PORTFOLIO

Pushpa

2022-11-27

Contents

1 Introduction

The vacation planning for India travel budget and expenses.Reducing the dimensionality of the Breast cancer factors by using principal component analysis and nonmetrics MDS. Predict the classification of the Kneighbors classifier, Gaussian classifier, and Neural network.machine learning algorithms will be able to detect the brain stroke effected or not effected. The prediction accuracy of the model will be analyzed by the end of this project. This project will provide insight of machine learning algorithms to perform classification on imbalanced dataset problems.

2 Vacation Planning for my Motherland this summer!!

2.1 Introduction

Since the pandemic struck, we could not go to my home country India, the vacation was long due and we planned the vacation in detail so as to be an experience that I will never forget, we planned to visit my parents, siblings, family, and friends and to explore the world of shopping and great street foods. The budget was laid out along with the timeline of the vacation.

2.2 Planning and Budget

Summer vacation is here it's coming, the planning of the summer vacation started way back in the spring. It was a time to enjoy with family and friends back in my hometown in India, meet my relatives' friends and enjoy some food which is unique in its taste and spicy levels. Before summer hits the meticulous planning of the summer vacation started with the below details

1.Departure timings of the flight, 2.Stay at Dubai 3.Departure from Dubai 4.Arrival in India 5.Stay at Bangalore 6.Going to my Hometown. 7.Making arrangements for the family functions 8.Shopping for the family functions 9.Travel of other cities 10.Flying back to Houston All this required proper planning, and a timeline budget to execute the vacation plans hence I detailed out the timeline on the budget required for the vacation. With limited time at my disposal, it was important for me to make the most of it and ensure you don't miss any once-in-a-lifetime experiences. A vacation itinerary helped me in planning out my vacation — to the day or even to the hour — just like a business travel itinerary helps you schedule your work trips. Since I was traveling in different international countries dealing with foreign currency, and expenditures would quickly go out of hand it was important for the proper budget to execute the vacation plan. Once I had my plan and budget ready, now it was important to pack my stuff including medicines which would make my vacation less worry. I prepared a list of the things that I need to carry on my vacation. A daily plan was created so that I can utilize the maximum time of the vacation, below is the template which I used. The daily plan enables you to plan your route in the way that makes the most sense and make sure you don't miss anything important. It also has space for meal planning, so no one gets hungry, and the weather forecast, so everyone dresses appropriately.

As there were some road trips during my vacation, I had planned out the trips and the below details • Name of the Taxi Company • Daily Schedule • Contact information • Other information such as restaurants, accommodation, and meeting points This ensures that everyone has the same information and minimizes the risk of the family getting separated.

I took the help of Monday.com to create my plan so that I could track the activities, some of the features are listed below, Monday.com's Work OS allows you to build and organize projects using customizable columns, statuses, automation, and more. Since each trip is unique, this makes it the perfect solution for planning and managing tailor-made travel itineraries. For example, you can use monday.com's project management features to manage each vacation, creating a workflow of tasks that span from the pre-departure preparations — such as visas or vaccinations — to the post-trip vacation feedback form. Suppose you're planning several trips at once. In that case, you can use the Board and Dashboard views to visualize the data of one or all your trips simultaneously, allowing you to track the expenses against the budget and reallocate resources where necessary.

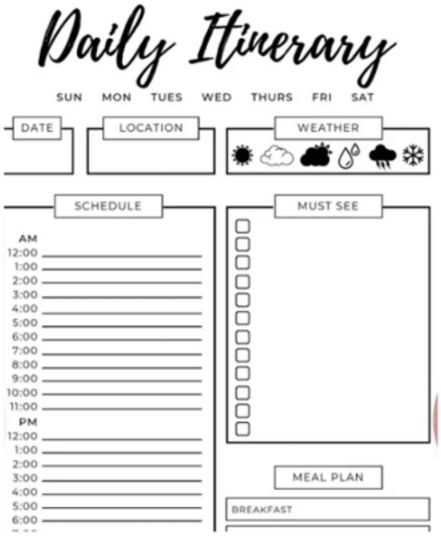
Once plotted, you can then view it on the Map view, so it's easy to reach all your destinations on time,

Budget for our International travel:

Budget for our travel at Dubai:

Budget for travel in India:

