A picture containing logo

Description automatically generatedLogo, company name

Description automatically generated

Addiction & Mental Health

Machine Learning

Logo, company name

Description automatically generated

Table of contents

[Core message 3](#_Toc64737519)

[Motivation & questions 3](#_Toc64737520)

[Data collection 3](#_Toc64737521)

[Machine learning models – Initial predictions 4](#_Toc64737522)

[Logistic Regression 4](#_Toc64737523)

[Random Forest 4](#_Toc64737524)

[Both models 5](#_Toc64737525)

[Category analysis – Initial predications 5](#_Toc64737526)

[Smoking 5](#_Toc64737527)

[Alcohol 6](#_Toc64737528)

[Drugs 7](#_Toc64737529)

[Mental health 8](#_Toc64737530)

[Categories 9](#_Toc64737531)

[Summery table prediction outcomes – Initial predications 10](#_Toc64737532)

[Category analysis – Final predication 11](#_Toc64737533)

[Smoking 11](#_Toc64737534)

[Alcohol 11](#_Toc64737535)

[Drugs 11](#_Toc64737536)

[Mental health 11](#_Toc64737537)

[Summery table prediction outcomes – Final predications 12](#_Toc64737538)

[Conclusion 13](#_Toc64737539)

[Limitations 13](#_Toc64737540)

[If I had more time 13](#_Toc64737541)

[Appendix 1 – Customer survey (drop down) 14](#_Toc64737542)

[Appendix 2 – Predictions 16](#_Toc64737543)

[Appendix 3 – Raw data summery 18](#_Toc64737544)

# Core message

Addiction and mental health are still taboo subjects in Australia. I hope to make it more available and up for discussion by bringing it to people’s attention.

Even though, alcohol is not considered an illegal substance, it has a big impact on Australians. It is part of the culture and socialising. I hope to provide more awareness what alcoholism is.

# Motivation & questions

thought it would be interesting to find out how much of our demographics is the cause for an addiction or mental health disorder.

I do think there are more factors, from personal experience, as I have a mild form of depression myself. Luckily, I can function normally most of the time,

To get people involved in the subject, it is not enough to just talk about it. There is plenty of easily available information out there, it is the will to take time out of your life to check it out. And most people are in denial. It is easier not to accept it, than to confront it.

People love talking about themselves, so I wanted to make this a proper interaction. What better way than to make a shore questionnaire and make a prediction on their substance abuse and mental health state.

This way, it will be an easier topic to talk about. Like, “have you seen that website with these ridiculous predictions about my future drug us?”

These are the predictions I want to answer:

1. Estimate days to smoke cigarettes in the next 30 days.
2. Estimate days to drink alcoholic beverages in the next 30 days.
3. How likely have you, or are you going to, use the following substances (%):

* Marijuana/ Hashish, Cocaine, Crack, Heroin, Hallucinogens, Inhalants, Methamphetamines, Pain relievers, Tranquilizers, Stimulants, Sedatives.

1. You have:

* no probable serious mental illness, or
* a probable serious mental illness.

# Data collection

Searching for data was easy. Finding publicly available data was not.

In Australia the only information about health that is publicly available is statistical information. Since I needed the raw data to apply machine learning, I could not use this. Fortunately, in the USA they are not as protective of their health data. The limitation is that the machine learning is based on behaviour of people in the USA. In Australia, we may, or may not, behave the same.

The data I found is a very long survey (about 2,000), the ‘National Survey on Drug Use and Health (NSDUH)’. This survey has been taken since 1979. I have used 2015 to 2019 as the basis for the machine learning. The raw datasets have over 56,000 rows of data with over 2,500 categories for each year.

I cleaned the data by:

* removing all rows with an error message
* removing all individuals aged under 18
* extracted the columns needed for this project
* removed answers that did not provide machine learning information, like ‘refused to answer’ or ‘not response’
* replaced some values with other values to get consistent data.

The cleaned dataset has 120,543 rows of data in 64 categories.

# Machine learning models – Initial predictions

The best models to work with categorical data are:

* Logistic Regression
* Random Forest
* Sequential
* SVC (sigmoid)
* Naïve Bayes

## Logistic Regression

I chose to start with the logistic regression because it is the benchmark model. It is the most common model as the logic behind it is explained more easily than the other available models. It uses a logarithmic transformation on the outcome variable which allows us to model a non-linear association in a linear way. The outcome is dichotomous (e.g. success/ failure, yes/ no, died/ lived, etc).

