Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Departamento de Electrónica, Sistemas e Informática

Maestría en Informática Aplicada

PREDICTION OF DEPRESSIVE BEHAVIORS IN SOCIAL NETWORKS WITH NAVE BAYES ALGORITHM

TRABAJO RECEPCIONAL que para obtener el **GRADO** de **MAESTRO EN INFORMÁTICA APLICADA**

Presenta: **JULIO ENRIQUE PATIÑO VILLAGRANA**

Asesor: **MTRO. VICTOR HUGO MARTÍNEZ SANCHEZ**

Tlaquepaque, Jalisco. Julio, 2021.

Acknowledgments

I would like to acknowledge:

To Mtro. Victor Martinez, who suggested me the topic for this scholar journey and like to give a big thank you for the guidance on my work graduation degree material, all of the books, articles, and fresh ideas time to time, also in the knowledge transfer of the programming language necessary for this to become a reality.

The Dr. Ivan Villalon Turrubiates, for giving me the opportunity to be part of this area of science investigation in the Computer Science area.

The company HP Inc. my previous employee who gave me the chance to course this master’s degree, supporting me with almost the full cost of scholarship.

The Instituto Tecnológico y de Estudios Superiores de Occidente (ITESO) for the resources provided for the development of this research. Additionally, my coworkers in my current company, DXC Technologies.

Dedication

I would like to dedicate this work to:

My children for always be there in the most complicated moments of this learning journey and new knowledge and experiences, for always waiting for me with a big smile despite my absence, to my wife for the unconditional support, my sister for the encourage, and never lose the faith on me, despite the fatigue and other road blocks in the way, my admiration to all. Thank you very much….

Abstracto

Hoy en día, la depresión se ha convertido en una de las enfermedades más graves y, con mucha frecuencia, fatales para cada uno de sus pacientes en todo el mundo que padecen esta afección. Incluso para miles y miles de seres humanos que la padecen, ni siquiera conocen su existencia, o no saben diferenciar entre lo que sienten y lo que sufren por ello.

Actualmente con todas las nuevas tecnologías que nos acompañan existen algunas dedicadas a la investigación, la gestión rápida de datos, algoritmos inteligentes y algunos lenguajes de programación que hacen que los estudios científicos sean más fáciles y maleables en su procesamiento.

Uno de ellos es muy innovador por el tiempo que se ha empleado, especialmente para bases de datos y todo lo que tiene que ver con parámetros que nos ayudan a medir, indagar, comparar. Además de esto, también tenemos variable aleatoria (RV) o aprendizaje bayesiano, las posibilidades de un evento A pueden ser determinantes, también con el conocimiento de que A opta por una cierta peculiaridad que determina sus posibilidades. El teorema de Bayes (BT) comprende posibilidades inversas al teorema de la posibilidad total. El teorema de posibilidades totales hace inferencia sobre un evento B, a partir de los resultados de los eventos A.

Por su parte, Bayes calcula la posibilidad de A condicional a B. Existen numerosas aplicaciones de las mismas, pero suelen estar más dedicadas a redes convolucionales, clasificatorias de imágenes, detección de otras enfermedades en el tiempo, procesamiento del lenguaje natural y también sistemas de pronóstico financiero, en el caso de la depresión como enfermedad o condición médica, ciencia de datos y uno de los programas de código más poderosos Python, y algunas bibliotecas existentes como Keras ya sincronizado con TensorFlow como marco, podríamos hacer numerosos hallazgos en muchas áreas no estudiadas relacionadas con la depresión como reconocimiento de texto (TR).

Muchos de los hallazgos más recientes relacionados con la depresión son muy fáciles de analizar en las redes sociales, siendo los más utilizados por 7 de cada 10 personas en todo el mundo. Cada uno de ellos deja una huella de emociones, sentimientos, reacciones, fotografías, incluso textos.

Python es uno de los lenguajes de programación más usados y amigables en su procesamiento y construcción de código, por lo que estas investigaciones y nuevos descubrimientos suceden de manera continua y fructífera al mismo tiempo.

Abstract

Depression today has become one of the most serious and very often fatal diseases for each of its patients worldwide who have this condition. Even for thousands and thousands of human beings who suffer from it, they do not even know of their existence, or they do not know how to differentiate between what they feel and what they suffer is about it.

Currently with all the new technologies that accompany us there are some dedicated to research, fast data management, intelligent algorithms and some programming languages that make scientific studies easier and more malleable in their processing.

One of them is very innovative because of the time that has been used, especially for databases and everything that has to do with parameters that help us measure, inquire, compare. In addition to this, we also have random variable (RV) or Bayesian learning, the possibilities of an event A can be determining, also with the knowledge that that A opt a certain peculiarity that determines its possibilities. Bayes' theorem (BT) comprehends possibilities inverse to the total possibility's theorem. The total possibilities theorem makes inference about an event B, from the results of events A. For his part, Bayes calculates the possibility of A conditional on B.

There are numerous applications of them, but they are usually more dedicated to convolutional networks, classificatory of images, detection of other illness in time, natural language processing, and also finance forecast systems, In the case of the depression as illness or medical condition, data science and one of the most powerful code programs Python, and some existing libraries like Keras already synchronized with TensorFlow as a framework, we could make numerous findings in many unstudied areas related to depression as text recognition (TR).

Many of the most recent findings related to the depression are very easy to analyze on social media, in the most used by 7 out of 10 people worldwide. Each of them leaves a print of emotions, feelings, reactions, photographs, even texts.

Python is one of the most used and friendly programming languages in its processing and code construction, so that these investigations and new discoveries happen in a continuous and fruitful way at the same time.

Table of Contents

LIST OF FIGURES

LIST OF ACRONYMS AND ABBREVIATIONS

AT		Automated Techniques
BM		Bayesian Mode
BT		Bayes Theorem
CNA		Canadian National Alliance
CP		Conditional Possibility
CSV		Comma Separated Values
DA		Data Analytic
DF		Data Frame
DMC		Depression as Illness or Medical Condition
DMD		Diagnosis Mental Disease
FB		Face Book
FN		False Negative
FP		False Positive
FS		Fundamentals of Statistics
iJIM		International Journal of Interactive Mobile Technologies
N-DA		Dimensional Array
RV		Random Variable
SM		Social Media
SSO		Single Sign On
SQL		Structured Query Language
SND		Social Network Dataset
SNP		Sequence Number Prediction
TN		True Negative
TP		True Positive
TR		Text Recognition
VIH		Human Immunodeficiency Virus
X-OBJ		X-Object

1. INTRODUCTION

Depression is an mental disfunction that it's caused by several feelings, to mention some; feeling sad, and also most of cases begin with losing interest of regular activities that was part of the life as basics, Depression can be also named in several ways, but medically speaking depends a lot of the severity of the problem.

In life Depression it's very dangerous and silent illness, many recent investigations have been based on the study of text, which have a common good in humanity. In the case of depression, text recognition plays a very important role, especially when doing a study of emotions on social networks today worldwide known in any language, location, religion, cultures, some of them like Facebook (FB). Instagram, Twitter, among others, where you can achieve image recognition and make smart classification of them. [1]

The aim of this work is to build and maximize a self-educated model based on Python programing language, also with the use of Bayesian Model (BM), depression text posts on social media from active users around the globe, compare depressing comments, from regular and also suicide posts.

With this, deep and automated processing and learning can also be achieved, reference models that can self-educate and yield immediate results for the search for new trends, and the inappropriate use of online information, in this case of text, posts on social networks. Many times, the same people leave their mark online and most leave wording and different expressions that can be processed instantly regarding the study that they want to achieve.

1.1. Background

This important research it’s based on several lectures, articles on important educational and investigational researches previously made, some of them very famous articles referred to psychiatric disorders, in this case the DMC (Depression as Illness or Medical Condition) as mortal disease.

Most common definition of Anxiety is a feel of worrying, disquietude, or concern, typically about a near event or something with an unsure upshot, usually showed in text post’s as emotions of sadness, and loss of control.

Advance research considering the deep study of depressive disorder and linking the different areas of text recognition research now days needed to develop more comprehensive theoretical models one of the most used in data analysis the BM, and new ways to identify depression on the worldwide social media or countries with high volume of detected cases.

1.2. Justification

The use of information technologies, in this case the implementation of intelligent algorithms capable of detecting negative, suspect behaviors or behaviors related to depression, specially text notes and post in Social Media (SM), play an important role not only for the community that suffers more from this disease and the help that can be granted.

To direct and / or indirect relatives, or even to those directly affected, the search for new solutions in relation to this evil and others that over the years have meant great human losses worldwide.

Modern medicine and many more alternative solutions, both spiritual and separate alternatives, have become so famous but mostly poorly focused on the cure or remedy of the disease, but without instead this project seeks to be more proactive and detect the problem long before the remedy.

1.3. Problem

The main problem is depression and how the early detection of the disease can be carried out in any type of person who uses social networks. In the day-to-day life of each user of any social network, depressive behaviors are left behind, in this case, there are many post’s and unexpected comments that are classified as part of the disease issue.

The aim is to attack the problem by creating an automated model that is capable of self-feeding not only from existing users’ text but key depression words from those that the users of the social networks publish in their personal profiles. Get probability of suicide and depressive emotions posted on the media.

The disease is a serious issue worldwide and proactivity is the key to be able to help in some way to proactively detect some depressive behavior and not focused on the solution after the problem.

1.4. Objective

1.4.1. General Objective:

Study the main approaches used for the extraction and implementation of prediction models based on text content data published on social networks, with the target to identify depression and potential risk of suicide within the media, using Bayesian Model.

1.4.2. Specific Objective:

Design and implement an algorithm on a frame of reference, which is executed on a web platform, which allows the analysis of profiles in social networks for the detection of depressive text behaviors.

Design and implement a web platform which executes the analysis of public and private profiles on social networks and determines, with a degree of certainty greater than 85%, the relationship that exists in their texting found in the profile and the relationship with symptoms of depression.

1.5. Scientific or technologic share / innovation

This document has the intention to demonstrate the different ways social network users behave in different depressive disorders in comments, large phrases and text in general, apart from this, it’s to have a technological help to reduce and detect the disease, by the use of programing algorithms in python codes and some libraries and frameworks available previously instructed for the same ambience.

El Informador

10 min ·

Hay muchas caras en la pandemia y una de las más feas es la depresión y el estrés que provoca.

Un niño de 11 años en California murió después de dispararse mientras tomaba clases a distancia 😔

INFORMADOR.MX

Niño se dispara mientras tomaba clases en línea

El menor de 11 años fue encontrado en su domicilio en California, ...

Figure 1: Children shoots himself while taking classes online

[https://www.informador.mx/internacional/Nino-se-dispara-mientras-tomaba-clases-en-linea-20201203-0053.html]

2. STATE OF ART OR THE TECHNIQUE

The early detection of depression these days, has become one of the greatest challenges of our time, especially in the most vulnerable community, the youth, according to F. Sadeque, in their last article [2], they mention that 12% of all disabilities in people are attributed to this condition, with the help of data science and current computational models this condition can be detected in various ways, from the recognition of images and texts and some other modalities [3].

Throughout this journey that has been studied on predictive models, several projects and universities have been given the task of demonstrating that texts and images can be traced for the analysis of public health, its factors, diseases such as obesity and that represents graphically all over the world, a very recent example has been [4].

The geotagged text ages of the Instagram social network, which like some other social networks are reliable data sources for this type of diagnosis, especially diseases that are associated with the different lifestyles of people globally, to mention a few, (Obesity, drinking or smoking), the text content of a lot of posting that show some meanings are usually archived for this type of studies, as well as the text suicide notes.

Starting from the main theme of this research, about depression and social networks have been the best dumbbell where adolescents have been the best actuaries, not leaving other generations behind, according to the IJIM article [5].

In this recent article mentions that fatigue in social networks and the psychological well-being of network users, however, currently the empirical relationships between psychosocial well-being and fatigue of social networks are not known, said study also mentions that the Single Sign On (SSO) to evaluate regardless psychosocial well-being quantity, such as overwhelming media use and fear of missing out, trigger fatigue and, furthermore social media fatigue causes depression and regularly anxiety.

In the meantime, it has been shown that following this and some other predictive reference frameworks through the use of data science, machine learning, and granularizing much of this with some programming languages like Python, could identify patterns of behavior, texts related to this suffering and be constantly fed back in their constant use.

2.1. Quick glance of the Depression

Depression[6] as of now in the majority of the globe is a reality and hurts our society very badly, especially in our adolescents, it can appear in several signs, this can also be misinterpreted with being lonely and sad, with the feeling that everything it's wrong, it can also be confuse by the need for something or someone to feel better or happy, a few of very obvious symptoms.

Being sad all the time for no reason, no feel the desire of eating, either going out as a regular person will feel, not having interest in what you do not normally have in life, most cases, suffering of insomnia, feeling lazy, sleepy , feel everything it's so wrong, world don't deserve me, some cases suicidal thoughts, the good news are that there is plenty of solutions, depending of the level of the depression, and that now days and on many places around the world have been successfully treated, most cases only depend on medical treatments, this last may represent an expense way to treat the illness but with less suffering.

