



FDA ADVERSE EVENT IN 2018

CONTENTS

Introduction.....	3
Literature review.....	5
Methodology.....	10
Data Dictionary.....	11
Statistical analysis:	26
Demographic.....	27
Drug:.....	28
Reaction:	28
Machine Learning analysis.....	30
Unsupervised Learning.....	35
Clustering.....	36
DBSCAN Clustering.....	37
Hierarchical clustering:	40
K-mean clustering.....	40
Supervised Learning.....	44
Data Preparation:	45
Cross-validation techniques:.....	47
Train test split – 5Kfold.....	49
SMOTE(Over Sampling).....	52
SMOTEENN.....	56
Evaluating.....	59
Conclusion.....	61
References.....	63

INTRODUCTION

Adverse Event Reporting System (FAERS) is a database that contains adverse event reports, medication error reports, and product quality complaints resulting in adverse events that were submitted to FDA. So, I'm seeking the reaction of drugs that made many people die. I expected with all my effort I will find some clues that what drugs have a negative effect on patients. So that drugs companies can fix their drug formulation or change their ingredient. I chose 2018 but the data was really big, and my computer can't handle the massive amount of data, so I decided to run one quarter instead. This data is public data so they were covering some important information such as patients addresses, patients' race, patients' behaviors, etc.... which can be helpful to find the pattern of patients. This data has a lot of Duplicates, Null values, and outliers. I need to work with it carefully because that really needs to be properly cleaned. I will apply some statistical methodologies and machine learning to this data set after finished the preprocessing step.

This study was conducted to analyze the reaction of Adverse events in 2018. We want to run some statistical tests and machine learning to find the answer that what drug caused the patient's death?

We will study how to separate queries choose what attributes are important and run some statistical tests on it such as odd ratio, Chi square, and fisher test. After doing that we will apply machine learning to it. I'm thinking about using unsupervised learning for patients to classify the patient's groups, but it depends on the result of preprocessing step. Finally, we will use a supervised study to build a predictive model on patients so we can understand and find any other reason why patients died.

Tools : Excel, SQL, Python, word, sklearn, Pandas, numpy, seaborn, collections, matplotlib.

Keywords: FDA FAERS, FDA 2018, Odd ratio, Chisquare, Fisher test, logistic regression, dbscan, K-Means, Hierarchical clustering, Decision Tree, random forest, KNN, XGBOOSTING

LITERATURE REVIEW

This document is intended for use by industry. It reflects the agency's current thinking on 1) the process for evaluating scientific evidence for a health claim, 2) the significance of the significant scientific agreement (SSA) standard in section 403(r)(3) of the Federal Food, Drug, and Cosmetic Act (the Act) (21 U.S.C. 343(r)(3)) and 21 CFR 101.14(c), and 3) credible scientific evidence to support a qualified health claim.

The FDA intends to adopt an evidence-based review approach to evaluate publicly available scientific data for SSA health claims or qualifying health claims on the association between a substance and a disease or health-related condition, as described in this guideline document.

(2) FDA's current perspective on the scientific review process it should adopt is explained in this guideline paper, which is designed to give assistance to health claim petitioners. (3)

The specific topics addressed in this guidance document are: (1) identifying studies that evaluate the substance/disease relationship, (2) identifying surrogate endpoints for disease risk, (3) evaluating the human studies to determine whether scientific conclusions can be drawn from them about the substance/disease relationship, (4)

assessing the methodological quality of each human study from which scientific conclusions about the substance/disease relationship can be drawn, (5) evaluating the totality of scientific evidence, (6) assessing significant scientific agreement, (7) specificity of claim language for qualified health claims, and (8) revaluation of existing SSA or qualified health claims.

The US Food and Drug Administration has asked Endo Pharmaceuticals to take its reformulated Opana ER (oxymorphone hydrochloride) opioid pain medicine off the market. The agency is requesting removal after thorough assessment, citing a fear that the drug's advantages may no longer exceed its hazards. This is the first time the FDA has taken action to stop the sale of a currently marketed opioid pain medicine owing to the dangers of misuse.

“We are facing an opioid epidemic – a public health crisis, and we must take all necessary steps to reduce the scope of opioid misuse and abuse,” said FDA Commissioner Scott Gottlieb, M.D. “We will continue to take regulatory steps when we see situations where an opioid product’s risks outweigh its benefits, not only for its intended patient population but also in regard to its potential for misuse and abuse.”

The FDA made its judgement after reviewing all available post marketing data, which showed a dramatic shift in the mode of abuse of Opana ER from nasal to injectable after the medication was reformulated. Injection usage of reformulated Opana ER has been linked to an HIV and hepatitis C outbreak, as well as

instances of a dangerous blood condition (thrombotic microangiopathy). Following a meeting of the FDA advisory committee in March 2017, a panel of 18 independent specialists decided 18-8 that the advantages of reformulated Opana ER no longer outweigh the dangers.

The FDA made its judgement after reviewing all available post marketing data, which showed a dramatic shift in the mode of abuse of Opana ER from nasal to injectable after the medication was reformulated. Injection usage of reformulated Opana ER has been linked to an HIV and hepatitis C outbreak, as well as instances of a dangerous blood condition (thrombotic microangiopathy). Following a meeting of the FDA advisory committee in March 2017, a panel of 18 independent specialists decided 18-8 that the advantages of reformulated Opana ER no longer outweigh the dangers.

“The abuse and manipulation of reformulated Opana ER by injection has resulted in a serious disease outbreak. When we determined that the product had dangerous unintended consequences, we made a decision to request its withdrawal from the market,” said Janet Woodcock, M.D., director of the FDA’s Center for Drug Evaluation and Research. “This action will protect the public from further potential for misuse and abuse of this product.”

The FDA has asked the firm to take reformulated Opana ER off the market voluntarily. If the corporation refuses to remove the product, the government plans to legally order it to be removed by withdrawing clearance. In the

meanwhile, the FDA is informing health-care providers and others about the extremely high hazards connected with this product's misuse.

As part of our response to the public health crisis, the FDA will continue to assess the risk-benefit balance of all authorized opioid analgesic medicines and take additional steps as needed.

The Food and Drug Administration (FDA), which is part of the US Department of Health and Human Services, promotes and protects public health by ensuring the safety, effectiveness, and security of human and veterinary drugs, vaccines, and other biological products for human use, as well as medical devices. The agency is also in charge of ensuring the safety and security of our nation's food supply, cosmetics, nutritional supplements, electronic radiation-emitting items, and tobacco products.

I don't think anyone use machine learning approach for the same criteria. They mostly using statistical methods because this dataset lacking information for running machine learning models Drug products with abuse potential generally contain drug substances that have central nervous system (CNS) activity and produce euphoria (or other changes in mood), hallucinations, and effects consistent with CNS depressants or stimulants. Thus, if a drug substance is CNS-active, the new drug product containing that drug substance will likely need to undergo a thorough assessment of its abuse potential and may be subject to control under the Controlled Substances

Act (CSA) (*see generally* 21 U.S.C. 811). The CSA contains five schedules of control: Schedules I, II, III, IV and V. Drugs or other substances with a high abuse potential, no

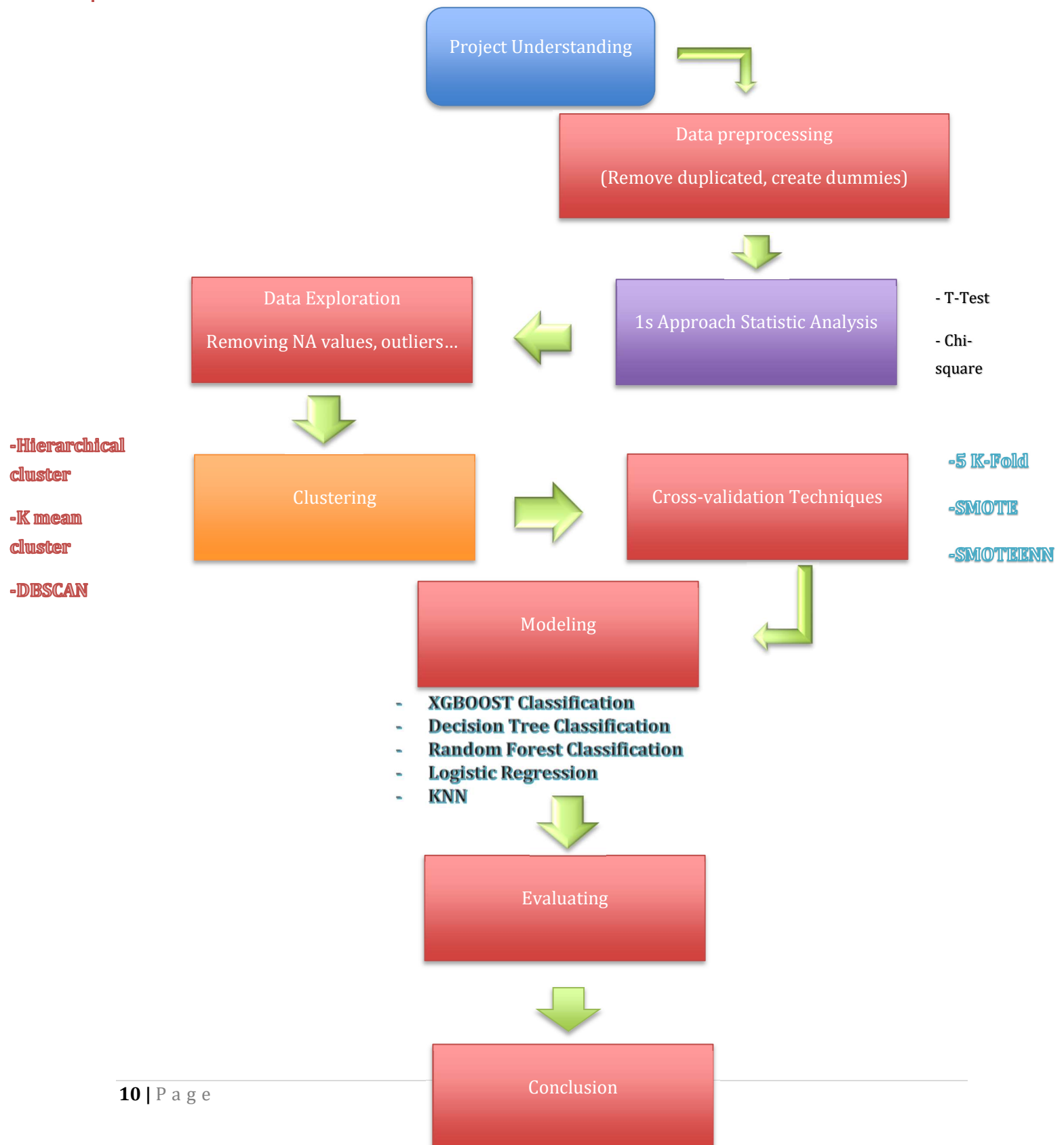
currently accepted medical use, and a lack of accepted safety for use under medical supervision

are controlled in Schedule I. Drugs or other substances with abuse potential that do have a currently accepted medical use (e.g., the drug or substance is in an FDA-approved product) are placed into Schedule II, III, IV, or V. The specific placement of a drug or other substance within Schedules II-V is determined by the relative abuse potential of the drug or substance and the relative degree to which it induces psychological or physical dependence (21 U.S.C. 812(b)).

In general, FDA's guidance documents do not establish legally enforceable responsibilities. Instead, guidance's describe the Agency's current thinking on a topic and should be viewed only as recommendations, unless specific regulatory or statutory requirements are cited. The use of the word *should* in Agency guidance's means that something is suggested or recommended, but not required.

Drug products that include CNS-active novel molecular entities (NMEs) as well as those that contain CNS-active drugs that are currently prohibited under the CSA are addressed in this advice. 2 Generic drug goods seeking clearance under section 505(j) of the Federal Food, Drug, and Cosmetic Act are normally not reviewed for abuse risk and placed on the same CSA schedule as the innovator drug product. Drug products with misuse potential may include chemicals or pharmacologically comparable compounds to existing banned drugs, or they may have unique chemical structures and/or methods of action in the brain.