Budget for the family functions



Daily Itinerary

SUN MON TUES WED THURS FRI SAT

DATE

LOCATION

WEATHER

SCHEDULE

AM
12:00
1:00
2:00
3:00
4:00
5:00
6:00
7:00
8:00
9:00
10:00
11:00
PM
12:00
1:00
2:00
3:00
4:00
5:00
6:00

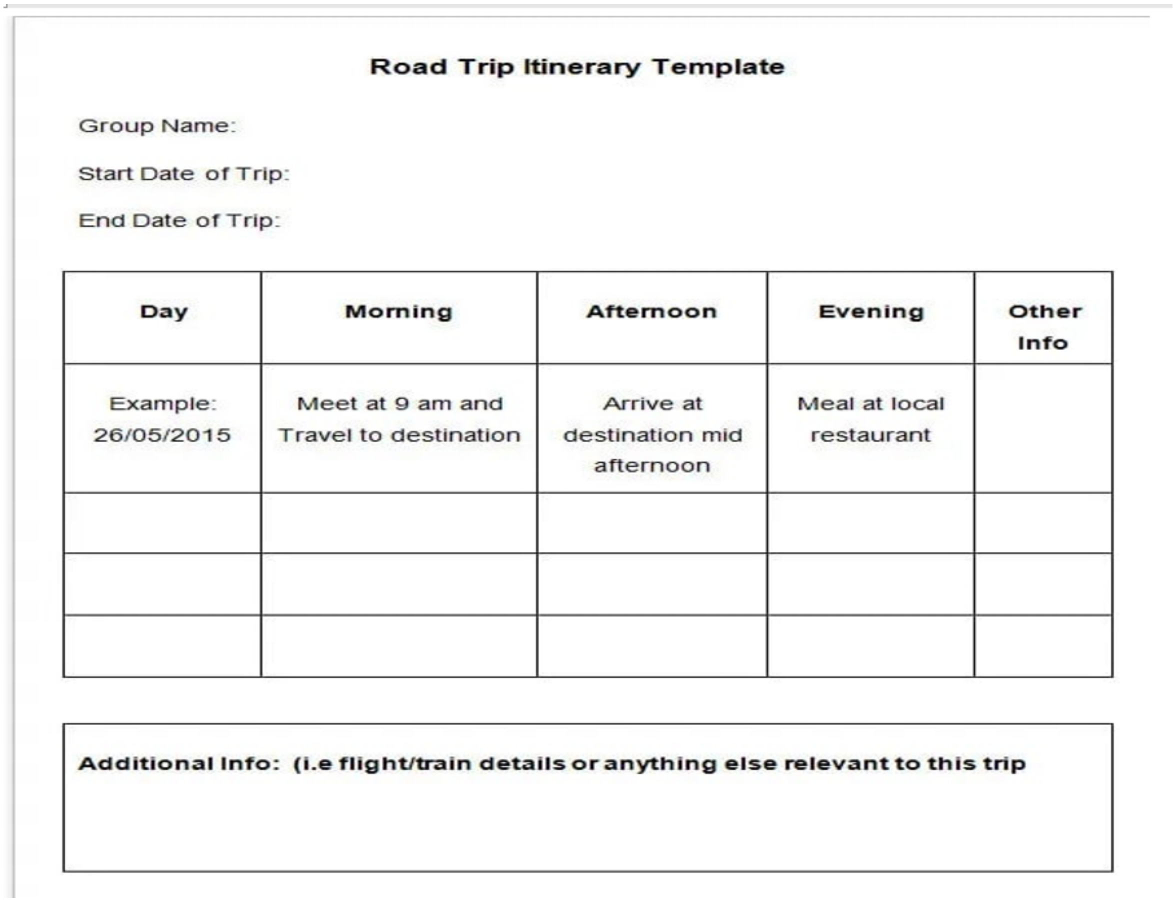
MUST SEE

□
□
□
□
□
□
□
□
□
□
□
□

MEAL PLAN

BREAKFAST

Figure 1: Figure



Road Trip Itinerary Template

Group Name:

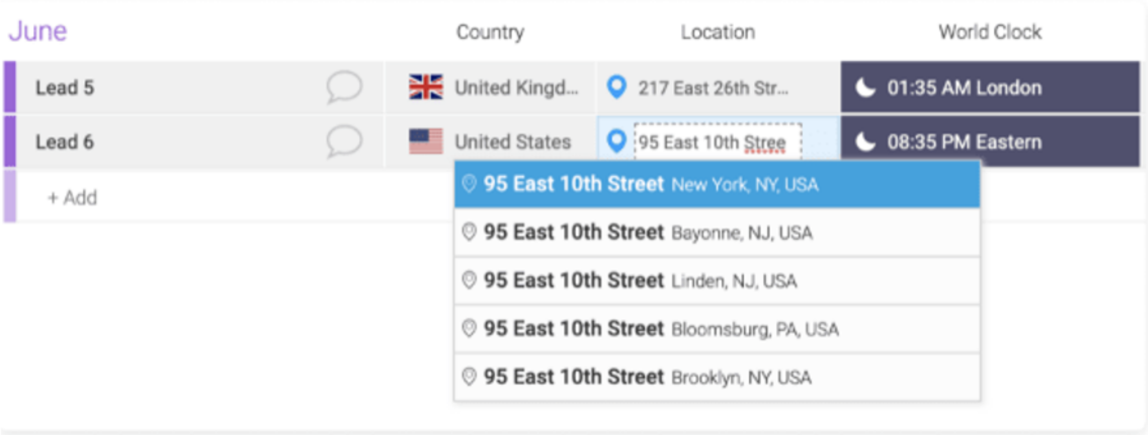
Start Date of Trip:

