value** | **Prob** |
| --- | --- | --- | --- |
| **Chi-Square** | 8 | 190.9291 | <.0001 |
| **Likelihood Ratio Chi-Square** | 8 | 212.3540 | <.0001 |
| **Mantel-Haenszel Chi-Square** | 1 | 15.7886 | <.0001 |
| **Phi Coefficient** |  | 0.3183 |  |
| **Contingency Coefficient** |  | 0.3033 |  |
| **Cramer's V** |  | 0.3183 |  |

|  |
| --- |
| ***Sample Size = 1885*** |

| **Table of Country by Cannabis** | | | |
| --- | --- | --- | --- |
| **Country** | **Cannabis** | | |
| **Frequency Expected Cell Chi-Square Row Pct** | **0** | **1** | **Total** |
| **Australia** | 5 17.761 9.1688 9.26 | 49 36.239 4.4938 90.74 | 54 |
| **Canada** | 20 28.615 2.5939 22.99 | 67 58.385 1.2713 77.01 | 87 |
| **New Zealand** | 0 1.6446 1.6446 0.00 | 5 3.3554 0.806 100.00 | 5 |
| **Other** | 18 38.812 11.16 15.25 | 100 79.188 5.4696 84.75 | 118 |
| **Republic of Ireland** | 4 6.5782 1.0105 20.00 | 16 13.422 0.4953 80.00 | 20 |
| **UK** | 543 343.38 116.04 52.01 | 501 700.62 56.873 47.99 | 1044 |
| **USA** | 30 183.2 128.12 5.39 | 527 373.8 62.792 94.61 | 557 |
| **Total** | 620 | 1265 | 1885 |

|  |
| --- |
| ***Statistics for Table of Country by Cannabis*** |

| **Statistic** | **DF** | **Value** | **Prob** |
| --- | --- | --- | --- |
| **Chi-Square** | 6 | 401.9358 | <.0001 |
| **Likelihood Ratio Chi-Square** | 6 | 460.7645 | <.0001 |
| **Mantel-Haenszel Chi-Square** | 1 | 0.0403 | 0.8408 |
| **Phi Coefficient** |  | 0.4618 |  |
| **Contingency Coefficient** |  | 0.4192 |  |
| **Cramer's V** |  | 0.4618 |  |

|  |
| --- |
| ***Sample Size = 1885*** |

| **Table of Ethnicity by Cannabis** | | | |
| --- | --- | --- | --- |
| **Ethnicity** | **Cannabis** | | |
| **Frequency Expected Cell Chi-Square Row Pct** | **0** | **1** | **Total** |
| **Asian** | 19 8.5517 12.765 73.08 | 7 17.448 6.2566 26.92 | 26 |
| **Black** | 23 10.854 13.591 69.70 | 10 22.146 6.6614 30.30 | 33 |
| **Mixed-Black/Asian** | 0 0.9867 0.9867 0.00 | 3 2.0133 0.4836 100.00 | 3 |
| **Mixed-White/Asian** | 4 6.5782 1.0105 20.00 | 16 13.422 0.4953 80.00 | 20 |
| **Mixed-White/Black** | 6 6.5782 0.0508 30.00 | 14 13.422 0.0249 70.00 | 20 |
| **Other** | 13 20.721 2.8773 20.63 | 50 42.279 1.4102 79.37 | 63 |
| **White** | 555 565.73 0.2035 32.27 | 1165 1154.3 0.0997 67.73 | 1720 |
| **Total** | 620 | 1265 | 1885 |

|  |
| --- |
| ***Statistics for Table of Ethnicity by Cannabis*** |

| **Statistic** | **DF** | **Value** | **Prob** |
| --- | --- | --- | --- |
| **Chi-Square** | 6 | 46.9174 | <.0001 |
| **Likelihood Ratio Chi-Square** | 6 | 45.2613 | <.0001 |
| **Mantel-Haenszel Chi-Square** | 1 | 23.9357 | <.0001 |
| **Phi Coefficient** |  | 0.1578 |  |
| **Contingency Coefficient** |  | 0.1558 |  |
| **Cramer's V** |  | 0.1578 |  |