METHODOLOGY



DATA DICTIONARY

Demographic		
Name	Description	Datatype
PRIMARYID	Unique number for identifying a FAERS report. This is the primary link field (primary key) between data files (example: 31234561). This is a concatenated key of Case ID and Case Version Number. It is the Identifier for the case sequence (version) number as reported by the manufacturer.	int64
CASEID	Number for identifying a FAERS case.	int64
CASEVERSION	Safety Report Version Number. The Initial Case will be version 1; follow-ups to the case will have sequentially incremented version numbers (for example, 2, 3, 4, etc.).	int64
I_F_COD	Code for initial or follow-up status of report, as reported by manufacturer. CODE MEANING_TEXT -----	object

Demographic		
	I Initial F Follow-up	
EVENT_DT	Date the adverse event occurred or began. (YYYYMMDD format) – If a complete date is not available, a partial date is provided. See the NOTE on dates at the end of this section.	float64
MFR_DT	Date manufacturer first received initial information. In subsequent versions of a case, the latest manufacturer received date will be provided (YYYYMMDD format). If a complete date is not available, a partial date will be provided. See the NOTE on dates at the end of this section.	float64
INIT_FDA_DT	Date FDA received first version (Initial) of Case (YYYYMMDD format)	int64
FDA_DT	Date FDA received Case. In subsequent versions of a case, the latest manufacturer received date will be provided (YYYYMMDD format).	int64

Demographic		
REPT_COD	<p>Code for the type of report submitted (See table below)</p> <p>Also, see Section E, End Note below.</p> <p>CODE MEANING_TEXT</p> <p>-----</p> <p>EXP Expedited (15-Day)</p> <p>PER Periodic (Non-Expedited)</p> <p>DIR Direct</p>	object
AUTH_NUM	<p>Regulatory Authority's case report number, when available.</p> <p>+ New tag added in 2014Q3 extract.</p>	object
MFR_NUM	Manufacturer's unique report identifier	object
MFR_SNDR	<p>Coded name of manufacturer sending report; if not found, then</p> <p>verbatim name of organization sending report.</p>	object
LIT_REF	<p>Literature Reference information, when available; populated</p> <p>with last 500 characters if >500 characters are available.</p> <p>+ New tag added in 2014Q3 extract.</p>	object

Demographic		
AGE	Numeric value of patient's age at event.	float64
AGE_COD	Unit abbreviation for patient's age (See table below) CODE MEANING_TEXT ---- DEC DECADE YR YEAR MON MONTH WK WEEK DY DAY HR HOUR	object
AGE_GRP	Patient Age Group code as follows, when available: CODE MEANING_TEXT ---- N Neonate I Infant C Child T Adolescent A Adult E Elderly + New tag added in 2014Q3 extract.	object

Demographic		
SEX	<p>Code for patient's sex (See table below)</p> <p>CODE MEANING_TEXT</p> <p>-----</p> <p>UNK Unknown</p> <p>M Male</p> <p>F Female</p>	object
E_SUB	<p>Whether (Y/N) this report was submitted under the electronic submissions procedure for manufacturers.</p>	object
WT	Numeric value of patient's weight.	float64
WT_COD	<p>Unit abbreviation for patient's weight (See table below)</p> <p>CODE MEANING_TEXT</p> <p>-----</p> <p>KG Kilograms</p> <p>LBS Pounds</p> <p>GMS Grams</p>	object
REPT_DT	<p>Date report was sent (YYYYMMDD format). If a complete date is not available, a partial date is provided. See the NOTE on dates at the end of this section.</p>	float64

Demographic		
TO_MFR	Whether (Y/N) voluntary reporter also notified manufacturer (blank for manufacturer reports).	object
OCCP_COD	Abbreviation for the reporter's type of occupation in the latest version of a case. CODE MEANING_TEXT ----- MD Physician PH Pharmacist OT Other health-professional LW Lawyer CN Consume	object
REPORTER_COUNTRY	The country of the reporter in the latest version of a case: NOTE: Country codes are available per the links below. http://estri.ich.org/icsr/ICH_ICSR_Specification_V2-3.pdf http://www.iso.org/iso/home/standards/country_codes/iso-3166-1_decoding_table.htm	object

Demographic		
OCCR_COUNTRY	The country where the event occurred.	object

DRUG datasetL

Drug		
Name	Description	datatype
PRIMARYID	Unique number for identifying a FAERS report. This is the primary link field (primary key) between data files (example: 31234561). This is a concatenated key of Case ID and Case Version Number. It is the Identifier for the case sequence (version) number as reported by the manufacturer.	int64
CASEID	Number for identifying a FAERS	int64

Drug		
	case.	
DRUG_SEQ	Unique number for identifying a drug for a Case. To link to the THERyyQq.TXT data file, both the Case number (primary key) and the DRUG_SEQ number (secondary key) are needed. (For an explanation of the DRUG_SEQ number, including an example, please see Section E, End Note 2, below.)	int64
ROLE_COD	Code for drug's reported role in event(See table below) CODE MEANING_TEXT ----- PS Primary Suspect Drug SS Secondary Suspect Drug C Concomitant I Interacting	object

Drug		
DRUGNAME	Name of medicinal product. If a "Valid Trade Name" is populated for this Case, then DRUGNAME = Valid Trade Name; if not, then DRUGNAME = "Verbatim" name, exactly as entered on the report.	object
PROD_AI	Product Active Ingredient, when available. + New tag added in 2014Q3 extract.	object
VAL_VBM	Code for source of DRUGNAME (See table below) CODE MEANING_TEXT ----- 1 Validated trade name used 2 Verbatim name used	int64
ROUTE	The route of drug	object

Drug		
	administration	
DOSE_VBM	Verbatim text for dose, frequency, and route, exactly as entered on report.	object
CUM_DOSE_CHR	Cumulative dose to first reaction	float64
CUM_DOSE_UNIT	Cumulative dose to first reaction unit CODE Meaning_Text ----- KG Kilogram(s) GM Gram(s) MG Milligram(s) UG Microgram(s) (µg) NG Nanogram(s) PG Picogram(s) MG/KG Milligram(s)/Kilogram UG/KG Microgram(s)/Kilogram (µG/KG) MG/M**2 Milligram(s)/Sq. Meter	object

Drug		
	UG/M**2 Microgram(s)/Sq. Meter ($\mu\text{G}/\text{M}^{**2}$) L Litre(s) ML Millilitre(s) UL Microlitre(s) (μL) BQ Becquerel(s) GBQ Gigabecquerel(s) MBQ Megabecquerel(s) KBQ Kilobecquerel(s) CI Curie(s) MCI Millicurie(s) UCI Microcurie(s) (μCI) NCI Nanocurie(s) MOL Mole(s) MMOL Millimole(s) UMOL Micromole(s) IU International Unit(s) KIU International Unit*(1000s) MIU International Unit*(1,000,000s) IU/KG IU/Kilogram MEQ Milliequivalent(s) PCT Percent (%) GTT Drop(s)	

Drug		
	<p>DF Dosage Form</p> <p>NOTE: The list below provides Dose codes which are commonly reported; however, dose codes are not limited to this list and other code values may be present.</p>	
DECHAL	<p>Dechallenge code, indicating if reaction abated when drug therapy was stopped (See table below)</p> <p>CODE MEANING_TEXT</p> <p>-----</p> <p>Y Positive dechallenge</p> <p>N Negative dechallenge</p> <p>U Unknown</p> <p>D Does not apply</p>	object
RECHAL	<p>Rechallenge code, indicating if reaction recurred when drug therapy was restarted (See table below)</p>	object

Drug		
	CODE MEANING_TEXT ----- Y Positive rechallenge N Negative rechallenge U Unknown D Does not apply	
LOT_NUM	Lot number of the drug (as reported).	object
EXP_DT	Expiration date of the drug. (YYYYMMDD format) - If a complete date is not available, a partial date is provided, See the NOTE on dates at the end of this section.	object
NDA_NUM	NDA number (numeric only)	float64
DOSE_AMT	Amount of drug reported	float64
DOSE_UNIT	Unit of drug dose	object

Drug		
DOSE_FORM	Form of dose reported	object
DOSE_FREQ	Code for Frequency CODE Meaning_Text ----- 1X Once or one time BID Twice a day BIW Twice a week HS At bedtime PRN As needed Q12H Every 12 hours Q2H Every 2 hours Q3H Every 3 hours Q3W Every 3 weeks Q4H Every 4 hours Q5H Every 5 hours Q6H Every 6 hours Q8H Every 8 hours QD Daily QH Every hour QID 4 times a day QM Monthly QOD Every other day QOW Every other week	object

Drug		
	QW Every week TID 3 times a day TIW 3 times a week UNK Unknown NOTE: The list below provides frequency codes which are commonly reported; however, dose frequency codes are not limited to this list and other code values may be present.	

For this specific Capstone Project I would like to use data from FDA Food and Drug Administration US) to find out what happened in 2018, the reason choose 2018 for analysis because I don't want my data messed with covid data in 2019. This dataset has a lot of missing values and also lack of necessary information such as race, home address, daily routine, income...etc. As I mention no one used this for applying machine learning but I want to make it become more understandable as much as I can. About what drugs make people died in 2018 and

what is age range of deceased people. Also compare the relation between weight, age to deceased people

For initial step, I Import 3 datasets Demographic, Drugs, Reaction:

- Demographic: the information of patients and information about what dates the data has published by reporter. Also, the reporter companies
- Drug: the Drug name, it connects to Demographic by PrimaryID (also is Patients ID)
- Reaction: the reaction after patients using that drug.

STATISTICAL ANALYSIS:

By using statistic analysis, we can get straight to the answer that how risky for patients using specific drug with other drugs. In this step I will apply hypothesis testing t-test, chi square test and find the odd ratio.

First step finds the most frequent reaction:

Table 1: top 5 reaction

1. Drug ineffective	19563
2. Death	17346
3. Toxicity to various agents	9383
4. Drug dose omission	9201
5. Alopecia	9078

So, in here I can't do anything with Drug Ineffective because that is another scenario and have to do a lot of research about drug ingredient so I will leave it and focus only the second top reaction "Death"

Second step I have to define the variables in the dataset which are useful for my analysis.

I chose

PrimaryID, age, weight, occur country.

Demographic

the reason why I chose only those variables below because for other variable they are mostly meaningless and regard to reporter companies. And for date time values those datetime values are some missing information and I also don't want to add it in to my dataset. So for my Initial Demographic variable I have 412702 records and 6 columns

Drug:

For Drug dataset, this Dataset Large because It's also included duplicate values and missing values, but we can't remove them right now. Because for each single patient they might use more than two drugs for their treatment so we will do cleaning part later when we merging them together.

Reaction:

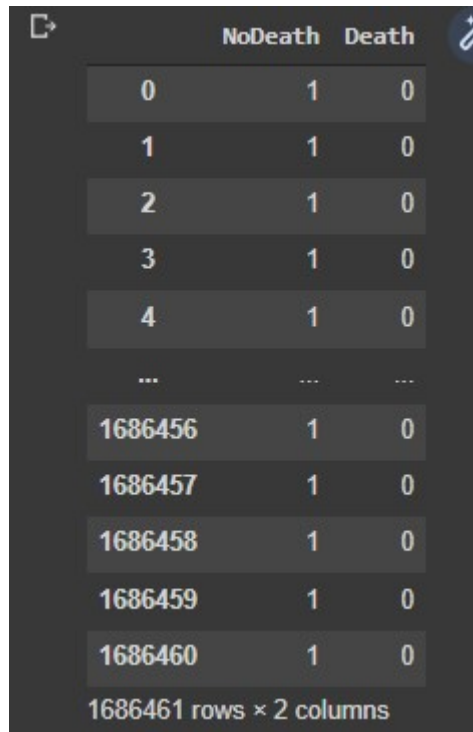
This dataset included all the reaction of patients after using drug.

Third step:

I merge the dataset of drug and reaction together to check the relation between drug use and reaction.

Forth step:

I create the dummy variable of reaction, but I only choose the reaction "Death" this is will be my target variable. So, in this column I will column death or not only. And 0 stands for not dead, 1 is dead.



