Universidad Nacional Autónoma de México

Facultad de Estudios Superiores FES ACATLÁN

A brief approach to alcoholism in the student population of F.E.S Acatlán

Luna Castañeda Abraham Iván
Guerrero López Luis Arturo
López Cuéllar Ricardo Giovanni
López Velasco José Armando

Profesor:
José Gustavo Fuentes Cabrera

Contents

1 Introduction

Definition

Alcoholism can be defined in several ways, depending on the approach and model used. Alcoholism is a chronic illness in which a person feels a desire to drink alcoholic beverages and can not control that desire[1]. On the other hand, the psychiatric diagnosis manual DSM-IV includes its definition-classification two main types of alcoholism: alcohol abuse and alcohol dependence; considers the problem as a group of cognitive symptoms, mental and physiological behaviors that indicate that the individual has not stopped consuming the substance, despite the appearance of serious problems related to it.

On the other hand, the International Classification of Diseases (ICD-10) of the World Health Organization (WHO) defines dependence as ”a pattern of physiological manifestations and mental and cognitive behaviors in which the consumption of a drug, or of a type of them, acquires the highest priority for the individual. ” It can be added that the characteristics of dependence, tolerance, abuse and withdrawal syndrome are fundamental elements for the diagnosis of alcoholism although these terms may have different definitions according to the author or source.

There are differences in the way in which this problem occurs, depending on the particular characteristics of each society, social group or geographic region.In this way we find that there are cultural, geographical, gender differences, or based on race or creed.

At what age do we start?

According to the different studies and more recent reports, it can be said that in Mexico the starting age of alcohol consumption is between 13 and 18 years of age for the population in genera[13]l; in those who become alcoholic, it can start at a little earlier (10 to 11 years old), although this is not a rule since other groups of patients that started between 18 and 24 years of age have been found (men and women respectively) with a very intense consumption until developing cirrhosis[9]. However, the most probable age at which alcohol consumption began was corroborated on the basis of studies in healthy young people in whom the age at which they started consuming some type of beverage at 12 years of age was identified[14].

Liqueur type

As for the type of drink ingested, it will depend on several factors, among which the geographical region as mentioned above and the socio-demographic and cultural characteristics of the population group analyzed stand out. According to the 2008 ENA, the most preferred beverages among the population aged

12 to 65 are: beer, distilled drinks, wine, prepared beverages, pulque and 96^o alcohol, in that order[7]. The order of preference by type of drink is similar in men and women. As for the pulque extracted from the pulque maguey, which is mentioned as the characteristic drink of the Mexican highlands, it has a concentration of 3 to 6g% alcohol[12]. Among young people, the order of preference is similar to adults, with beer in the first place, drinks prepared in 3^o and wine in 4^o [7].

Are young students drinking?

According to the surveys for Ciudad de México conducted by the SEP-INPRF it was found that 61.4 % of adolescents have used alcohol at some time in their life and a 31.9 % have consumed it in the last month [15]. In this population, men are observed to have a discreetly higher frequency (34.0 %) than women (29.9 %). Regarding the educational level, at the junior high school level 22.6 % of adolescents have consumed alcohol in the last month. For high schools, this percentage is doubled, so that in technical schools 50.1 % of adolescents have drunk alcohol in the last month, and in high schools 43.4 % [15]. In relation to the age of adolescents, the percentage of consumers aged 14 was between 20 and 28 %, which is about half of those who are 18 or older. It is also noticeable that half of 17-year-olds have drunk alcohol in the last month, even when they are minors.

Epidemiology

According to the World Health Report of the World Health Organization (WHO), alcohol is among the first three risk factors for diseases in countries like ours and, at the same time, alcohol use diseases are among the first 10 major importance diseases[2].

The number of consults granted for drug problems in the two main public institutions of social security (medical insurance) in Mexico that covered more than half of the population (50 million Instituto Mexicano Social Security, IMSS and 7 million Instituto de Seguridad Social para los Trabajadores del Estado, ISSSTE) alcohol ranked first and second place in frequency as a reason for consultation in these institutions. In the mentioned year, 61,527 consultations were granted for reasons of abuse of alcohol and substances in the IMSS; Of that total, 48,115 consultations (40,759 for men and 7,356 for women) were due to alcohol problems[3, 4].

An important methodological aspect to take into account to study the epidemiology of alcoholism is the great limitation with which we find ourselves to study it, since most of the people or alcoholic patients come to request attention or help, in any of its modalities, when they are already in some clinical stage with well-established clinical or psychological involvement or even with the presence of alcoholism complications. Therefore, the need to investigate this problem

using strategies to study the clinical, psychological or medical problem in apparently healthy groups (students, homes), special groups (patients with alcoholic cirrhosis) and even those alcoholics who come to the emergency medical services problems related to an acute intoxication problem(accidents, injuries).
Alcoholism is one of the five main public health problems in Mexico, an example of this is the high proportion of deaths, both in forensic medical services and the Epidemiological and statistical system of deaths, which were associated with alcohol consumption[5] . In a review of cases of the Forensic Medical Service in the city of Guadalajara, Jal. of 2364 cases, 450 corpses (19 %) showed evidence of some addictive substance in body fluids, and of these 94.2 % tested positive for alcohol. Among the bodies positive for alcohol, the most frequent causes of death were: run over (20 %), motor vehicle crash (16 %), asphyxia (15.3 %) and firearm projectile (12 %)[6] .

The pattern in México

The distribution of alcohol consumption is not homogeneous in the population. On the one hand, 63 % of the rural adult population and 44 % of the urban population are abstemious. According to the Encuesta Nacional de Adicciones 2002 (ENA) in Mexico 72 % of adult men consume some type of alcoholic beverage while this happens in 43 % of adult women; according to the survey, 12.4 % of men are customary drinkers and 9.3 % meet the criterion of dependence, while in women 0.7 % are customary drinkers and 0.7 % have some dependence[8].
According to the 2008 ENA, customary consumption is more frequent in men at a ratio of 5.8 men per woman[7]. According to the same survey, in Mexico there are 27 million people who drink large quantities when consuming, something that has been considered the typical pattern of Mexicans[9, 10]. If we take into account that each of these subjects directly or indirectly affects four ascendants or descendants, this represents 108 million people affected by the immoderate consumption of alcohol; In other words, practically all Mexicans have someone close to them who has an alcohol problem.
Another characteristic of the pattern of consumption of episodic type that occurs more frequently in Mexicans[10], is that it is not performed daily, probably being the most frequent weekends[9], payday or end of the month, with large amounts of alcohol per occasion of consumption[10]. As for young people, the most recent ENA shows that they are copying the pattern of consumption of adults. A study conducted in high risk areas in the city of Guadalajara, the places of greatest consumption in order of frequency were, apart from the houses: street, parties, empty lots and parks. The pattern of highest consumption was also found on weekends and at night[11]. In these places there are high levels of tolerance and indifference.
In the same federative entity, the Epidemiological Surveillance System of Addictions that reports statistical information of patients who go to treatment, shows that the pattern of alcoholism in patients with addictions shows that the most consumed beverages are pure alcohol (46 %), beer (33 %), distillates (17.8

%), table wine (1.3 %); the starting age of alcohol consumption in most cases is between 10 and 24 years of age.

2 Problem approach

According to the ENA of 2002, per capita alcohol consumption was 2.79 liters for the population between 12 and 65 years old[8]. However, this number increases to 3.48 liters when considering the urban population between 18 and 65 years; If the men of the cities are examined exclusively, the number reaches 7.13 liters, while in rural environment males it is 5.91 liters.

These numbers contrast with the consumption in women that reaches 0.648 liters in urban adult women and 0.211 in rural women [9] and it is mentioned that they begin consumption at slightly later ages[8]. In men, consumption reaches its peak between 30 and 39 years of age in the urban population and between 40 and 49 years in the rural population and drops below 50 years of age [9]. A study conducted in the Clínica de Atención de Problemas Relacionados con el Alcohol (CAPRA) of the Hospital General de México in 1000 subjects of both sexes, found that 82.2 % of patients ingested more than 160 g of alcohol daily, 13.8 % ingested between 80 and 160 g and 3.4 % ingested less than 80g[12]. Those amounts are in the range that causes damage to the body.

As students we find normal to see our colleagues going once or twice a week to any of the bar that is near to the University campus, sometimes not even to a bar, but to a clandestine place that serves a strange kind of liquor. Certainly we can not say, yet, that they are alcoholics, but they may be starting to walk a very dangerous path and it is the objective of this machine learning project to be able to identify in time those who have great possibilities of becoming alcoholics, so that awareness strategies can be implemented, which, despite not being enough, are the only vaccine for this disease.

3 Theoretical framework

The chosen language and tools

There are multiple programing languages that allow us to work with machine learning algorithms, for this project, we are using Python, the Scikit-Learn library and Jupyter notebook.
The Python programming language is establishing itself as one of the most popular languages for scientific computing. Thanks to its high-level interactive nature and its maturing ecosystem of scientific libraries, it is an appealing choice for algorithmic development and exploratory data analysis. Yet, as a general purpose language, it is increasingly used not only in academic settings but also in industry.[18]

Project Jupyter

Project Jupyter's mission is to create open source tools for interactive scientific computing and data science in research, education and industry, with an emphasis on usability, collaboration and reproducibility. For the first decade, Python focused strictly on scientific and interactive computing in the Python language, providing a rich interactive shell well suited to the workflow of everyday research, as well as tools for parallel computing. It was part of an organic ecosystem of open source projects for scientific computing in Python, informally known as the "SciPy Stack". [19]

Scikit-Learn

This library is specifically designed to work with machine learning using python language. Scikit-Learn is an open source python library of popular machine learning algorithms that will allow us to build models.
The project was started in 2007 as a google summer of code project by David Cournapeau, lather that year Matthieu Bruncher started working on this proyect as a part of his thesis. In 2010 Fabian Pedregosa, Gael Varoquaux , Alexander Gramfort, and Vincet Michel of INRIA took the project leadership and produced the first public release. Nowdays the project is developed very actively by an enthusiastic community of contributors.[20]

Machine Learning

These days there is a lot of information in the world, all this information is daily collected in many ways and by different means. This information has a great value since with it a great variety of events and behaviors can be predicted. For this complicated task today there is Machine Learning. This term was coined by Albert Samuel in 1959, Samuel was an American pioneer in the field of gaming and artificial intelligence.[16]

Machine learning is the scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as ”training data”, in order to make predictions or decisions without explicitly programmed to perform the task. [17]

Model training

There are three ways to train a model:

- Supervised Learning: Is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples.[24]
- Unsupervised Learning: Unsupervised machine learning infers patterns from a dataset without reference to known, or labeled, outcomes. Unlike supervised machine learning, unsupervised machine learning methods cannot be directly applied to a regression or a classification problem because we have no idea what the values for the output data might be, making it impossible to train the algorithm the way it is normally done. Unsupervised learning can instead be used for discovering the underlying structure of the data.[22]
- Semi-Supervised Learning: Is a class of machine learning tasks and techniques that make use small amount of labeled data with a large amount of unlabeled data. Many machine-learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy over unsupervised learning, but without the time and costs needed for supervised learning.[23]

Machine learning classification models

A machine learning model is a mathematical model capable of generating predictions by detecting patterns in the data that is entered. There are multiple types of models that can be used, including classification models and regression models.

- Binary classification models: They have the ability to predict a binary result, that is, if something happens or not, these types of models can answer questions as if a customer bought a product or not, an email is useful or is spam among many other unknowns. To form binary classification models machine learning uses different algorithms, one of the most efficient and in some way the "standard" algorithm, is the logistic regression.

- Logistic regression models: A widely used model in different branches of knowledge such as medical and social sciences. Also called logistic model, it uses a logistic function to model a binary dependent variable. Mathematically speaking, a binary logistic model has dependent variables with two possible values, such as: fail / pass, true / false, dead / alive etc. these are represented by a variable indicator, also called dummy variable, this indicates the presence or absence of categorical effect, where the two variables are marked with 0 and 1. In the logistic model the logarithm of the probabilities for the value labeled with 1 is a linear combination of one or more independent variables that work as predictors, the independent variables can be binary variables or continuous variables.

- Multiclass classification models: These models allow to generate predictions for several classes, this means one or more than two results, a multiclass classification model can answer questions that involve differentiating between several options, for example, which genre has a movie, which categories of products are more important for a client, what kind of factors affect a population etc. Similar to those of binary classification, the "standard" algorithm used to generate these models is the multinomial logisitc regression.

- Roc curve: The curve roc works from a matrix of confusion, this matrix classifies the four possible states that can be obtained in a binary prediction, the states are the following:

  - True positive: This occurs when the value of the prediction is true or positive and also the given data.
  - False positive: Occurs when the value of the prediction is negative and the given value is positive.
  - True negative: Occurs when the value of the prediction is negative as well as the given value.
  - False Negative: Occurs when the value of the prediction is negative and the given value is positive.[21]

- Random Forests: Random Forests add an additional layer of randomness to bagging. In addition to constructing each tree using a different bootstrap sample of the data, random forests change how the classification or regression trees are constructed.