|  |
| --- |
| ***Sample Size = 1885*** |

|  |  |
| --- | --- |
| **Selection Method** | Simple Random Sampling |
| **Strata Variable** | Cannabis |

|  |  |
| --- | --- |
| **Input Data Set** | DRUG\_USAGE1 |
| **Random Number Seed** | 1256345 |
| **Stratum Sampling Rate** | 0.6667 |
| **Number of Strata** | 2 |
| **Total Sample Size** | 1258 |
| **Output Data Set** | DRUG\_USAGE\_SPLIT |



| **Obs** | **Variable** | **scorr** | **P\_scorr** | **hcorr** | **P\_hcorr** | **ranksp** | **rankho** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | Agreeableness | 0.17596 | <.0001 | 0.00414 | <.0001 | 3 | 4 |
| **2** | Conscientiousness | 0.27510 | <.0001 | 0.01206 | <.0001 | 2 | 2 |
| **3** | Extraversion | 0.04322 | 0.1255 | 0.00002 | 0.3853 | 7 | 7 |
| **4** | Impulsiveness | 0.17226 | <.0001 | 0.00587 | <.0001 | 4 | 3 |
| **5** | Neuroticism | 0.16259 | <.0001 | 0.00405 | <.0001 | 5 | 5 |
| **6** | Openness | 0.33202 | <.0001 | 0.01738 | <.0001 | 1 | 1 |
| **7** | Sensation | 0.15924 | <.0001 | 0.00367 | <.0001 | 6 | 6 |

| **Analysis Variable : Cannabis** | | |
| --- | --- | --- |
| **Rank for Variable Agreeableness** | **N Obs** | **Sum** |
| 0 | 28 | 24.0000000 |
| 1 | 14 | 11.0000000 |
| 2 | 41 | 34.0000000 |
| 3 | 26 | 23.0000000 |
| 4 | 28 | 19.0000000 |
| 6 | 32 | 29.0000000 |
| 7 | 47 | 38.0000000 |
| 9 | 62 | 43.0000000 |
| 12 | 51 | 37.0000000 |
| 14 | 68 | 49.0000000 |
| 17 | 63 | 43.0000000 |
| 19 | 70 | 48.0000000 |
| 22 | 69 | 45.0000000 |
| 25 | 64 | 43.0000000 |
| 27 | 68 | 42.0000000 |
| 30 | 81 | 50.0000000 |
| 33 | 63 | 51.0000000 |
| 36 | 72 | 37.0000000 |
| 38 | 65 | 41.0000000 |
| 41 | 60 | 34.0000000 |
| 43 | 46 | 29.0000000 |
| 45 | 32 | 19.0000000 |
| 46 | 23 | 8.0000000 |
| 47 | 25 | 17.0000000 |
| 48 | 38 | 21.0000000 |
| 49 | 22 | 9.0000000 |

|  |
| --- |
| 0.670906 |



| **Analysis Variable : Cannabis** | | |
| --- | --- | --- |
| **Rank for Variable Openness** | **N Obs** | **Sum** |
| 0 | 27 | 11.0000000 |
| 1 | 22 | 7.0000000 |
| 2 | 30 | 12.0000000 |
| 3 | 22 | 10.0000000 |
| 4 | 34 | 13.0000000 |
| 6 | 44 | 18.0000000 |
| 7 | 35 | 18.0000000 |
| 9 | 55 | 30.0000000 |
| 11 | 54 | 31.0000000 |
| 13 | 55 | 30.0000000 |
| 16 | 57 | 34.0000000 |
| 18 | 68 | 45.0000000 |
| 21 | 75 | 42.0000000 |
| 24 | 98 | 72.0000000 |
| 28 | 64 | 42.0000000 |
| 30 | 76 | 55.0000000 |
| 33 | 67 | 51.0000000 |
| 36 | 57 | 48.0000000 |
| 38 | 58 | 48.0000000 |
| 40 | 57 | 50.0000000 |
| 42 | 50 | 45.0000000 |
| 44 | 37 | 29.0000000 |
| 46 | 40 | 33.0000000 |
| 47 | 29 | 26.0000000 |
| 48 | 21 | 20.0000000 |
| 49 | 26 | 24.0000000 |