The dataset already came with binary encoding. Because the fitting of the data to the model didn’t work as it was too large, I realised I also had to encode the data. Once this was completed, there was no more issue. It did take 1h20 to complete the fitting of the largest dataset to the model using the parameter optimisation Grid Search CV.

## Random Forest

As my dataset consists of a lot of classifications, it is important to know which ones contribute most to the prediction. This is important because even tough people want to talk about themselves and be involved, they don’t want to spend hours answering questions. This is why I chose the Random Forest model as the second model to ensure I can ask the least number of questions to get the most accurate prediction.

## Both models

To get an idea on how much data is needed to be as accurate as possible, I started with all the categories and worked my way down.

* The initial prediction is based on all 63 categories with category 64 as the y-value.
* The second prediction is based on all demographic categories and the selected categories related to the prediction.
* The third prediction is based on only the demographic categories.
* The fourth prediction is based on the most important demographic categories and the single most important related category.

# Category analysis – Initial predications

## Smoking

This is to accurately predict the estimated number of days that the subject will smoke cigarettes in the next 30 days.

*Initial prediction*

The initial prediction based on the 63 categories is almost 100% (0.999) for both models and whether or not the classifications have been optimised.

*Second prediction*

The second prediction based on 20 categories just as accurate at almost 100% (0.999) for both models and whether or not the classifications have been optimised.

The category that influences the prediction the most is ‘CIG100LF’, which is if they have smoked 100 cigarettes in their life. This category will be added to the questions to the subject. As such I will also use it in the predications of the models for the other predications (alcohol, drug and mental health).

*Third prediction*

The accuracy of the third prediction based on 16 categories is a lot less. This is to be expected as there is no information about any smoking habit from the subject.

The initial R2 value in the Logistic Regression model is 0.601 for the training data and 0.596 for the test data. The exact same score is achieved after optimising the classifications. The values are very close, so the model is successful.

The outcome of the prediction of the Random Forest model is not as good as the Logistic regression model. The R2 for the training data is 0.796 but the R2 for the test data is 0.578 which is not very close together. This means the prediction is not going to be very accurate. After optimising the parameters, the outcome is nearly the same as the initial model.

The demographic categories that had the least amount of influence on the predication (<0.02) and may be removed depending on the outcome of the other predications are:

* 0.003559328298874442, 'DIFFDRESS'
* 0.011260526233340732, 'DIFFWALK'
* 0.011853741454514339, 'DIFFERAND'
* 0.012038787040043929, 'SERVICE'

## Alcohol

This is to accurately predict the estimated number of days that the subject will drink an alcoholic beverage in the next 30 days.

For the smoking predication there was a question in the survey about estimated smokes per 30 days using bins. This option is also available for drinks but more than 50% skipped the question making the dataset significantly smaller. That is why I decided not to use the same category.

*Initial prediction*

As expected, due to the number of available options (30), the prediction value is quite low at approximately 0.54. I didn’t expect it to be this low though due to still having 4 other alcohol related categories.

The random Forest model has a very high R2 value for the training data but the same low value for the test data. This indicates that the model is not working well.

*Second prediction*

The second prediction has almost the same values as the first prediction, including the high difference in training and test data by Random Forest.

The category that influences the prediction the most is ‘ALCUS30D, which is the number of alcoholic drinks each day that you had a drink in the last 30 days. This category will be added to the questions to the subject. As such I will also use it in the predications of the models for the other predications (drug and mental health).

*Third prediction*

As expected, the Logistic regression prediction value is even less at 0.37. The Random Forest predication for the training data is significantly more than the test data which we have seen before for this model.

The feature importance is showing that there are 7 categories under 0.02. To keep it consistent with the predication for smoking, the bottom 4 categories are:

* 0.0011064271469488575, 'DIFFDRESS'
* 0.003918203159664558, 'DIFFWALK'
* 0.004167867517782593, 'DIFFERAND
* 0.010210052727015927, 'SERVICE'

## Drugs

This is to accurately predict how likely it is that the subject will use any of these drugs in their lifetime.