  In standard trees, each node is split using the best split among all variables. In a random forest, each node is split using the best among a sub-set of predictors randomly chosen at that node.This somewhat counter intuitive strategy turns out to perform very well compared to many other classifiers, including discriminant analysis, support vector machines and neural networks, and is robust against over fitting.[25]

Ordinal encoders

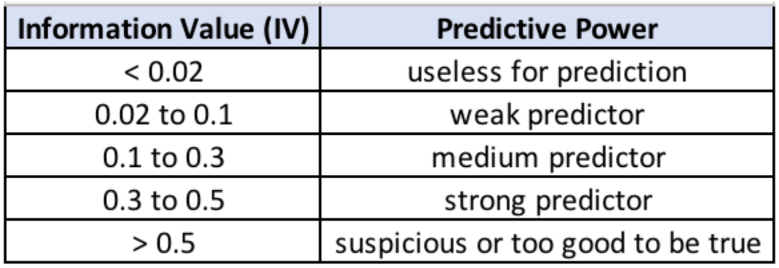
Ordinal encoders,given a dataset, encode categorical fields using ordinal encoding, which uses a single column of integers to represent field classes. It then creates a new dataset, with additional fields containing ordinal encodings of the categorical fields.
If classes have a known order (such as Like, Somewhat Like, Neutral, Somewhat Dislike, and Dislike), the integer mapping can be supplied; otherwise, integers are assigned by class count, in descending order (in the case of ties, classes are ordered alphabetically).

Weight of evidence and Information Value

Weight of evidence (WOE) and Information value (IV) are techniques to perform variable transformation and selection. The formulas to calculate WOE and IV are this:

$$WOE = \ln \frac{Event\%}{NoEvent\%}$$
$$IV = \sum (Event\% - NoEvent\%) * (WOE)$$

We are using WOE because it can handle missing values and outliers. The transformation is based on logarithmic value of distributions. Also, there's no need for dummy variables.
On the other hand, IV value can be used to select variables quickly. The selection is made based on the table

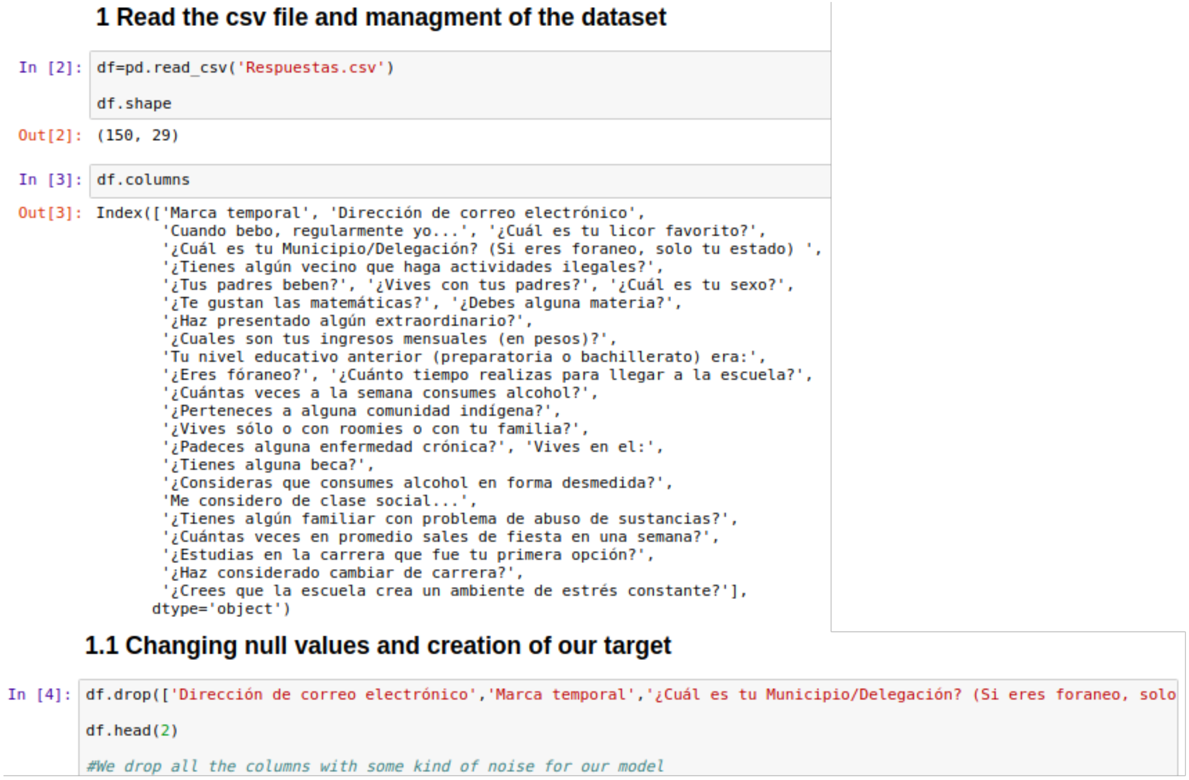


Information Value (IV)	Predictive Power
< 0.02	useless for prediction
0.02 to 0.1	weak predictor
0.1 to 0.3	medium predictor
0.3 to 0.5	strong predictor
> 0.5	suspicious or too good to be true

4 Development

As said during our introduction, we're using Jupyer notebook along with python to extract data, treat it and feed it to the model.The structure of our notebook is defined so that, before putting blocks of code, it's briefly described what will that block perform.

Now, let's see the process:

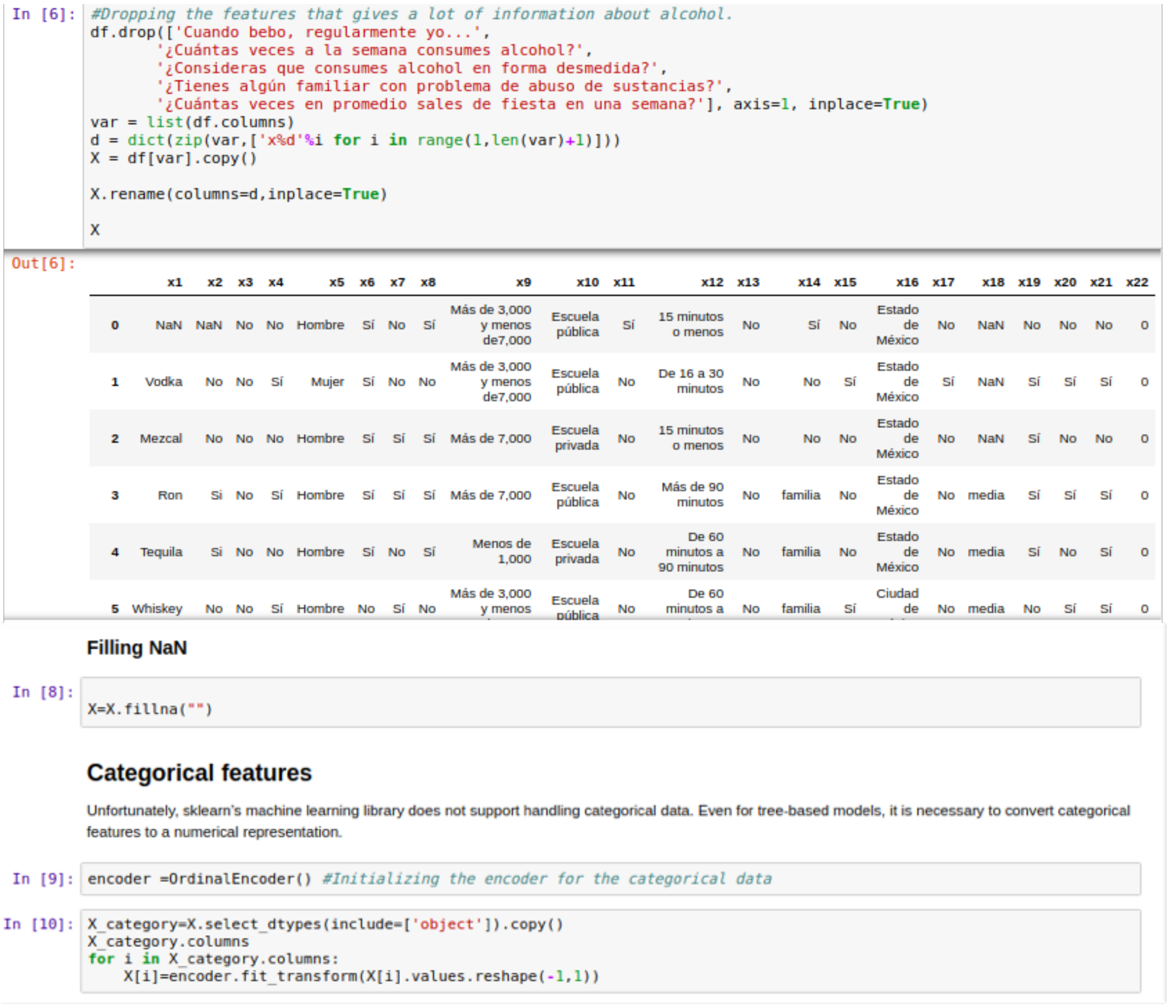


1 Read the csv file and managment of the dataset

```
In [2]: df=pd.read_csv('Respuestas.csv')

        df.shape

Out[2]: (150, 29)

In [3]: df.columns

Out[3]: Index(['Marca temporal', 'Dirección de correo electrónico',
               'Cuando bebo, regularmente yo...', '¿Cuál es tu licor favorito?',
               '¿Cuál es tu Municipio/Delegación? (Si eres foraneo, solo tu estado) ',
               '¿Tienes algún vecino que haga actividades ilegales?',
               '¿Tus padres beben?', '¿Vives con tus padres?', '¿Cuál es tu sexo?',
               '¿Te gustan las matemáticas?', '¿Debes alguna materia?',
               '¿Haz presentado algún extraordinario?',
               '¿Cuales son tus ingresos mensuales (en pesos)?',
               'Tu nivel educativo anterior (preparatoria o bachillerato) era:',
               '¿Eres fóraneo?', '¿Cuánto tiempo realizas para llegar a la escuela?',
               '¿Cuántas veces a la semana consumes alcohol?',
               '¿Perteneces a alguna comunidad indígena?',
               '¿Vives sólo o con roomies o con tu familia?',
               '¿Padeces alguna enfermedad crónica?', 'Vives en el:',
               '¿Tienes alguna beca?',
               '¿Consideras que consumes alcohol en forma desmedida?',
               'Me considero de clase social...',
               '¿Tienes algún familiar con problema de abuso de sustancias?',
               '¿Cuántas veces en promedio sales de fiesta en una semana?',
               '¿Estudias en la carrera que fue tu primera opción?',
               '¿Haz considerado cambiar de carrera?',
               '¿Crees que la escuela crea un ambiente de estrés constante?'],
              dtype='object')
```

1.1 Changing null values and creation of our target

```
In [4]: df.drop(['Dirección de correo electrónico','Marca temporal','¿Cuál es tu Municipio/Delegación? (Si eres foraneo, solo

        df.head(2)

        #We drop all the columns with some kind of noise for our model
```



```
In [6]: #Dropping the features that gives a lot of information about alcohol.
        df.drop(['Cuando bebo, regularmente yo...',
                '¿Cuántas veces a la semana consumes alcohol?',
                '¿Consideras que consumes alcohol en forma desmedida?',
                '¿Tienes algún familiar con problema de abuso de sustancias?',
                '¿Cuántas veces en promedio sales de fiesta en una semana?'], axis=1, inplace=True)
        var = list(df.columns)
        d = dict(zip(var,['x%d'%i for i in range(1,len(var)+1)]))
        X = df[var].copy()

        X.rename(columns=d,inplace=True)

        X
```

Out[6]:

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16	x17	x18	x19	x20	x21	x22
0	NaN	NaN	No	No	Hombre	Sí	No	Sí	Más de 3,000 y menos de7,000	Escuela pública	Sí	15 minutos o menos	No	Sí	No	Estado de México	No	NaN	No	No	No	0
1	Vodka	No	No	Sí	Mujer	Sí	No	No	Más de 3,000 y menos de7,000	Escuela pública	No	De 16 a 30 minutos	No	No	Sí	Estado de México	Sí	NaN	Sí	Sí	Sí	0
2	Mezcal	No	No	No	Hombre	Sí	Sí	Sí	Más de 7,000	Escuela privada	No	15 minutos o menos	No	No	No	Estado de México	No	NaN	Sí	No	No	0
3	Ron	Si	No	Sí	Hombre	Sí	Sí	Sí	Más de 7,000	Escuela pública	No	Más de 90 minutos	No	familia	No	Estado de México	No	media	Sí	Sí	Sí	0
4	Tequila	Si	No	No	Hombre	Sí	No	Sí	Menos de 1,000	Escuela privada	No	De 60 minutos a 90 minutos	No	familia	No	Estado de México	No	media	Sí	No	Sí	0
5	Whiskey	No	No	Sí	Hombre	No	Sí	No	Más de 3,000 y menos	Escuela pública	No	De 60 minutos a	No	familia	Sí	Ciudad de	No	media	No	Sí	Sí	0