| **Analysis Variable : Cannabis** | | |
| --- | --- | --- |
| **Rank for Variable Conscientiousness** | **N Obs** | **Sum** |
| 0 | 27 | 24.0000000 |
| 1 | 15 | 14.0000000 |
| 2 | 43 | 35.0000000 |
| 3 | 30 | 24.0000000 |
| 5 | 23 | 17.0000000 |
| 6 | 32 | 30.0000000 |
| 7 | 38 | 35.0000000 |
| 8 | 36 | 25.0000000 |
| 10 | 45 | 36.0000000 |
| 12 | 65 | 54.0000000 |
| 15 | 55 | 43.0000000 |
| 17 | 60 | 52.0000000 |
| 19 | 63 | 40.0000000 |
| 22 | 63 | 40.0000000 |
| 25 | 69 | 48.0000000 |
| 27 | 55 | 35.0000000 |
| 30 | 76 | 56.0000000 |
| 32 | 65 | 30.0000000 |
| 35 | 78 | 40.0000000 |
| 38 | 64 | 36.0000000 |
| 41 | 63 | 34.0000000 |
| 43 | 56 | 31.0000000 |
| 45 | 33 | 23.0000000 |
| 46 | 34 | 13.0000000 |
| 47 | 23 | 6.0000000 |
| 48 | 17 | 4.0000000 |
| 49 | 30 | 19.0000000 |



| **Analysis Variable : Cannabis** | | |
| --- | --- | --- |
| **Rank for Variable Neuroticism** | **N Obs** | **Sum** |
| 0 | 20 | 12.0000000 |
| 1 | 29 | 11.0000000 |
| 2 | 18 | 13.0000000 |
| 3 | 38 | 27.0000000 |
| 4 | 25 | 13.0000000 |
| 5 | 34 | 21.0000000 |
| 7 | 34 | 21.0000000 |
| 8 | 42 | 26.0000000 |
| 10 | 51 | 36.0000000 |
| 12 | 42 | 26.0000000 |
| 13 | 35 | 24.0000000 |
| 15 | 57 | 34.0000000 |
| 18 | 56 | 34.0000000 |
| 20 | 46 | 24.0000000 |
| 21 | 52 | 30.0000000 |
| 23 | 45 | 24.0000000 |
| 25 | 52 | 29.0000000 |
| 27 | 48 | 32.0000000 |
| 29 | 49 | 33.0000000 |
| 31 | 45 | 27.0000000 |
| 33 | 47 | 32.0000000 |
| 35 | 36 | 32.0000000 |
| 36 | 47 | 31.0000000 |
| 38 | 35 | 27.0000000 |
| 39 | 30 | 21.0000000 |
| 40 | 25 | 22.0000000 |
| 42 | 50 | 36.0000000 |
| 43 | 18 | 16.0000000 |
| 44 | 33 | 30.0000000 |
| 45 | 28 | 26.0000000 |
| 46 | 18 | 13.0000000 |
| 47 | 22 | 21.0000000 |
| 48 | 24 | 19.0000000 |
| 49 | 27 | 21.0000000 |



| **Analysis Variable : Cannabis** | | |
| --- | --- | --- |
| **Impulsiveness** | **N Obs** | **Sum** |
| 0.37% | 4 | 4.0000000 |
| 1.06% | 14 | 5.0000000 |
| 5.52% | 78 | 67.0000000 |
| 7.85% | 101 | 89.0000000 |
| 0.34% | 128 | 110.0000000 |
| 1.46% | 142 | 98.0000000 |
| 3.63% | 161 | 112.0000000 |
| 4.64% | 184 | 81.0000000 |
| 6.29% | 199 | 113.0000000 |
| 8.83% | 247 | 165.0000000 |