2.2. Depression cross cultures

There are several articles, all about depression and the different cultures who play the biggest number of cases of this illness, [7], clearly the probability of having depression relates more in quantity in the white community for some reason, while the black people suffer from a feeling or state of despair, they don't really understand what depression it's all about, depression it's less significant for them apparently because they are behind in resources to obtain medical assistance, neither physiological treatments.

Natives of the European Union culture besides the medical help they can get, they suffer more from loss of life sense, they can be rich but empty in the inside , this gave you a clue of why some countries have the biggest number of Depression in their communities, sadly poor countries in difference, the number it's not even accountable, meaning we can't even have a number to start with and compare with rich countries.

2.3. Depression and the different ages

Depression illness by no excuse can be underestimated in any of the victim ages, from the youngest age, and also the most youngest, baby born even to the senior community, each one expresses in some way or another, some of the very occasional symptoms in recent born may be that It is a lot of irritation, and also that they do not eat anything but liquids, somehow usually any young can consume instead of regular food.

There are many ways also this disease can be alert on the oldest victims, usually the most common feeling it's the feel of self-isolation, looking for always to be alone, they look for independence away from their relatives, even for their couple, the most worry out of this, is that they feel this is a normal feeling to be apart from everything and everywhere, every person and each age it's way different and each shows the different levels of depression, each age has its risk factors and psychiatrist most likely don't find rest or peace on babies and oldest people. [8].

2.4. Depression in different social status communities

Through the years[9], most of the population that suffered from depression was the one with the lowest income, today if we compare the current economic level, more patients with this condition are medicated but at a more than the regular people with regular lives medium-high level, according to many investigators and some records found. back then, they cited that the most regular patients were women, now the same level in percentage are also man, with the impression of having this first option as more communicative of feelings, the disease metrics no longer depend only of their gender but of high levels of lucidity, maybe people who do not have a good job and good pay they stop feeling depressed, or maybe they simply do not say it, or another option is that they do not make their illness public, or simply the way of living in their life is that way, they temped to live like this, they simply do not have time to make this public.

2.5. Relation of people depress and their mood favorite colors in life.

Coloring in life and people it’s something very usual, on clothes, on all of the things in life, what we use, what we see, every person have their own favorite, since we were children, colors are always everywhere, and have our favorites, weird how very young people already have a favorite color.

This includes the mood, black represents death, sadness, loneliness and so on, white represents peace, life, happiness and many more examples on how the colors play important role in our lives.

In a recent study made in Manchester European Union (MEU), it was demonstrated that people who suffer from anxiety, healthy people, and even who are depressed, their agony of depression and mood changes can be express in few colors.

Healthy people, based on random selection of items in a basket, they tempted to select the yellow colors.

Anxious people went more for gray, light dark, depressed people with no doubt selected the dark colors, and noticed too that their outfit was also with obscure selection of clothes and hair colors.

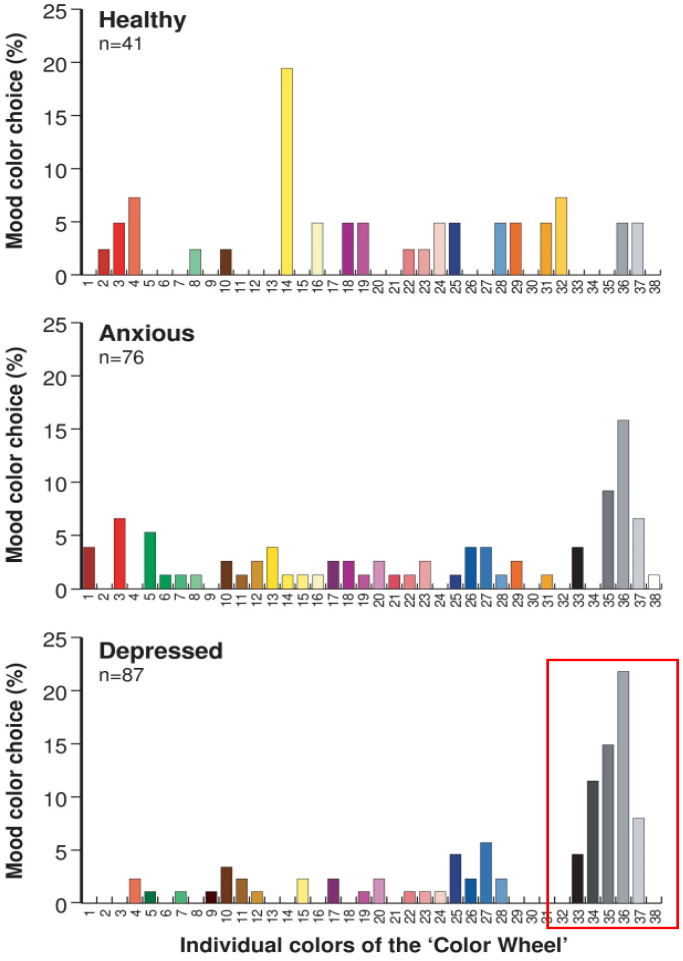


Figure 2: Individual colors of the 'Color Wheel'

[https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-10-12]

2.6. Depression and Addictions to Drugs, Alcohol.

The connection between addiction and depression, and the body mood connection, the fact that the quickest way to change your state of mind on people with depression, is to change your physiology.

Doing or abuse from drugs, what these drugs do, they affect the center in the brain having to do with pleasure, and they release a chemical called dopamine [10] which is the field of chemical and they make you feel high instantly.

For example, there are two basic groups of these chemicals:

1) First is called Stimulus you have cocaine, crack, methamphetamine, they get you up and you can imagine that a person with depression would be attracted to these drugs.

2) Downers the opiates, and alcohol, and to often some people cannabis, all these calm people down and of course people with anxiety would be likely to use those diapers medications, the results feel really good but in the long run first able to figure this out with any of these drugs are short acting, they wear off and then you're worse off than you were before, there is another problem as you take more and more of these drugs the brain gets used to them and you have a phenomenon called dependence for tolerance, in which you need more and more of the chemical to get the same high.

After all, you get dependent intolerant of the drugs and so that's why they're really not good in the long term.

People who were both depressed directions on one hand and are taking drugs and the other you have something called a dual diagnosis or co-occurring disorders the Canadian National Alliance (CNA) of mental illness says 80% [11] of people who suffer from addiction also suffer from a mood disorder.

2.7. Depression and the current medical techniques using technologies.

At present, many of the studies have been based on programs that analyze the history of different things, such as studies of the mind, electronic signals of the brain, some of the heart, among others, one of the closest to what this article refers to, is relative to electroencephalography.

But this is more related through the cure and not the on-time detection of depression, which gave more unreality of the risk this represent.

2.7. Detecting depression on social media with Bayes

Despite from percentage of diagnosing mental disease (DMD) have recovered over the past recent decades, several cases remain hidden, like no one notice either the affected person. Indicators related with mental disease are very obvious on different social media channels, example Twitter, Facebook, and plenty of web chats, and new automated techniques (AT) are increasing trough time to time, usually they are capable to identify depression and other type of mental disease.

Mental affected users have been spotted and ask to participate in random questionaries, this publish info in the media can be a clear diagnostic of them by detecting key words.

The early discovery of this techniques can easily define as super-efficient tools, can also find and identified ill users or red flag individuals in risk through the large volume of chats and posts on the media, receptive data and supervise info of social network, and in the near time may will see plenty of new techniques apart from what already exist.

Some of the usages of bayes in the text recognition are, email content, tags within search engines such as Google, Categorization of products and so on, to make Bayes model efficient, same as in this work, there are a prerequisites for BM Classifier, first of all we need to identify what is need as the mission of BM, calculate the file or files that will integrate the matrix on each class, after this the calculation of frequencies on each class, so the recall elementary probability it's expose and the tagging can start classifying key elements of the model.[12]

Believe it or not in this age of actual technological new era, people interactions are consider as gold, specially text notes, chats, blogs, etc., users get used to express more easily their emotions, by sharing their way of thinking in social media pages. The objective of several investigations and plenty of recent papers express and demonstrate the importance of this exiting investigations, the main purpose is to develop a data-analytic (DA) and establish model to interpretate, detect the depression of any user on the media.

Like in several other investigations, the proposed plan it's to model a data is collection from the participants in social media, posts of two popular social media sites: twitter and FB. Depression rankings of a participant has been detected based on the severity of the comment, how violent the message can be classified, by using profanity on the message, acting weird, by telling suicide notes, attacking others often, lonely feelings and so on. The tags, alerts, and data from the daily messages on the media can interoperate and demonstrate the damage that user it's presenting, what depression disorder symptoms are common of certain users, all this can be tagged and separated using Support Vector Separation and Naïve Bayes algorithm to notice depression more effectively. [13]

3. THEORIC/CONCEPTUAL FRAMEWORK

3.1. Bayes' theorem.

One of the most debated issues throughout history has been whether probability and statistics are better or worse than the Bayes theorem.

Known as Bayes' rule, it is what you get from the probability theory that links conditional probabilities, If A and B express two events, $P\,(A \mid B)$ expresses the conditional probability of A happening, given that B, The two conditional probabilities $P\,(A \mid B)$ and $P\,(B \mid A)$are usually different outcomes.

Bayes theorem gives a relation between $P\,(A \mid B)$ and $P\,(B \mid A)$ as the main formula.

One of the most important ways to use Bayes' theorem is to update or revise each strength based on the certainty in the light and related to the new later certainty.

Formally as a theorem, it is validated in many ways, and expressed in the same way.

As a formal theorem, Bayes' theorem is valid in all probability deductions. However, there is much debate regarding the fundamentals of statistics (FS), the frequentist and Bayesian expressions differ on the different modes and modes to which probabilities should be applied and replicated in applications.

Whereas frequentist probabilities fix probabilities to events in different order, according to their continuity of occurrence or subsets of groups as proportions of the total, Bayesians suggest probabilities to suggestions that are not true. [14]

3.2. Bayes’ Theorem Manifesto

Narrate the conditional and on the edge of possibility of stochastic events A and B as we can see on equation **(1)**:

$$P(A\,|\mathrm{B}) = \frac{P(A\mathrm{IB})\ P(A)}{\mathrm{P(B)}} \quad \textbf{(1)}$$

Every expression in Bayes’ theorem interpreted a unique name:

- P(A) is the before possibility or possibility of **A**. It is officially call ”prior” in the understanding that doesn’t get into any information about the next possibility B.
- P(A|B) is the conditional possibility of A, given B. It is also called the after possibility because it is derived from or depends upon the specified value of B.
- P(B|A) is the conditional possibility of B given A.
- P(B) is the before possibility of B, and act as a normal constant.

3.3. Bayes’ theorem & likelihood

It can be also be described and specified of likelihood: $\mathrm{P(A|B)} \ \propto\ \mathrm{L(A|B)\ P(A)}$

L(A|B) s the likelihood of A given unalterable B. The norm is then an immediate result out of the relation $\mathrm{P(B|A)} \ =\ \mathrm{L(A|B)}$ in several conditions the likelihood function L can be reproduce by a continues factor, so that it is equivalent to, but it doesn’t equally get the conditional possibility(CP) P.

This nomenclature gives the theorem the possibility and may be rephrase as below equation **(2).**

$$POSTERIOR = \frac{\text{LIKELIHOOD X PRIOR}}{\text{NORMALIZING CONSTANT}} \ \textbf{(2)}$$

In simple terms: the after possibility is relative to the product of the before possibility and the likelihood. Besides and plus this, the ratio $\mathrm{L(A|B)/P(B)}$sometimes it’s named the normal likelihood or normalized likelihood, so the theorem may also be rephrased as after = normalized likelihood × before.