Before predicting the likelihood of drug use, I changed the dataframe slightly. I had to summarise some of the drug categories into one category as they were of the same type of drugs (hallucinogens (10 types) and inhalants (13 types)).

*Initial prediction*

As there are so many drugs to predict I am using an excerpt for the Proof-of-concept prediction. I’m using one of the groups and two other drugs. That way I can see if the prediction is consistent or totally different. I have randomly chosen to start with Marijuana, Cocaine and inhalants.

In all 3 predictions the category ‘CIGTRY’, the age when they first smoked a cigarette, is in the top 2 of feature importance. I will add this category to the questions of the survey and it will be added to the mental health prediction.

*Second prediction*

The predictions of the 3 drug types varies by about 20%. For two of the drug types the difference of the prediction between the first and second dataset is negligible. For one drug type the difference is quite significant from a 100% accuracy to a 72% accuracy.

The feature importance is almost identical for all 3 drug types. I will use these predictions without running all the drug types.

For cocaine and marijuana these are the bottom 4 categories:

* 'DIFFDRESS'
* 'DIFFERAND’
* 'DIFFWALK'
* ‘SEXIDENT’
* ‘SERVICE’ - at no 6 from the bottom, still under 0.02

For inhalants the bottom 4 categories are:

* 'DIFFDRESS'
* 'DIFFWALK’
* 'DIFFERAND’
* 'DIFFTHINK’ - very high with mental health
* ‘SEXIDENT’ - at no 5 from the bottom, still under 0.02
* ‘SERVICE’ - at no 6 from the bottom, still under 0.02

## Mental health

This is to accurately predict how likely it is that the subject has a probable serious mental illness.

Before predicting the likely mental illness, I calculated the total of the 6 mental health category outcomes in order to predict the final total result.

As the values in the category are 0 and 1, there is no need to encode these values.

*Initial prediction*

The prediction value for the Logistic Regression is very high at just above 0.92.

The Random Forest predication again has some difference between the training and test data. Not as much though staying within 0.1.

The category that influences the prediction the most is ‘ALCYRTOT, which is the total number of days alcohol was used in the past 12 months. I am not adding this category to the questions as it is a big guess for a quick survey.

*Second prediction*

As there are no other mental health categories, this prediction is similar to the third prediction of the smoking and alcohol predictions. The prediction is performed on all demographic categories with the addition of the most important smoking and alcohol categories.

The outcome of this prediction is almost as high as the initial prediction at 0.92 for the training data and 0.919 for the test data. This also indicates that the accuracy is not really affected by the lack of the category that most influences the outcome in the first prediction.

The demographic categories that had the least amount of influence on the predication (<0.02) and may be removed depending on the outcome of the other predications are:

* 0.004089621921158707, 'DIFFDRESS'
* 0.00877968472434257, 'DIFFWALK'
* 0.009145889493849356, 'SERVICE'
* 0.01925657313909633, 'SEXIDENT'

## Categories

To get a balanced outcome of the predictions, the following questions will be added to the survey:

* Have you smoked at least a 100 cigarettes in your lifetime?
* What age were you when you first smoked a cigarette?
* In the last 30 days, on days you had an alcoholic beverage, how many drinks did you have?

Some categories did not contribute much to the outcome of the prediction. As it appears that some of the categories are not important for any of the predications I’m trying to make with this study.

To balance out the number of questions to ask in the survey, the following categories will be removed:

* 'DIFFDRESS'
* 'DIFFWALK'
* 'SERVICE'
* 'DIFFERAND'

## Summery table prediction outcomes – Initial predications

Table

Description automatically generated

# Category analysis – Final predication

These predications were done after the other 1st, 2nd and 3rd predications (smoking, alcohol and mental health) were finalised so this will be the model for the website using all the categories that will be requested from the subject.

After the initial prediction results, it appears that the random forest does not perform well with a limited number of categories. The values for the training and test data vary significantly. I have added the Sequential model to add another option to find the best model for the most accurate predictions.

The final columns used in the predication are shown in Appendix 1.