End Date of Trip:

Day	Morning	Afternoon	Evening	Other Info
Example: 26/05/2015	Meet at 9 am and Travel to destination	Arrive at destination mid afternoon	Meal at local restaurant	

Additional Info: (i.e flight/train details or anything else relevant to this trip

Figure 2: Figure



June	Country	Location	World Clock
Lead 5	United Kingd...	217 East 26th Str...	01:35 AM London
Lead 6	United States	95 East 10th Stree	08:35 PM Eastern
+ Add			

95 East 10th Street New York, NY, USA

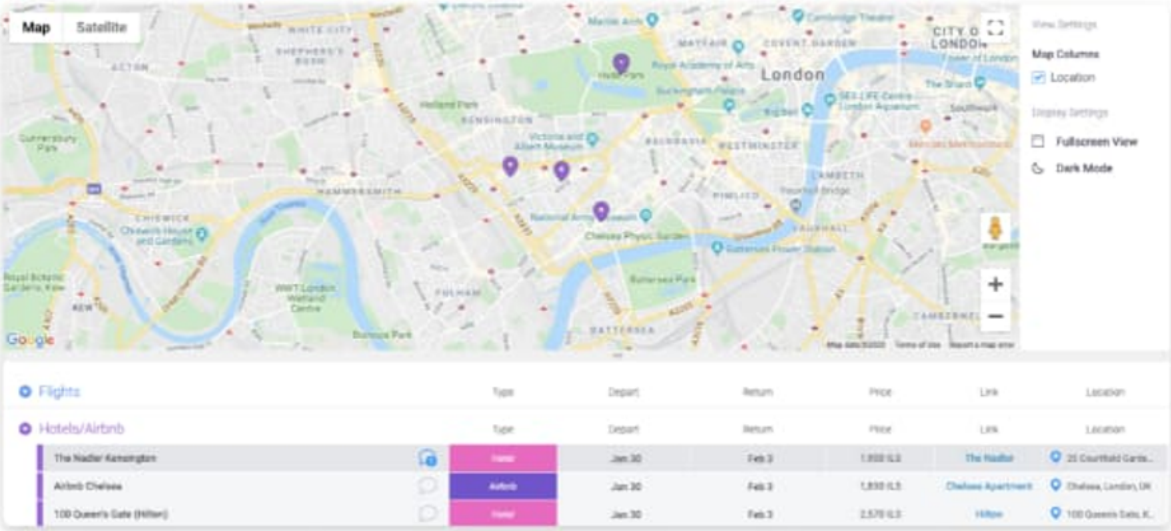
95 East 10th Street Bayonne, NJ, USA

95 East 10th Street Linden, NJ, USA

95 East 10th Street Bloomsburg, PA, USA

95 East 10th Street Brooklyn, NY, USA

Figure 3: Figure



View Settings

Map Columns

☑ Location

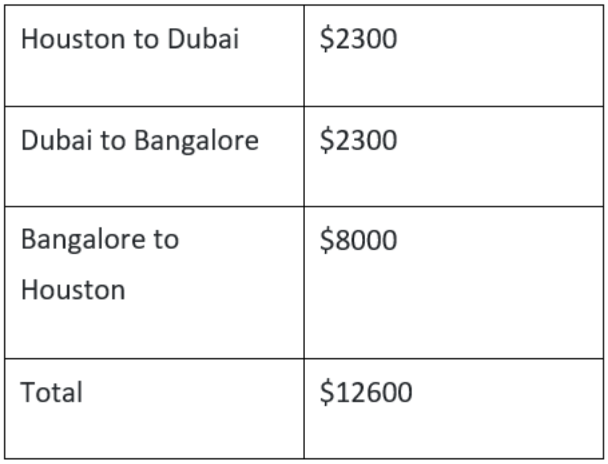
Display Settings