3.4. Origin of conditional probabilities

To arise from the theorem, the initial statement of what it means conditional possibility. The probability of event A given event B is interpretated as equation **(3)**:

$$P = (\mathrm{A} \mid \mathrm{B}) = \frac{\mathrm{P(A \cap B)}}{\mathrm{P(B)}} \quad \textbf{(3)}$$

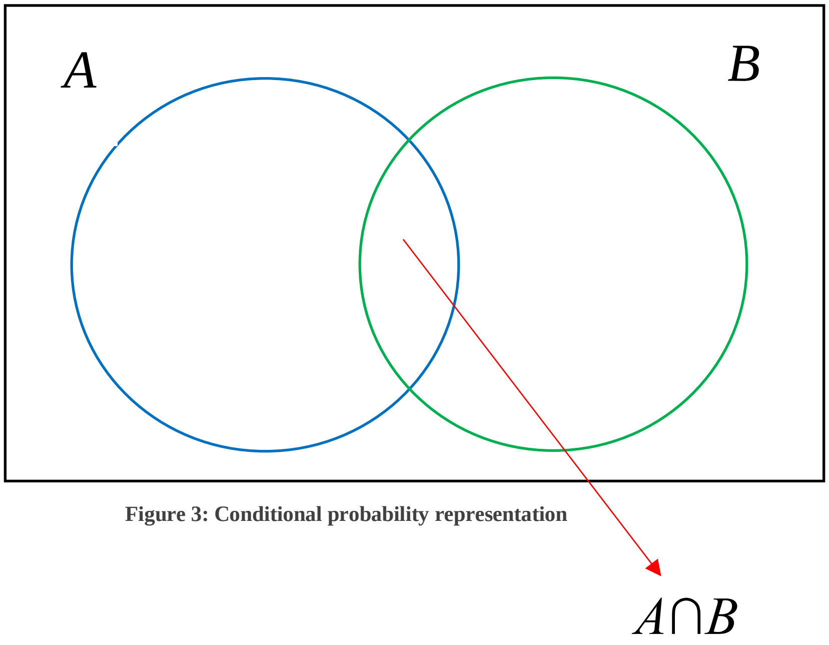


Figure 3: Conditional probability representation

Moreover, the probability of event B given event A is as below equation **(4):**

$$P = (\mathrm{B} \mid \mathrm{A}) = \frac{\mathrm{P(A \cap B)}}{\mathrm{P(A)}} \quad \textbf{(4)}$$

Reorganizing and mixing these two equations, find as below expression:

$$P\ (A \mid B) P\ (B) = \mathrm{P\ (A\ \cap B)} = P(B \mid A)\ P\ (A) \quad \textbf{(5)}$$

This lemma is most likely named as the result rule for possibility. Braking apart both angles by P(B), assuming that it is non-zero, we get as result Bayes' theorem as shown on equation (6):

$$P\ (\mathrm{A} \mid \mathrm{B}) = \frac{\mathrm{P(B \mid A)\ P\ (A)}}{\mathrm{P(B)}} \quad \textbf{(6)}$$

3.5. Alternate forms of Bayes’ theorem

Bayes’ theorem is frequently spruce by noting as the following equation (7):

$$P(B) = P(A \setminus B) + P(AC \setminus B) = P(B|A)P(A) + P(B|AC)P(AC) \quad \textbf{(7)}$$

where AC is the correlative event of A (frequently called ”not A”).

Therefore, the theorem can be restated as equation **(8)**:

$$\mathrm{P(A|B)} = \frac{\mathrm{P(B|A)\ P(A)}}{P(B|A)P(A)+P(B|AC)P(AC)} \quad \textbf{(8)}$$

More typically, where Ai term a partition of the occurrence space, any Ai in the separation. equation **(9)**:

$$\mathrm{P(Ai|B)} = \frac{P(B|Ai)P(Ai)}{\sum P(B|Aj\ )\ P(Aj\ )} \quad \textbf{(9)}$$

3.6. Bayes’ model and likelihood

Bayes’ theorem additionally can also be transcribed methodically in terms of a likelihood ratio and odds O as equation **(10).**

$$\mathrm{O(A|B)} = \mathrm{O(A)} \cdot \_\mathrm{(A|B)} \quad \textbf{(10)}$$

Where O(A|B) = $\frac{\mathrm{P(A|B)}}{P(AC|B)}$ are the odds of A given B,

And O(A) = $\frac{\mathrm{P(A)}}{P(AC)}$ are the likelihood of A by itself, while $\Lambda\mathrm{(A|B)} = \frac{\mathrm{L\ (A|B)}}{P(AC|B)} = \frac{\mathrm{P\ (B\ |\ A)}}{P(B\ |AC)}$ **(11)**

is the likelihood ratio as equation **(11).**

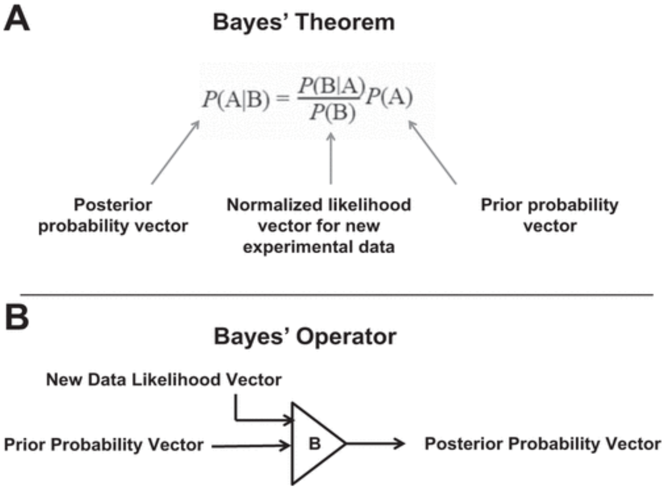

A
Bayes' Theorem
$P(A|B) = \frac{P(B|A)}{P(B)} P(A)$
Posterior probability vector
Normalized likelihood vector for new experimental data
Prior probability vector
B
Bayes' Operator
New Data Likelihood Vector
Prior Probability Vector
B
Posterior Probability Vector

Figure 3: Bayes Theorem Figure

[https://journals.physiology.org/doi/full/10.1152/physiolgenomics.00120.2016]

3.7. Bayes' theorem for probability densities

Also, there is a different version of Bayes' theorem for endless allocation. It is slightly harder to arise from because possibility densities, rigorously articulation, are not possibility, Bayes' theorem has to be proven by a limit method. as equation **(12)** $\mathrm{f}\,(\mathrm{x} \mid \mathrm{y}) = \frac{f(x,y)}{\mathrm{f(y)}} = \frac{f(y,x)\, f\,(x)}{\mathrm{f(y)}}$ **(12).**

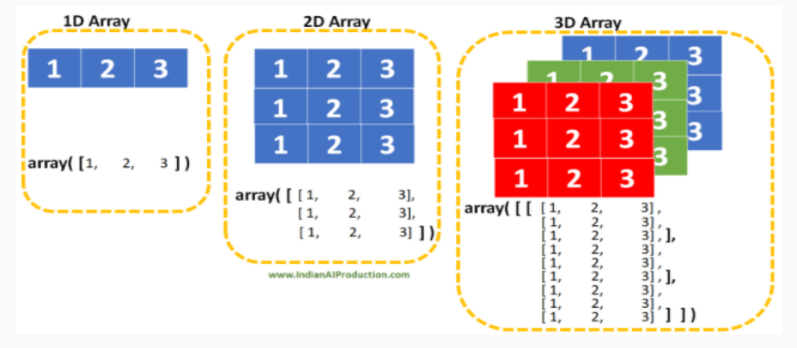
And there is an equivalent term of the law of total possibility $\mathrm{f}\,(\mathrm{x} \mid \mathrm{y}) = c^2 \frac{f(y|x)\, f(x}{\int_{\_}^{\infty} \infty\ \mathrm{f(f(y|x)\, f(x)\, dx)}}$

f(x, y) is the collective allocation of X and Y, f(x—y) is the after allocation of X given Y=y, f(y—x) = L(x—y) is (as a function of x) the likelihood function of X given Y=y, and f(x) and f(y) are the marginal allocations of X and Y singly, with f(x) being the before allocation of X. Here we have fulfilled in a customary misuse of notation, on the other side f for each one of these conditions, despite each one is really a nonidentical role; the functions are characterized by the singular names of their dispute.

3.8. Libraries quick glance

3.8.1. NumPy

NumPy is the central packet for research computer science in language programing specially in Python programing. This package has multidimension array object, several deduced items (such as cover arrays and matrix development see (Figure 3), also variety of patterns for quick direction on arrays, this includes mathematical, logical, shape math, sorting, selecting, I/O, discrete calculations and build codes.



```
array( [1,    2,    3 ])
```

```
array( [ [ 1,    2,      3],
         [ 1,    2,      3],
         [ 1,    2,      3] ])
```

```
array( [ [  [1,    2,      3] ,
            [1,    2,      3] ,
            [1,    2,      3] , ],
            [1,    2,      3] ,
            [1,    2,      3] ,
            [1,    2,      3] , ],
            [1,    2,      3] ,
            [1,    2,      3] ,
            [1,    2,      3] ] ])
```

Figure 4: Representation of NumPy ndarray (multidimensional array)

[https://indianaiproduction.com/python-numpy-array/]

NumPy packet, is the ND array item, this wraps the n-dimention of uniform array data styles, within several functions and options of operations, being the best option of compilation code for performance. There are several important differences between NumPy arrays within standardization of Python process flows:

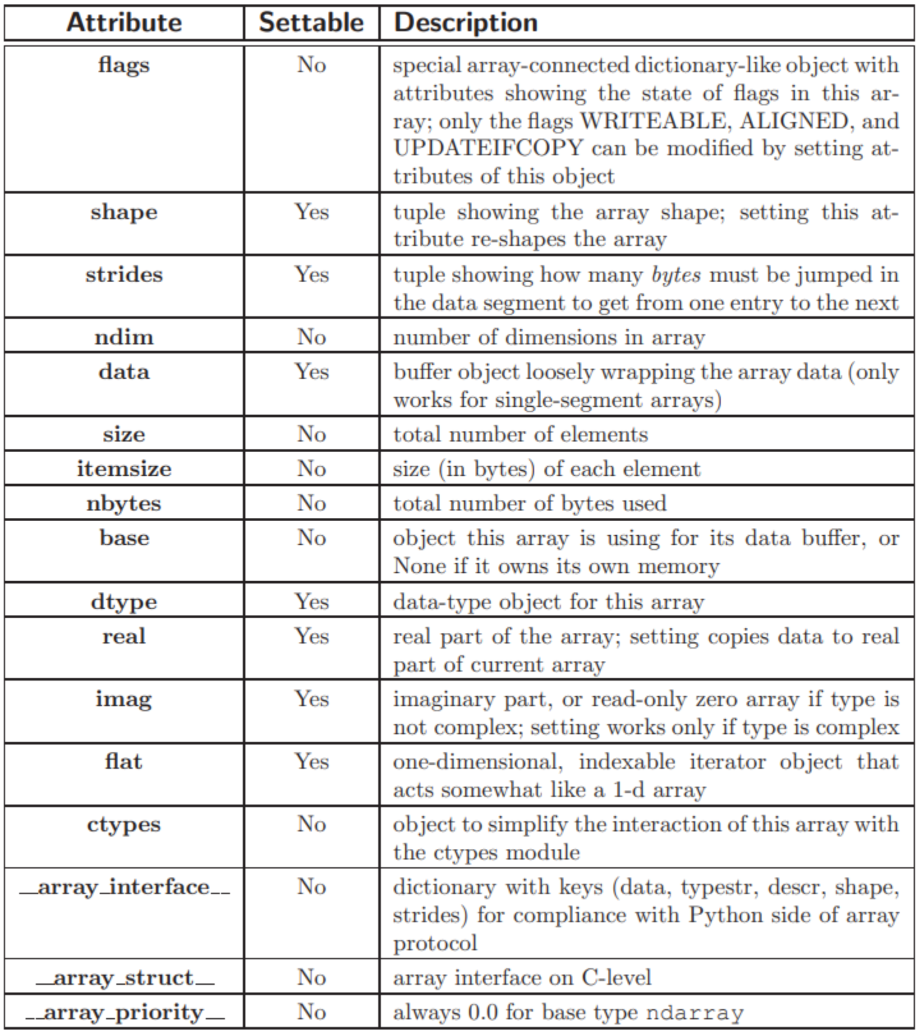
Arrays within NumPy are already fixed in dimensions at the time they are first created, in contrast Python transmitted as list (can get big with very dynamic movements). Swapping how big is the ND array, it will start from creating a brand-new array and get rid of the initial. Components in a NumPy array are basic requirements to the equal data style and will occupied the equal space in the memory. The exclusion: Only one can be part of the items in Python and NumPy, all can be as different as it wish in size.

NumPy arrays make things easier, complex math problems and operations performance, and so many other different types of functions on huge number of data. Usually, this type of process and operations are taking off from the own operation, with more and more efficiency using the less of programing as possible.

Important topics in size and the flow of information have very fast speed in processing and they are enormously important in science of the new era of computers. [15]

3.8.2. NDarray Attributes

Array characteristic bounce back flow of info that is innate to the array itself. typically, getting into the same array spot so properties allows you to retrieve, usually the set innate characteristic of the array lacking and setting completely new array. The disguised characteristics of the arrays and the own structure of an array can only get reset a few out of the complete lists of arrays in memory, Table 1 below shows all the characteristics and the meaning of each plus the description.