Filling NaN

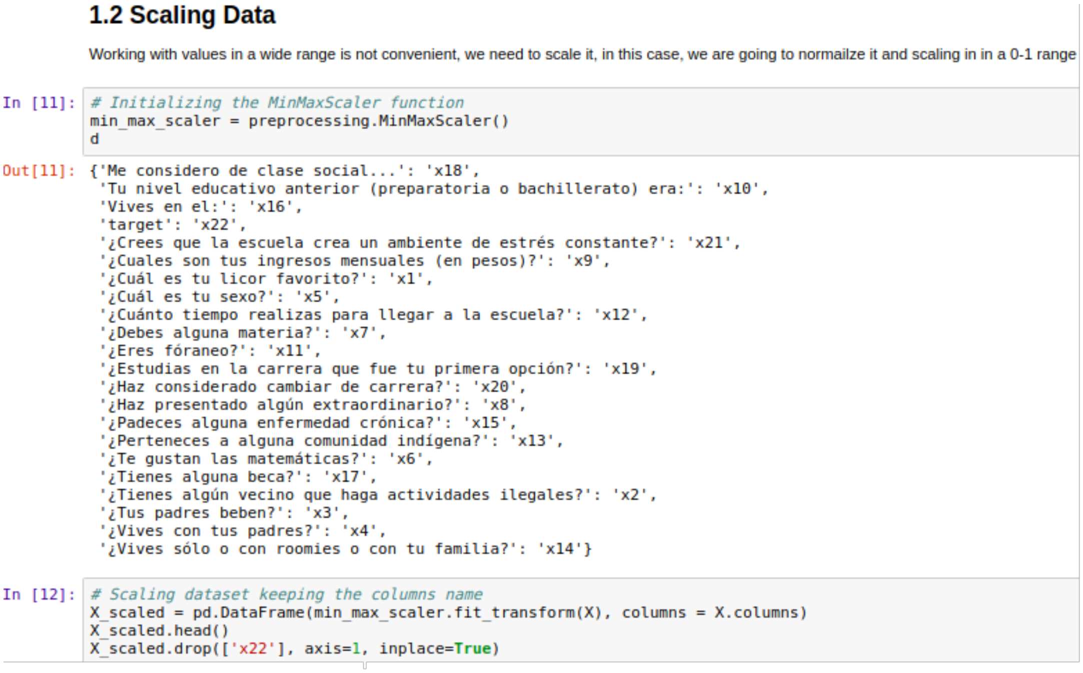
```
In [8]: X=X.fillna("")
```

Categorical features

Unfortunately, sklearn's machine learning library does not support handling categorical data. Even for tree-based models, it is necessary to convert categorical features to a numerical representation.

```
In [9]: encoder =OrdinalEncoder() #Initializing the encoder for the categorical data
```

```
In [10]: X_category=X.select_dtypes(include=['object']).copy()
         X_category.columns
         for i in X_category.columns:
             X[i]=encoder.fit_transform(X[i].values.reshape(-1,1))
```



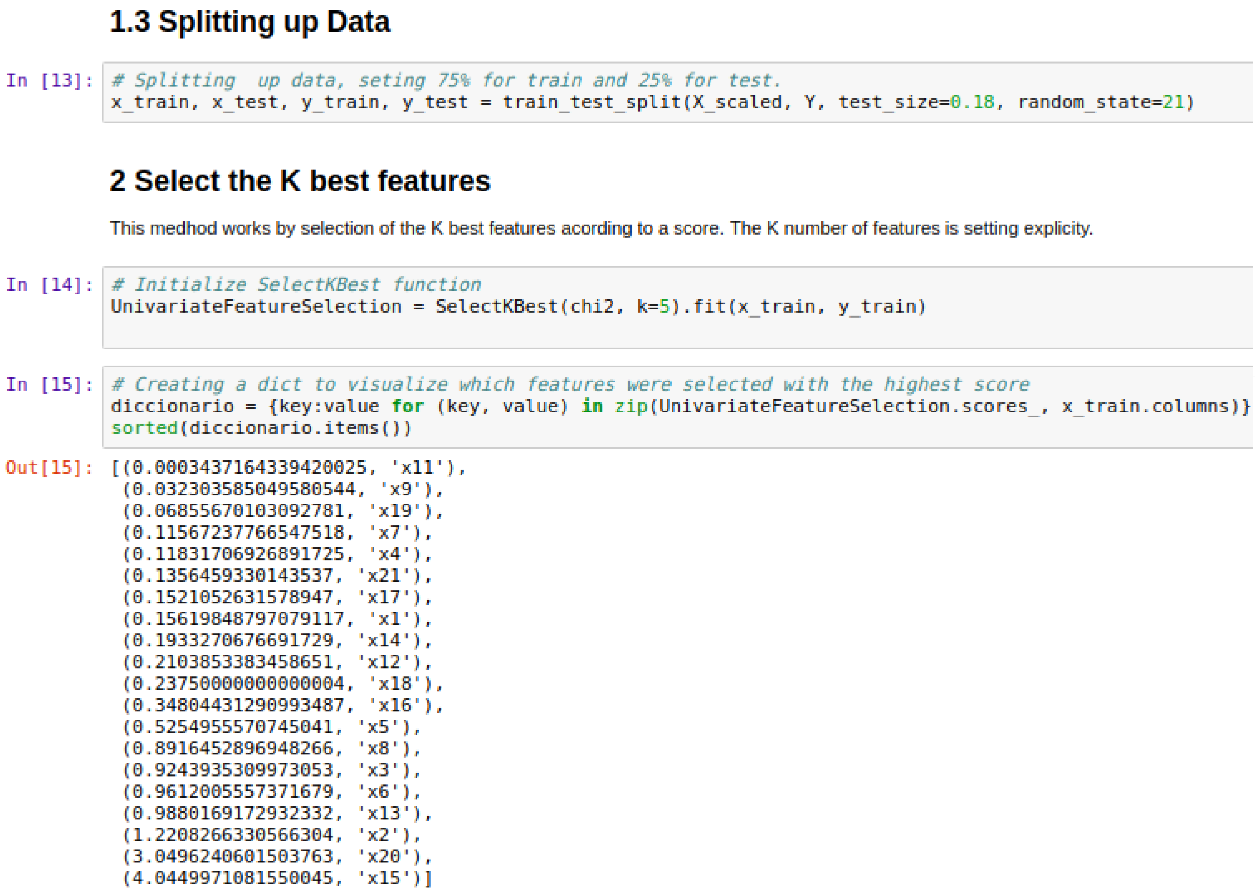
1.2 Scaling Data

Working with values in a wide range is not convenient, we need to scale it, in this case, we are going to normailze it and scaling in in a 0-1 range

In [11]:
```
# Initializing the MinMaxScaler function
min_max_scaler = preprocessing.MinMaxScaler()
d
```

Out[11]:
```
{'Me considero de clase social...': 'x18',
 'Tu nivel educativo anterior (preparatoria o bachillerato) era:': 'x10',
 'Vives en el:': 'x16',
 'target': 'x22',
 '¿Crees que la escuela crea un ambiente de estrés constante?': 'x21',
 '¿Cuales son tus ingresos mensuales (en pesos)?': 'x9',
 '¿Cuál es tu licor favorito?': 'x1',
 '¿Cuál es tu sexo?': 'x5',
 '¿Cuánto tiempo realizas para llegar a la escuela?': 'x12',
 '¿Debes alguna materia?': 'x7',
 '¿Eres fóraneo?': 'x11',
 '¿Estudias en la carrera que fue tu primera opción?': 'x19',
 '¿Haz considerado cambiar de carrera?': 'x20',
 '¿Haz presentado algún extraordinario?': 'x8',
 '¿Padeces alguna enfermedad crónica?': 'x15',
 '¿Perteneces a alguna comunidad indígena?': 'x13',
 '¿Te gustan las matemáticas?': 'x6',
 '¿Tienes alguna beca?': 'x17',
 '¿Tienes algún vecino que haga actividades ilegales?': 'x2',
 '¿Tus padres beben?': 'x3',
 '¿Vives con tus padres?': 'x4',
 '¿Vives sólo o con roomies o con tu familia?': 'x14'}
```

In [12]:
```
# Scaling dataset keeping the columns name
X_scaled = pd.DataFrame(min_max_scaler.fit_transform(X), columns = X.columns)
X_scaled.head()
X_scaled.drop(['x22'], axis=1, inplace=True)
```



1.3 Splitting up Data

```
In [13]: # Splitting  up data, seting 75% for train and 25% for test.
         x_train, x_test, y_train, y_test = train_test_split(X_scaled, Y, test_size=0.18, random_state=21)
```

2 Select the K best features

This medhod works by selection of the K best features acording to a score. The K number of features is setting explicity.

```
In [14]: # Initialize SelectKBest function
         UnivariateFeatureSelection = SelectKBest(chi2, k=5).fit(x_train, y_train)
```

```
In [15]: # Creating a dict to visualize which features were selected with the highest score
         diccionario = {key:value for (key, value) in zip(UnivariateFeatureSelection.scores_, x_train.columns)}
         sorted(diccionario.items())

Out[15]: [(0.0003437164339420025, 'x11'),
          (0.032303585049580544, 'x9'),
          (0.06855670103092781, 'x19'),
          (0.11567237766547518, 'x7'),
          (0.11831706926891725, 'x4'),
          (0.1356459330143537, 'x21'),
          (0.1521052631578947, 'x17'),
          (0.15619848797079117, 'x1'),
          (0.1933270676691729, 'x14'),
          (0.2103853383458651, 'x12'),
          (0.23750000000000004, 'x18'),
          (0.34804431290993487, 'x16'),
          (0.5254955570745041, 'x5'),
          (0.8916452896948266, 'x8'),
          (0.9243935309973053, 'x3'),
          (0.9612005557371679, 'x6'),
          (0.9880169172932332, 'x13'),
          (1.2208266330566304, 'x2'),
          (3.0496240601503763, 'x20'),
          (4.0449971081550045, 'x15')]
```

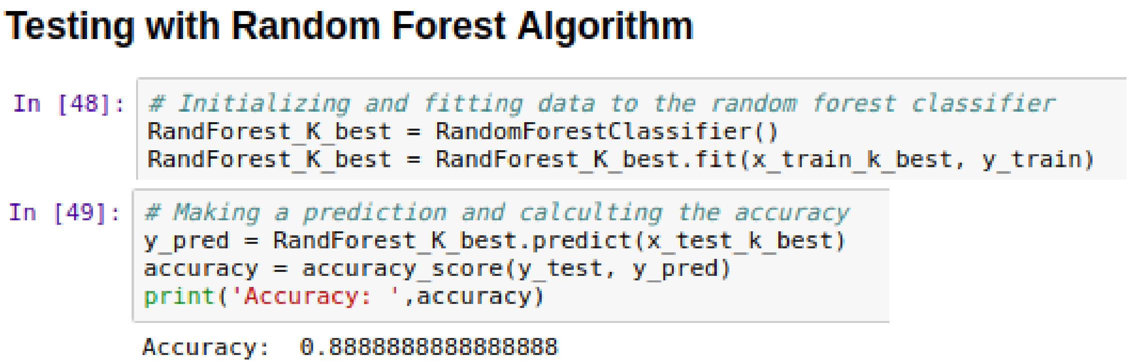
At this point, we should point that our best features are:

- x_{15} :¿Padeces alguna enfermedad crónica?
- x_{20} :¿Haz considerado cambiar de carrera?
- x_2 :¿Tienes algún vecino que haga actividades ilegales?

- x_{13} :¿Perteneces a alguna comunidad indigena?
- x_6 :¿Te gustan las matemáticas?

Training different models

Just for academic purposes, we tested different classification models in our data. This is nor practical for a bigger project, on a bigger scale you should already know which model suits best your data.