□ Fullscreen View

Dark Mode

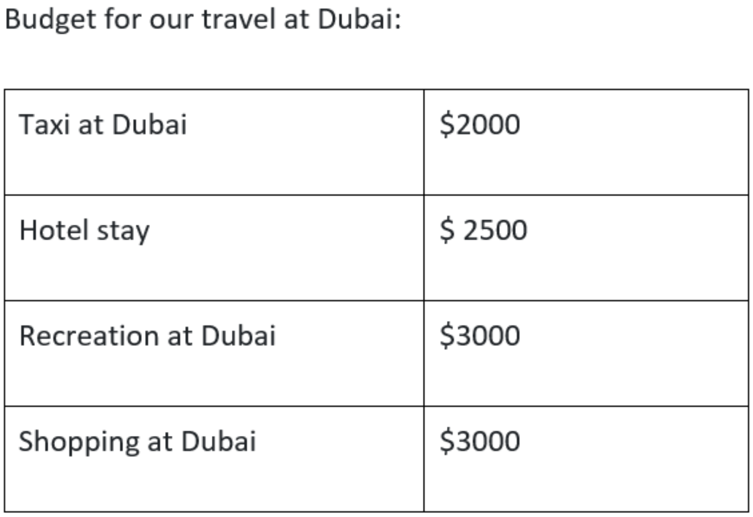
Flights	Type	Depart	Return	Price	Link	Location
Hotels/Airbnb	Type	Depart	Return	Price	Link	Location
The Nadler Kensington	Hotel	Jan 30	Feb 3	1,800 ILS	The Nadler	25 Courtfield Gards...
Airbnb Chelsea	Airbnb	Jan 30	Feb 3	1,830 ILS	Chelsea Apartment	Chelsea, London, UK
100 Queen's Gate (Hilton)	Hotel	Jan 30	Feb 3	2,570 ILS	Hilton	100 Queen's Gate, K...

Figure 4: Figure



Houston to Dubai	$2300
Dubai to Bangalore	$2300
Bangalore to Houston	$8000
Total	$12600

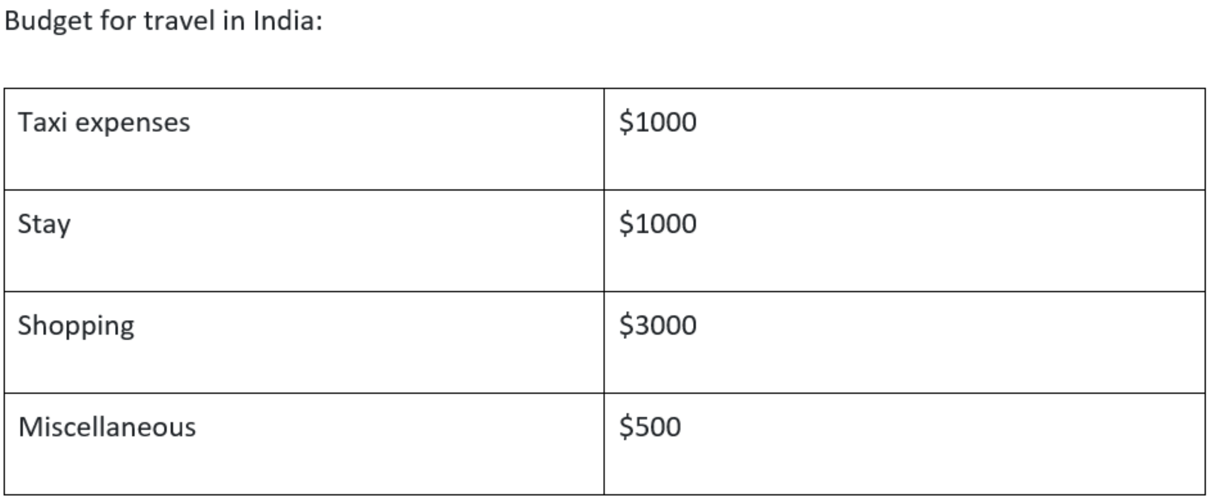
Figure 5: Figure



Budget for our travel at Dubai:

Taxi at Dubai	$2000
Hotel stay	$ 2500
Recreation at Dubai	$3000
Shopping at Dubai	$3000

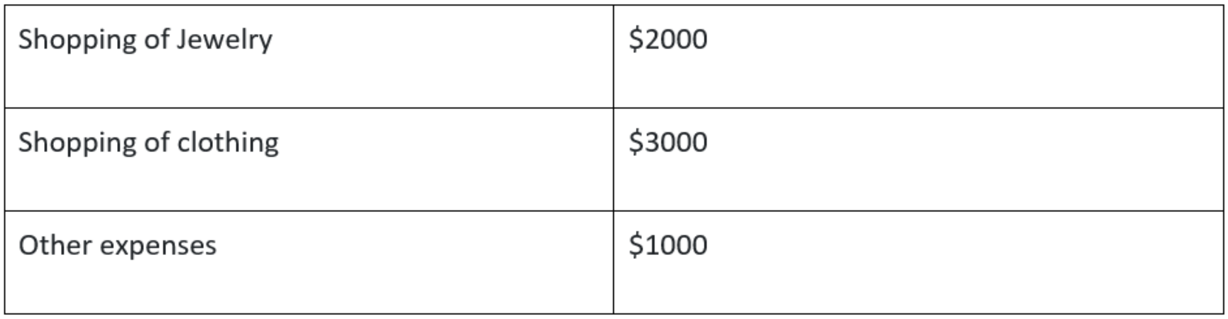
Figure 6: Figure



Budget for travel in India:

Taxi expenses	$1000
Stay	$1000
Shopping	$3000
Miscellaneous	$500

Figure 7: Figure

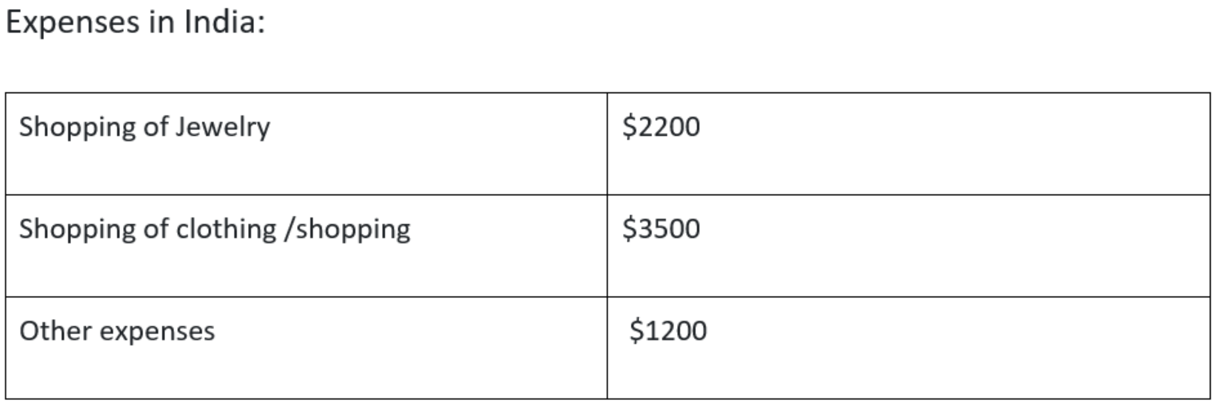


Shopping of Jewelry	$2000
Shopping of clothing	$3000
Other expenses	$1000

Figure 8: Figure

With all the plan , schedule and budget ready, now the D day came to travel. we had all our documents and necessary covid related information ready to arrive at the airport. It was an emotional farewell for the family to fly from Houston to Dubai. we enjoyed every bit of the journey. When landed in Dubai, we stayed in Dubai experienced the sand dunes of Dubai, skydived at Dubai, and stayed at Burj Khalifa, Rome around the lavish malls and the beautiful city of Dubai. As we came to end to the end of the tour. Good thing was that I was logging all my travel expenses and experiences on my tablet. We came to the important leg of our vacation, we landed in India (@ Bangalore), and we started off with sip of coffee at the airport which was refreshing to the soul.
This trip in India was a memorable one with lots of fun-filled meetups, shopping, eating and celebrations with my family and friends. We did shopping of traditional Indian clothes for the 50th marriage Anniversary of my parents and the family ceremony of my cousin. we had blast at the street foods, and enjoyed every bit of it. At last, the vacation was coming to an end which was an emotional one for me and my kids and family in India. Finally, we boarded our flight back to Houston.

Some of the expenses and comparisons are listed below :



Expenses in India:

Shopping of Jewelry	$2200
Shopping of clothing /shopping	$3500
Other expenses	$1200

Figure 9: Expenses in India

When compared to Dubai, we had more fun to shop and haggle around to get more variety of the clothes especially in the traditional wears.

Comparison between expenses in India and Dubai are listed below :

2.3 Conclusion

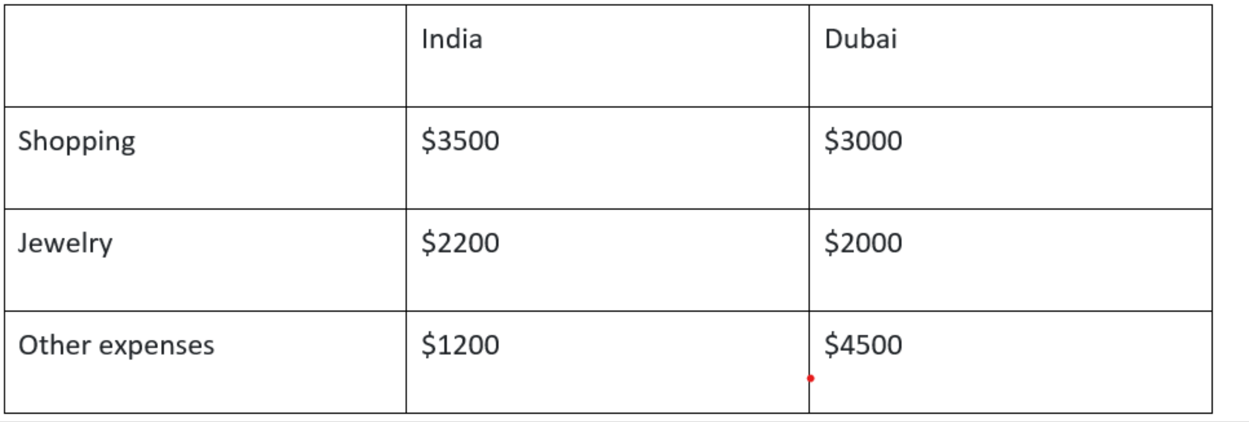
When we compare the expenses in India and Dubai, as data suggest Dubai is costlier compared to India on the flip side, we get items all over the world and branded in Dubai but in India, we get more traditional, handmade items which are very popular. The trip to India was more emotional and a good experience in which the connections very made stronger whereas the Dubai vacation for more fun-oriented.