Attribute	Settable	Description
flags	No	special array-connected dictionary-like object with attributes showing the state of flags in this array; only the flags WRITEABLE, ALIGNED, and UPDATEIFCOPY can be modified by setting attributes of this object
shape	Yes	tuple showing the array shape; setting this attribute re-shapes the array
strides	Yes	tuple showing how many *bytes* must be jumped in the data segment to get from one entry to the next
ndim	No	number of dimensions in array
data	Yes	buffer object loosely wrapping the array data (only works for single-segment arrays)
size	No	total number of elements
itemsize	No	size (in bytes) of each element
nbytes	No	total number of bytes used
base	No	object this array is using for its data buffer, or None if it owns its own memory
dtype	Yes	data-type object for this array
real	Yes	real part of the array; setting copies data to real part of current array
imag	Yes	imaginary part, or read-only zero array if type is not complex; setting works only if type is complex
flat	Yes	one-dimensional, indexable iterator object that acts somewhat like a 1-d array
ctypes	No	object to simplify the interaction of this array with the ctypes module
__array_interface__	No	dictionary with keys (data, typestr, descr, shape, strides) for compliance with Python side of array protocol
__array_struct__	No	array interface on C-level
__array_priority__	No	always 0.0 for base type `ndarray`

Figure 5: Attributes of the ndarray.

[https://numpy.org/doc/stable/reference/generated/numpy.ndarray.html#numpy.ndarray]

3.8.3. Array indexing

Overpowering array indexing was a very significant part introduced by numarray, and as a result a significant piece of the proactivity of NumPy. Particularly, the wish to pick a random component located on its own location in the spot of the confusion matrix and understanding to a certain spot was desired.

Two different types of guiding in the index area handy using the X[obj] syntax: basic slicing, and advanced indexing. For the representation of this syntax given below, X is the array to-be-sliced and obj is the selection object. Moreover, define N ≡X.ndim. These two methods of slicing have nonidentical behavior and are triggered depending on obj. Appending additional functionalism yet remaining congruent with old uses of slicing complexing the rules in small proportion. [16]

3.8.4. Pandas: a python data analysis library

Pandas gives special-level data assemblage and applications strategically build to work
with composition or tabular data fast, easy, and expressive at all levels. Since the launched of this library back in 2010, it has been a great key player in conjunction with Python and enable to be an overpowering and profitable data analytics environment.

The main objects in panda's library that will be used in often for data analytics are the Data Frame (DF),
a tabular, column-mapped data sets with both of them row and column labels or tags, and the
Series, a one-dimension tagged array object. [17]

In conjunction of another library NumPy with the malleable data manipulation mix of possibilities of excel documents, in many formats, in example CSV files, besides this very friendly environment with relational databases (as example SQL). Make this easy to reshape, wedge and slice, roll up and dice contents, perform collections of enormous amounts of data, and made selections of subsets of data.

3.8.5. Colum's, Variables, Splits, Normalization in Pandas

Datasets in general can have several columns and can contain values instead of variables, special characters, and so on, depending on the source of the information, in this work the datasets used were specially scramble of content, several movements were need it, with the help of Pandas the work was very fruitfully, the initial work it's to have one column fixed, when presenting data in a table, but for data analytics, this content need to be rework and shape, so Python can minimize the time utilize on the data initial examination, Plenty of times the column content in a data set may have multiple variables.

This content format is commonly viewed in Social Network Datasets (SND), when working with Facebook for example.

One of the uncomplicated forms of identifying whether multiple units are represented in a table is by detecting in each of the rows and look for repeated values from one to another. Since the Facebook content specially text it's made by several users around the globe, and in many languages, English should be the main for this type of work.

3.9. Accuracy

Dealing with categorization problems there is always a temp to predict a binary conclusion. How you can predict if a person in the bank who is asking for a home credit loan? Will this have a risk of fraud? Will this person end with the loan until the end? The overall of number of false predictions is the ratio that we care most at the end, classifying falsely, or positive, falsely negative, or positive.

The formula of this prediction it’s represented as follows:

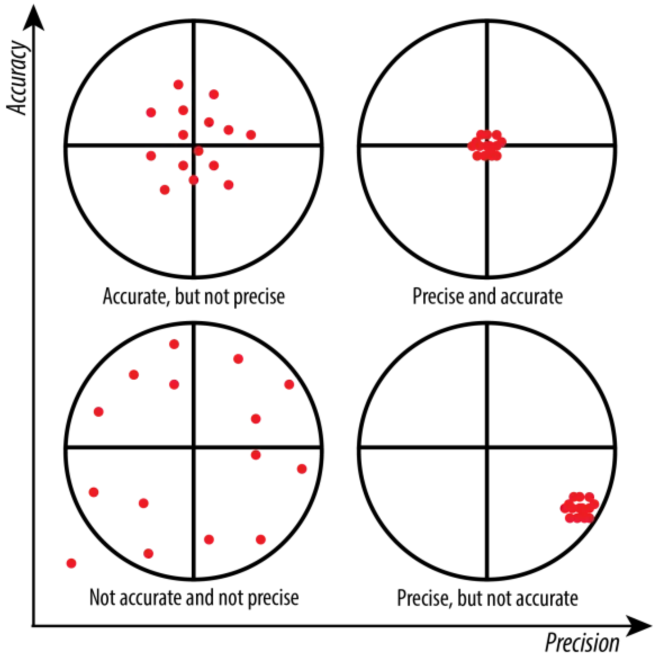

Accuracy
Accurate, but not precise
Precise and accurate
Not accurate and not precise
Precise, but not accurate
Precision

Figure 6: Accuracy and Pression representation

[https://wp.stolaf.edu/it/gis-precision-accuracy/]

$$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

3.10. Precision and Recall

A lot of content can be found within this topic, a head of time, we need to have a quick glance into what is Type I and Type II mistakes. These two conditions that are not matchless to complex operations in classifications, they are more like important, also related very close to statistical theory testing.

Type I Error: False positive (rejection of a true null hypothesis)

Type II Error: False negative (non-rejection of a false null hypothesis)

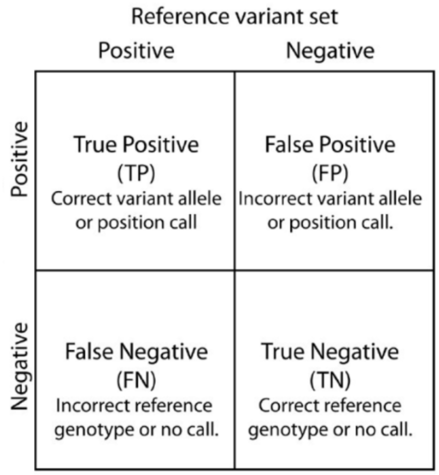
So, with that in mind we can define precision as the percentage of relevant results, in the meantime recall is distinguish as the proportion of pertinent outcomes from the correct classification of this model running smoothly. [18]

3.11. True Positives, Negative, False Positives, Negative

A TP is a result in which the model correctly predicts the positive class. Similarly, a TN is a result in which the model correctly predicts the N class.

A FP is the outcome in which the model incorrectly predicts the P class. And a FN is a result where the model incorrectly predicts the negative class, few examples, alive or not alive, VIH / not VIH, Loan fraud/not fraud, virus/not virus, etc. this is not only the result of happening or not happening hence it can be also, horse or cow, dry and wet, female or male, only one can be consider as positive or negative, The output really depends from the topic it's in place to be studied, there is no good P, or bad N option. [19]

TP, FP, FN, and TN are basically characterized functions based by the relation in the middle of alternatives called by the S.N.P calling the algorithm and known dissimilarity between the reference point and the demo point that was previously analyzed.



	Reference variant set	
	Positive	Negative
Positive	True Positive (TP) Correct variant allele or position call	False Positive (FP) Incorrect variant allele or position call.
Negative	False Negative (FN) Incorrect reference genotype or no call.	True Negative (TN) Correct reference genotype or no call.

Figure 7:Confusion Matrix with TP, FP, FN, TN

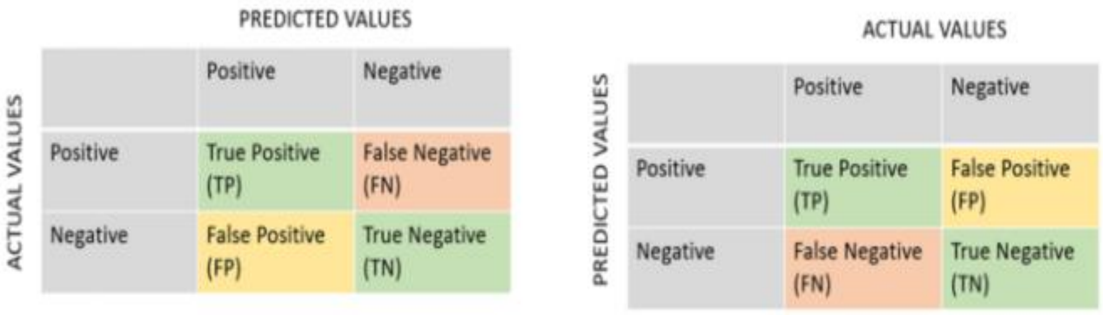
[https://www.researchgate.net/figure/Contingency-table-True-Positive-False-Positive-False-Negative-and-True-Negatives-are_fig5_280535795]

3.12. Confusion Matrix

What is the Confusion Matrix? It’s basically a performance estimation of machine learning categorization where a single problem can be represented in several ways, one insight can become two or more classes, is really valuable for measuring Accuracy, Recall, Specificity and Precision.

It can be deduced that distributions of any achievement indicator computed in the confusion matrix let you measure the variability of a sign and to rate the significance of a spot difference between two achievement indicators. The values or indicators of a confusion matrix are remarks upcoming from a multiform, in the case of Bayesian road where the not known parameters of the multiform parameters in which the unknown parameters of the multinomial likelihood function, they are called themselves and presumed to be produce from an unplanned vector.

Example below represent a simple confusion matrix, the location of the predicted values and actual values moves from the position of False negative (FN) and False positive (FP) but True positive (TP) and True negative (TN) keeps the position in the same spot of the matrix.



		PREDICTED VALUES	
		Positive	Negative
ACTUAL VALUES	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

		ACTUAL VALUES	
		Positive	Negative
PREDICTED VALUES	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Figure 8: Confusion Matrix

https://towardsdatascience.com/baffling-concept-of-true-positive-and-true-negative-bffbc340f107

Speaking of Normalization, an important point it’s to define the content, it’s truly necessary to have a normalize data content, otherwise the predicted values can show some inconsistencies, as you can see on the below example on the left side the entered number it’s a bit confused since the calculation goes from 0 to 1 only, as much the after 0 gets close to the 1 number, means your prediction it’s close to a good result, now if you see the right Confusion Matrix, the survived result it’s 0.92 which brings a great example of an efficient result was made after normalization.

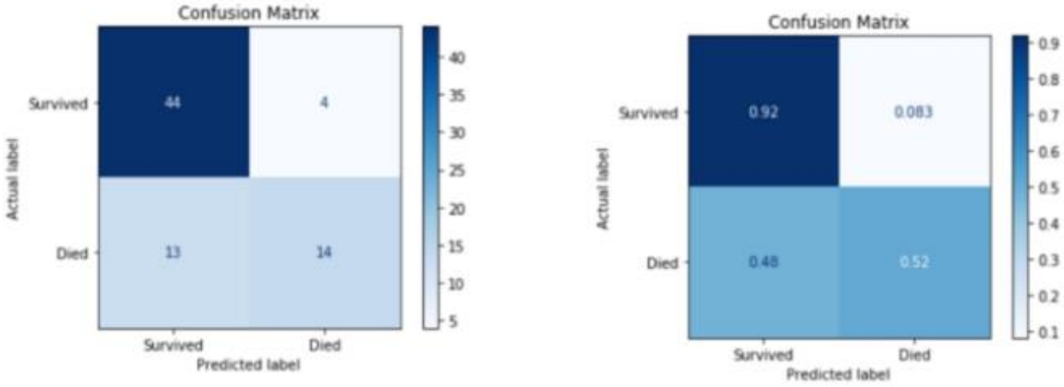


Figure 9: Confusion Matrix (2)

[https://towardsdatascience.com/baffling-concept-of-true-positive-and-true-negative-bffbc340f107]

4. DEVELOPMENT AND METODOLOGY

4.1. Requirements

System/SW Requirement	**Description: This table represent all of the requirements need in order to develop the proper function of Bayes Model.**
SRS-1	Must look for several datasets for demo
SRS-2	Most be a popular social media dataset such as Facebook in example
SRS-3	Format must be CSV File
SRS-4	Columns from the file must be a few
SRS-5	Must have at most 3 columns per sheet in CSV File
SRS-6	Text Comments must be filtered
SRS-7	Columns must have description header
SRS-8	List of depressed comments must be at least 15 thousand
SRS-9	List of suicide comments must be at least 15 thousand
SRS-10	List of regular comments must be at least 15 thousand
SRS-11	Pandas Library must be available
SRS-12	Pandas Library must be imported
SRS-13	NumPy Library must be available
SRS-14	NumPy Library must be imported
SRS-15	Non-ascii characters shall be removed before start working with the model
SRS-16	Non-alpha-numeric characters shall be removed before start working with the model
SRS-17	All text should be in lower-case format
SRS-18	Messages with key unusual words shall be remove from the data set
SRS-19	Sentences must be less than 24 words
SRS-20	Sentences messages shall include tags classifications
SRS-21	Non suicide tags shall be part of the classificatory
SRS-22	Suicide tags shall be part of the classificatory
SRS-23	Instance Creation must be part of the Count Vectorizer
SRS-24	Message Column with legend TRAIN shall be part of the final clasificator
SRS-25	Message Column with legend TEST shall be part of the final clasificator
SRS-26	Wrapper function must be part of the clasificator code
SRS-27	Likelihood function must be part of the clasificator code
SRS-28	Confusion Matrix shall be used for final results demonstration

4.2. Dataset: Nearby Social Network - All Posts

First with the impression that this dataset was going to be useful for this project, since the description was obvious, total of 48.6 Gigas of content and over 1M lines, unfortunately the material inside the dataset it’s completely coded and no text notes were found apart from inexplicable numbers, the information inside it’s completely unusual. https://www.kaggle.com/brianhamachek/nearby-social-network-all-posts sample of content below.