|  |
| --- |
| 0.670906 |



| **Analysis Variable : Cannabis** | | |
| --- | --- | --- |
| **Sensation** | **N Obs** | **Sum** |
| 3.77% | 48 | 18.0000000 |
| 4.62% | 51 | 14.0000000 |
| 5.46% | 67 | 63.0000000 |
| 7.00% | 91 | 39.0000000 |
| 8.97% | 114 | 48.0000000 |
| 1.14% | 138 | 123.0000000 |
| 1.19% | 287 | 205.0000000 |
| 1.62% | 143 | 108.0000000 |
| 1.83% | 151 | 96.0000000 |
| 3.21% | 168 | 130.0000000 |

|  |
| --- |
| 0.670906 |



Now we will handle the Categorical variables and solve the problem with Quasi complete separation with the categorical variables Ethnicity and Country by the method of Smoothed weight of evidence.

In this method, we replace the levels of Categorical variables with adjusted log odds with a smoothing factor c=24)

The formula for smoothed weight of evidence = #ofevents+ c\*rho1/#ofnon-events + c\*rho0

GREENACCRES METHOD TO REDUCE THE NUMBER OF LEVELS OF CATEGORICAL VARIABLE:

Education has way too many levels. This can lead to problems with high dimensionality as a large number of dummy variables will need to be created for Education. Therefore, we will use Proc cluster to cluster the levels based on minimum reduction in chi-square value

To find which cluster to stop, we will multiply the proportion of chi-square value remaining after clustering with the overall chi-square value of association. This will give us the chi-square value of association for different clusters. This can then be used to find the log of p value for each of those chi-square values based on chi-square distribution curve. The cluster that has lowest value of log of p value is selected as the cluster to stop.

From the scatter plot it is clear that the 4th cluster shows the lowest P-value and therefore, highest chi-square association with the response variable. Now we can also find out which level belongs to which cluster using proc tree.

Now we will write set of rules to a file to create clusters in our original dataset: drug.drug\_usage\_n

filename rclus "/folders/myfolders/project-drug\_usage/clus.sas";

data \_null\_;

file rclus;

set work.clus end=last;

if \_n\_=1 then put "select (Education);";

put " when ('" Education +(-1) "') Education\_clus = '" Cluster +(-1) "';";

if last then do;

put " otherwise Education\_clus = 'U';"/ "end;";

end;

run;

data drug.drug\_usage\_n;

set drug.drug\_usage\_n;

%include rclus/ source2;

run;

ods rtf close;

| **Age** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **18-24** | 425 | 33.78 | 425 | 33.78 |
| **25-34** | 326 | 25.91 | 751 | 59.70 |
| **35-44** | 237 | 18.84 | 988 | 78.54 |
| **45-54** | 192 | 15.26 | 1180 | 93.80 |
| **55-64** | 66 | 5.25 | 1246 | 99.05 |
| **65+** | 12 | 0.95 | 1258 | 100.00 |

| **Gender** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **Female** | 635 | 50.48 | 635 | 50.48 |
| **Male** | 623 | 49.52 | 1258 | 100.00 |

| **Education** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **Doctorate degree** | 56 | 4.45 | 56 | 4.45 |
| **Left school at 16 years** | 67 | 5.33 | 123 | 9.78 |
| **Left school at 17 years** | 22 | 1.75 | 145 | 11.53 |
| **Left school at 18 years** | 68 | 5.41 | 213 | 16.93 |
| **Left school before 16 years** | 16 | 1.27 | 229 | 18.20 |
| **Masters degree** | 201 | 15.98 | 430 | 34.18 |
| **Professional certificate/ diploma** | 179 | 14.23 | 609 | 48.41 |
| **Some college or university, no certificate or degree** | 326 | 25.91 | 935 | 74.32 |
| **University degree** | 323 | 25.68 | 1258 | 100.00 |