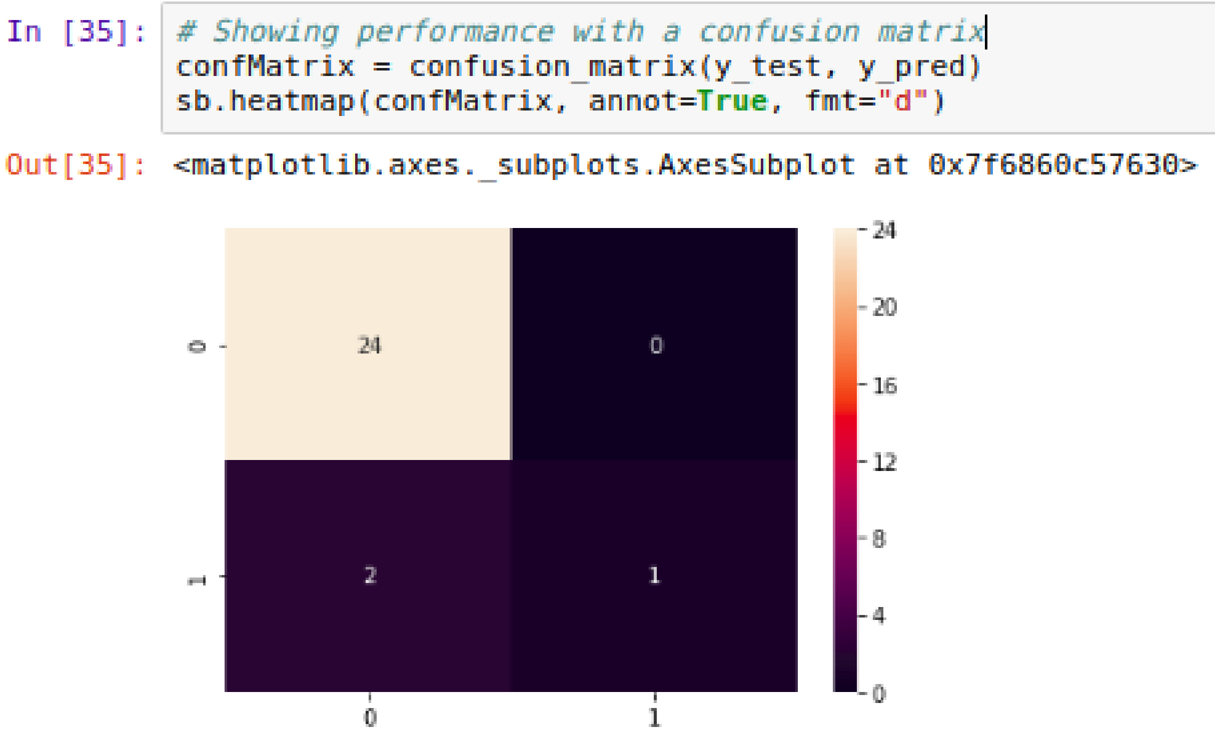


Testing with Random Forest Algorithm

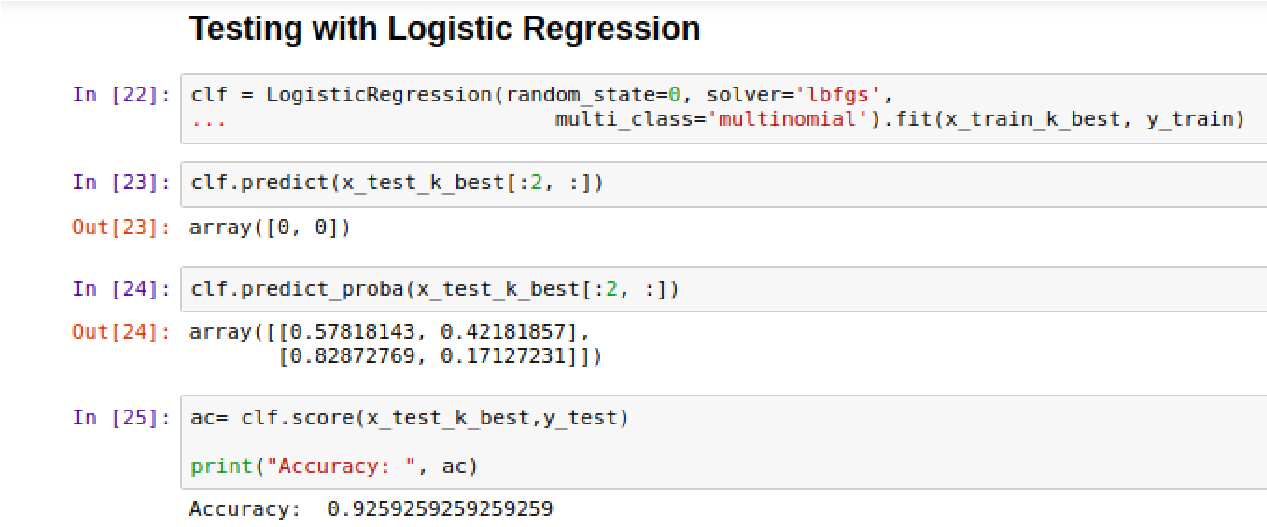
In [48]:
```
# Initializing and fitting data to the random forest classifier
RandForest_K_best = RandomForestClassifier()
RandForest_K_best = RandForest_K_best.fit(x_train_k_best, y_train)
```

In [49]:
```
# Making a prediction and calculting the accuracy
y_pred = RandForest_K_best.predict(x_test_k_best)
accuracy = accuracy_score(y_test, y_pred)
print('Accuracy: ',accuracy)
```

```
Accuracy:  0.8888888888888888
```



```
In [35]: # Showing performance with a confusion matrix
         confMatrix = confusion_matrix(y_test, y_pred)
         sb.heatmap(confMatrix, annot=True, fmt="d")

Out[35]: <matplotlib.axes._subplots.AxesSubplot at 0x7f6860c57630>
```



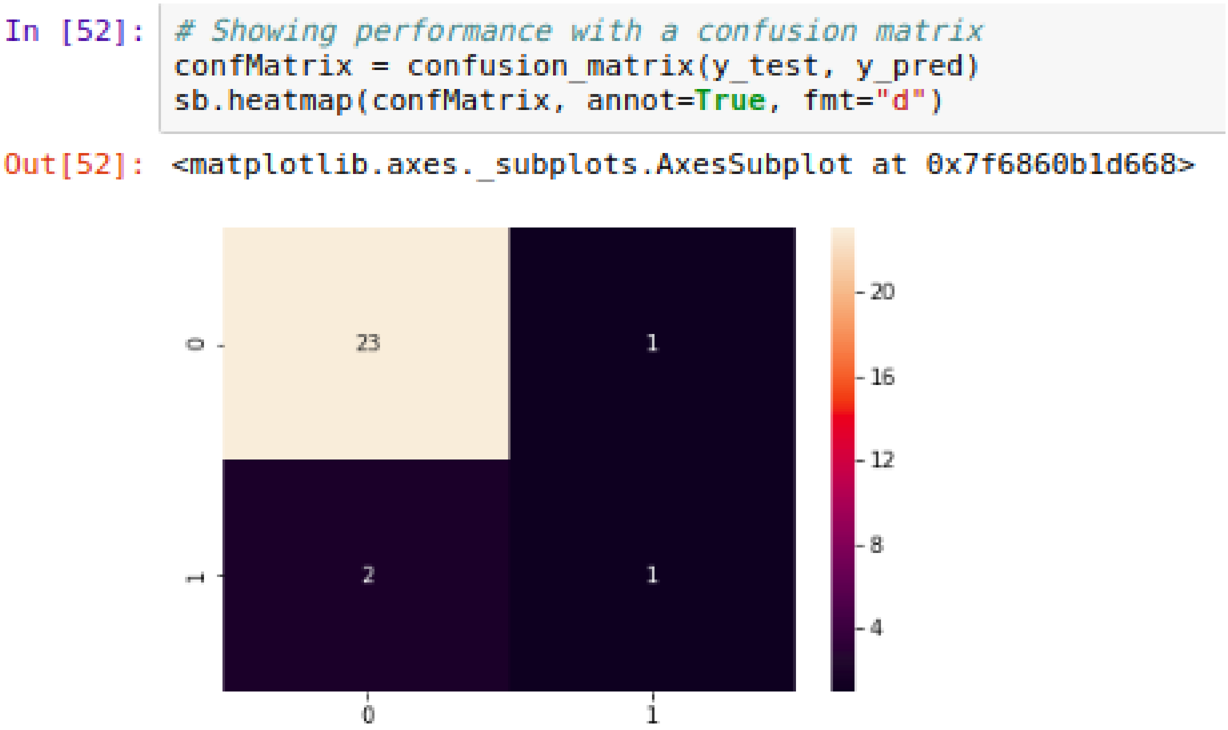
Testing with Logistic Regression

```
In [22]: clf = LogisticRegression(random_state=0, solver='lbfgs',
         ...                          multi_class='multinomial').fit(x_train_k_best, y_train)
```

```
In [23]: clf.predict(x_test_k_best[:2, :])

Out[23]: array([0, 0])
```

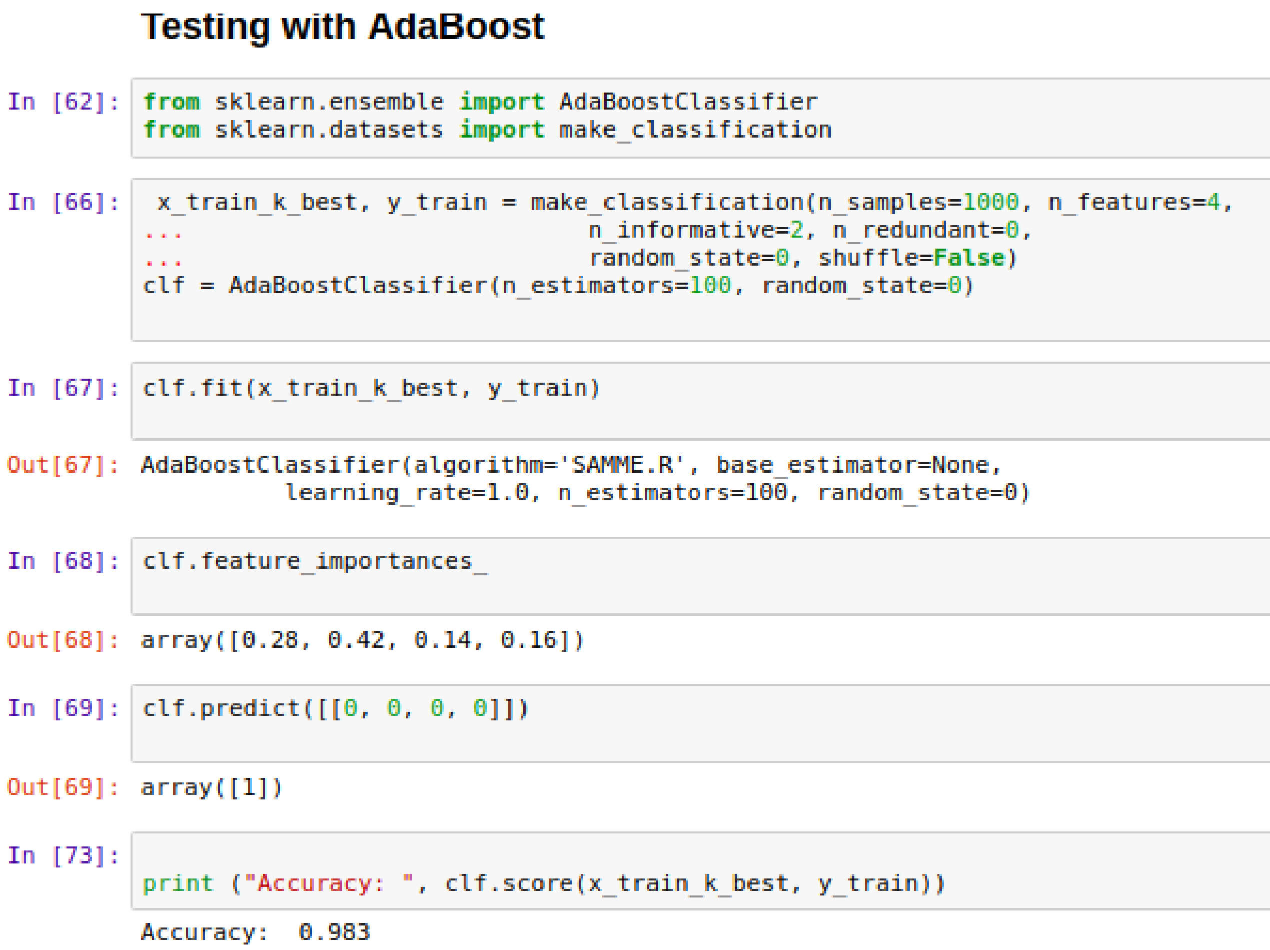
```
In [24]: clf.predict_proba(x_test_k_best[:2, :])

Out[24]: array([[0.57818143, 0.42181857],
                [0.82872769, 0.17127231]])
```

```
In [25]: ac= clf.score(x_test_k_best,y_test)

         print("Accuracy: ", ac)

         Accuracy:  0.9259259259259259
```



```
In [52]: # Showing performance with a confusion matrix
         confMatrix = confusion_matrix(y_test, y_pred)
         sb.heatmap(confMatrix, annot=True, fmt="d")
```

```
Out[52]: <matplotlib.axes._subplots.AxesSubplot at 0x7f6860b1d668>
```



Testing with AdaBoost

```
In [62]: from sklearn.ensemble import AdaBoostClassifier
         from sklearn.datasets import make_classification
```