4.3. Dataset: Textism in total collected authentic chats and posts.

This collection of information was taken from used millennial interactions in different social media applications, this data aims to focus on different targets, one of them was the influence the young users have among a lot of others, for the way they express themselves, slangs used, and so other acronymous filtered on the media, the unfortunate detail about the content inside of this collection of data is the images instead of text. https://data.mendeley.com/datasets/z4r3bvwsbb/2.

4.4. Dataset: Suicide posts Posted on Social Media.

This data set was originally perfect for this research, by downloading the file there was a lot of comments with not sense of anything, comments without tags, and the worst that there were only about 500 lines with don't met the criteria for our Bayesian model we were trying to deploy.

Besides that, half the content had several special characters in between, which will make the work even worst of cleaning, at the end the decision was made, leave this option.

https://www.kaggle.com/mohanedmashaly/suicide-notes?select=test.csv.

4.5. Selected Dataset: Suicide and Depression Detection.

This dataset was selected for this investigation, since has data about depression for over 10 years in Reddit which is a social news aggregation, web content rating, and discussion website, and it claims to be "the front-page of the internet" as its moniker, recently including livestream content through Reddit Public, big differentiator with famous social media apps.

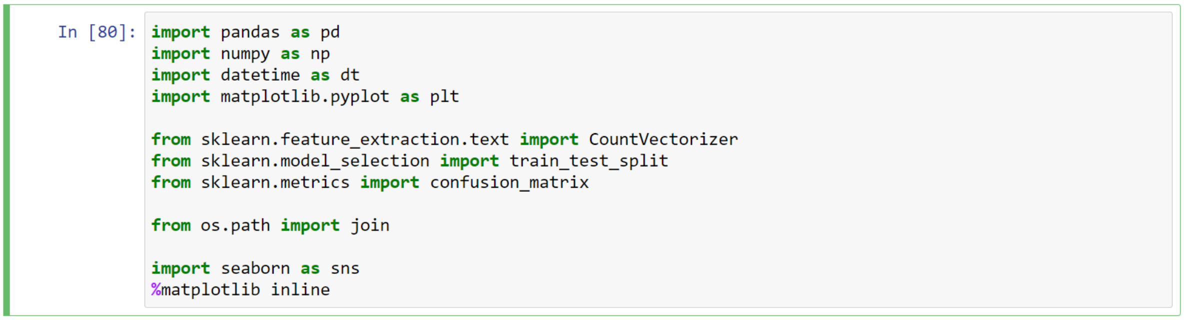
The content has suicide, disappointed comments separated and selected from others, the intention of having this detail it's to easily educate the algorithm with real people way of thinking.

Over 12 thousand lines of comments good and bad, the result will represent unique and detail expressions of many different users on the Reddit media, and can be used to identify on other social media apps. https://www.kaggle.com/nikhileswarkomati/suicide-watch/version/2:

4.5. The process

Importing libraries, adding the Count Vectorizer to mention it is one of the greatest tools from the (scikit-learn) library in Python. This one it’s used to remodel text in a vector on the principal of the frequency (count) of all the text content in a formal text. This is one of the best uses of this function avoids several manual works and (for using in further analysis of text).

A library is an external package that's created usually by a third party or some other Python developer and it's made widely available for the public use, an example of where you can find a lot of libraries such as NumPy, Pandas, GitHub [20] and some others for example, they provide prewritten sections of code and we can actually import them and then, have all of this pre written code available to anyone, once the library it’s imported, you can have access to anything within the whole content, and use it to simplify things a lot, getting random numbers or doing clots, Python environment enables the friendly use of code like in this project.

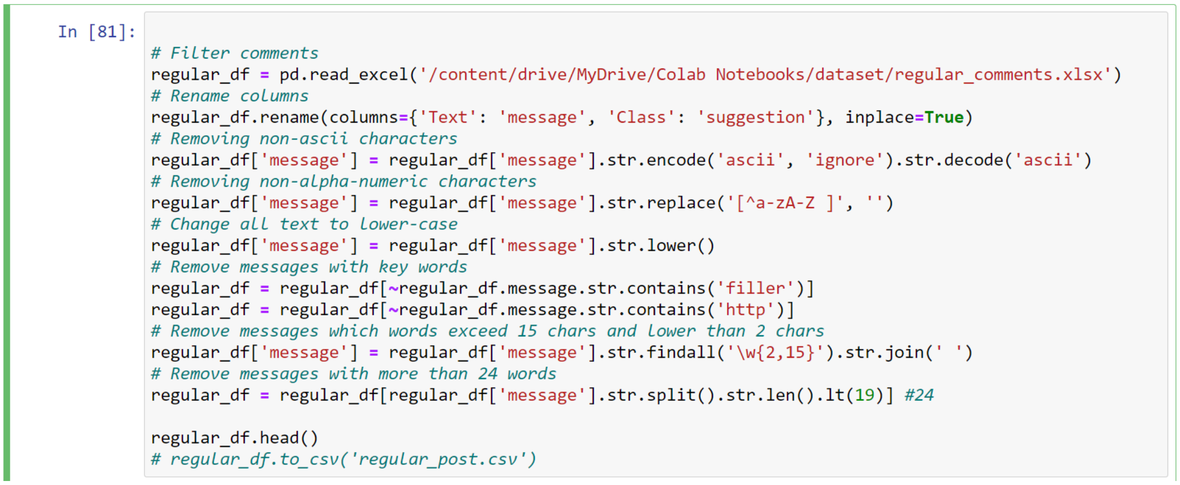


```
In [80]: import pandas as pd
         import numpy as np
         import datetime as dt
         import matplotlib.pyplot as plt

         from sklearn.feature_extraction.text import CountVectorizer
         from sklearn.model_selection import train_test_split
         from sklearn.metrics import confusion_matrix

         from os.path import join

         import seaborn as sns
         %matplotlib inline
```

Figure 10: Import of libraries

Cleaning the content it’s one of the most important steps on the process, first of all the content needs to be located, looking at the Filtering the comments, renaming the columns, cleaning the non-characters (ascii), removing all the non-alpha characters, all moved to lower case, removing also key words with no meaning, also removing words with more than 15 characters and last but not least removing also sentences with more than 24 words to make final model more efficient with the result.

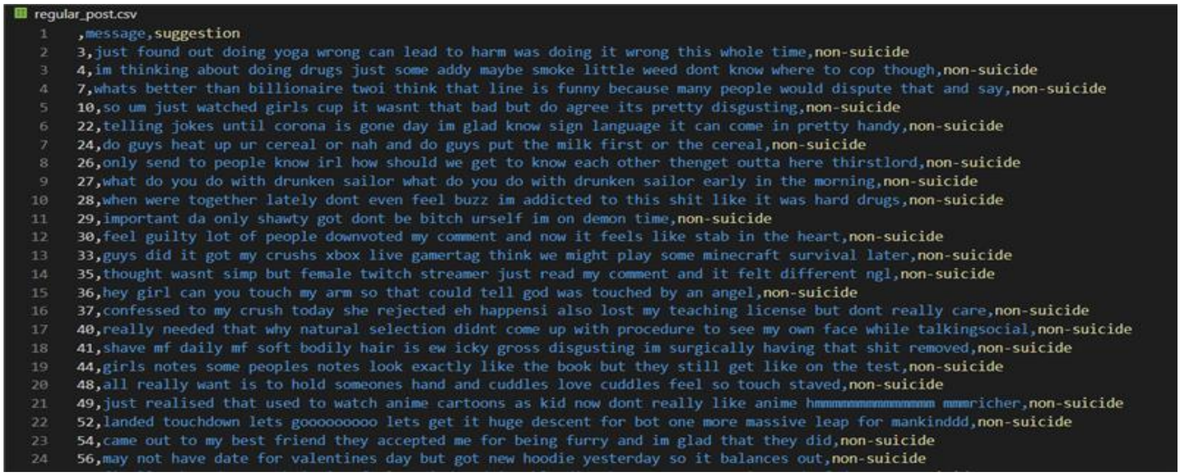
At the end the final result will appear as CSV file which will be the final set of data for analysis, this will integrate short sentences as figure 12.



```
In [81]:
# Filter comments
regular_df = pd.read_excel('/content/drive/MyDrive/Colab Notebooks/dataset/regular_comments.xlsx')
# Rename columns
regular_df.rename(columns={'Text': 'message', 'Class': 'suggestion'}, inplace=True)
# Removing non-ascii characters
regular_df['message'] = regular_df['message'].str.encode('ascii', 'ignore').str.decode('ascii')
# Removing non-alpha-numeric characters
regular_df['message'] = regular_df['message'].str.replace('[^a-zA-Z ]', '')
# Change all text to lower-case
regular_df['message'] = regular_df['message'].str.lower()
# Remove messages with key words
regular_df = regular_df[~regular_df.message.str.contains('filler')]
regular_df = regular_df[~regular_df.message.str.contains('http')]
# Remove messages which words exceed 15 chars and lower than 2 chars
regular_df['message'] = regular_df['message'].str.findall('\w{2,15}').str.join(' ')
# Remove messages with more than 24 words
regular_df = regular_df[regular_df['message'].str.split().str.len().lt(19)] #24

regular_df.head()
# regular_df.to_csv('regular_post.csv')
```

Figure 11: Cleaning the content

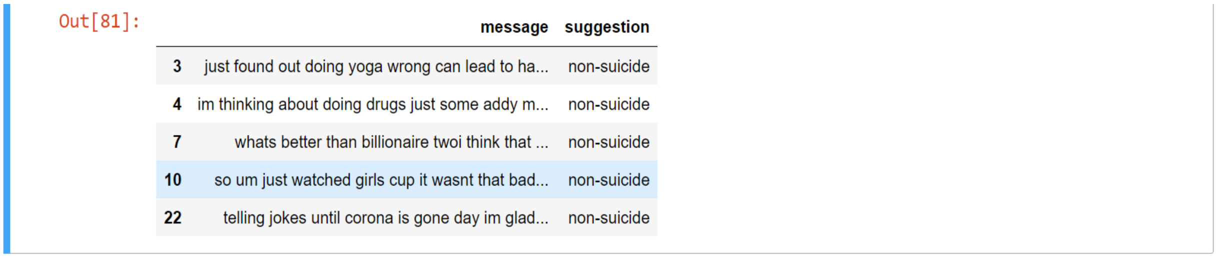
Every sentence must be checked and approved by the function given by the code, respect the parameters we will use, in the case of example below, this is a small extraction of the CSV file, as notice all are classified within a badge, non-suicide refers to the comment, if have bad words, bad or negative expressions, the rule will remain as if the code finds a sad, mad, desperate comments with certain words, the model will classify the sentence or comment as non, or suicide note.



```
,message,suggestion
3,just found out doing yoga wrong can lead to harm was doing it wrong this whole time,non-suicide
4,im thinking about doing drugs just some addy maybe smoke little weed dont know where to cop though,non-suicide
7,whats better than billionaire twoi think that line is funny because many people would dispute that and say,non-suicide
10,so um just watched girls cup it wasnt that bad but do agree its pretty disgusting,non-suicide
22,telling jokes until corona is gone day im glad know sign language it can come in pretty handy,non-suicide
24,do guys heat up ur cereal or nah and do guys put the milk first or the cereal,non-suicide
26,only send to people know irl how should we get to know each other thenget outta here thirstlord,non-suicide
27,what do you do with drunken sailor what do you do with drunken sailor early in the morning,non-suicide
28,when were together lately dont even feel buzz im addicted to this shit like it was hard drugs,non-suicide
29,important da only shawty got dont be bitch urself im on demon time,non-suicide
30,feel guilty lot of people downvoted my comment and now it feels like stab in the heart,non-suicide
33,guys did it got my crushs xbox live gamertag think we might play some minecraft survival later,non-suicide
35,thought wasnt simp but female twitch streamer just read my comment and it felt different ngl,non-suicide
36,hey girl can you touch my arm so that could tell god was touched by an angel,non-suicide
37,confessed to my crush today she rejected eh happensi also lost my teaching license but dont really care,non-suicide
40,really needed that why natural selection didnt come up with procedure to see my own face while talkingsocial,non-suicide
41,shave mf daily mf soft bodily hair is ew icky gross disgusting im surgically having that shit removed,non-suicide
44,girls notes some peoples notes look exactly like the book but they still get like on the test,non-suicide
48,all really want is to hold someones hand and cuddles love cuddles feel so touch staved,non-suicide
49,just realised that used to watch anime cartoons as kid now dont really like anime hmmmmmmmmmmmmmmm mmmricher,non-suicide
52,landed touchdown lets goooooooooo lets get it huge descent for bot one more massive leap for mankinddd,non-suicide
54,came out to my best friend they accepted me for being furry and im glad that they did,non-suicide
56,may not have date for valentines day but got new hoodie yesterday so it balances out,non-suicide
```

Figure 12: Cleaning the content as small sentences

On the figure below number 5, it is after the code was executed, you can see the message into one column and right after the suggestion, as regular comment, suicide, or non-suicide risk of depression illness detection.