## Smoking

## Alcohol

## Drugs

## Mental health

## Table Description automatically generatedSummery table prediction outcomes – Final predications

# Conclusion

# Limitations

* In Australia the only information about health that is publicly available are statistics. Since I need the raw data to apply machine learning, I could not use this. Fortunately, in the USA they are not as protective of their health data. The limitation is that the machine learning is based on behaviour of people in the USA. In Australia, we may, or may not, behave the same.

# If I had more time

* I would have liked to have looked at the impact of substance abuse on mental health.
* I would have liked to see if the accuracy of the alcohol predictions would have improved using the similar question as I did for smoking with an answer using bins for alcohol usage in the past 30 days. And the impact of significantly reducing the training data on the other predictions.

# Appendix 1 – Customer survey (drop down)

|  |  |
| --- | --- |
| **Category (code)** | **Code - response** |
| Age (AGE2) | 7 – 18 years old  8 – 19 years old  9 – 20 years old  10 – 21 years old  11 – 22 or 23 years old  12 – 24 or 25 years old  13 – between 26 and 29 years old  14 – between 30 and 34 years old  15 – between 35 and 49 years old  16 – between 50 and 64 years old  17 – 65 years or older |
| Marital status (IRMARIT) | 1 – Married  2 – Widowed  3 – Divorced or Separated  4 – Never been married |
| Overall Health (HEALTH) | 1 – Excellent  2 – Very good  3 – Good  4 – Fair  5 – Poor  94 – Don’t know |
| Times moved in the past 12 months (MOVSINPYR2) | 0 – 0 times  1 – 1 time  2 – 2 times  3 – 3 or more times |
| Sexual attraction (SEXATRACT) | 1 – Only attracted to opposite sex  2 – Mostly attracted to opposite sex  3 – Equally attracted to males and females  4 – Mostly attracted to same sex  5 – Only attracted to same sex  6 – Not sure  98 - Skip |
| Sexual identity (SEXIDENT) | 1 – Heterosexual  2 – Lesbian or Gay  3 – Bisexual  94 – Don’t know  98 - Skip |
| Serious difficulty concentrating, remembering, making decisions? (DIFFTHINK) | 1 – Yes  2 - No |
| Highest education (I REDUHIGHST2) | 1 – 5th grade or less grade completed  2 – 6th grade completed  3 – 7th grade completed  4 – 8th grade completed  5 – 9th grade completed  6 – 10th grade completed  7 – 11th or 12th grade completed, no diploma  8 - High school diploma/GED  9 - Some college credit, but no degree  10 – Associate degree (for example, AA, AS)  11 - College graduate or higher |
| Work situation in the past week (WRKSTATWK2) | 1 – Worked full-time  2 – Worked part-time  3 – Has job or volunteer worker, did not work  4 – Unemployed/on layoff, looking for work  5 – Disabled  6 – Keeping house full-time  7 – In school/ training  8 - Retired  9 - Does not have a job, some other reason |
| How many employers in the past 12 months (WRKNUMJOB2) | 1 – 1  2 – 2  3 – 3  4 – 4 or more |
| Employment status (IRWRKSTAT) | 1 – Employed full-time  2 – Employed part-time  3 – Unemployed  4 – Other (incl. not in labour force) |
| Income (IRPINC3) | 1 – Less than $10,000  2 – $10,000 - $19,999  3 – $20,000 - $29,999  4 – $30,000 - $39,999  5 - $40,000 - $49,999  6 - $50,000 - $74,999  7 - $75,000 or more |
| Smoked 100 cig in lifetime (CIG100LF) | 1 – yes  2 – no  91 – never used cigarettes |
| Age when first smoked cigarette (CIGTRY) | RANGE 1 – 65  991 – Never used cigarettes |
| Estimate # drinks on days that you drink in last 30 days (ALCUS30D) | RANGE 1 – 85  991 – Never used alcohol  993 – Did not use alcohol in last 30 days |