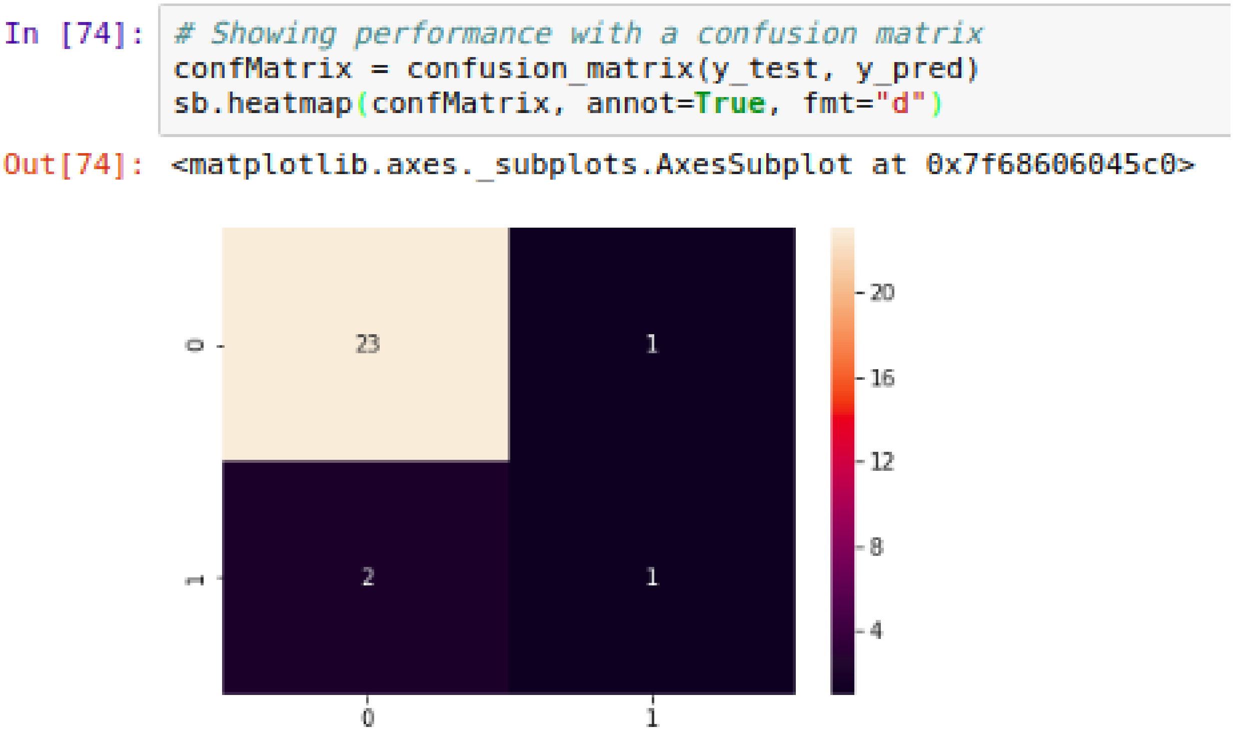
```
In [66]:  x_train_k_best, y_train = make_classification(n_samples=1000, n_features=4,
         ...                              n_informative=2, n_redundant=0,
         ...                              random_state=0, shuffle=False)
         clf = AdaBoostClassifier(n_estimators=100, random_state=0)
```

```
In [67]: clf.fit(x_train_k_best, y_train)

Out[67]: AdaBoostClassifier(algorithm='SAMME.R', base_estimator=None,
                   learning_rate=1.0, n_estimators=100, random_state=0)
```

```
In [68]: clf.feature_importances_

Out[68]: array([0.28, 0.42, 0.14, 0.16])
```

```
In [69]: clf.predict([[0, 0, 0, 0]])

Out[69]: array([1])
```

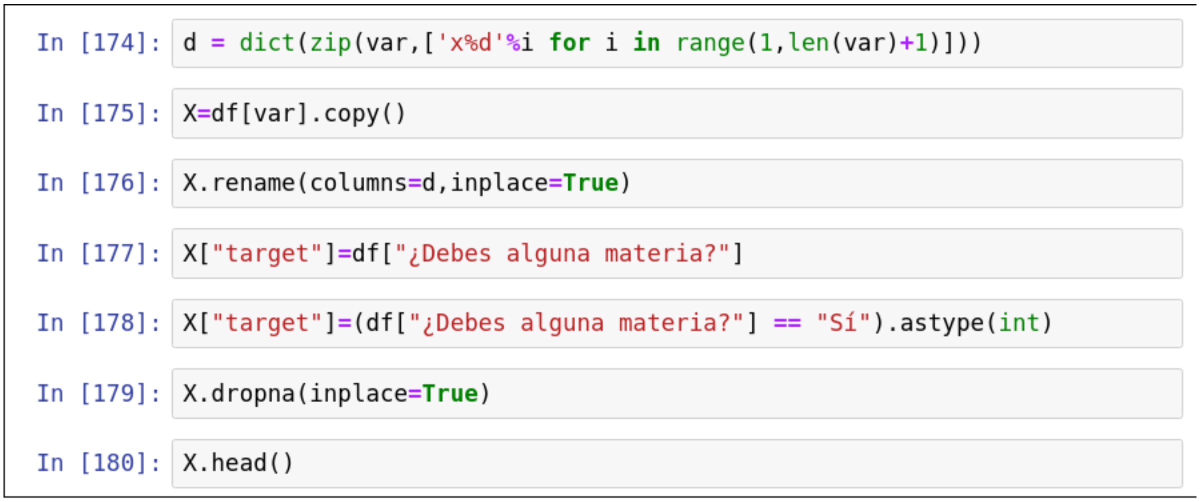
```
In [73]: 
         print ("Accuracy: ", clf.score(x_train_k_best, y_train))
         Accuracy:  0.983
```



```
In [74]: # Showing performance with a confusion matrix
         confMatrix = confusion_matrix(y_test, y_pred)
         sb.heatmap(confMatrix, annot=True, fmt="d")

Out[74]: <matplotlib.axes._subplots.AxesSubplot at 0x7f68606045c0>
```

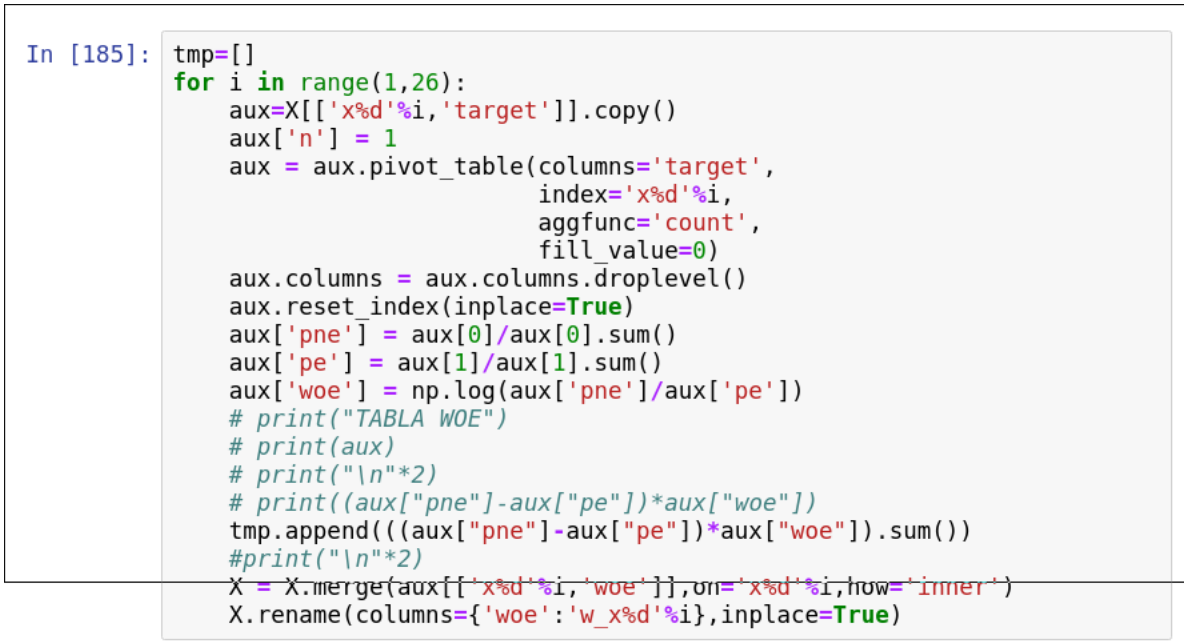
Second Model

first we select the variables with which we are going to work and we rename them to be able to work with WOE //

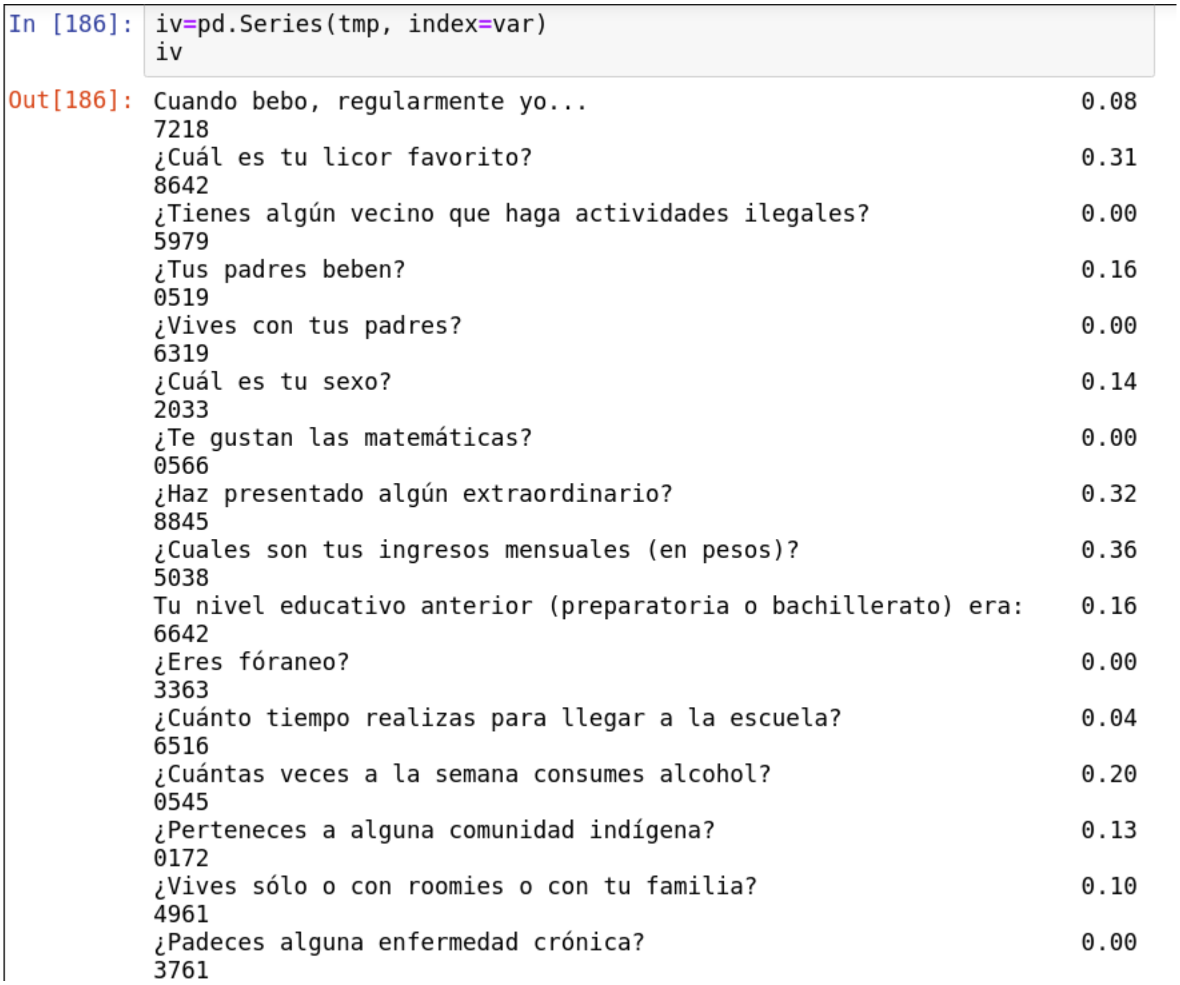


```
In [174]: d = dict(zip(var,['x%d'%i for i in range(1,len(var)+1)]))

In [175]: X=df[var].copy()

In [176]: X.rename(columns=d,inplace=True)

In [177]: X["target"]=df["¿Debes alguna materia?"]

In [178]: X["target"]=(df["¿Debes alguna materia?"] == "Sí").astype(int)

In [179]: X.dropna(inplace=True)

In [180]: X.head()
```

After selecting, we will calculate WOE and IV, this will give us the variables which give us more information