	NoDeath	Death
0	1	0
1	1	0
2	1	0
3	1	0
4	1	0
...
1686456	1	0
1686457	1	0
1686458	1	0
1686459	1	0
1686460	1	0

1686461 rows × 2 columns

Fifth step:

I merge my dummy column death to my drug dataset, and I pick up the most frequent drug causes patient died, and the answer is **OPANA ER**

1.Opana ER	4871
2.EYLEA	2748
3.PERCOCET	2661
4.OXYCONTIN	2651
5.Opana	2639

```
[22] # count the top 5 drugs which most frequency cause ppl died
s = Counter(DiedDummies['drugname'])
s.most_common(5)
print ("",s.most_common(5))

[('OPANA ER', 4871), ('EYLEA', 2748), ('PERCOCET', 2661), ('OXYCONTIN', 2651), ('OPANA', 2639)]
```

Final step:

I tabulate the cross table to calculate odd ratio and chi square-test:

Odd ratio: 8.07539446

Chi square: <0.05

So that I can conclude Opana Drug highly cause to patients die and people who were using it is higher than other drug 8 times.

MACHINE LEARNING ANALYSIS.

For machine Learning Analysis. I will focus on other factors. I aim to the Patients, so as we know at the previous step OHPANA ER is the main factor cause patients dying but why they

die? Maybe they have other reason, so we need to study about patient what their age, are they underweight, or they are obese? Where are they living ?etc. I will apply Clustering method to group patients. So that we have better view to the picture. What group has the most effective and which lower risk?

The other tools I will add that is prediction tool which give me the guess to the future so we can determine patient have high risk or not and that we can use to help doctor saving their time and money. Also, I can help patients who are in high risk stay away of being dead by give them a sign or alert if they are in the risk.

Before to start, I have to make sure my dataset is clean, no outliers and no missing value. In order to make it cleaned. I have to check all of my variables. First thing I will go with variable Age.

Variable Age (before Cleaning outliers)

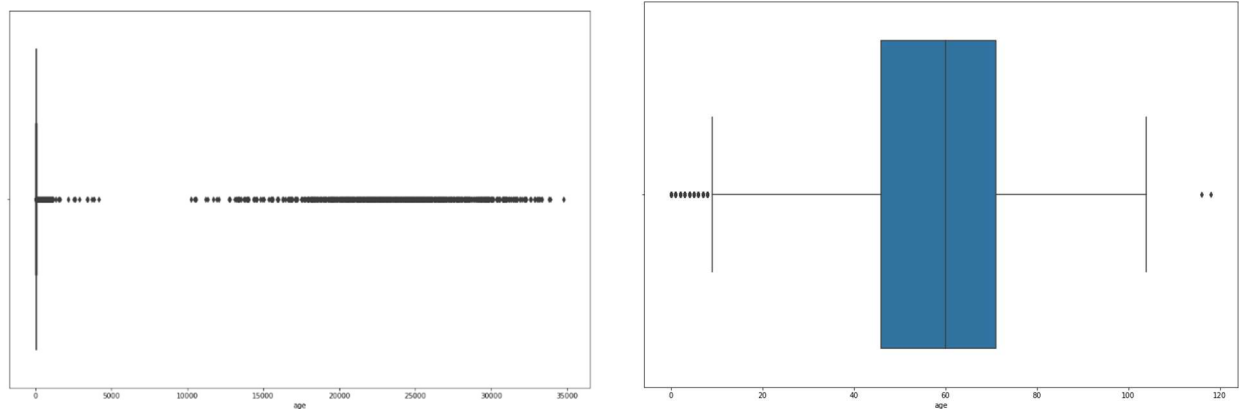


Figure 1: Age variable before and after remove outliers

After Cleaning

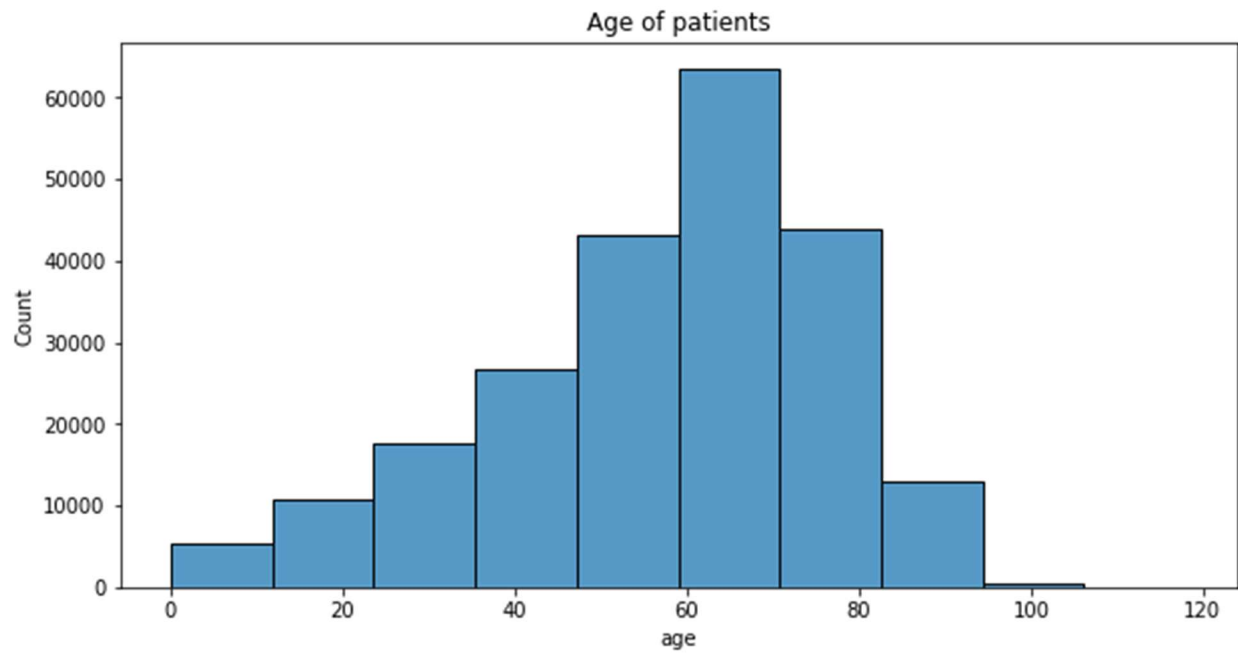


Figure 2:Age variable distribution.

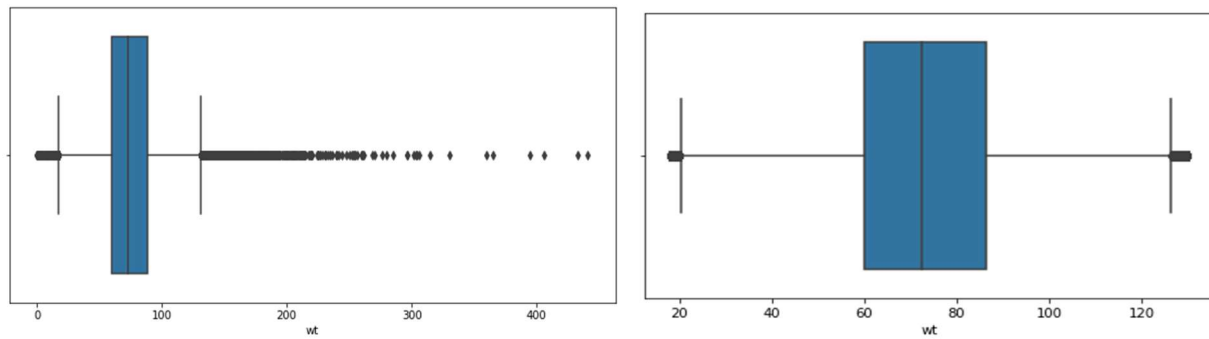


Figure 3: before and after cleaning

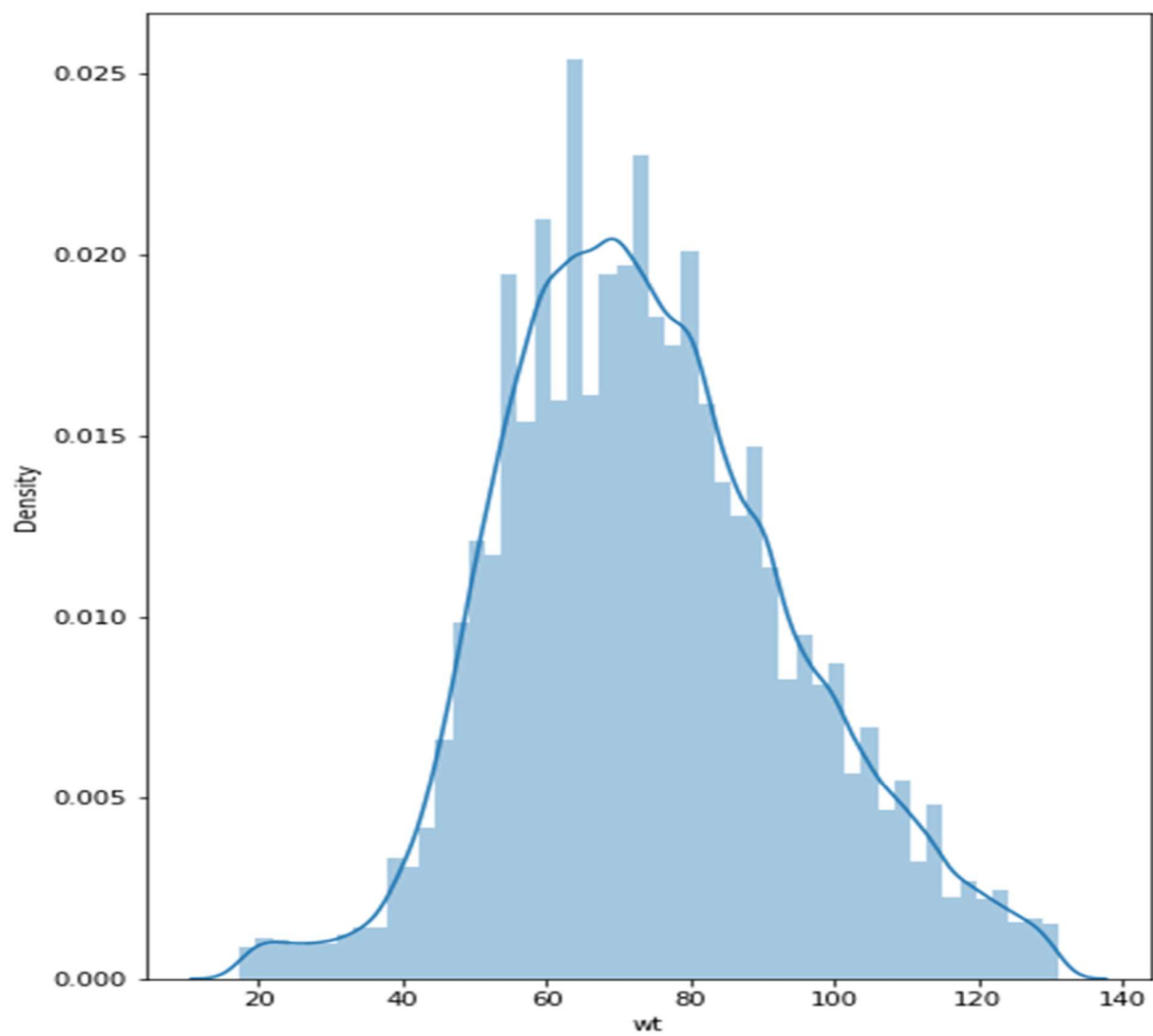


Figure 4:weight variable distribution

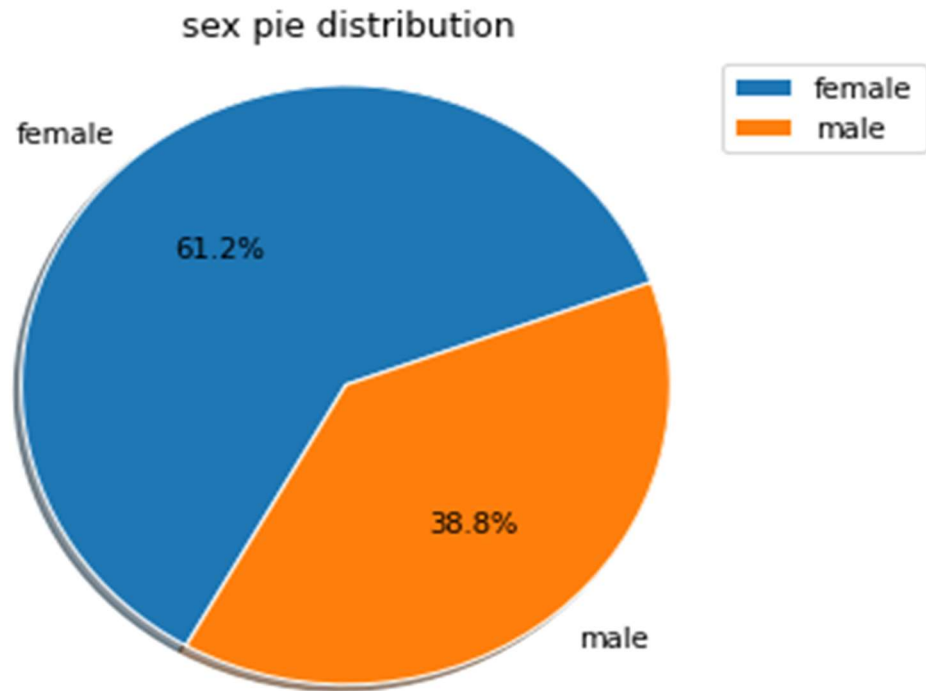


Figure 5:Gender pie chart

Variable Countries (after cleaning outliers): there are a lot of countries and the graph is hard to show those countries so I only pick top 10 of them.