/

3 Breast cancer Wisconsin

3.1 Introduction

Breast cancer is a disease in which cells in the breast grow out of control. There are different kinds of breast cancer. The kind of breast cancer depends on which cells in the breast turn into cancer. Breast cancer is a type of cancer that starts in the breast. It can start in one or both breasts. Breast cancer as basiclly



	India	Dubai
Shopping	$3500	$3000
Jewelry	$2200	$2000
Other expenses	$1200	$4500

Figure 10: Compre between India and Dubai

two categories of breast cancer That is * Malignant type breast cancer * Benign type breast cancer It's important to understand that most breast lumps are benign and not cancer . Non-cancer breast tumors are abnormal growths, but they do not spread outside of the breast. They are not life threatening, but some types of benign breast lumps can increase a woman's risk of getting breast cancer. Any breast lump or change needs to be checked by a health care professional to find out if it is benign or malignant (cancerous) and if it might affect your future cancer risk. The main aim of this project is Breast cancer cells are Benign or Malignant. ..which are non-cancerous.Reducing the dimensionality of the Breast cancer factors by using principal component analysis and nonmetrics MDS. Predict the classification of the Kneighbors classifier, Gaussian classifier, and Neural network.

3.2 Literature Review

Cancer is a broad term for a class of diseases characterized by abnormal cells that grow and invade healthy cells in the body. Breast cancer starts in the cells of the breast as a group of cancer cells that can then invade surrounding tissues or spread (metastasize) to other areas of the body. A tumor is a mass of abnormal tissue. There are two types of breast cancer tumors: those that are non-cancerous, or 'benign', and those that are cancerous, which are 'malignant'. Benign Tumors When a tumor is diagnosed as benign, doctors will usually leave it alone rather than remove it. Even though these tumors are not generally aggressive toward surrounding tissue, occasionally they may continue to grow, pressing on other tissue and causing pain or other problems. In these situations, the tumor is removed, allowing pain or complications to subside. Malignant tumors are cancerous and may be aggressive because they invade and damage surrounding tissue. When a tumor is suspected to be malignant, the doctor will perform a biopsy to determine the severity or aggressiveness of the tumor.

Numerous studies have shown that occurrence of breast cancer is due to combination of lot of factors. The most prominent factor is getting older i.e. age for women. Most breast cancers appear after the age of 40 or beyond that. Significant risk factors for breast cancer are – age, genetic mutations, reproductive history, race/ethnicity, family history, personal history of breast cancer, obesity, drinking alcohol, exposure to radiation, hormone therapy, sedentary lifestyle etc. Family history is one of the most prevalent genetic risk factors for developing breast cancer.

BRCA1 & BRCA2 (Breast Cancer genes 1 & 2) are the best-known genes linked to breast cancer risk.

Every human being has the genes, but some people have an inherited mutation in one or both the genes that increase the risk for breast cancer. After BRCA1 &BRCA2, PALB2 is the third most prevalent breast cancer gene. In addition to BRCA1/2, gene mutations there are these types gene mutations which put increased risk-ATM, BARD1,CDH1,CHEK2,NBN,NF1,PALB2,PTEN etc.PALB2 is short for "Partner And Localizer of BRCA2." In other words, it works in partnership with the BRCA2 gene to repair DNA damage and thereby prevent breast cancer from developing. Having a mutated CHEK2 gene doubles the risk of breast cancer in women. In men, it makes male breast cancer 10 times more likely to occur. CDH1 is a tumor suppression gene but a mutation can also make it easier for individual cancer cells to break off from a breast tumor and metastasize, or spread to other parts of the body. In this study, relationships of various genes with the breast cancer stages have been evaluated.