| **Country** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **Australia** | 35 | 2.78 | 35 | 2.78 |
| **Canada** | 54 | 4.29 | 89 | 7.07 |
| **New Zealand** | 3 | 0.24 | 92 | 7.31 |
| **Other** | 80 | 6.36 | 172 | 13.67 |
| **Republic of Ireland** | 15 | 1.19 | 187 | 14.86 |
| **UK** | 703 | 55.88 | 890 | 70.75 |
| **USA** | 368 | 29.25 | 1258 | 100.00 |

| **Ethnicity** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **Asian** | 13 | 1.03 | 13 | 1.03 |
| **Black** | 15 | 1.19 | 28 | 2.23 |
| **Mixed-Black/Asian** | 3 | 0.24 | 31 | 2.46 |
| **Mixed-White/Asian** | 13 | 1.03 | 44 | 3.50 |
| **Mixed-White/Black** | 14 | 1.11 | 58 | 4.61 |
| **Other** | 47 | 3.74 | 105 | 8.35 |
| **White** | 1153 | 91.65 | 1258 | 100.00 |

| **Eigenvalues of the Covariance Matrix** | | | | |
| --- | --- | --- | --- | --- |
|  | **Eigenvalue** | **Difference** | **Proportion** | **Cumulative** |
| **1** | 0.01225609 |  | 1.0000 | 1.0000 |

|  |  |
| --- | --- |
| **Root-Mean-Square Total-Sample Standard Deviation** | 0.110707 |

|  |  |
| --- | --- |
| **Root-Mean-Square Distance Between Observations** | 0.156564 |

| **Cluster History** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Number of Clusters** | **Clusters Joined** | | **Freq** | **Semipartial R-Square** | **R-Square** | **Tie** |
| **9** | Left school at 16 years | Professional certificate/ diploma | 246 | 0.0000 | 1.00 |  |
| **8** | Left school before 16 years | University degree | 339 | 0.0001 | 1.00 |  |
| **7** | Doctorate degree | Masters degree | 257 | 0.0001 | 1.00 |  |
| **6** | Left school at 17 years | Left school at 18 years | 90 | 0.0025 | .997 |  |
| **5** | CL9 | CL8 | 585 | 0.0048 | .993 |  |
| **4** |  | CL6 | 1348 | 0.0250 | .968 |  |
| **3** | CL7 | CL5 | 842 | 0.0649 | .903 |  |
| **2** | CL4 | CL3 | 2190 | 0.2020 | .701 |  |
| **1** | CL2 | Some college or university, no certificate or degree | 2516 | 0.7007 | .000 |  |



| **CLUSNAME** | **Education** | **CLUSTER** |
| --- | --- | --- |
| **CL4** | Left school at 17 years | 3 |
| Left school at 18 years | 3 |

| **CLUSNAME** | **Education** | **CLUSTER** |
| --- | --- | --- |
| **CL5** | Left school at 16 years | 1 |
| Professional certificate/ diploma | 1 |
| Left school before 16 years | 1 |
| University degree | 1 |

| **CLUSNAME** | **Education** | **CLUSTER** |
| --- | --- | --- |
| **CL7** | Doctorate degree | 2 |
| Masters degree | 2 |

| **CLUSNAME** | **Education** | **CLUSTER** |
| --- | --- | --- |
| **Some college or university, no certificate or degree** | Some college or university, no certificate or degree | 4 |

Next we will determine multi collinearity among different numeric variables and try to find out if any input variables that shows high level of multicollinearity can be eliminated from the model selection process.

For this we use **Proc Varclus** to cluster the variable. This algorithm uses iterative algorithm- **Principal Component Analysis** for clustering different input variables. In each cluster, the 1-Rsquare ratio is calculated. This is the ratio of 1-Rsquare within cluster/ 1- Rsquare with next cluster. This ratio must be small for an input variable within a cluster as it shows high level of correlation with the variables within cluster and low level of correlation with variables in next cluster.