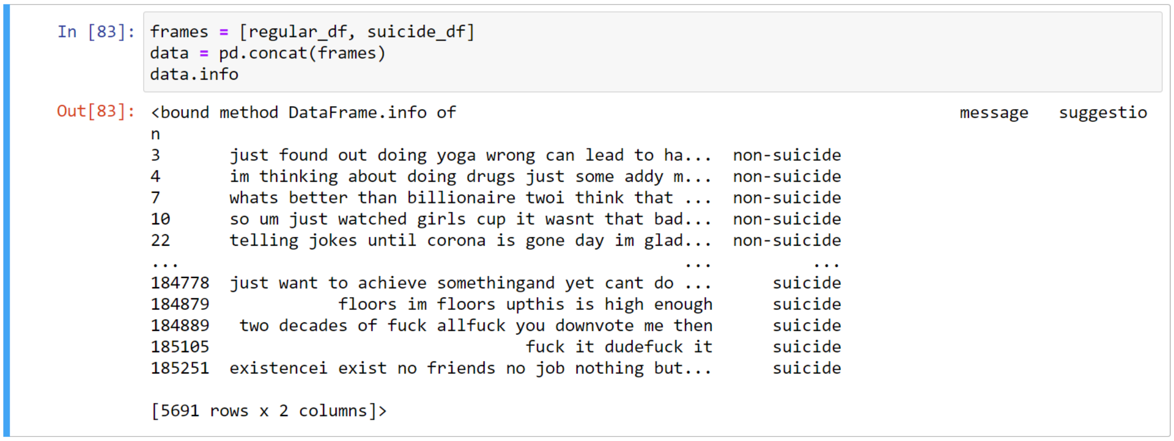


Out[81]:

	message	suggestion
3	just found out doing yoga wrong can lead to ha...	non-suicide
4	im thinking about doing drugs just some addy m...	non-suicide
7	whats better than billionaire twoi think that ...	non-suicide
10	so um just watched girls cup it wasnt that bad...	non-suicide
22	telling jokes until corona is gone day im glad...	non-suicide

Figure 13:The sample of result after cleaning

Creating the final dataset to work with, at the end with less you get more, only two columns are key in the dataset, instead of having several, the inside content can be easily granulated, in terms of Python the code it's not that complicated in difference of other programing languages, very few lines coded with spectacular results. In the journey of study, the statistical variations in characteristic of various items, this work model is capable to accomplish an efficient text search for a specified target item of object, same as categorize target and depression words, meaning, hate, fear, self-injuries, and so on.

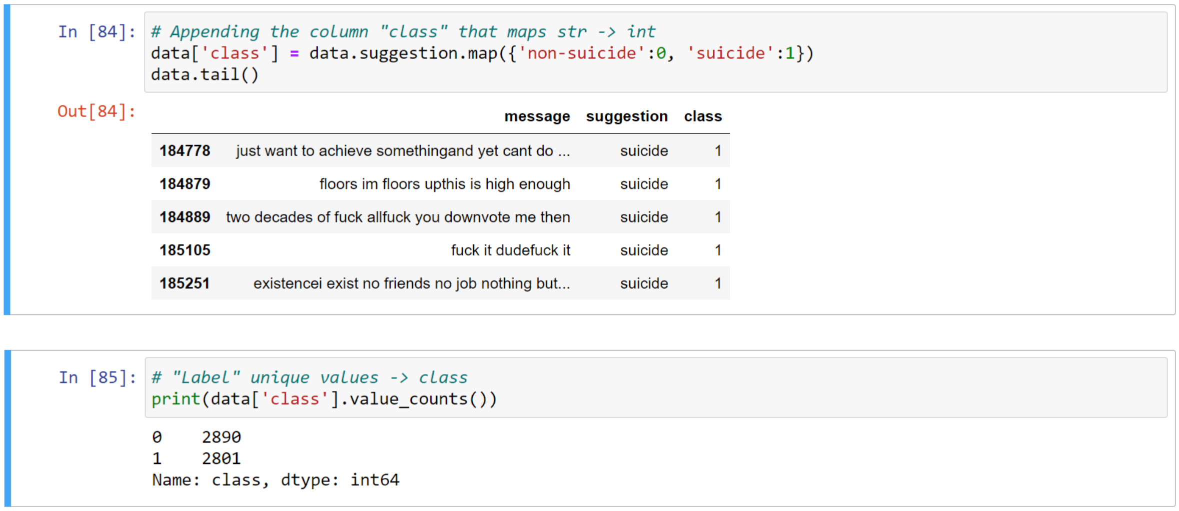


```
In [83]: frames = [regular_df, suicide_df]
         data = pd.concat(frames)
         data.info

Out[83]: <bound method DataFrame.info of                                                    message    suggestio
         n
         3       just found out doing yoga wrong can lead to ha...  non-suicide
         4       im thinking about doing drugs just some addy m...  non-suicide
         7       whats better than billionaire twoi think that ...  non-suicide
         10      so um just watched girls cup it wasnt that bad...  non-suicide
         22      telling jokes until corona is gone day im glad...  non-suicide
         ...                                                   ...          ...
         184778  just want to achieve somethingand yet cant do ...      suicide
         184879             floors im floors upthis is high enough      suicide
         184889   two decades of fuck allfuck you downvote me then      suicide
         185105                               fuck it dudefuck it      suicide
         185251  existencei exist no friends no job nothing but...      suicide

         [5691 rows x 2 columns]>
```

Figure 14: Data set result sample

Loading and getting things prepare for the model, separating within to indicators, 0 for non-suicide comments, and 1 for suicide comments,

Meantime we can use incidences to calculate possibilities of occurrence for unqualified attributes, we can't use the equal route for endless attributes. In place of, we first need to compute the significant of and variance for x in this case (0) in each class and then calculate (1) using this formula:



```
In [84]: # Appending the column "class" that maps str -> int
         data['class'] = data.suggestion.map({'non-suicide':0, 'suicide':1})
         data.tail()
```

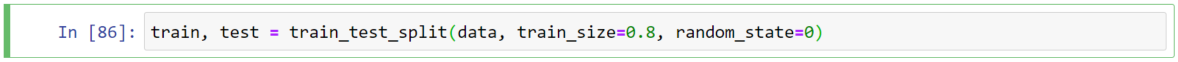
Out[84]:

	message	suggestion	class
184778	just want to achieve somethingand yet cant do ...	suicide	1
184879	floors im floors upthis is high enough	suicide	1
184889	two decades of fuck allfuck you downvote me then	suicide	1
185105	fuck it dudefuck it	suicide	1
185251	existencei exist no friends no job nothing but...	suicide	1

```
In [85]: # "Label" unique values -> class
         print(data['class'].value_counts())

         0    2890
         1    2801
         Name: class, dtype: int64
```

Figure 15: Sample of result after loading and classifying model

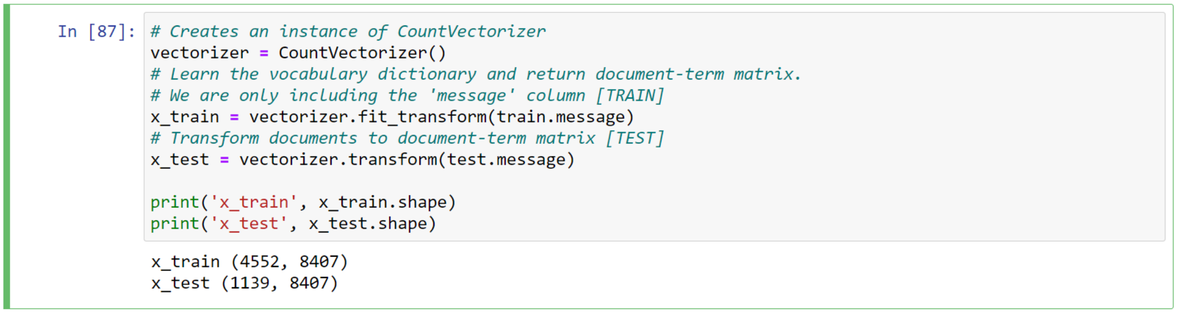
The use of this function, splitting datasets with the Sklearn train_test_split, we splitted the dataset so the 0.8 = 80% to train and the rest 0.2 = 20% to validate the model, this quick function that enclose input validation and output next and operation to input data into a unit call to divide (and have an option for subsampling) data in a one-liner.



```
In [86]: train, test = train_test_split(data, train_size=0.8, random_state=0)
```

Figure 16: Split arrays function

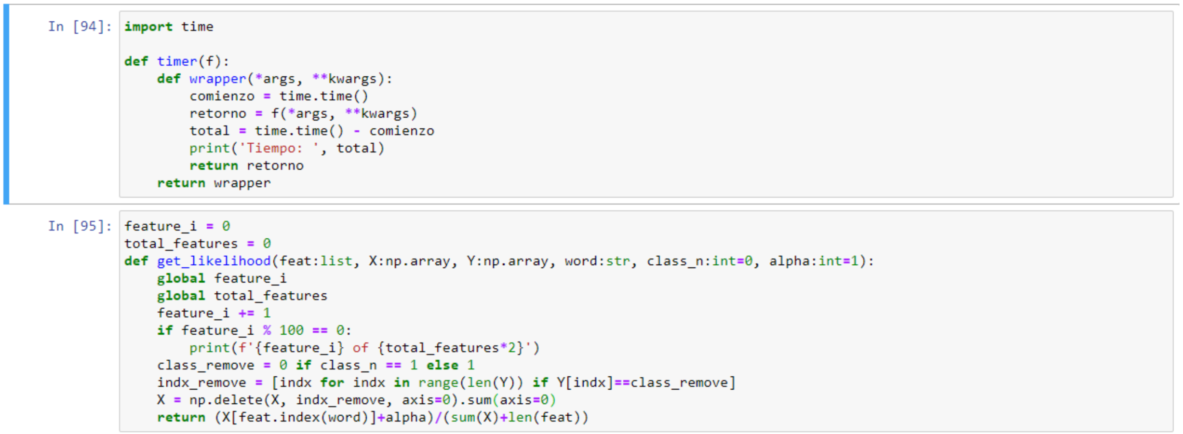
CountVectorizer the instance was created, after the vocabulary dictionary was learn the result it's returns as document-term matrix, another column was inserted as [TRAIN] and [TEST], so the 80% and 20% result can be classified. This function basically converts a collection of data set text files, to a matrix of token total, the execution of this process produces a rare characterization of the counts using CountVectorizer, if you don't add a prior-library or dictionary and you don't use the convert that does this particular process then the feature of numbers will be the same to the vocabulary size found by pre-analyzing the data like raw distribution.



```
In [87]: # Creates an instance of CountVectorizer
         vectorizer = CountVectorizer()
         # Learn the vocabulary dictionary and return document-term matrix.
         # We are only including the 'message' column [TRAIN]
         x_train = vectorizer.fit_transform(train.message)
         # Transform documents to document-term matrix [TEST]
         x_test = vectorizer.transform(test.message)

         print('x_train', x_train.shape)
         print('x_test', x_test.shape)

         x_train (4552, 8407)
         x_test (1139, 8407)
```