Breast cancer is the most prevalent cancer in the world (4.4 million survivors up to 5 years following diagnosis) and the second most common cause of cancer-related mortality in women's wide world (Parkin et al., 2005). It also accounts for 23% (1.38 million) of the total new cancer cases and 14% (458,400) of the total cancer deaths in 2008 and ranks second most common cancer overall (10.9% of all cancers) but ranks fifth as the cause of death (Ferlay et al., 2010). 1.15 million new breast cancer cases were recorded in 2004 and over 500,000 deaths were reported around the world, and more than half of all cases occurred in industrialized countries (Parkin and Fernandez, 2006). Breast cancer incidence rates vary from 19.3 per 100,000 women in Eastern Africa to 89.7 per 100,000 women in Western Europe. They are normally high in developed regions of the world (except Japan) and low in most of the developing regions. Due to more favorable survival of breast cancer in developed regions, the range of mortality rates is very much less, approximately 6-19 per 100,000. Notwithstanding, it is still the most frequent cause of cancer death in women in both developing (269 000 deaths, 12.7% of total) and developed regions, where the estimated 189 000 deaths are almost equal to the estimated number of deaths from lung cancer (188 000 deaths) (Ferlay et al., 2010) Breast cancer is common in women both in developed and developing countries, comprising 16% of all female cancers. Although it is thought to be common cancer in developed countries, a majority (69%) of all breast cancer deaths occur in the developing world. Indeed, increase life expectancy, increased urbanization, and the adoption of western lifestyles have increased the incidence of breast cancer in developing countries (Kanavos, 2006). Even though it is now the most common cancer both in developed and developing regions with around 690 000 new cases estimated in each region, much of the burden of incidence, morbidity, and mortality will occur in the developing world with a population ratio of 1:4 (Ferlay et al., 2010). As developing countries succeed in achieving lifestyles like those in advanced economies, they will also encounter much higher cancer rates, particularly cancers of the breast. This forms part of a larger epidemiological transition in which the burden of the chronic, non-communicable disease once limited to industrialized nations, is now increasing in less developed countries (Kanavos, 2006). There are about 15 to 20 sections called lobes in a female's breast and each lobe is made of many smaller sections known as lobules, which in turn have groups of tiny glands or milk-producing glands that can make milk. It is also made up of ducts, tiny tubes that carry the milk from the lobules to the nipple, and stroma fatty tissue and connective tissue surrounding the ducts and lobules, blood vessels, and lymphatic vessels (American Cancer Society booklet). The lymph system is very important in breast cancer research in that, it is one-way breast cancers can spread and has several parts. Lymph nodes are small, bean-shaped collections of immune system cells that are connected by lymphatic vessels. These vessels are like small veins, except that they carry a clear fluid called lymph in place of blood away from the breast. They also contain Lymph tissue fluid and waste products, in 7 additions to immune system cells. Breast cancer cells can enter lymphatic vessels and begin to grow in lymph nodes. Most lymphatic vessels in the breast connect to lymph nodes under the arm (axillary nodes). Some lymphatic vessels that connect to lymph nodes inside the chest are called internal lymph nodes, and those either above or below the collarbone are called supraclavicular or infraclavicular nodes (American Cancer Society booklet)

3.3 Materials and methods

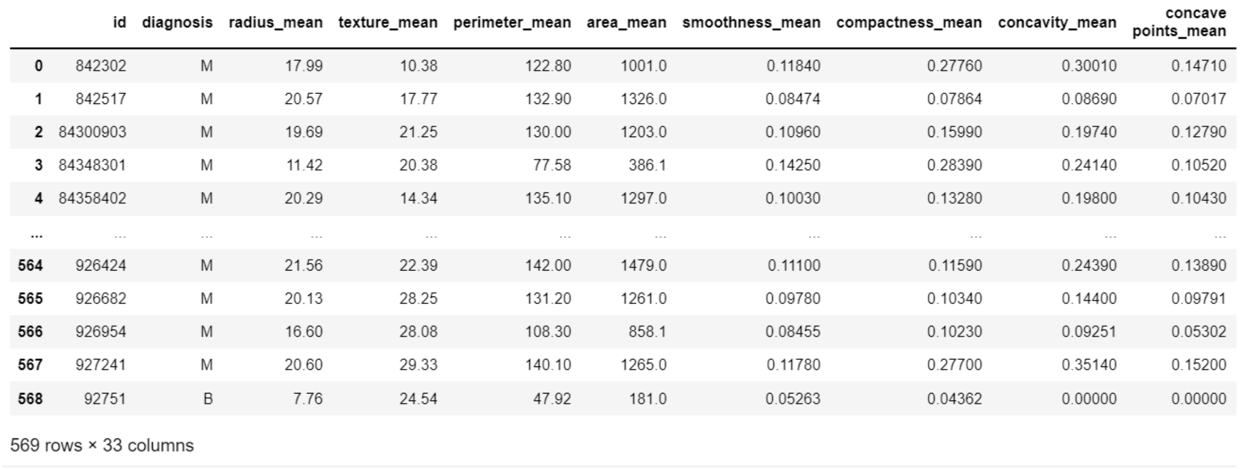
3.3.1 PROBLEM, PURPOSE

This analysis aims to determine which Breast cancer cells are Benign or Mligant which non-cancerous. This is a multivariate classification to reduce the dimensionality of Breast cancer factors by using PCA and nonmetric MDS techniques. predict the Knn classifier, Gaussian classifier, and neural network.