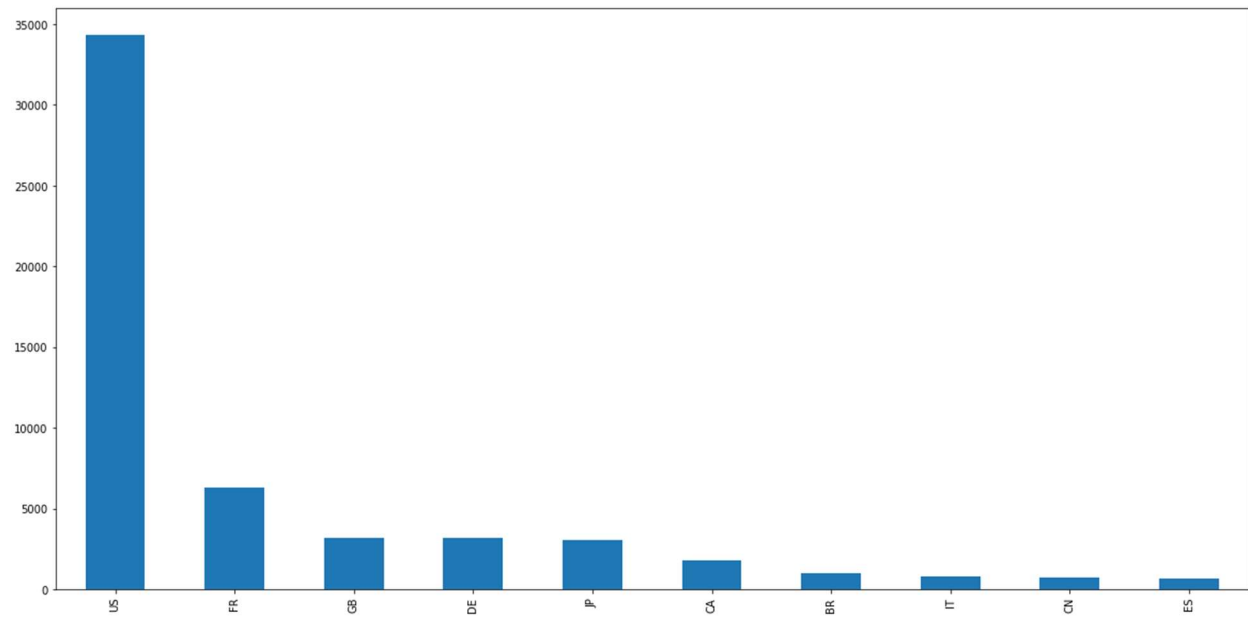


Figure 6: Occurred countries

Unsupervised Learning

CLUSTERING

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

With the Clustering techniques I can showing the other aspect in my project. Imagine we can classify patients just in short time and even they don't need to come to the hospital. All the patients need is filling survey, talking to AI nurses or doctors. After that all the patient's information will be classified as a patient groups. With that group we can determine appropriately treatments or drugs by older patient dataset. It helps hospital times and moneys.

For this project, in order to cluster the patients, I will need to apply two techniques:

1. Hierarchical clustering
2. K-mean clustering
3. DBSCAN

Data preparation:

For Clustering techniques, the data must be numerical. So, I translate all value yes or no into numeric. But they are not just numeric they are category variables. Also due to technical problem, my computer can't process the large number of observations, so I have to take randomly 15000 samples in my dataset and use that for Clustering. And this is my plot.

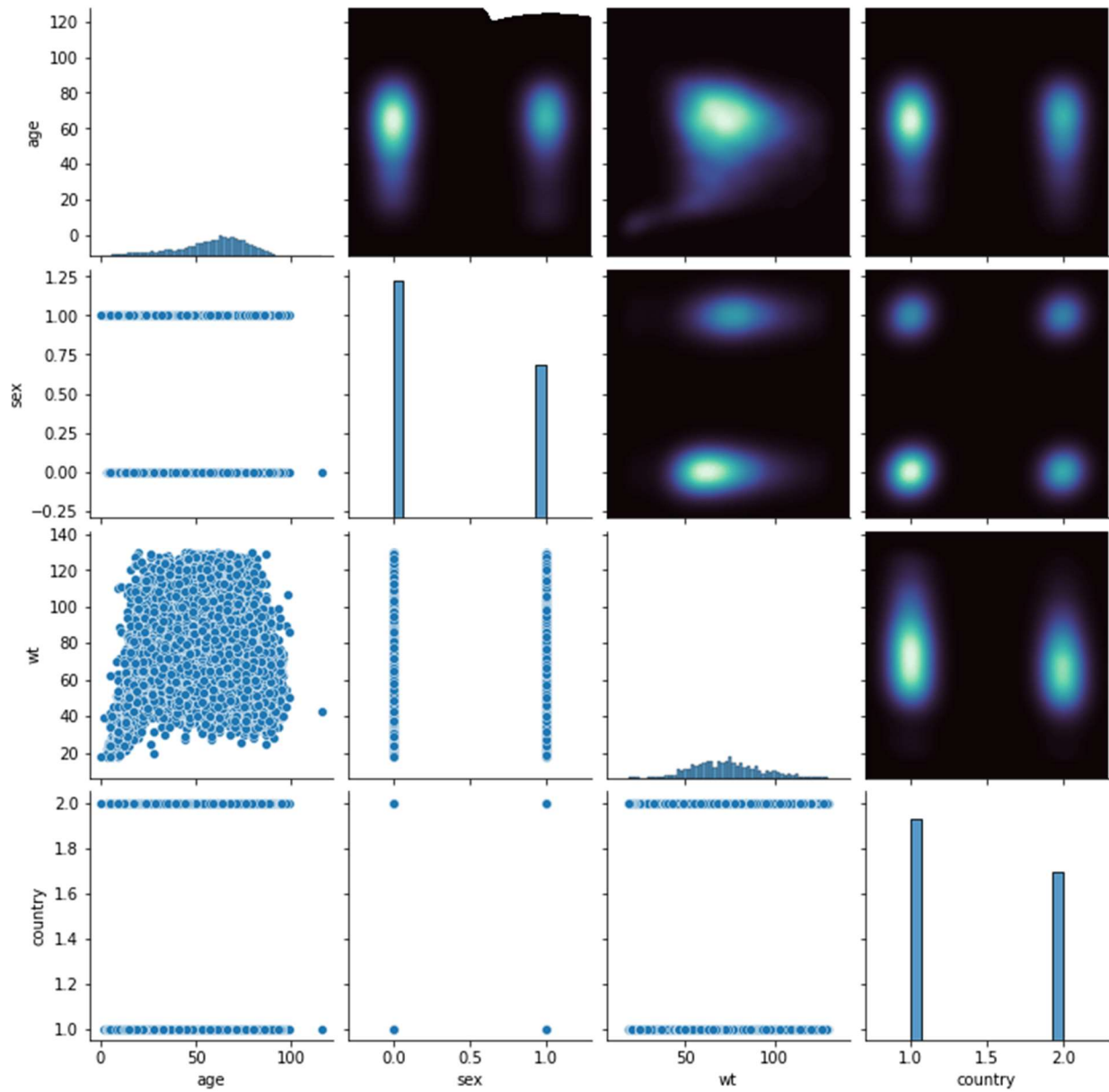


Figure 7: Variables pair plot

DBSCAN CLUSTERING

Density-Based Clustering is a term used to describe unsupervised learning approaches for identifying unique groups/clusters in data. It is based on the premise that a cluster in data space is

a continuous region of high point density that is separated from other clusters by contiguous regions of low point density.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering base technique. It can find clusters of various forms and sizes in a vast quantity of data that is noisy and contains outliers.

minPts: The minimum number of points (a threshold) clustered together for a region to be considered dense.

eps (ϵ): A distance measure that will be used to locate the points in the neighborhood of any point.

For this approach we have to define the parameter. For my project I will have to calculate the Epsilon via KNN method.

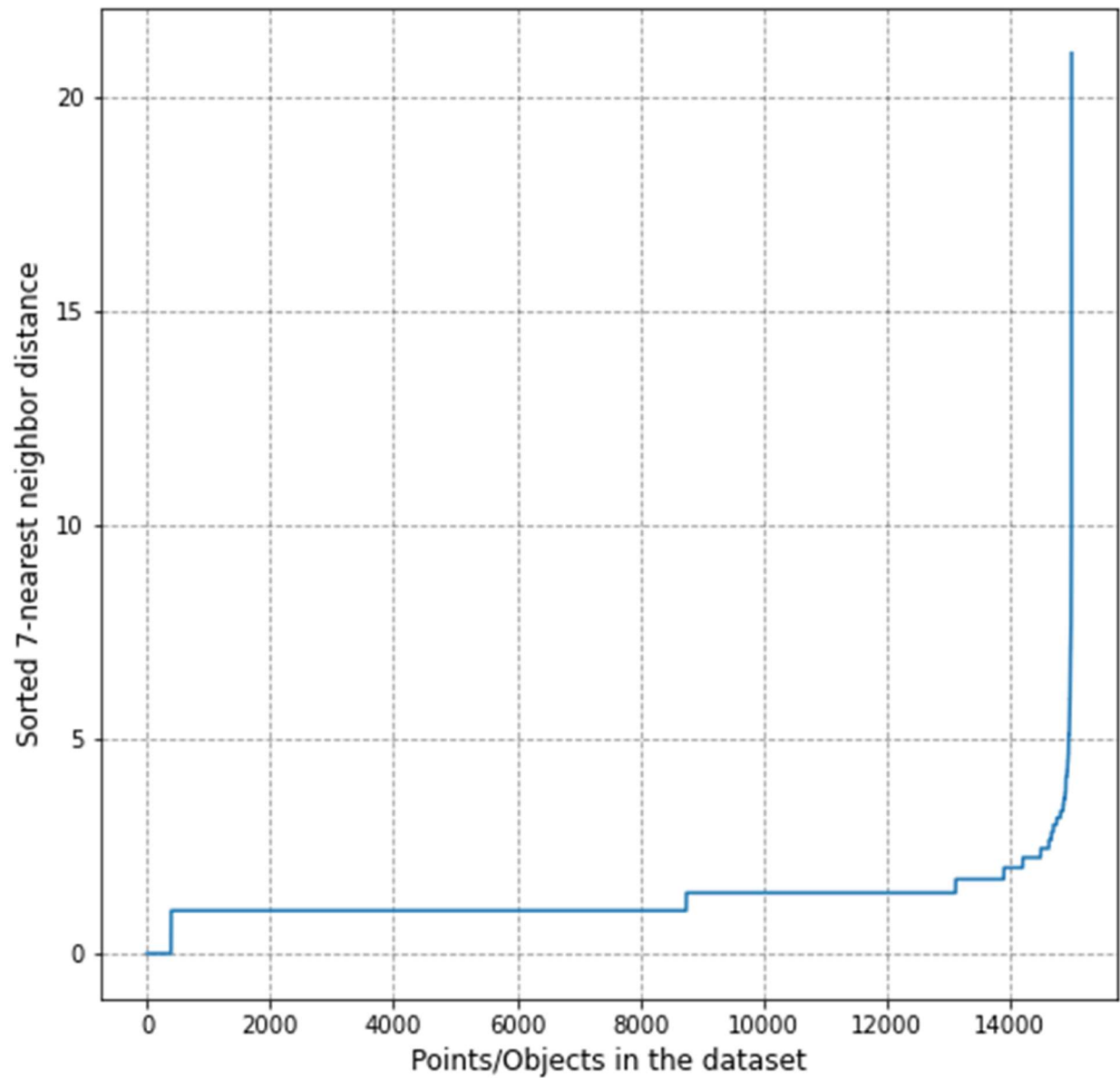


Figure 8:DBSCAN hyperparameter

After defined the Epsilon I input min points as default (5) and I got 3 cluster: $\{-1,0,1,2\}$
-1 is noise cluster, and we don't use that.

