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](#_Toc64204323)

[Motivation & questions 3](#_Toc64204324)

[Data collection 3](#_Toc64204325)

[Machine learning 4](#_Toc64204326)

[Conclusion summery 4](#_Toc64204327)

[Implications 4](#_Toc64204328)

[Limitations 5](#_Toc64204329)

[Appendix 1 – Customer survey (drop down) 6](#_Toc64204330)

[Appendix 2 – Predictions 8](#_Toc64204331)

[Appendix 3 – Available info in addition to appendix 1 & 2 10](#_Toc64204332)

# 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. Estimate of number of alcoholic drinks on those days.
4. 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

I have categorical data:

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

# Conclusion summery

# Implications

# 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 there was more time, I could have looked at the impact of substance abuse on mental health.

# 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 |
| Ever been in the armed forces? (SERVICE) | 1 – Yes  2 – No  99 – Skip |
| 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 |
| Serious difficulty walking (DIFFWALK) | 1 – Yes  2 - No |
| Serious difficulty dressing or bathing (DIFFDRESS) | 1 – Yes  2 - No |
| Serious difficulty doing errands alone (DIFFERAND) | 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 |
| 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 - $39,9994  5 - $40,000 - $49,999  6 - $50,000 - $74,999  7 - $75,000 or more |

# 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) | 1 – 1 or 2 days  2 – 3 to 5 days  3 – 6 to 9 days  4 – 10 to 19 days  5 – 20 to 29 days  6 – All 30 days  91 – Never used cigarettes  93 – Did not use cig in last 30 days  94 – Don’t know |

|  |  |
| --- | --- |
| 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  91 – Never used alcohol  93 – Did not use alcohol in past 30 days |

|  |  |
| --- | --- |
| 1. **Estimate of number of alcoholic drinks on those days.** | |
| *Category (code)* | *Code - response* |
| Usual # drinks per day past 30 days (ALCUS30D) | RANGE – 1 to 85  991 – Never used alcohol  993 – Did not use alcohol in last 30 days |

|  |  |
| --- | --- |
| 1. **How likely have you, or are you going to use the following substances (%)** | |
| *Category (code)* | *Code - response* |
| Marijuana/ Hashish (MJEVER) | 1 – Yes  2 - No |
| Cocaine (COCEVER) | 1 – Yes  2 - No |
| Crack (CRKEVER) | 1 – Yes  2 - No |
| Heroin (HEREVER) | 1 – Yes  2 - No |
| Methamphetamines (METHAMEVR) | 1 – Yes  2 - No |
| Pain relievers (PNRANYLIF) | 1 – Yes  2 - No |
| Tranquilizers (TRQANYLIF) | 1 – Yes  2 - No |
| Stimulants (STMANYLIF) | 1 – Yes  2 - No |
| Sedatives (SEDANYLIF) | 1 – Yes  2 - No |
| Hallucinogens = group |  |
| Inhalants = group |  |

|  |  |
| --- | --- |
| 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 |
| Score 6-18 No probable serious mental illness  Score 19-30 Probable serious mental illness | |

# Appendix 3 – Available info in addition to appendix 1 & 2

|  |  |
| --- | --- |
| **Category (code)** | **Code - response** |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

# Appendix 4 – 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

(data\_complete\_df)

Final total dataframe after merging:

Rows/ columns: 120,543/ 64