Figure 17: Train and Test representation

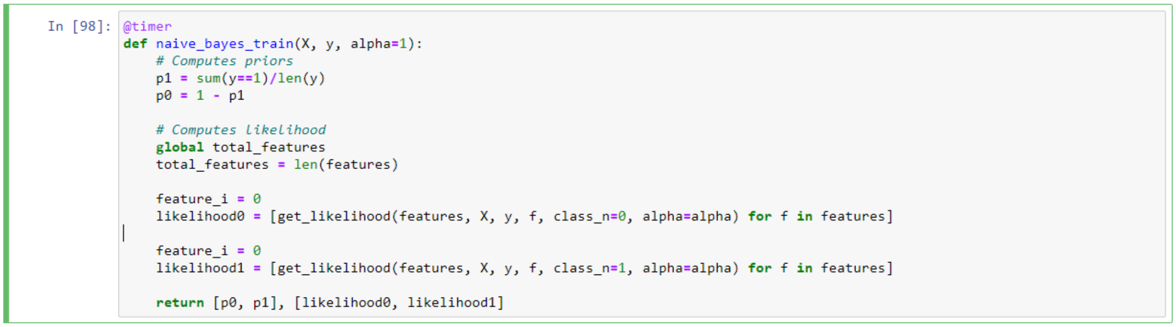
In picture 12 you can observe a Wrapper it's basically what measure the time that one particular function takes to get execute, it's auxiliar function in simple words, hence likelihood it's used to get into NumPy Arrays with a single or multiple feature lists, now speaking about wrappers in python and the applications are also established or known as **decorators** [21] and this are very robust and valuable tool in Python since it gives programmers the benefit of modify the behavior of function or class. Decorators let us wrap any other functions in order to stretch-out the conduct of the wrapped function, without doing any permanent changes, decorators also are known as self-efficient constructors, functions are reserved as the argument into other function and then called inside the wrapper process.



```
In [94]: import time

         def timer(f):
             def wrapper(*args, **kwargs):
                 comienzo = time.time()
                 retorno = f(*args, **kwargs)
                 total = time.time() - comienzo
                 print('Tiempo: ', total)
                 return retorno
             return wrapper

In [95]: feature_i = 0
         total_features = 0
         def get_likelihood(feat:list, X:np.array, Y:np.array, word:str, class_n:int=0, alpha:int=1):
             global feature_i
             global total_features
             feature_i += 1
             if feature_i % 100 == 0:
                 print(f'{feature_i} of {total_features*2}')
             class_remove = 0 if class_n == 1 else 1
             indx_remove = [indx for indx in range(len(Y)) if Y[indx]==class_remove]
             X = np.delete(X, indx_remove, axis=0).sum(axis=0)
             return (X[feat.index(word)]+alpha)/(sum(X)+len(feat))
```

Figure 18: Import time, Timer and Wrapper

On figure 14, it can be notice that this is principal algorithm to train the model, the previous function created it's our guide to obtain the likelihood feature, lets remark that this particular function often in bayes model describe the use and final result that results in the data that is manipulated in example, this model can perfectly work also for Netflix viewers, predict who will cancel subscription from a service depending from parameters used on each participant, each model trained contains its own set of parameters that at the end will describe what the model will look like.

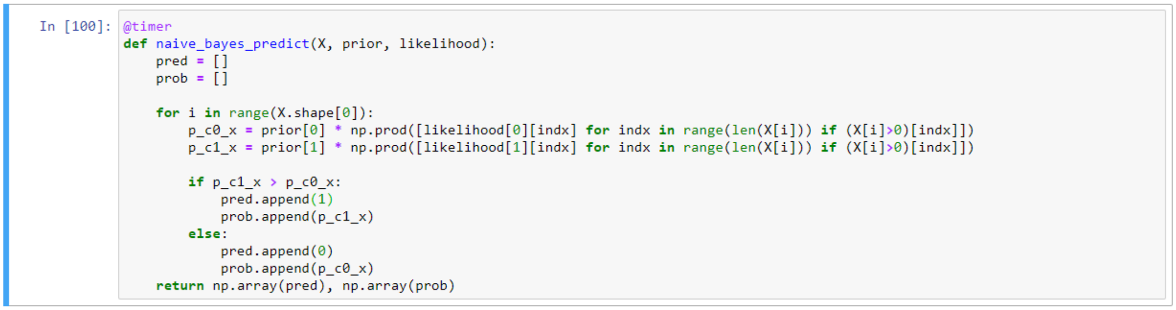


```
@timer
def naive_bayes_train(X, y, alpha=1):
    # Computes priors
    p1 = sum(y==1)/len(y)
    p0 = 1 - p1

    # Computes likelihood
    global total_features
    total_features = len(features)

    feature_i = 0
    likelihood0 = [get_likelihood(features, X, y, f, class_n=0, alpha=alpha) for f in features]

    feature_i = 0
    likelihood1 = [get_likelihood(features, X, y, f, class_n=1, alpha=alpha) for f in features]

    return [p0, p1], [likelihood0, likelihood1]
```

Figure 19: Putting all together in an algorithm

This function Works to know each and all the prediction values, to begin with you need an array of likelihood data, that previously you work with, Naive Bayes is a categorization code for binary (two-class) and multiclass categorization obstacles or barriers. It is called Naive Bayes or idiot Bayes this is basically the estimation of the possibility for all of the classes that are simplified to make their estimation susceptible. [22]

Instead of trying to figure out the possibility of each characteristic values, all are presumed to be hypothetically self-determining given the class value, this is a very robust supposition that is most improbable in real life or real sets of data.



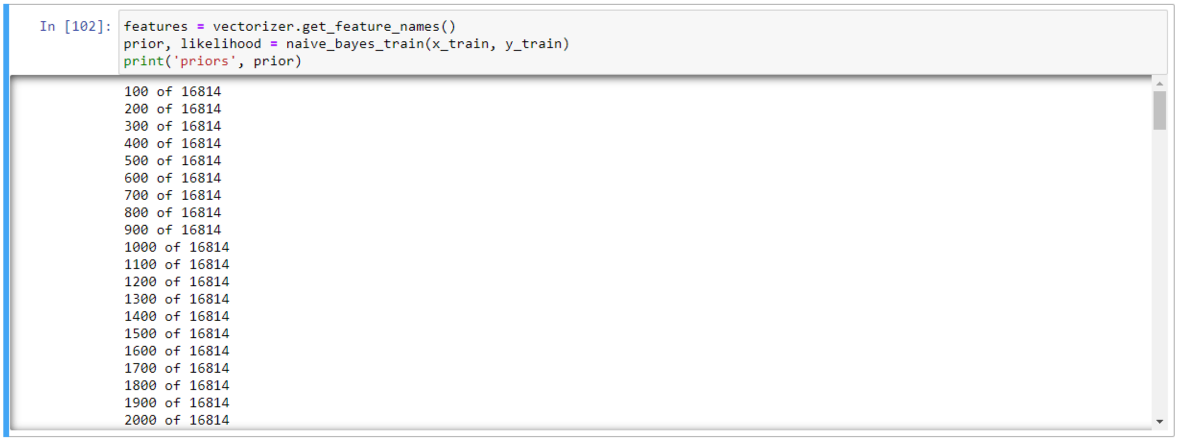
```
@timer
def naive_bayes_predict(X, prior, likelihood):
    pred = []
    prob = []

    for i in range(X.shape[0]):
        p_c0_x = prior[0] * np.prod([likelihood[0][indx] for indx in range(len(X[i])) if (X[i]>0)[indx]])
        p_c1_x = prior[1] * np.prod([likelihood[1][indx] for indx in range(len(X[i])) if (X[i]>0)[indx]])

        if p_c1_x > p_c0_x:
            pred.append(1)
            prob.append(p_c1_x)
        else:
            pred.append(0)
            prob.append(p_c0_x)
    return np.array(pred), np.array(prob)
```

Figure 20: Implement prediction function

As represented into picture 18, our model learns the prior probability of an object' textual aspect and content, having a distance of characters within a unique feature map, in the visual search, the model have influence over various different feature maps by computing the possibility of a given spotted object for each indicator inside of a feature map.

As outcome, positions in the maps with the biggest possibilities will be look around first, as they denote likely locations for the spotted object.

Both of them the prior and likelihood possibilities can be educated from any train side of the object and the conditions. Same way the speed and simplicity will remain as one of the most exiting features around this process avoiding considerably time-consuming recognition algorithms.



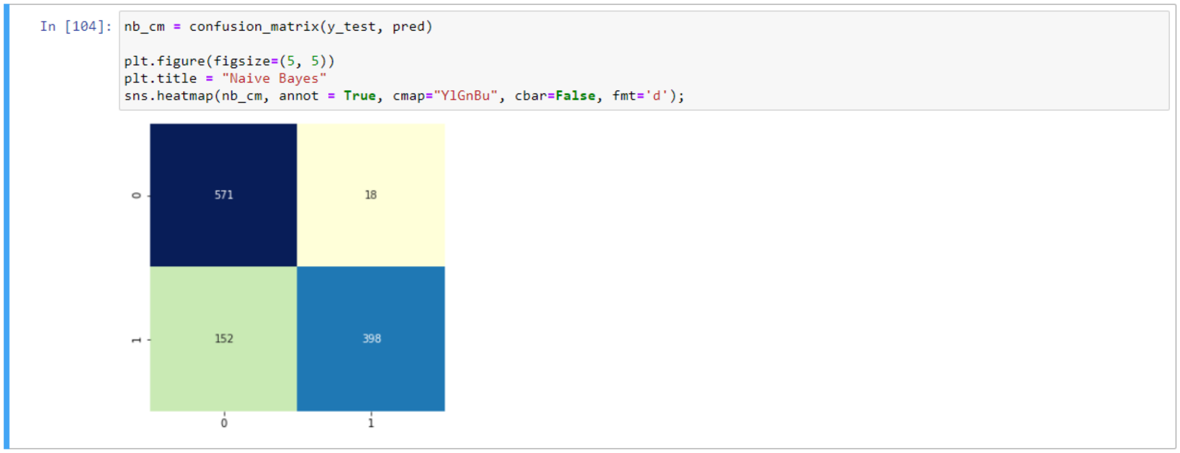
```
In [102]: features = vectorizer.get_feature_names()
          prior, likelihood = naive_bayes_train(x_train, y_train)
          print('priors', prior)
```

```
          100 of 16814
          200 of 16814
          300 of 16814
          400 of 16814
          500 of 16814
          600 of 16814
          700 of 16814
          800 of 16814
          900 of 16814
          1000 of 16814
          1100 of 16814
          1200 of 16814
          1300 of 16814
          1400 of 16814
          1500 of 16814
          1600 of 16814
          1700 of 16814
          1800 of 16814
          1900 of 16814
          2000 of 16814
```

Figure 21: Training the non-suicide / suicide model

Depression illness detection in social networks media is a multidisciplinary area where bayes model and the different features offered, detect indicator of depression in the users of social media.

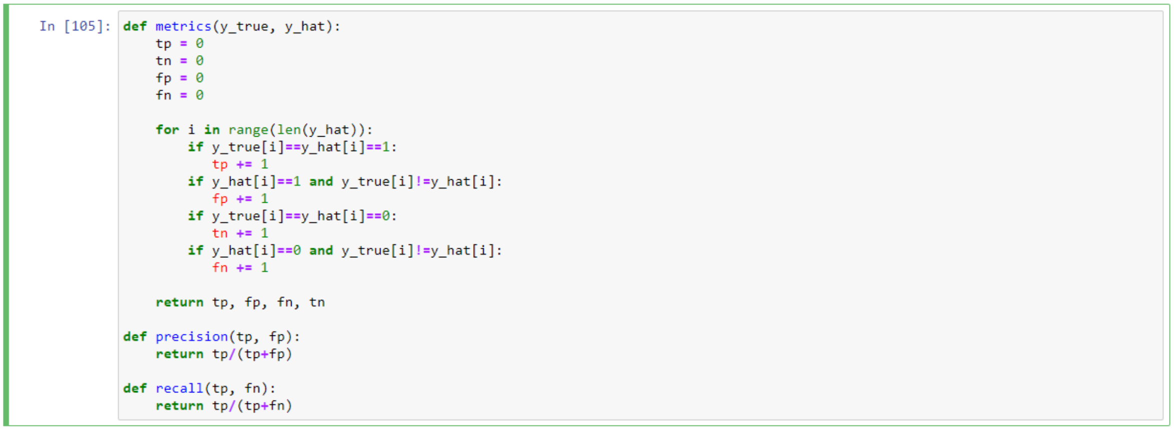
In representation of figure 19 the confusion matrix was made by using an inventory that has made methodical remarks plus records of the behaviors and symptoms of depressed users in the media, developed model was designed that contains features and vectors with characteristics from both sides depressed and non-depressed classes = Suicide or Non Suicide text.



```
In [104]: nb_cm = confusion_matrix(y_test, pred)

          plt.figure(figsize=(5, 5))
          plt.title = "Naive Bayes"
          sns.heatmap(nb_cm, annot = True, cmap="YlGnBu", cbar=False, fmt='d');
```

Figure 22: Evaluating text classification and the confusion matrix

In pattern recognition modeling, information recovery and categorization, precision is the portion of relevant occurrence in between the recovery occurrences, hence in recall is the portion of pertinent instances that were recovered. The two of them precision and recall are therefore based on relevancy.