HIERARCHICAL CLUSTERING:

Hierarchical clustering, also known as hierarchical cluster analysis, is an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.

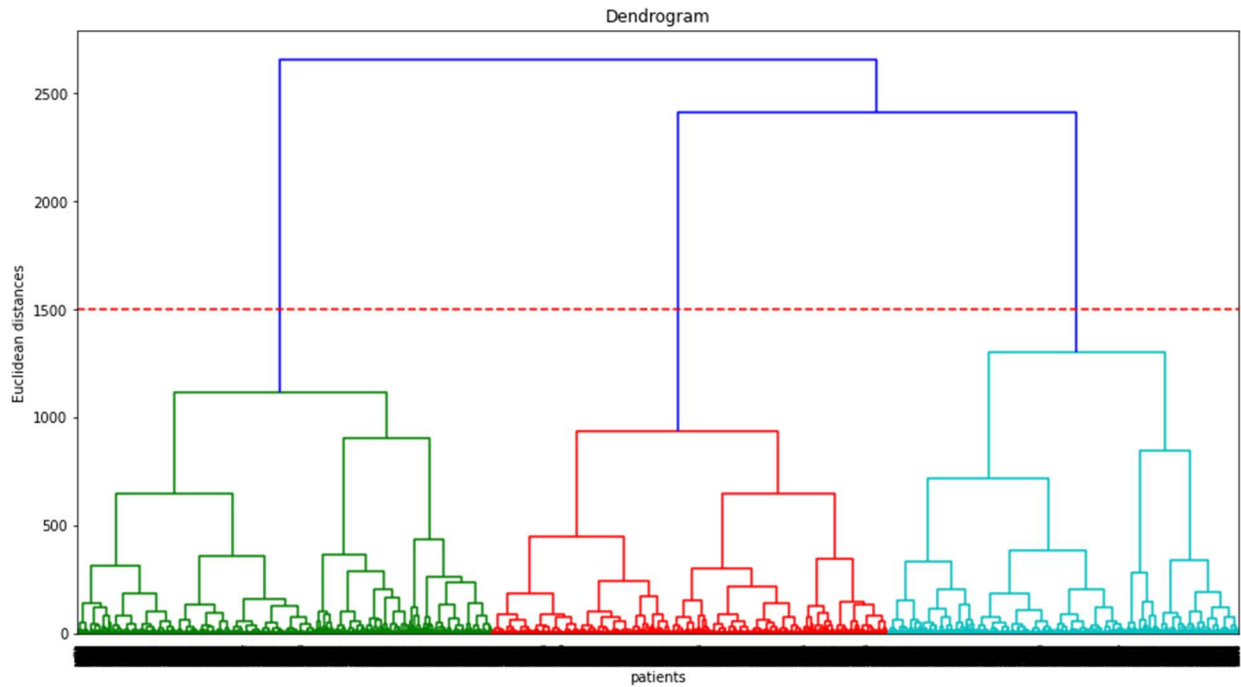


Figure 9: Hierarchical Clustering

Hierarchical Cluster data by best matched pair. that means the smaller pair the bigger similar each of them so in order to get the number of clusters we must check the farthest distance it can go.

We can see clearly 3 clusters in the figure and let move to another step K-means cluster.

K-MEAN CLUSTERING

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. k-means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. For instance, better Euclidean solutions can be found using k-medians and k-medoids.

The problem is computationally difficult (NP-hard); however, efficient heuristic algorithms converge quickly to a local optimum. These are usually similar to the expectation-maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both k-means and Gaussian mixture modeling. They both use cluster centers to model the data; however, k-means clustering tends to find clusters of comparable spatial extent, while the Gaussian mixture model allows clusters to have different shapes.

The unsupervised k-means algorithm has a loose relationship to the k-nearest neighbor classifier, a popular supervised machine learning technique for classification that is often confused with k-means due to the name. Applying the 1-nearest neighbor classifier to the cluster centers obtained by k-means classifies new data into the existing clusters. This is known as nearest centroid classifier or Rocchio algorithm.

For K-means cluster in order to define the number of cluster we have to apply elbow techniques to our dataset.

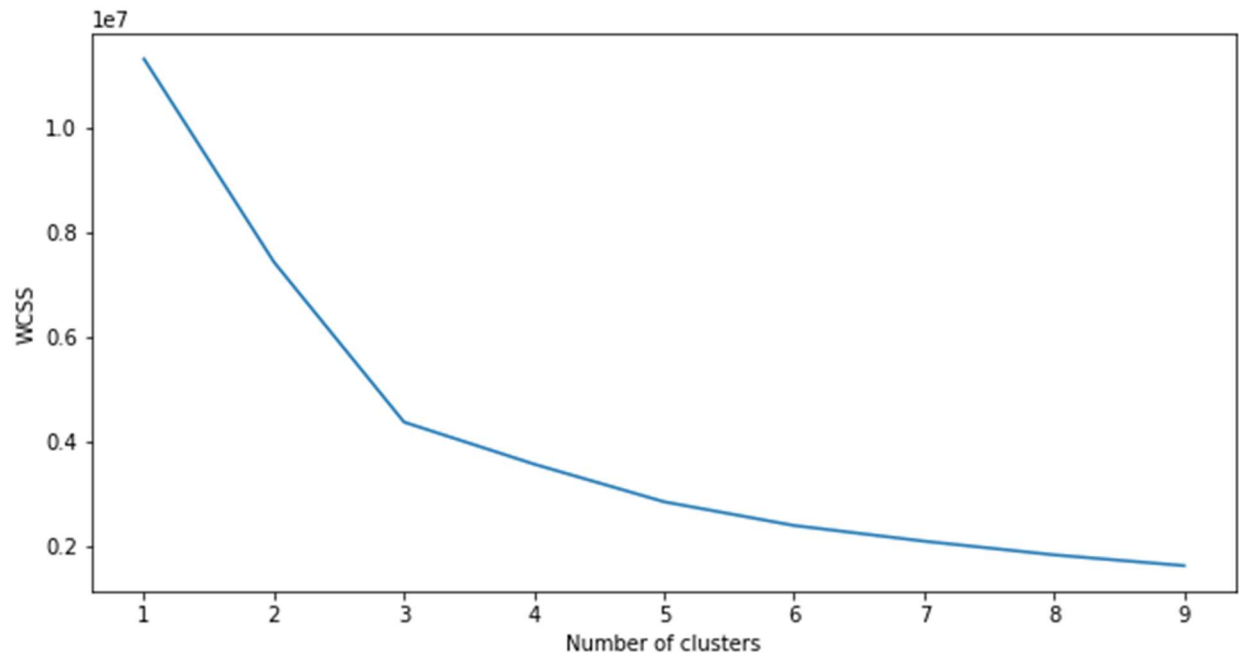


Figure 10: K-mean Elbow method

So, the answer of elbow also same with Hierarchical technique so I conclude my dataset has 3 cluster groups

-1 is noise cluster, and we don't use that

Table 2:Cluster Centers

	Age	sex	weight	Country
Cluster0	69.43821544	0.33563252	64.054887	1.47886704
Cluster1	30.90448286	0.3105772	61.21828304	1.48256666
Cluster2	58.57867169	0.52964592	96.12109784	1.32369579

This is how 3 cluster separate my dataset in variable Age & weight graph (the reason I use age and weight because they are numerical variable to compare other binary values).

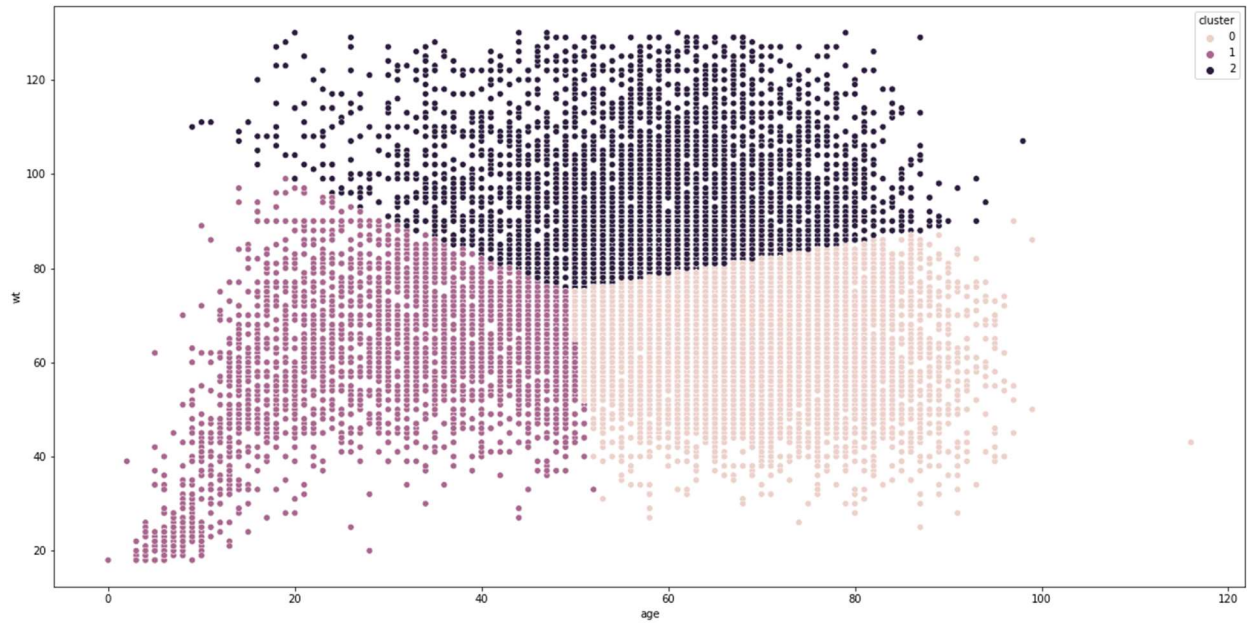


Figure 11: Age and weight scatter plot separated by Cluster

To see other variables, I need to use 3-dimension plot:

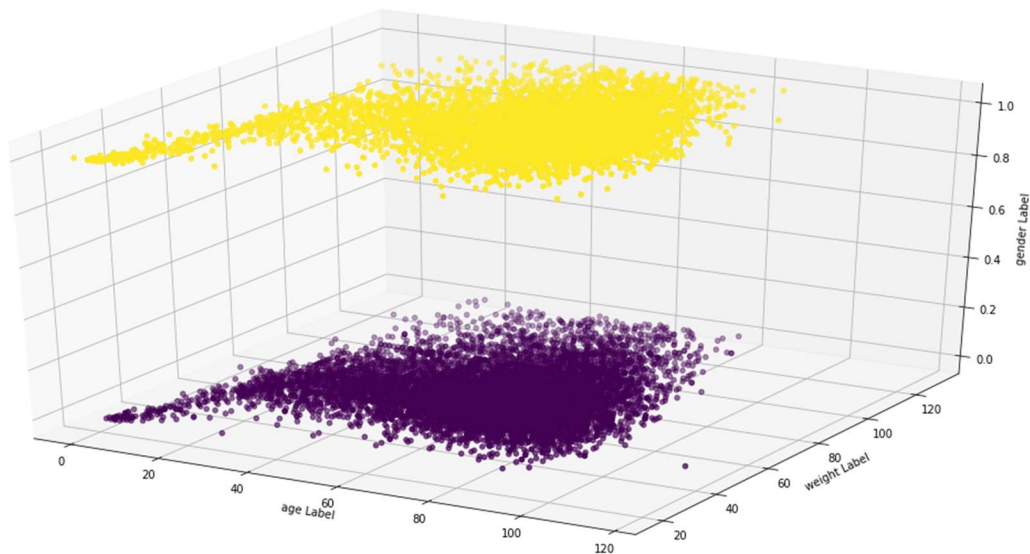


Figure 12: 3 Dimension plot for age, weight and gender

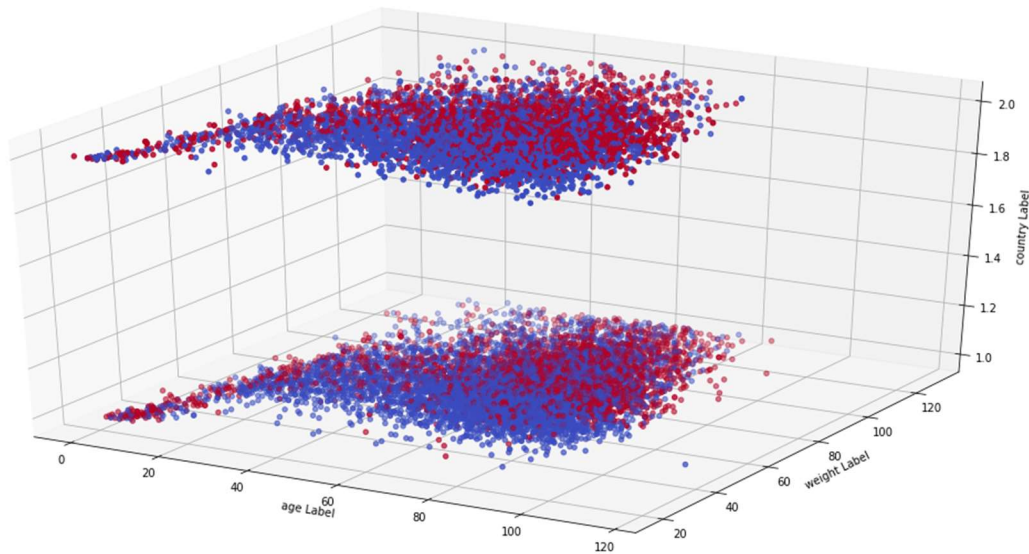


Figure 13: Plot for age, weight, country

Because my dataset is lacking a lot of information so I couldn't provide you a better view of the patient's demographic. I was expecting I can find another dataset similar to this one and match them together, but they are all confidential. So, I have to exploit as much as I can for this one.

Supervised Learning

Supervised learning (SL) is the machine learning task of learning a function that maps an input to an output based on example input-output pairs.[1] It infers a function from labeled training data consisting of a set of training examples.[2] In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way

(see inductive bias). This statistical quality of an algorithm is measured through the so-called generalization error.

I will apply 4 models to this supervised learning techniques:

- Decision Tree classification
- Random Forest Classification
- Logistic Regression
- K Nearest Neighbor

Data Preparation:

So, for this specific modeling I will need to prepare my datasets. I merged 3 different datasets into one and use that to training.

Variable Name	Definition	Value	Description
primary	Patient ID	Unstructured values	
age	age	Numerical	
sex	gender	Categorical	
wt	weight	Numerical	
country	Patient's Country	Categorical	

cluster	Patient group cluster	Categorical	
drugname	Drug name	Categorical	
prod_ai	Drug ingredient	Categorical	
Death	binary	Categorical	

Drug name and prod_ai were come from drug dataset

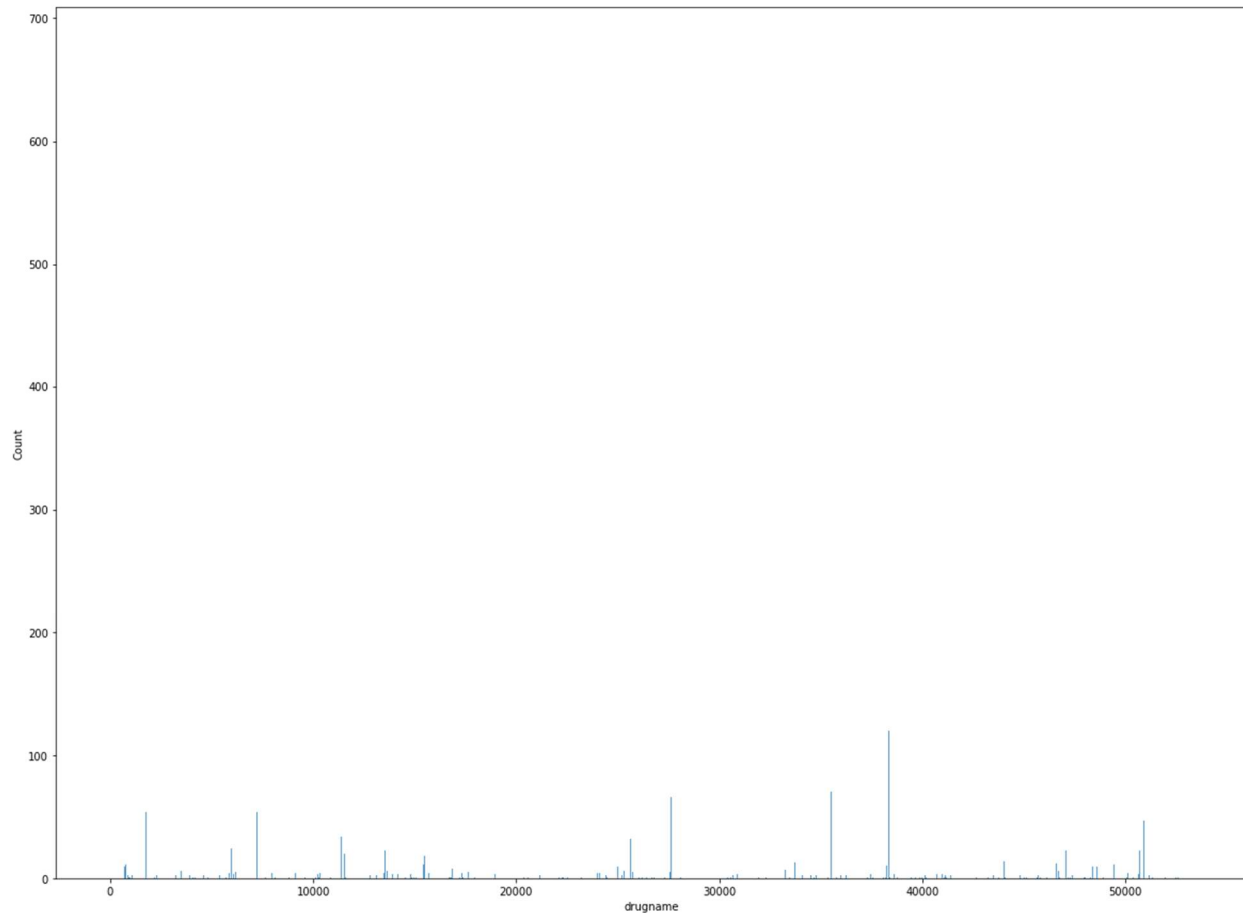
Death comes from death event (dummy variable)

The reason I put drugname and prod_ai as categorical because I will need that to train my logistic classification they are not accept string.

Also, I used “cat. Categories” from pandas library to translate them into the number

For Drug categorical it has 53332 category values

And Prod_ai has 5334 category values.



Drug name plot.

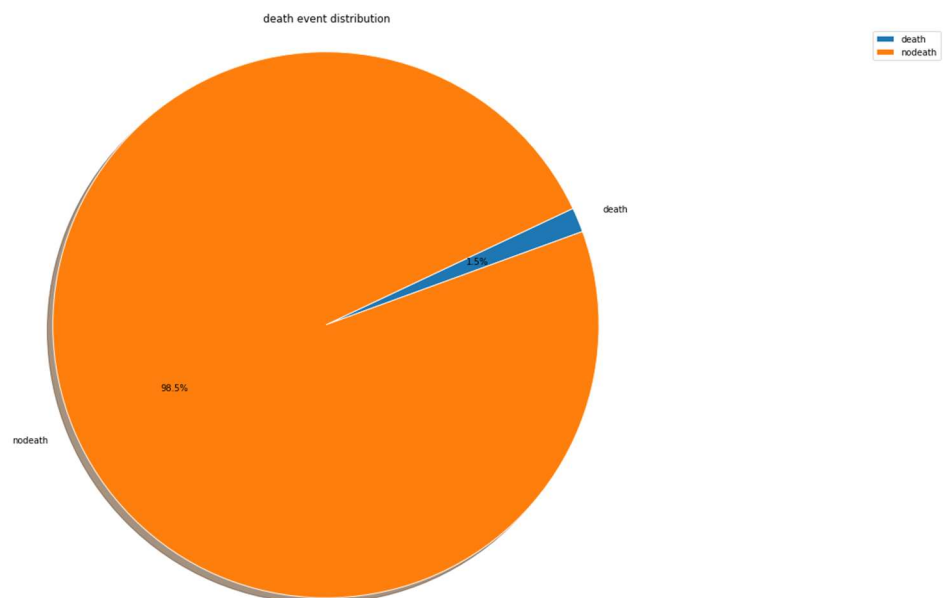
Cross-validation techniques:

Cross validation is a model evaluation method that is better than residuals. The problem with residual evaluations is that they do not give an indication of how well the learner will do when it is asked to make new predictions for data it has not already seen. One way to overcome this problem is to not use the entire data set when training a learner. Some of the data is removed before training begins. Then when training is done, the data that was removed can be used to test

the performance of the learned model on ``new" data. This is the basic idea for a whole class of model evaluation methods called cross validation.

I have a burden on this step because my dataset is imbalanced. The patients dead is only 1.5% of the data so I will need to over sampling techniques (I will explain more about them in below) but I also use Train test split techniques(traditional one) for comparison

A/Train test split:



TRAIN TEST SPLIT – 5KFOLD

*****Logistic Regression

The accuracy: 0.9838310509816862

The Classification Report:

	precision	recall	f1-score	support
0	0.98	1.00	0.99	17889
1	0.00	0.00	0.00	294
accuracy			0.98	18183
macro avg	0.49	0.50	0.50	18183
weighted avg	0.97	0.98	0.98	18183

*****KNeighbors

The accuracy: 0.9835560688555244

The Classification Report:

	precision	recall	f1-score	support
0	0.98	1.00	0.99	17889
1	0.27	0.01	0.02	294

accuracy		0.98	18183
macro avg	0.63	0.50	0.51 18183
weighted avg	0.97	0.98	0.98 18183

*****DecisionTree

The accuracy: 0.9784414013089149

The Classification Report:

	precision	recall	f1-score	support
0	0.99	0.99	0.99	17889
1	0.29	0.23	0.26	294

accuracy		0.98	18183
macro avg	0.64	0.61	0.62 18183
weighted avg	0.98	0.98	0.98 18183

*****RandomForest

The accuracy: 0.9797063190892592

The Classification Report:

	precision	recall	f1-score	support
0	0.99	0.99	0.99	17889
1	0.28	0.16	0.21	294
accuracy			0.98	18183
macro avg	0.63	0.58	0.60	18183
weighted avg	0.97	0.98	0.98	18183

*****XGBoost

	precision	recall	f1-score	support
0	0.98	1.00	0.99	17889
1	0.34	0.06	0.11	294
accuracy			0.98	18183
macro avg	0.66	0.53	0.55	18183
weighted avg	0.97	0.98	0.98	18183

For Imbalance dataset We need to use F1 score or ROC, AUC scores to evaluate models.