# Appendix 2 – Predictions

|  |  |
| --- | --- |
| 1. **Estimate days to smoke cigarettes in the next 30 days.** | |
| *Category (code)* | *Code - response* |
| Best estimate # of days smoked cig past 30 days (CG30EST) | 0 – 1 or 2 days  1 – 3 to 5 days  2 – 6 to 9 days  3 – 10 to 19 days  4 – 20 to 29 days  5 – All 30 days  6 – Never used cigarettes  7 – Did not use cig in last 30 days |

|  |  |
| --- | --- |
| 1. **Estimate days to drink alcoholic beverages in the next 30 days.** | |
| *Category (code)* | *Code - response* |
| # Days had one or more drinks past 30 days (ALCDAYS) | RANGE – 1 to 30  31 – Never used alcohol |

|  |  |
| --- | --- |
| 1. **How likely have you, or are you going to use the following substances** | |
| *Category (code)* | *Code - response* |
| Marijuana/ Hashish (MJEVER) | 0 – Yes  1 - No |
| Cocaine (COCEVER) | 0 – Yes  1 - No |
| Crack (CRKEVER) | 0 – Yes  1 - No |
| Heroin (HEREVER) | 0 – Yes  1 - No |
| Methamphetamines (METHAMEVR) | 0 – Yes  1 - No |
| Pain relievers (PNRANYLIF) | 0 – Yes  1 - No |
| Tranquilizers (TRQANYLIF) | 0 – Yes  1 - No |
| Stimulants (STMANYLIF) | 0 – Yes  1 - No |
| Sedatives (SEDANYLIF) | 0 – Yes  1 - No |
| Hallucinogens = group | 1 – Yes  0 - No |
| Inhalants = group | 1 – Yes  0 - No |

|  |  |
| --- | --- |
| 1. **You have % chance at having a probable serious mental illness.** | |
| *Category (code)* | *Code - response* |
| During the past 30 days, how often did you feel: | |
| nervous (DSTNRV30) | 1 – All of the time  2 – Most of the time  3 – Some of the time  4 – A little of the time  5 – None of the time |
| hopeless (DSTHOP30) | 1 – All of the time  2 – Most of the time  3 – Some of the time  4 – A little of the time  5 – None of the time |
| restless/ fidgety (DSTRST30) | 1 – All of the time  2 – Most of the time  3 – Some of the time  4 – A little of the time  5 – None of the time |
| so sad or depressed that nothing could cheer you up (DSTCHR30) | 1 – All of the time  2 – Most of the time  3 – Some of the time  4 – A little of the time  5 – None of the time |
| that everything was an effort (DSTEFF30) | 1 – All of the time  2 – Most of the time  3 – Some of the time  4 – A little of the time  5 – None of the time |
| worthless (DSTNGD30) | 1 – All of the time  2 – Most of the time  3 – Some of the time  4 – A little of the time  5 – None of the time |
| Total score <= 18 (No probable serious mental illness) | 1 – Yes  0 - No |
| Score 6-18 No probable serious mental illness  Score 19-30 Probable serious mental illness | |

# Appendix 3 – Raw data summery

2019 (data1\_df\_cp)

Total rows/ columns: 56,136/ 2,741

Columns used: 64

Rows after age removal: 42,739

Rows after error 85 removal: 41,950

Rows after removal of some responses:

Final dataframe rows/ columns: 24,584/ 64

2018 (data2\_df\_cp)

Total rows/ columns: 56,313/ 2,691

Columns used: 64

Rows after age removal: 43,026

Rows after error 85 removal: 42,252

Rows after removal of some responses:

Final dataframe rows/ columns: 24,575/ 64

2017 (data3\_df\_cp)

Total rows/ columns: 56,276/ 2,668

Columns used: 64

Rows after age removal: 42,554

Rows after error 85 removal: 41,761

Rows after removal of some responses:

Final dataframe rows/ columns: 23,961/ 64

2016 (data4\_df\_cp)

Total rows/ columns: 56,897/ 2,668

Columns used: 64

Rows after age removal: 42,625

Rows after error 85 removal: 41,844

Rows after removal of some responses:

Final dataframe rows/ columns: 23,613/ 64

2015 (data5\_df\_cp)

Total rows/ columns: 57,146/ 2,679

Columns used: 64

Rows after age removal: 43,561

Rows after error 85 removal: 42,752

Rows after removal of some responses:

Final dataframe rows/ columns: 23,810/ 64

Final total dataframe after merging (data\_complete\_df):

Rows/ columns: 120,543/ 64

Dataframe for the final prediction models:

Rows/ columns: 120,543/ 15