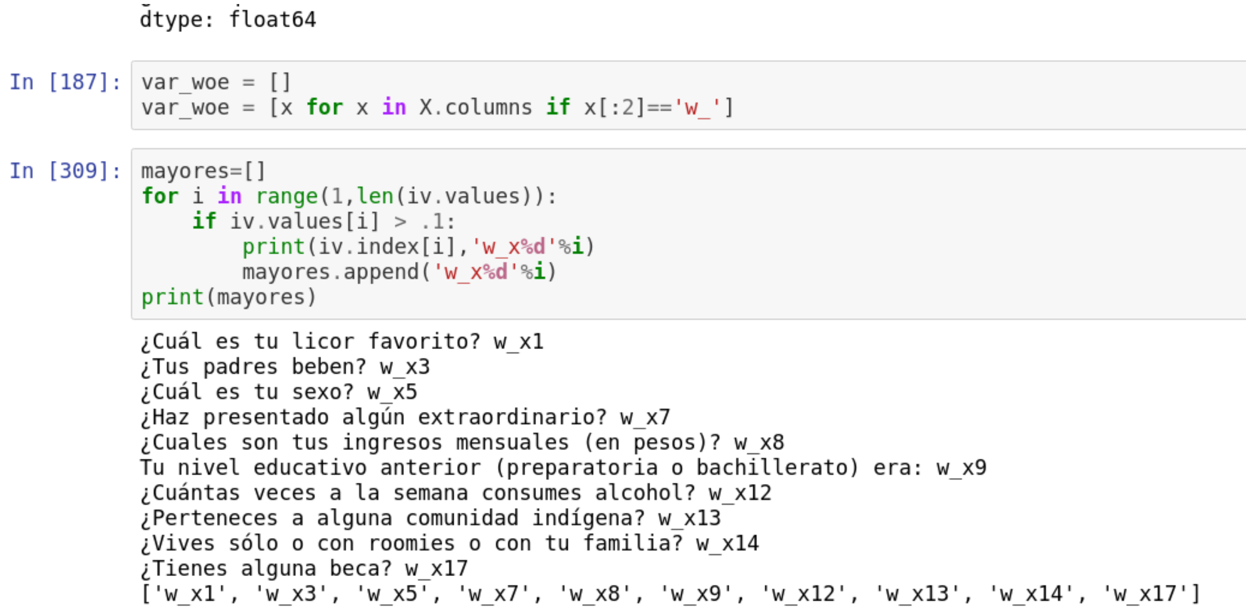
```
In [185]: tmp=[]
          for i in range(1,26):
              aux=X[['x%d'%i,'target']].copy()
              aux['n'] = 1
              aux = aux.pivot_table(columns='target',
                                    index='x%d'%i,
                                    aggfunc='count',
                                    fill_value=0)
              aux.columns = aux.columns.droplevel()
              aux.reset_index(inplace=True)
              aux['pne'] = aux[0]/aux[0].sum()
              aux['pe'] = aux[1]/aux[1].sum()
              aux['woe'] = np.log(aux['pne']/aux['pe'])
              # print("TABLA WOE")
              # print(aux)
              # print("\n"*2)
              # print((aux["pne"]-aux["pe"])*aux["woe"])
              tmp.append(((aux["pne"]-aux["pe"])*aux["woe"]).sum())
              #print("\n"*2)
              X = X.merge(aux[['x%d'%i,'woe']],on='x%d'%i,how='inner')
              X.rename(columns={'woe':'w_x%d'%i},inplace=True)
```



```
In [186]: iv=pd.Series(tmp, index=var)
          iv
```

```
Out[186]: Cuando bebo, regularmente yo...                                        0.08
          7218
          ¿Cuál es tu licor favorito?                                            0.31
          8642
          ¿Tienes algún vecino que haga actividades ilegales?                    0.00
          5979
          ¿Tus padres beben?                                                     0.16
          0519
          ¿Vives con tus padres?                                                 0.00
          6319
          ¿Cuál es tu sexo?                                                      0.14
          2033
          ¿Te gustan las matemáticas?                                            0.00
          0566
          ¿Haz presentado algún extraordinario?                                  0.32
          8845
          ¿Cuales son tus ingresos mensuales (en pesos)?                         0.36
          5038
          Tu nivel educativo anterior (preparatoria o bachillerato) era:        0.16
          6642
          ¿Eres fóraneo?                                                         0.00
          3363
          ¿Cuánto tiempo realizas para llegar a la escuela?                      0.04
          6516
          ¿Cuántas veces a la semana consumes alcohol?                           0.20
          0545
          ¿Perteneces a alguna comunidad indígena?                               0.13
          0172
          ¿Vives sólo o con roomies o con tu familia?                            0.10
          4961
          ¿Padeces alguna enfermedad crónica?                                    0.00
          3761
```

an algorithm was implemented to select the questions that give us more information, We put the value to .1 because we believe it is the value with which we have the best predictive power



```
dtype: float64
```

```
In [187]: var_woe = []
          var_woe = [x for x in X.columns if x[:2]=='w_']
```

```
In [309]: mayores=[]
          for i in range(1,len(iv.values)):
              if iv.values[i] > .1:
                  print(iv.index[i],'w_x%d'%i)
                  mayores.append('w_x%d'%i)
          print(mayores)
```

```
¿Cuál es tu licor favorito? w_x1
¿Tus padres beben? w_x3
¿Cuál es tu sexo? w_x5
¿Haz presentado algún extraordinario? w_x7
¿Cuales son tus ingresos mensuales (en pesos)? w_x8
Tu nivel educativo anterior (preparatoria o bachillerato) era: w_x9
¿Cuántas veces a la semana consumes alcohol? w_x12
¿Perteneces a alguna comunidad indígena? w_x13
¿Vives sólo o con roomies o con tu familia? w_x14
¿Tienes alguna beca? w_x17
['w_x1', 'w_x3', 'w_x5', 'w_x7', 'w_x8', 'w_x9', 'w_x12', 'w_x13', 'w_x14', 'w_x17']
```

Once the data has been obtained, which have already gone through a treatment, we will apply the logistic regression algorithm

```
In [312]: from sklearn.linear_model import LogisticRegression
          from sklearn.metrics import roc_auc_score
          from sklearn.metrics import accuracy_score
          from sklearn.model_selection import train_test_split
```

```
In [324]: y = X['target'].copy()
          Xw = X[mayores].copy()
```

```
In [325]: modelo = LogisticRegression()
```

```
In [326]: Xt,Xv,yt,yv = train_test_split(Xw,y,train_size=0.7)

/home/arturo/virtualenv/unam/lib/python3.7/site-packages/sklearn/model_selection/_split.py
m version 0.21, test_size will always complement train_size unless both are specified.
  FutureWarning)
```

```
In [327]: modelo.fit(Xt,yt)

/home/arturo/virtualenv/unam/lib/python3.7/site-packages/sklearn/linear_model/logistic.py
lt solver will be changed to 'lbfgs' in 0.22. Specify a solver to silence this warning.
  FutureWarning)

Out[327]: LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                    intercept_scaling=1, max_iter=100, multi_class='warn',
                    n_jobs=None, penalty='l2', random_state=None, solver='warn',
                    tol=0.0001, verbose=0, warm_start=False)
```

```
In [328]: modelo.coef_

Out[328]: array([[-0.25638459, -0.25638459,  0.28456453,  0.28456453,  0.12375487,
                   0.12375487, -0.0081188 , -0.0081188 , -0.467535  , -0.467535  ,
                  -0.55590293, -0.55590293, -0.15963492, -0.15963492, -0.1755246 ,
                  -0.1755246 , -0.44336573, -0.44336573, -0.09773637, -0.09773637]])
```

```
In [329]: modelo.intercept_

Out[329]: array([-0.10590526])
```

and as we can see we get a prediction power of 73

In [330]:

```
print(accuracy_score(y_true=yt,y_pred=modelo.predict(Xt)))
print(accuracy_score(y_true=yv,y_pred=modelo.predict(Xv)))
```

```
0.7227722772277227
0.7045454545454546
```

In [331]:

```
print(roc_auc_score(y_true=yt,y_score=modelo.predict_proba(Xt)[:,1]))
print(roc_auc_score(y_true=yv,y_score=modelo.predict_proba(Xv)[:,1]))
```

```
0.7759026687598115
0.7136842105263159
```

In [332]:

```
aux = pd.DataFrame({'y^':modelo.predict_proba(Xt)[:,1],'y':yt})
aux['y^'] = pd.cut(aux['y^'],bins=np.arange(0,1,0.1),include_lowest=True).astype(str)
aux['n'] = 1
aux = aux.pivot_table(index='y^',columns='y',aggfunc='count',values='n',fill_value=0)
```

In [333]:

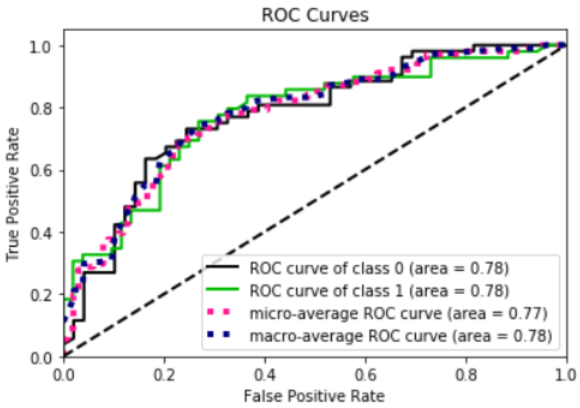
```
aux
```

Out[333]:

y	0	1
y^		
(-0.001, 0.1]	1	0
(0.1, 0.2]	10	2
(0.2, 0.3]	8	3
(0.3, 0.4]	10	3
(0.4, 0.5]	9	6
(0.5, 0.6]	5	12
(0.6, 0.7]	6	7
(0.7, 0.8]	2	7
(0.8, 0.9]	1	9

```
In [334]: import scikitplot as skplt
          import matplotlib.pyplot as plt
          %matplotlib inline

          y_true = yt
          y_probas = modelo.predict_proba(Xt)

          skplt.metrics.plot_roc_curve(y_true,y_probas)
          plt.show()
```

```
/home/arturo/virtualenv/unam/lib/python3.7/site-packages/sklearn/utils/deprecation.py:77:
ion plot_roc_curve is deprecated; This will be removed in v0.5.0. Please use scikitplot.me
  warnings.warn(msg, category=DeprecationWarning)
```



5 Description of the data

In this section we will detail what were the options that the respondents had. Each one of the questions was of multiple choice, and in this way we will study the relationship between reproving matters and alcohol. Below is what were the options to answer

Cuando bebo, regularmente yo...

1. Me tomo un par de cervezas/cocteles
2. Me pongo "happy"
3. No tomo
4. Me emborracho
5. Me pierdo totalmente

¿Cuál es tu licor favorito?

1. Cerveza

2. Tequila
3. Whiskey
4. Vodka
5. otros
6. Ninguno
7. Ron
8. Vino

¿Tienes algún vecino que haga actividades ilegales?

1. No
2. Si

¿Tus padres beben?

1. No
2. Si

¿Vives con tus padres?

1. Sí
2. No

¿Cuál es tu sexo?

1. Hombre
2. Mujer

¿Te gustan las matemáticas?

1. Sí
2. No

¿Debes alguna materia?

1. Sí
2. No

¿Haz presentado algún extraordinario?

1. Sí

2. No

¿Cuales son tus ingresos mensuales (en pesos)?

1. Más de 1,000 y menos de 3,000
2. Más de 3,000 y menos de7,000
3. Más de 7,000
4. Menos de 1,000

Tu nivel educativo anterior (preparatoria o bachillerato) era:

1. Escuela pública
2. Escuela privada

¿Eres fóraneo?

1. No
2. Sí

¿Cuánto tiempo realizas para llegar a la escuela?

1. De 31 a 60 minutos
2. De 60 minutos a 90 minutos
3. Más de 90 minutos
4. De 16 a 30 minutos
5. 15 minutos o menos

¿Cuántas veces a la semana consumes alcohol?

1. Menos de 1
2. Sólo 1
3. No consumo
4. 2 o más veces

¿Perteneces a alguna comunidad indígena?

1. No
2. Sí

¿Vives sólo o con roomies o con tu familia?

1. familia
2. roomies
3. solo

¿Padeces alguna enfermedad crónica?

1. No
2. Sí

Vives en el:

1. Estado de México
2. Ciudad de México

¿Tienes alguna beca?

1. No
2. Sí

¿Consideras que consumes alcohol en forma desmedida?

1. No
2. Tal vez
3. Sí

Me considero de clase social...

1. media
2. baja
3. alta

¿Tienes algún familiar con problema de abuso de sustancias?

1. No
2. Sí

¿Cuántas veces en promedio sales de fiesta en una semana?

1. Menos de 1
2. Al menos 2
3. Más de 2

¿Estudias en la carrera que fue tu primera opción?

1. Sí
2. No

¿Haz considerado cambiar de carrera?

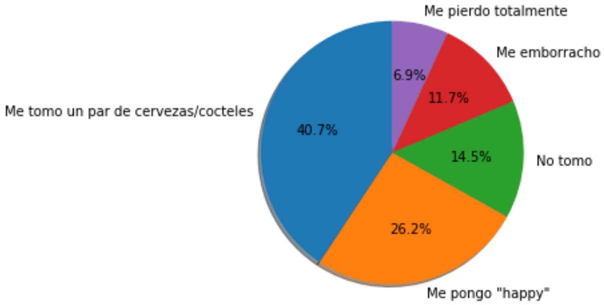
1. No
2. Sí

¿Crees que la escuela crea un ambiente de estrés constante?