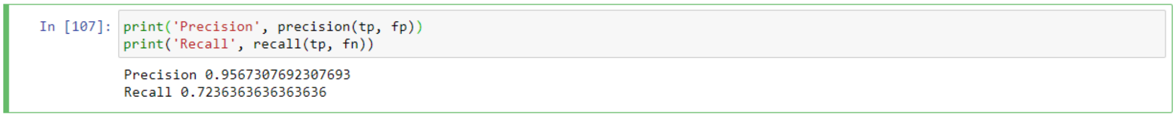
```
In [105]: def metrics(y_true, y_hat):
              tp = 0
              tn = 0
              fp = 0
              fn = 0

              for i in range(len(y_hat)):
                  if y_true[i]==y_hat[i]==1:
                     tp += 1
                  if y_hat[i]==1 and y_true[i]!=y_hat[i]:
                     fp += 1
                  if y_true[i]==y_hat[i]==0:
                     tn += 1
                  if y_hat[i]==0 and y_true[i]!=y_hat[i]:
                     fn += 1

              return tp, fp, fn, tn

          def precision(tp, fp):
              return tp/(tp+fp)

          def recall(tp, fn):
              return tp/(tp+fn)
```

Figure 23: Precision and Recall

In a this result, you can tell that True Negatives are the highest score true negatives which means that the model gives this number to the wrong area negatively, passing the best score with less count to the False Positives to the positive class.

Results:

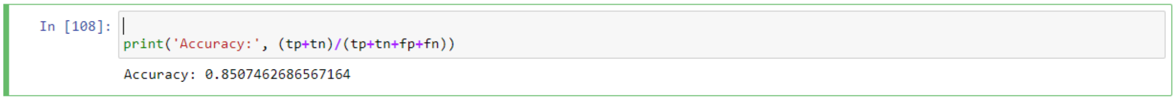
True Positives: 398
True Negatives: 571
False Positives: 18
False Negatives: 152



```
In [107]: print('Precision', precision(tp, fp))
          print('Recall', recall(tp, fn))

          Precision 0.9567307692307693
          Recall 0.7236363636363636
```

Figure 24: Precision and Recall part 2

Results:
Precision 0.9567307692307693
Recall 0.7236363636363636



```
In [108]: 
          print('Accuracy:', (tp+tn)/(tp+tn+fp+fn))

          Accuracy: 0.8507462686567164
```

Figure 25: Putting together the accuracy

This Bayes model developed in this work is victoriously functioning, in classifying the users of social network media giving result for depressed and also non-depressed users, achieving the accuracy score 0.85%, the result it’s good but not great, with a true positive of this score means that there is this possibility of users have or are in the process of suffer depression.

Results:
Accuracy: **0.8507462686567164**

5. RESULTS AND DISCUTIONS

5.1. Results

This work presents results to align with medical conditions with the highest possible accuracy, this model it's based with build a judgement power source for text posts-outcomes adapted to the Bayesian social media classifiers.

This prediction model analysis that was trained on data of over 500 thousand comments on the beginning, users from all over united states with sensitive expressions related to depression, and so many more with regular comments.

This Bayes Classifier was trained with a high content of media comments, in the search of qualified text structure that can be measured by maximum length of 24 words.

The probability performance proved here, was that this Bayesian Classifier with selected arrays and split data by wrapper characteristic selection can definitely predict up to 0.85% of depression in users participating in any social media.

Count vectorization was performed as follows, x_train 4552, 8407 words, on the other way around x_test 1139, 8407 separating messages as expected, depress or not depress, suicide no suicide, normal or no normal message from the user. Speaking about the False Positive Range on this result came across with 18 points, with precision of 0.95% and recall of 0.73%, all this in about 4 minutes of processing the algorithm.

All this results translated to real life can be the human pregnancy test, True Positive it's when a result it's given as positive but the woman it's not pregnant, only the blood test can give the highest percentage of accuracy, and in contrary, if the test it's giving as no pregnancy but the woman it is pregnant, the result with Bayes will translated as False Positive.

In the Confusion Matrix positives are most likely in light colors and negatives with dark colors to guide you easily to the nearest point of accuracy.

5.2. Discussion

In this work there are two different slopes that we face in the middle of the process, both are related to the libraries we used, first NumPy [22], an example it's how easy this library can squeeze the content of each file with CSV format in different segments, as well as masked arrays, all the logical operations that happened in between to map the different arrays in the confusion matrix.

On the other side, Pandas, [24] library that basically wraps previews work from NumPy and structure and manipulate numerical tables with the formal intention of tagging all the content and the columns on each table, this was very helpful on this investigation because of the time consuming benefit saved, and avoiding complex manual work.

The benefit out of this final result will help our society, and the vulnerable population using social network media, or at some point having to do with it directly or indirectly, Bayes algorithm it's a friendly and unique technique, that can be use code with very few lines of programing, that come across with functional development, the community won't even notice that their prints in the media can be audit for their own wellbeing, any other developer or person can adapt this code within numerous complex problems.

On the other side of the innumerable benefits, we cannot forget about developers, this can be a suitable guide and have reference for future investigations, avoiding starting from scratch, Bayes Model it's now winning the war versus probability and frequentist.

6. CONCLUSIONS

In general terms, depression and many other mental illnesses can be detected using modern technologies, same way it was used to be detected several years ago, after first sign of the disease by studying depression with bayes model gets precision on the final probability, depression it's the silent mental illness, most times with deathly result also, can be detected easily through the social media.

Detecting the proper channel of communication, a lot of data can help to retrieve bulks of information., with the proper social media depressed users can be detected easily besides the data found on social media, the cleaning code it's most likely need it in all the bayes model technique, with Python programing language remains the top ranking of languages in data analysis.

Several datasets located in the web don't have the definition well done described, likewise the filtering phase it's most necessary for the Bayes Classifier to work directly within probability, more like Bag of Words, ignoring set of words depending position and taking more in consideration frequency Bayes Rule recall conditional probability in random variables.

NumPy provides powerful data structures by using multidimensional arrays similar with Pandas library, it helps a lot with access data through indexes or names for datasets, likewise Pandas library provides methods for reordering, dividing, and combining datasets.

Likelihood is the same number of certain examples of class spitted by the number of entries and helps the Confusion Matrix, by telling easily true positive vs true negatives True positives are equal results of final detail you are looking for in your model, by training bayes classifier model any set of data can help better than plenty of probabilities techniques.

In the next coming months after I end up with my master's degree, i would like to try on publishing this work into an article, also implement this model using other brands of social media such as Instagram or Twitter.

Another ambitious plan it's to keep digging into Bayes Model and other different usages, maybe more medial purposes, in example attention deficit disorder. Getting to know this new model with my current employee and find uses within the organization with Microsoft (Yammer) as example, to detect certain behaviors. Also explore more into details of NumPy and Pandas libraries, what other usages I can get out of them, since they have very extensive material to learn.

7. BIBLIOGRAPHY

[1] V. R. K. Garimella, V. R. K. Garimella, A. Alfayad, I. Weber, Aalto University, Aalto University, Abdulrahman Alfayad Carnegie Mellon University, Carnegie Mellon University, Ingmar Weber Qatar Computing Research Institute, Qatar Computing Research Institute, Yahoo, University of Maryland / National Park Service, University of Michigan, Microsoft, and University of Iowa, “Social Media Image Analysis for Public Health,” Social Media Image Analysis for Public Health | Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 01-May-2016. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/2858036.2858234. [Accessed: 04-Jun-2020].

[2] Depression (major depressive disorder). (2018, February 03). Retrieved December 03, 2020, from https://www.mayoclinic.org/diseases-conditions/depression/symptoms-causes/syc-20356007

[3] V. R. K. Garimella, V. R. K. Garimella, A. Alfayad, I. Weber, Aalto University, Aalto University, Abdulrahman Alfayad Carnegie Mellon University, Carnegie Mellon University, Ingmar Weber Qatar Computing Research Institute, Qatar Computing Research Institute, Yahoo, University of Maryland / National Park Service, University of Michigan, Microsoft, and University of Iowa, “Social Media Image Analysis for Public Health,” Social Media Image Analysis for Public Health | Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 01-May-2016. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/2858036.2858234. [Accessed: 04-Jun-2020].

[4] A. Dhir, Y. Yossatorn, P. Kaur, and S. Chen, “Online social media fatigue and psychological wellbeing-A study of compulsive use, fear of missing out, fatigue, anxiety and depression,” International Journal of Information Management, 23-Feb-2018. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0268401217310629. [Accessed: 04-Jun-2020].

[5] F. Stallone, G. J. Huba, W. G. Lawlor, and R. R. Fieve, “Longitudinal Studies of Diurnal Variations in Depression: A Sample of 643 Patient Days: The British Journal of Psychiatry,” Cambridge Core, 29-Jan-2018. [Online]. Available: https://www.cambridge.org/core/journals/the-british-journal-of-psychiatry/article/longitudinal-studies-of-diurnal-variations-in-depression-a-sample-of-643-patient-days/4CD5EB550DC850A28CC02A26409CBD8D. [Accessed: 04-Jun-2020].

[6] F. T. Torres, What Is Depression?, 01-Oct-2020. [Online]. Available: https://www.psychiatry.org/patients-families/depression/what-is-depression. [Accessed: 23-Feb-2021].

[7] J. E. Wiley, “Cross-cultural comparison of symptom networks in late-life major depressive disorder: Yoruba Africans and the Spanish Population,” International journal of geriatric psychiatry, 20-Sep-2020. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/32394534/. [Accessed: 23-Feb-2021].

[8] D. M. D. Becker, “Home,” Sutter Health, 16-Aug-2017. [Online]. Available: https://www.sutterhealth.org/health/mental/depression-by-age. [Accessed: 23-Feb-2021].

[9] [9] P. D. Jane M. Murphy, “Depression and Anxiety in Relation to Social Status,” Archives of General Psychiatry, 01-Mar-1991. [Online]. Available: https://jamanetwork.com/journals/jamapsychiatry/article-abstract/495253. [Accessed: 23-Feb-2021].

[10] O. Hornykiewicz, “DOPAMINE (3-HYDROXYTYRAMINE) AND BRAIN FUNCTION,” Pharmacological Reviews, 01-Jun-1966. [Online]. Available: https://pharmrev.aspetjournals.org/content/18/2/925. [Accessed: 25-Feb-2021].

[11] S. Brien, “Mental Illness and Addiction: Facts and Statistics,” CAMH, 02-Jul-1996. [Online]. Available: https://www.camh.ca/en/driving-change/the-crisis-is-real/mental-health-statistics. [Accessed: 25-Feb-2021].

[12] Chávez, G. (2019, March 6). *Implementing a Naive Bayes classifier for text categorization in Five Steps*. Medium. https://towardsdatascience.com/implementing-a-naive-bayes-classifier-for-text-categorization-in-five-steps-f9192cdd54c3.

[13] Depression Detection by Analyzing Social Media Posts of User. IEEE Xplore. (n.d.). https://ieeexplore.ieee.org/document/9065101.

[14] Joyce, J. (2003, June 28). Bayes' Theorem. Stanford Encyclopedia of Philosophy. https://stanford.library.sydney.edu.au/archives/sum2016/entries/bayes-theorem/.

[15] What is NumPy?¶. What is NumPy? - NumPy v1.21 Manual. (n.d.). https://numpy.org/doc/stable/user/whatisnumpy.html

[16] Kazarinoff, P. D. (n.d.). *Array Indexing*. Array Indexing - Problem Solving with Python. https://problemsolvingwithpython.com/05-NumPy-and-Arrays/05.05-Array-Indexing/#:~:text=Indexing%20is%20an%20operation%20that,is%20stored%20in%20an%20array.

[17] Mueller, J., & Massaron, L. (2019). Python for data science. John Wiley & Sons.

[18] Storey, J. D. (n.d.). The positive false discovery rate: a Bayesian interpretation and the q-value. The Annals of Statistics. https://projecteuclid.org/journals/annals-of-statistics/volume-31/issue-6/The-positive-false-discovery-rate--a-Bayesian-interpretation-and/10.1214/aos/1074290335.full.

[19] Mishra, S. (2021, June 4). Baffling Concept of True Positive and True Negative. Medium. https://towardsdatascience.com/baffling-concept-of-true-positive-and-true-negative-bffbc340f107.

[20] Victoria, E. K. U. of, Kalliamvakou, E., Victoria, U. of, Georgios Gousios Delft University of Technology, Gousios, G., Technology, D. U. of, Victoria, K. B. U. of, Blincoe, K., Victoria, L. S. U. of, Singer, L., Daniel M. German University of Victoria, German, D. M., Victoria, D. D. U. of, Damian, D., Davis, U. of C. at, Hong Kong University of Science and Technology, Klagenfurt, U. of, Contributor MetricsExpand All Eirini Kalliamvakou University of Victoria Publication Years2008 - 2019Publi, Eirini Kalliamvakou University of Victoria Publication Years2008 - 2019Publication counts9Available for Download6Citation count430Downloads (cumulative)7, & Authors: Eirini Kalliamvakou University of Victoria. (2014, May 1). The promises and perils of mining GitHub. The promises and perils of mining GitHub | Proceedings of the 11th Working Conference on Mining Software Repositories. https://dl.acm.org/doi/abs/10.1145/2597073.2597074.

[21] Function Wrappers in Python. GeeksforGeeks. (2020, June 22). https://www.geeksforgeeks.org/function-wrappers-in-python/.

[22] Brownlee, J. (2019, October 24). Naive Bayes Classifier From Scratch in Python. Machine Learning Mastery. https://machinelearningmastery.com/naive-bayes-classifier-scratch-python/.