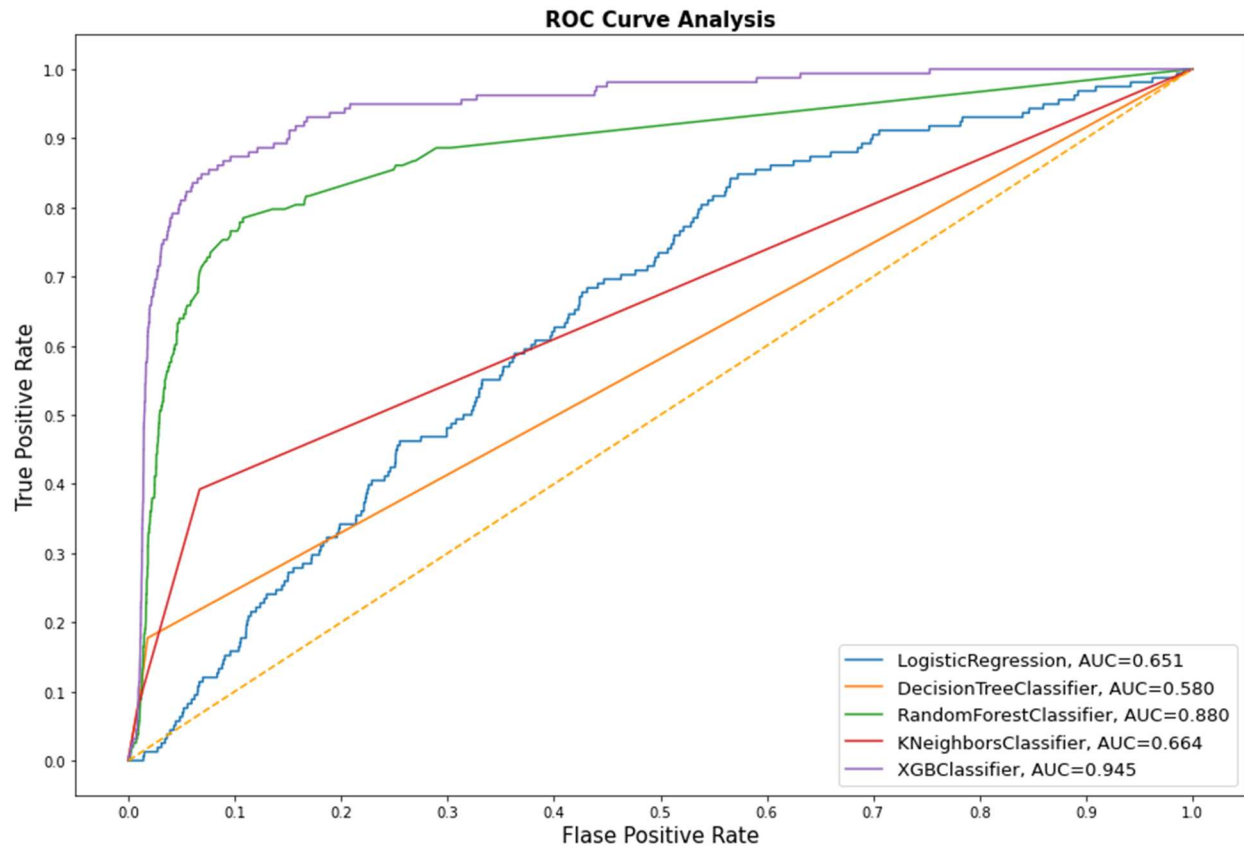


Figure 14:AUC scores of 5-K Fold cross validation

SMOTE(OVER SAMPLING)

*****Logistic Regression

precision recall f1-score support

0	0.99	0.60	0.74	8861
1	0.03	0.61	0.05	158

accuracy			0.60	9019
macro avg	0.51	0.60	0.40	9019
weighted avg	0.97	0.60	0.73	9019

*****KNN :3 K- neighbor

	precision	recall	f1-score	support
0	0.99	0.91	0.95	8861
1	0.08	0.40	0.13	158
accuracy			0.91	9019

macro avg	0.53	0.66	0.54	9019
weighted avg	0.97	0.91	0.94	9019

*****Random Forest

	precision	recall	f1-score	support
0	0.98	0.98	0.98	8861
1	0.11	0.15	0.13	158
accuracy			0.97	9019
macro avg	0.55	0.56	0.56	9019
weighted avg	0.97	0.97	0.97	9019

*****Decision Tree

	precision	recall	f1-score	support
0	0.98	0.97	0.98	8861
1	0.09	0.16	0.12	158
accuracy			0.96	9019
macro avg	0.54	0.57	0.55	9019
weighted avg	0.97	0.96	0.96	9019

*****XGBoost

	precision	recall	f1-score	support
0	0.99	0.98	0.99	17889
1	0.32	0.48	0.39	294
accuracy			0.98	18183
macro avg	0.66	0.73	0.69	18183
weighted avg	0.98	0.98	0.98	18183

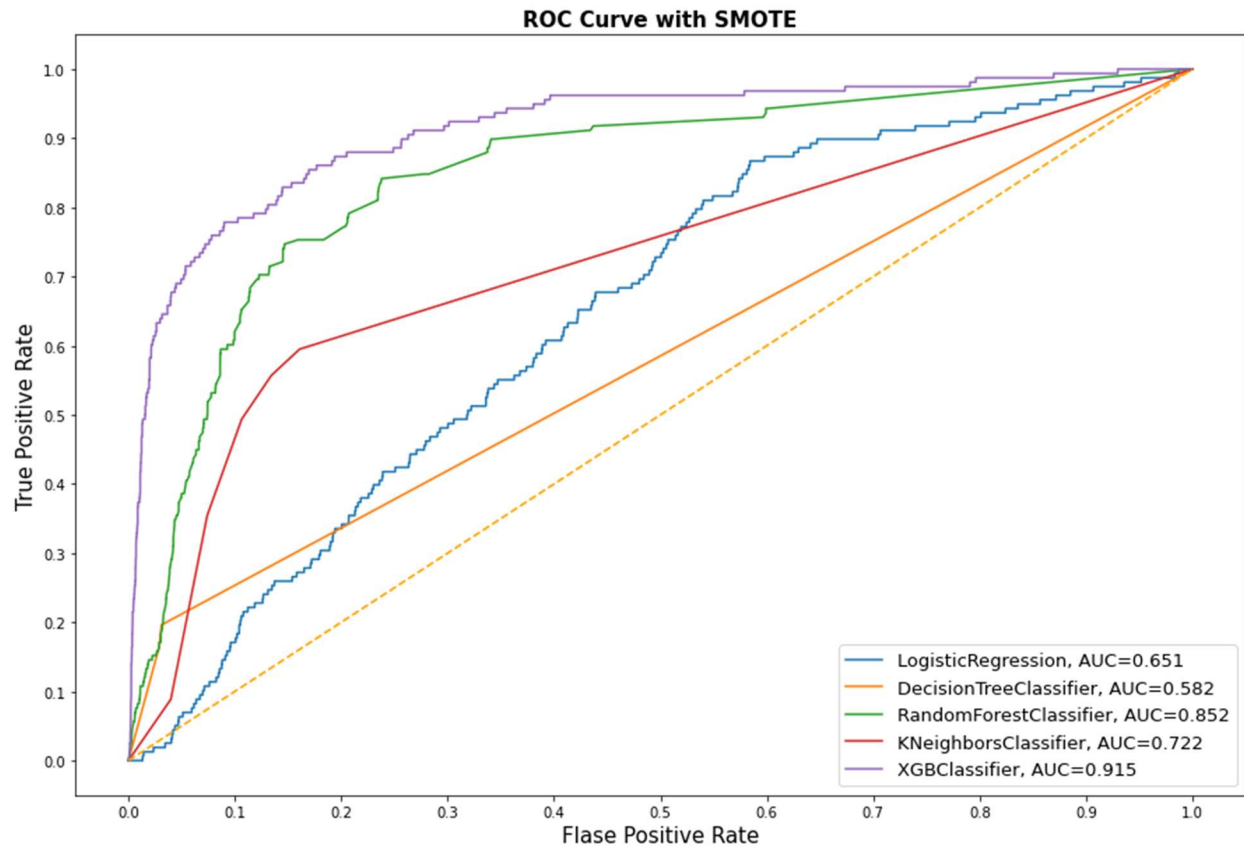


Figure 15: SMOTE ROC_AUC scores

SMOTEENN

*****Logistic Regreeion

	precision	recall	f1-score	support	
0		0.99	0.56	0.71	8861
1		0.03	0.67	0.05	158
accuracy					0.56 9019
macro avg	0.51	0.61	0.38		9019
weighted avg	0.97	0.56	0.70		9019

*****Decision Tree

	precision	recall	f1-score	support
0	0.99	0.95	0.97	8861
1	0.16	0.52	0.24	158
accuracy		0.94		9019
macro avg	0.57	0.73	0.60	9019
weighted avg	0.98	0.94	0.96	9019