1. Sí
2. No

6 Univariate analysis

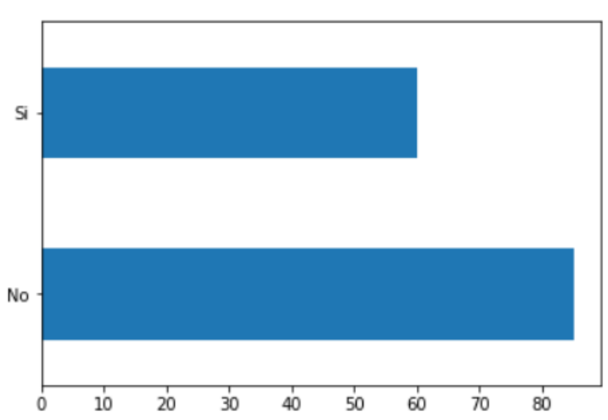
Cuando bebo, regularmente yo...
como podemos observar los encuestados toman solo un par de cervezas



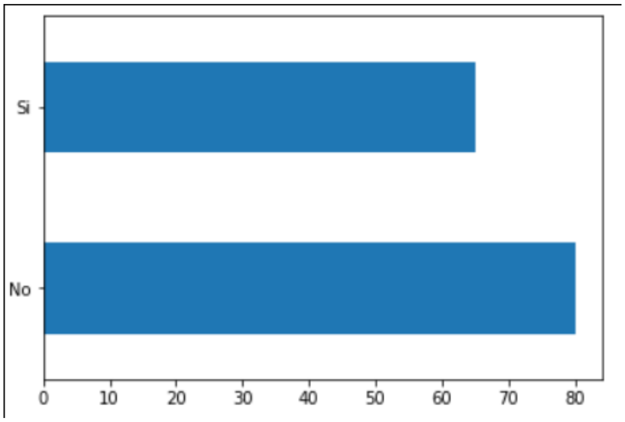
¿Cuál es tu licor favorito?
in this graph we can appreciate that the favorite drink is beer

img/2.png

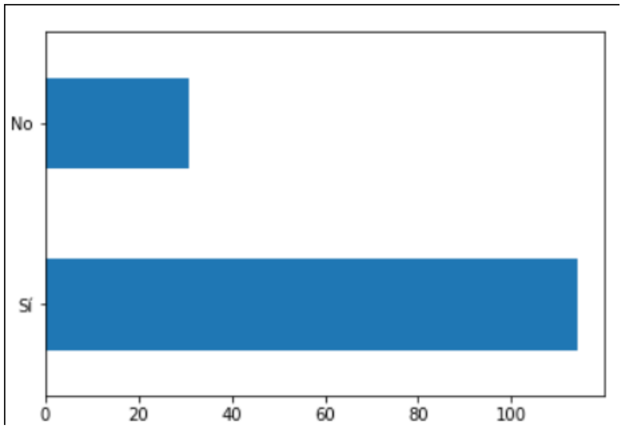
¿Tienes algún vecino que haga actividades ilegales?



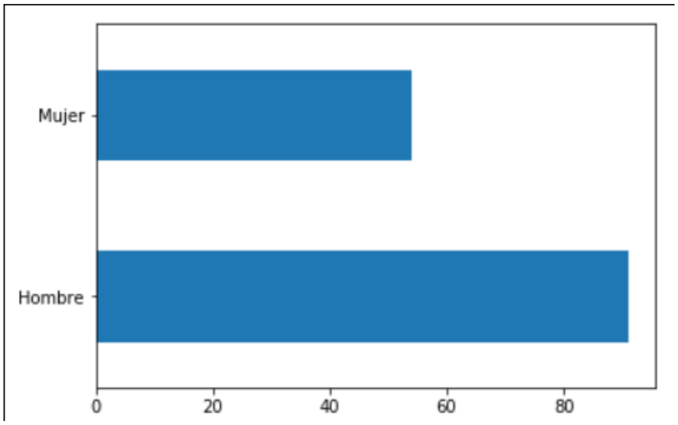
¿Tus padres beben?



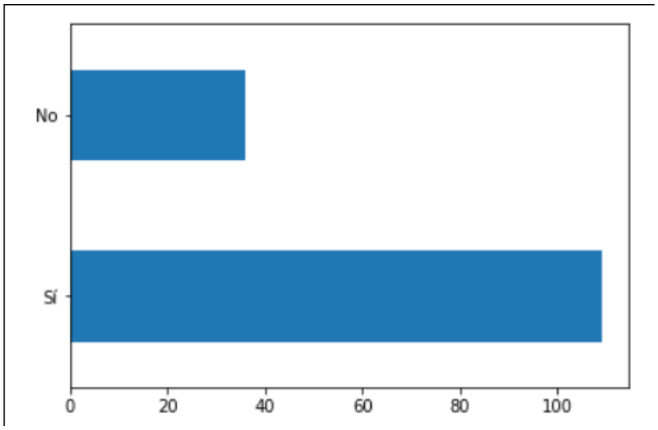
¿Vives con tus padres?



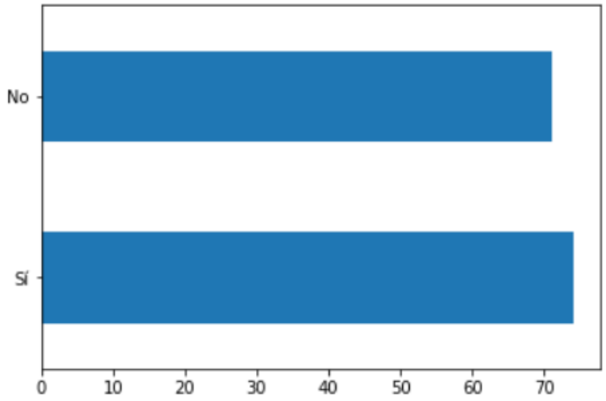
¿Cuál es tu sexo?



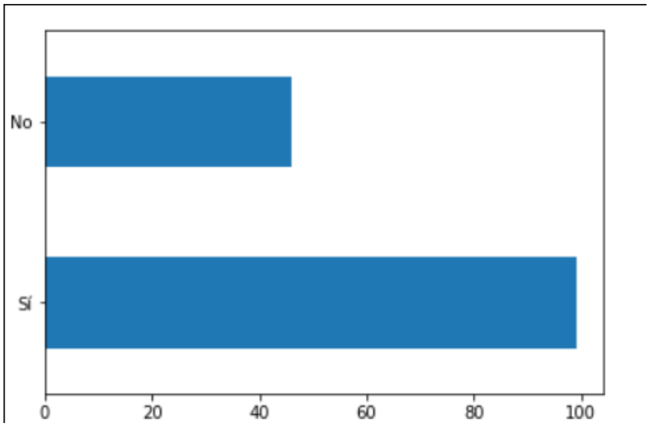
¿Te gustan las matemáticas?



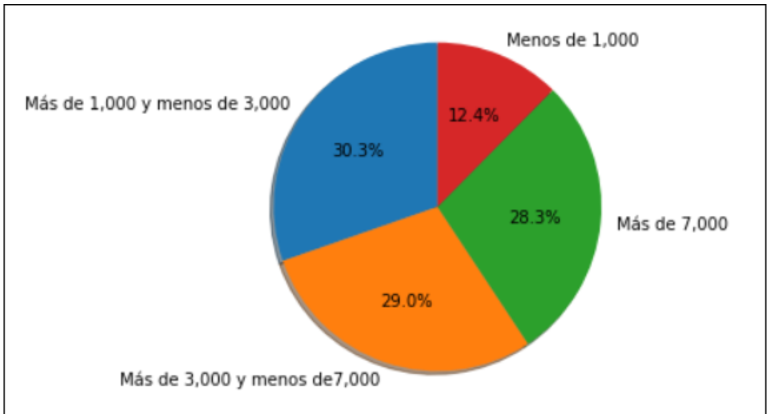
¿Debes alguna materia?



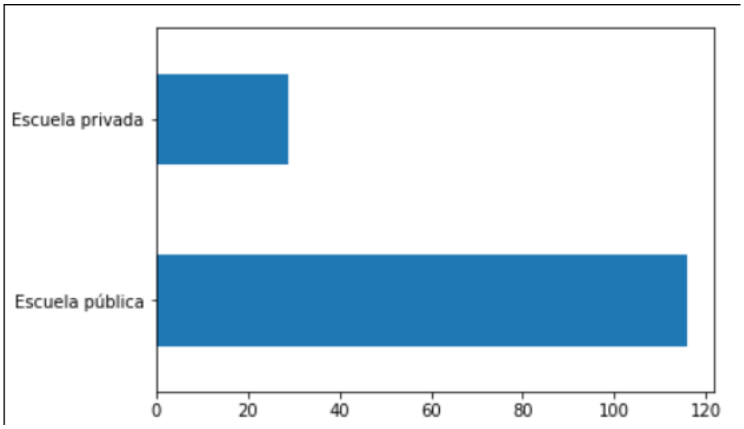
¿Haz presentado algún extraordinario?



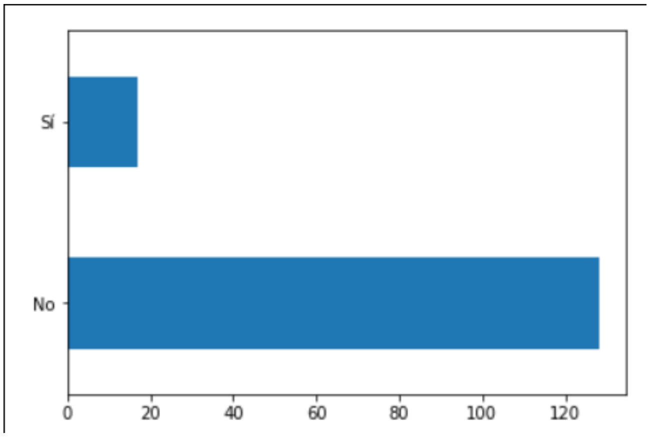
¿Cuales son tus ingresos mensuales (en pesos)?



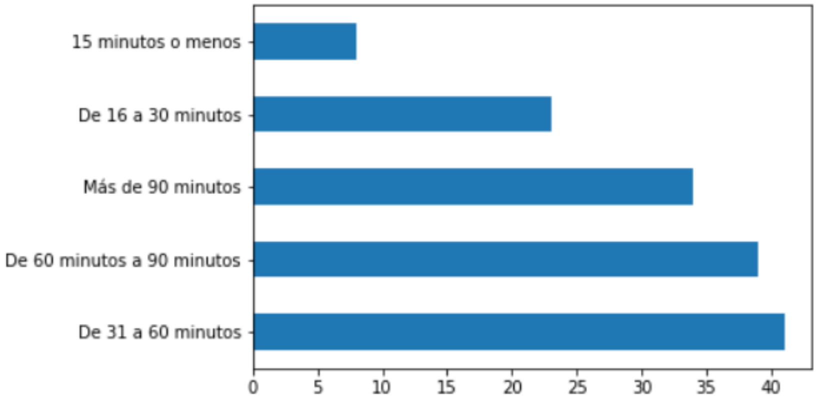
Tu nivel educativo anterior (preparatoria o bachillerato) era:



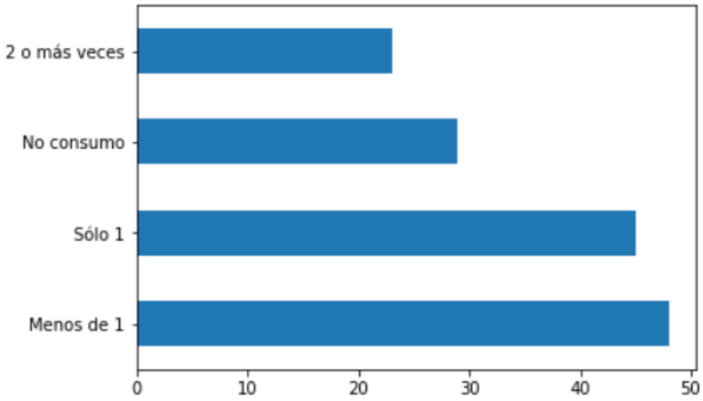
¿Eres fóraneo?



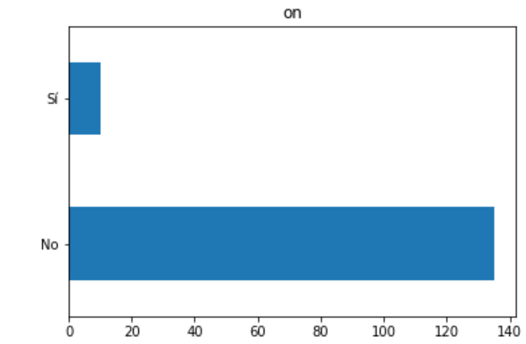
¿Cuánto tiempo realizas para llegar a la escuela?



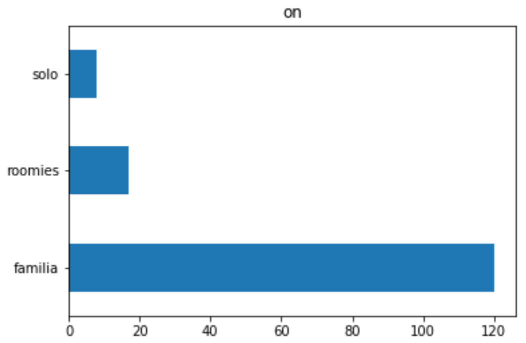
¿Cuántas veces a la semana consumes alcohol?