The variables with the smallest 1- Rsquare ratio is selected as the chosen variable in that cluster.

In Principal component analysis, eigenvectors are formed which are linear combinations of input variables. As many eigenvectors are formed as there are variables in the cluster.

We split the input variables into clusters based on the second eigenvalue. If the threshold for second eigenvalue is below the clusters 2nd eigenvalue, then the cluster will split. If we choose a higher threshold, we will maintain lesser variance in our input variables. However if we choose a smaller threshold, we maintain greater level of variance in the model.

proc varclus data=drug.drug\_usage\_n maxeigen=0.7 hi;  
var &interval;  
run;

From proc varclus results, we can see that the last cluster that is formed is the 6 cluster solution.  
The first cluster contains 2 variables, B\_Neuro and B\_Cons. B\_Neuro shows slightly lower 1-Rsquare Ratio. Similarly B\_Open and Country\_Swoe shows slight difference in 1-Rsquare ratio with Openness being slightly lower.  
  
Therefore, difference is not very large. And based on understanding of the subject, Neuroticism and Conscientiousness can both be equally important as predictors of being Cannabis user.

Similarly both Country and Openness can be potential predictors. Therefore, we will keep all the variables in the model development and validation process.

Now we will use **Proc Logistics** to determine which interactions in the model will be influential enough or significant enough to keep. We will use Forward Selection method that includes all the main effects in the final selected model and selects only those interactions in the final model which are significant. We choose the significance level based on the value of BIC (Bayesian Information criteria). The significance level that lowers BIC is calculated by determining the 1 minus p-value for chi-square statistic of ln of n where n is the number of cases.

proc logistic data=drug.drug\_usage\_n namelen=50;

class &categorical/ param=ref ref=first;

model Cannabis(event='1')= &interval &categorical B\_neuro\*B\_agree/selection=backward clodds=pl slstay=&sl

hier=single fast;

run;

The variables Impulsiveness and Sensation were removed from the final model. Again we can see from Type 3 Analysis of Effects table that B\_agree is not significant but still it remains in final model to maintain the hierarchy of the model as the interaction effect B\_neuro\*B\_agree is in the final model. The Global Null Hypothesis table shows that at least one of parameters are statistically significant and different from zero.

The c statistic of 0.883 shows that model has good predictive ability. We also make an oddratio plot for the interaction effect B\_neuro at different values of B\_agree. The Oddratio plot shows that the oddratios for B\_neuro are statistically significant.

proc logistic data=drug.drug\_usage\_n namelen=50;

class &categorical/ param=ref ref=first;

model Cannabis(event='1')= B\_neuro B\_cons B\_open B\_agree Ethnic\_swoe Country\_swoe &categorical B\_neuro\*B\_agree;

units B\_neuro=10/ default=5;

oddsratio B\_neuro/ at (B\_agree=5,20,50) cl=pl;

run;

Therefore based on the model selected, we include these variables for final validation of the model.

Now we use the all possible **subset regression technique** where all the possible models with different number of variables are selected and rank ordered based on the value of their score-chi-square statistic.

We use the best option to select only 1 model from every single model size comprising of different number of variables based on the highest score chi square value. Since score chi square value simply increases as the number of variables increases, we instead use fitstat option by scoring the same dataset that use to build the model. We use different fit statistics on the insample evaluation performed to determine the best model to use for further validation. For this we use macro function to automate this selection process.

We see that model with 11 variables has the highest AUC which is equivalent to the c statistic that shows the predictive ability of the model. Higher the c statistics, better is our model fit. However there are also other fit statistics such as BIC and misclassification (number of false positives and number of false negatives divided by total sample size using a central cutoff of 0.5 for the probability). We will sort the dataset by BIC to see which model gives lowest BIC as low BIC shows better fit of the model.