*****Random Forest

precision recall f1-score support

0 0.99 0.95 0.97 8861

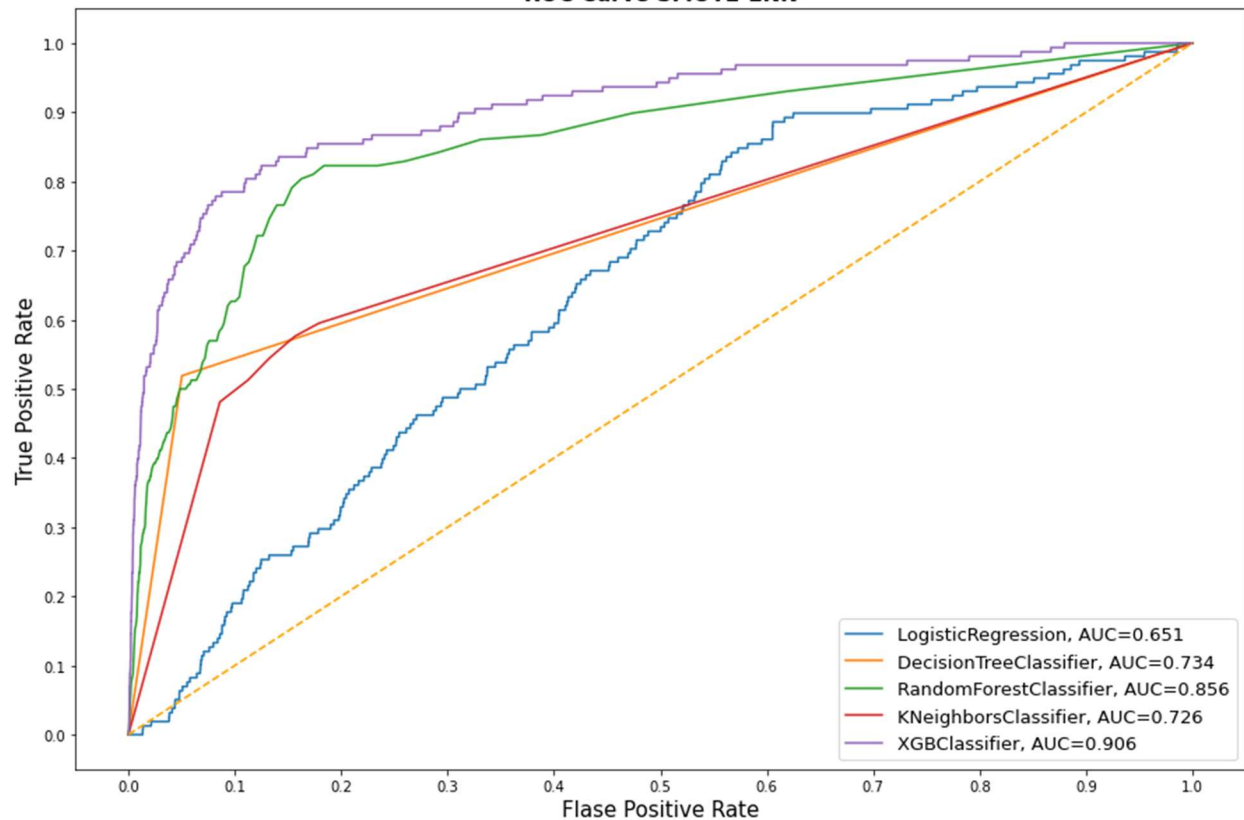
1 0.15 0.46 0.23 158

accuracy 0.95 9019

macro avg 0.57 0.71 0.60 9019

weighted avg 0.98 0.95 0.96 9019

ROC Curve SMOTE-ENN



EVALUATING

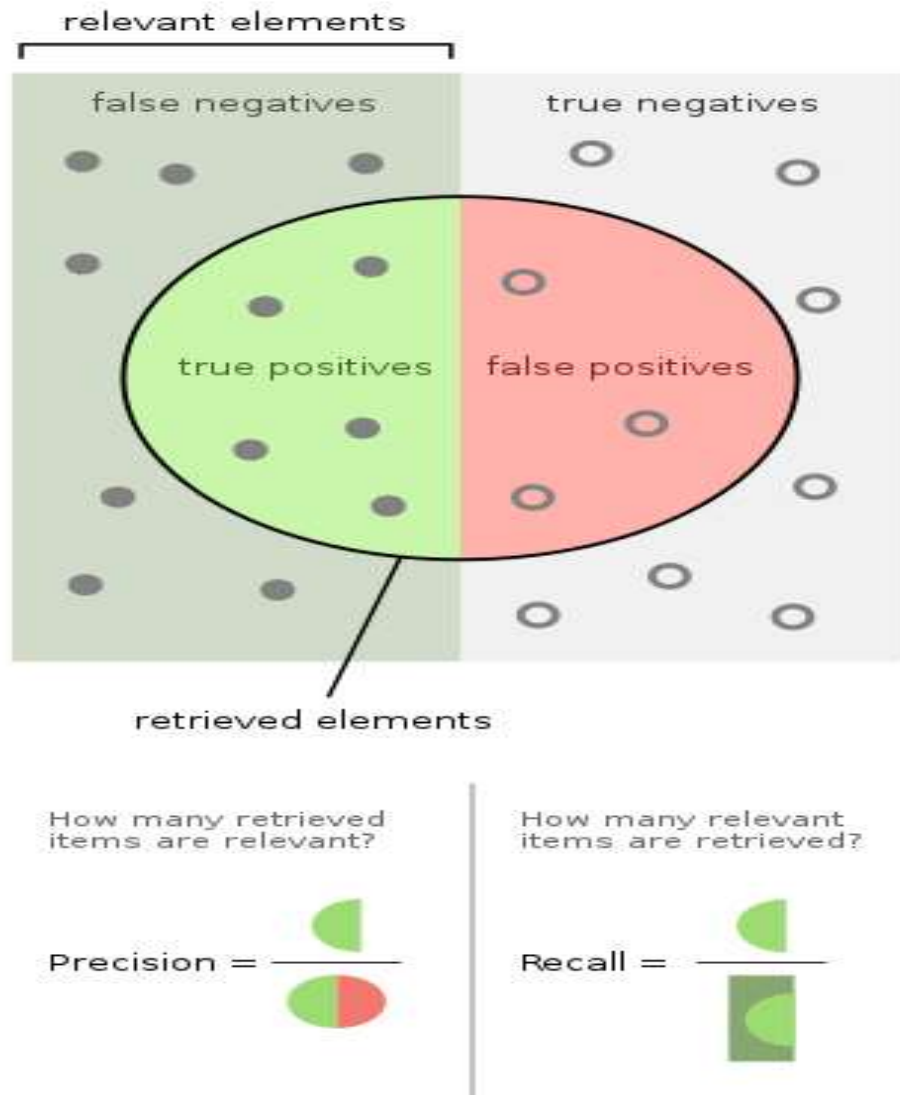


Figure 16: Precision & Recall

Precision is the ratio between the True Positives and all the Positives. how many correctly predicting we have made in our Model.

The recall is the measure of our model correctly identifying True Positives

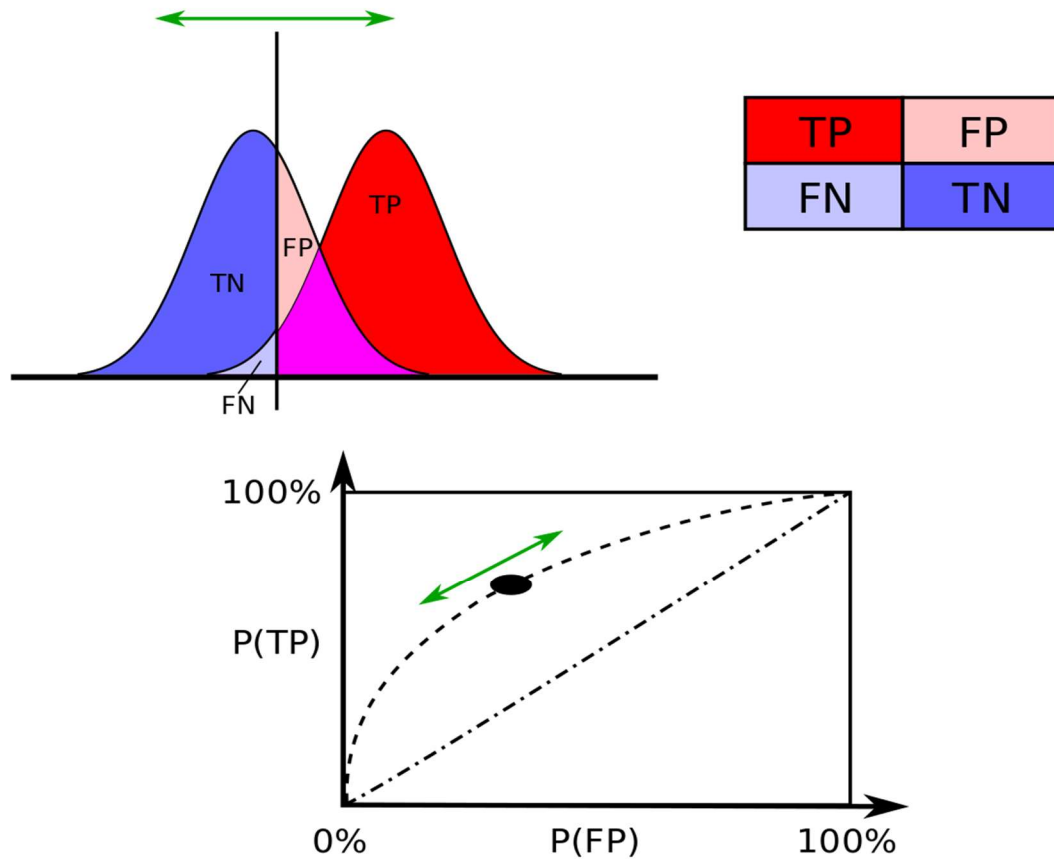


Figure 17: ROC & AUC scoring

AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1. By analogy, the Higher the AUC, the better the model is at distinguishing between patients with the disease and no disease.

CONCLUSION

In this conclusion I want to finalize my project by answer 2 questions I have made from Methodologies step.

First question was finding the most side effect reaction after patients consumed drug and the drug was cause that reaction?

- The Most frequently reaction: Death
- Name of the Drug was causing patients Death: Opana ER.

Second question what other causes lead to that reaction is: age, weight, drug name, drug materials:

In order to answer this question, we need to build a prediction model which can help us to predict the future patients. We want to predict that what patients are on the risk of death and give them the appropriate treatment or drugs. For answering the 2nd question, I will give you my best prediction model I have made at the modeling part.

ROC Curve Analysis

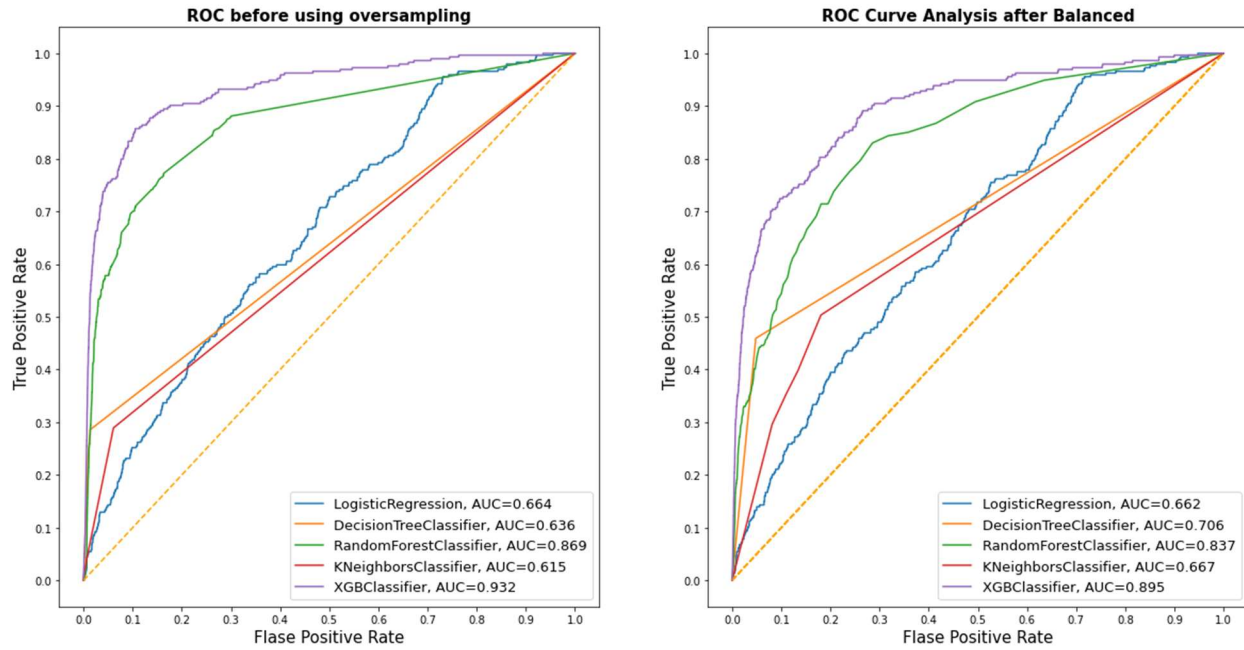


Figure 18: Cross validation techniques compare between 5 K fold and Oversampling techniques.

The Extremely Gradient boosting classification combine with 5-k fold was the best model. It gives me 93 AUC score.

REFERENCES

- (n.d.). US Food and Drug Administration. Retrieved April 21, 2022, from <https://www.fda.gov>
- Chauhan, N. S. (2020, April 27). *DBSCAN Clustering Algorithm in Machine Learning*. KDnuggets. Retrieved April 21, 2022, from <https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html>
- Clustering | Types Of Clustering | Clustering Applications*. (2016, November 3). Analytics Vidhya. Retrieved April 21, 2022, from <https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/>
- Does an unbalanced sample matter when doing logistic regression?* (2011, January 7). Cross Validated. Retrieved April 21, 2022, from <https://stats.stackexchange.com/questions/6067/does-an-unbalanced-sample-matter-when-doing-logistic-regression>
- FDA Adverse Event Reporting System (FAERS) Public Dashboard | FDA*. (2021, October 22). US Food and Drug Administration. Retrieved April 21, 2022, from <https://www.fda.gov/drugs/questions-and-answers-fdas-adverse-event-reporting-system-faers/fda-adverse-event-reporting-system-faers-public-dashboard>
- FDA requests removal of Opana ER for risks related to abuse | FDA*. (2017, June 8). US Food and Drug Administration. Retrieved April 21, 2022, from <https://www.fda.gov/news-events/press-announcements/fda-requests-removal-opana-er-risks-related-abuse>
- Getting Started with Hierarchical Clustering in Python | Engineering Education (EngEd) Program*. (2021, December 15). Section.io. Retrieved April 21, 2022, from <https://www.section.io/engineering-education/hierarchical-clustering-in-python/>

- Gould, D. (2021, April 7). *Beginner's Guide to XGBoost for Classification Problems*. Towards Data Science. Retrieved April 21, 2022, from <https://towardsdatascience.com/beginners-guide-to-xgboost-for-classification-problems-50f75aac5390>
- k-means clustering*. (n.d.). Wikipedia. Retrieved April 21, 2022, from https://en.wikipedia.org/wiki/K-means_clustering
- Magee, J. F. (n.d.). *Decision Trees for Decision Making*. Harvard Business Review. Retrieved April 21, 2022, from <https://hbr.org/1964/07/decision-trees-for-decision-making>
- Potential Signals of Serious Risks/New Safety Information Identified from the FDA Adverse Event Reporting System (FAERS) | FDA*. (2022, April 6). US Food and Drug Administration. Retrieved April 21, 2022, from <https://www.fda.gov/drugs/questions-and-answers-fdas-adverse-event-reporting-system-faers/potential-signals-serious-risksnew-safety-information-identified-fda-adverse-event-reporting-system>
- sklearn.cluster.KMeans — scikit-learn 1.0.2 documentation*. (n.d.). Scikit-learn. Retrieved April 21, 2022, from <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>
- sklearn.neighbors.KNeighborsClassifier — scikit-learn 1.0.2 documentation*. (n.d.). Scikit-learn. Retrieved April 21, 2022, from <https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html>
- Yadav, D. (2020, April 14). *Weighted Logistic Regression for Imbalanced Dataset | by Dinesh Yadav*. Towards Data Science. Retrieved April 21, 2022, from <https://towardsdatascience.com/weighted-logistic-regression-for-imbalanced-dataset-9a5cd88e68b>