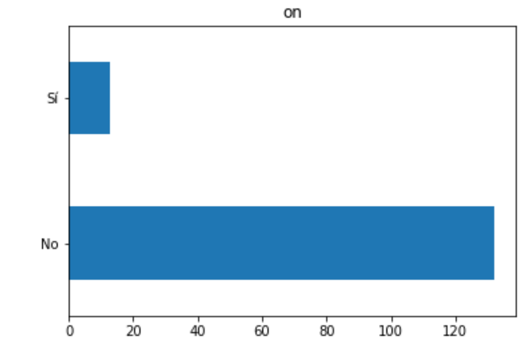
¿Perteneces a alguna comunidad indígena?



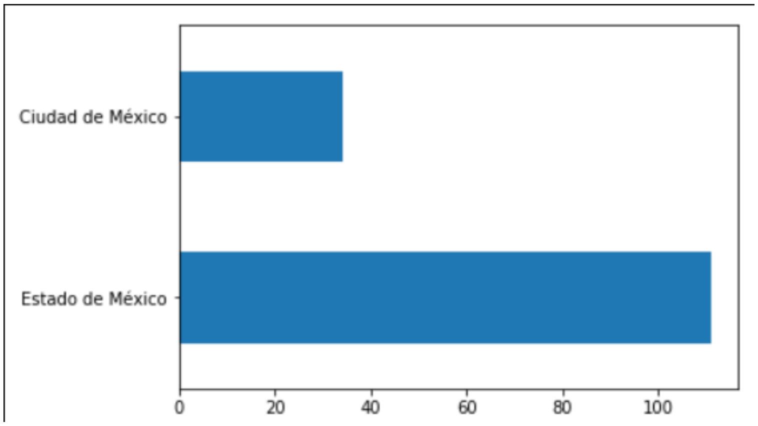
¿Vives sólo o con roomies o con tu familia?



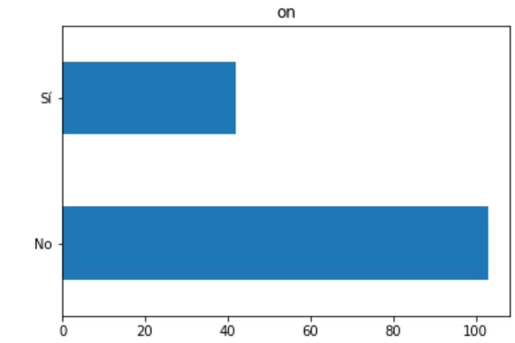
¿Padeces alguna enfermedad crónica?



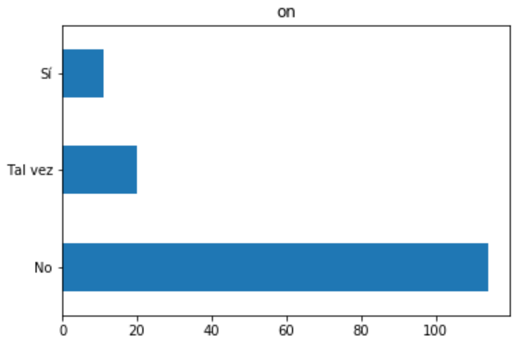
Vives en el:



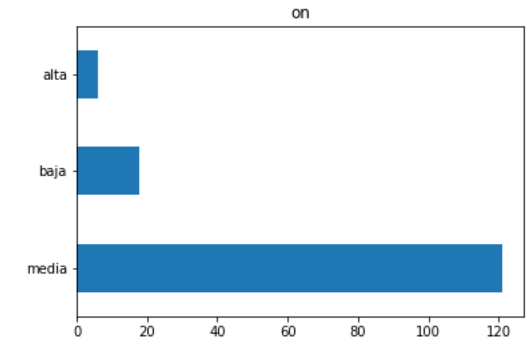
¿Tienes alguna beca?



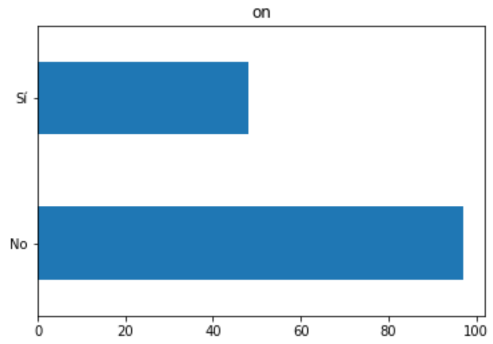
¿Consideras que consumes alcohol en forma desmedida?



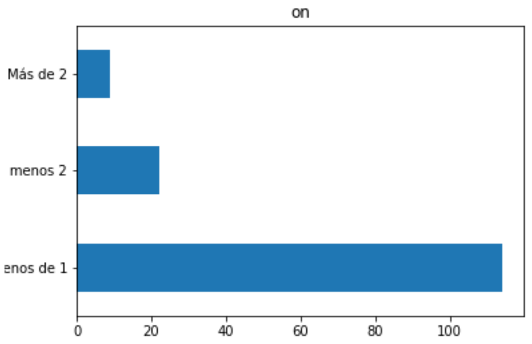
Me considero de clase social...



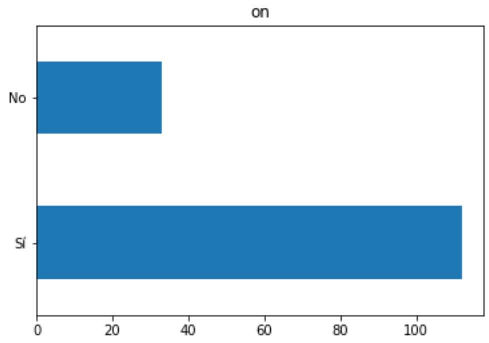
¿Tienes algún familiar con problema de abuso de sustancias?



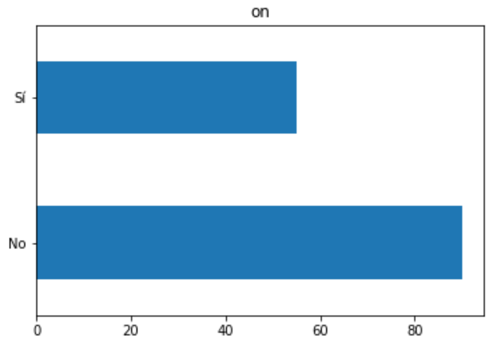
¿Cuántas veces en promedio sales de fiesta en una semana?



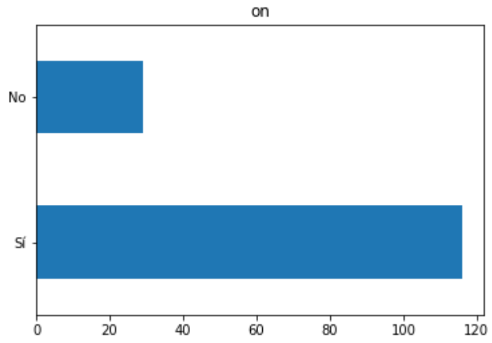
¿Estudias en la carrera que fue tu primera opción?



¿Haz considerado cambiar de carrera?



¿Crees que la escuela crea un ambiente de estrés constante?



7 Multivariate analysis

8 Results

9 Application strategies for generated knowledge

We have decided to split the solution into 4 stages:

- 1.- Reformulation of the instrument: For a project to have certain guarantees, it is necessary for experts in the subject to be involved, so we decided to ask for guidance in COESI for this task because they have experience and a more oriented formation to this type of needs.
- 2.- Use of the instrument: The coordination of the career will apply the test to new students.
- 3.- Identification: Based on the above, students with a high probability of suffering from the problem will be identified.
- 4.- Follow-up: The coordinator proposed a possible protocol, which consists of psychological support given at COESI followed up throughout the career, monitoring academic performance. If the student has economic school support, he will be asked for proof of expenses.

10 Conclusions

Throughout the elaboration of this brief study we faced several obstacles, which, with deeper knowledge we could have solved without problem. The most significant happened at the moment of studying the dataset, specifically the target, since this variable is very skewed towards no: 0, so, our model will predict quite easily, since it's enough to say that all "are not alcoholics" and the model will guess with a high probability.
Since this is not convenient, we have found that these types of datasets require a different treatment from the one we have given, however, our abilities do not yet have the potential to do so. In some way, the knowledge generated has been extensive and it was very convenient to have faced a dataset of this type at such an early stage in our training.

After all this process we obtained a model capable of detecting the situation when a student might need to intervene to neglect the school according to the patterns that it manifests.

in this way we can achieve that they do not fail so many students per semester and help them to release the pressure in a different way than alcohol

References

- 1.- [Diccionario de cáncer]. (s.f.). Recuperado 12 mayo, 2019, de *https : //www.cancer.gov/espanol/publicaciones/diccionario/def/alcoholismo*

- 2.- Organización Mundial de la Salud. Reporte mundial de salud 2013. Consultado en: https://www.who.int/whr/es/

- 3.- Córdova A, Muñoz O, Guarneros A, Rosales R, Camarena E. Instituto Mexicano del Seguro Social. Información 2001. En: Observatorio mexicano en tabaco, alcohol y otras drogas. Consejo Nacional contra las Adicciones 2002,pp 83-86

- 4.- Arenas A, Castillo G, López Alvarez ME. Instituto de Seguridad Social de los Trabajadores del Estado. Concentrado Nacional de Adicciones 2001. En: Observatorio mexicano en tabaco, alcohol y otras drogas. Consejo Nacional contra las Adicciones 2002,pp 87-90

- 5.- Kuri P, Alvarez C, Cravioto P, García E, Garlván F, Tapia C.R. Sistema Epidemiológico y Estadístico de las Defunciones. En: Observatorio mexicano en tabaco, alcohol y otras drogas. Consejo Nacional contra las Adicciones 2002, pp111-116

- 6.- Gomez G, Robles LJ. Mortalidad asociada a sustancias adictivas en cadáveres del servicio médico forense. Anuario de Investigación en Adicciones 2002; 3:44-51

- 7.- Encuesta Nacional de Adicciones 2008. Consejo Nacional contra las adicciones. México: Secretaría de salud.

- 8.- Encuesta Nacional de Adicciones 2002. Consejo Nacional contra las adicciones. México: Secretaría de salud.

- 9.- Campollo O, Martinez MD, Valencia JJ, Segura J. Drinking patterns and beverage preferences of liver cirrosis patients in México. Substance use and misuse 2001; 36: 387-398

- 10.- Medina Mora ME, Natera G, Borges G. Alcoholismo y abuso de bebidas alcohólicas. En: Observatorio mexicano en tabaco, alcohol y otras drogas. Consejo Nacional contra las adicciones 2002, pp 15-25.

- 11.- Martínez A, Alvarez AL. Estudio básico de comunidad objetivo (EBCO) de la ciudad de Guadalajara. Anuario de investigación de adicciones 2002; 3: 4-19

- 12.- Diaz Belmont A. Introducción. Alcoholismo. Beneficios y efectos deletéreos del etanol. Ed. Piensa. México 1997.

- 13.- Campollo O. El alcoholismo en México. En: Muñoz Espinoza L. Hepatología. McGraw Hill México, 2007, pg. 181-187

- 14.- Campollo O, Alvarez C, Sánchez H, Toro J. Patrón de alcoholismo en estudiantes de educación media superior. Rev de gastroenterol mex 2003; 68: 183-4

- 15.- Villatoro, J., Medina-Mora, M.E., Rojano, C., y cols. (2001). Consumo de Drogas, Alcohol y Tabaco en Estudiantes del Distrito Federal: medición otoño 2000. Reporte Global del Distrito Federal. INP-SEP. México.

- 16.- E. A. Weiss (1992). "Arthur Lee Samuel (1901-90)". IEEE Annals of the History of Computing. 14 (3): 55–69. doi:10.1109/85.150082.

- 17.- Kotsiantis, S. B., Zaharakis, I., and Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 160, 3-24.

- 18.-Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... and Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.

- 19.-Guttag, John V. (12 August 2016). Introduction to Computation and Programming Using Python: With Application to Understanding Data. MIT Press. ISBN 978-0- 262-52962-4.

- 20.- Garreta, R., and Moncecchi, G. (2013). Learning scikit-learn: machine learning in python. Packt Publishing Ltd.

- 21.- Amazon. (s.f.). Amazon Machine Learning - Guía del desarrollador. Recuperado 7 mayo, 2019, de $https : //docs.aws.amazon.com/es_es/machine-learning/latest/dg/machinelearning-dg.pdf\#types-of-ml-models$

- 22.- Unsupervised learning. (s.f.). Recuperado 7 mayo, 2019, de $https : //en.wikipedia.org/wiki/Unsupervised_learning$

- 23.- Semi-supervised learning. (s.f.). Recuperado 7 mayo, 2019, de $https : //en.wikipedia.org/wiki/Semi-supervised_learning$

- 24.- Supervised learning. (s.f.). Recuperado 7 mayo, 2019, de $https : //en.wikipedia.org/wiki/Supervised_learning$

- 25.- Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.