3.3.2 Data source

Our data set contains 569 patiens samples that are represented by IDs.Each sample is an instance of a proportion of a tumor that is exampled for mutations.The name of samples in many cases originates from the cell line name.There are two classes in the dataset that represent two cells of cancer. cancer data source is collect from https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

3.3.3 Dataset information



	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.30010	0.14710
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.08690	0.07017
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.19740	0.12790
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.24140	0.10520
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.19800	0.10430
...	...	...	...	...	...	...	...	...	...	...
564	926424	M	21.56	22.39	142.00	1479.0	0.11100	0.11590	0.24390	0.13890
565	926682	M	20.13	28.25	131.20	1261.0	0.09780	0.10340	0.14400	0.09791
566	926954	M	16.60	28.08	108.30	858.1	0.08455	0.10230	0.09251	0.05302
567	927241	M	20.60	29.33	140.10	1265.0	0.11780	0.27700	0.35140	0.15200
568	92751	B	7.76	24.54	47.92	181.0	0.05263	0.04362	0.00000	0.00000

569 rows × 33 columns

3.4 Results and discussion

3.4.1 Exploratory Data Analysis

Exploratory Data Analysis refers to the vital process of performing initial investigations on data to discover patterns, to spot anomalies, To check the correlation Visual representation.It is a good practice to understand the data First and try to gather as many insights as possible from it. Exploratory Data Analysis is all about making sense of data in hand, before getting them dirty with it. Exploratory Data Analysis is an approach to data analysis that postpones the usual assumptions about what kind of model the data follow with the more direct approach of allowing the data itself to reveal its underlying structure and model.

The initial dataset included two diagnosis stages of breast cancer but the dataset is highly balanced data. In the exploration, we choose eight features to do for the analysis of the correlation between the variables.

Check the correlation between the variables

Figure :11-The above visual show these features are highly correlated and dependent on each other. we can see that the radius mean and perimeter mean attributes are hinting at the presence of multicollinearity between the variables.

3.4.2 Dimension Reduction

In machine learning classification problems, there are often too many factors on the basis of which the final classification is done. These factors are basically variables called features. The higher the number of features, the harder it gets to visualize the training set and then work on it. Sometimes, most of these features are correlated, and hence redundant. This is where dimensionality reduction algorithms come into play.

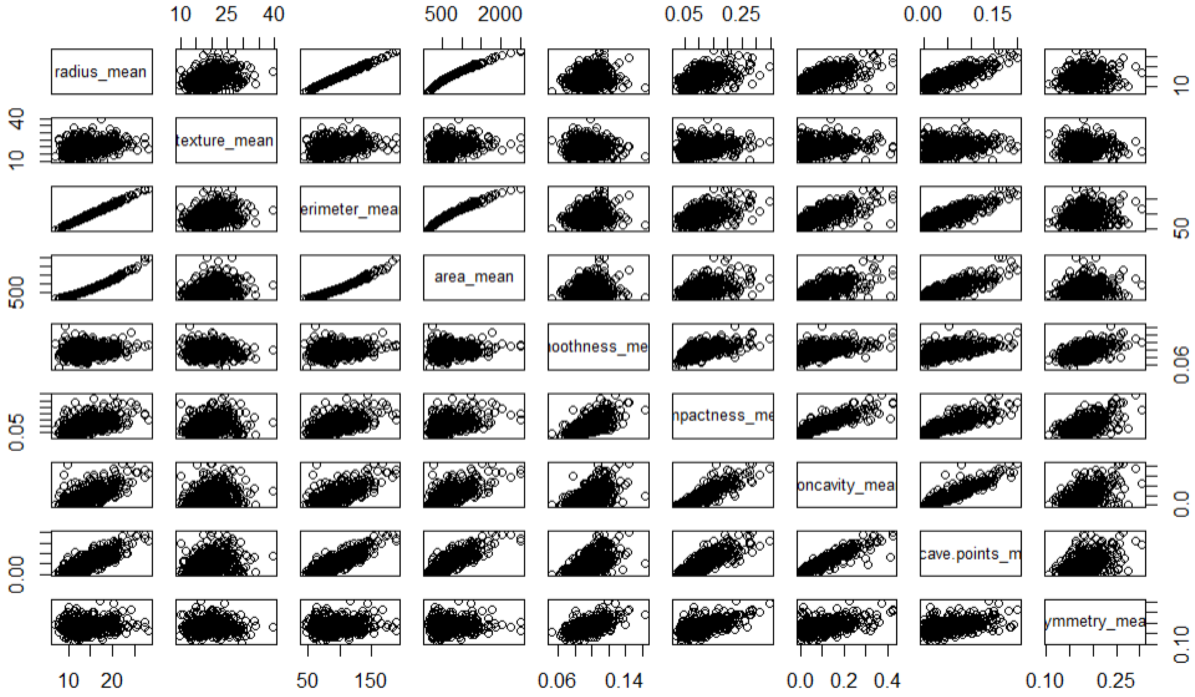


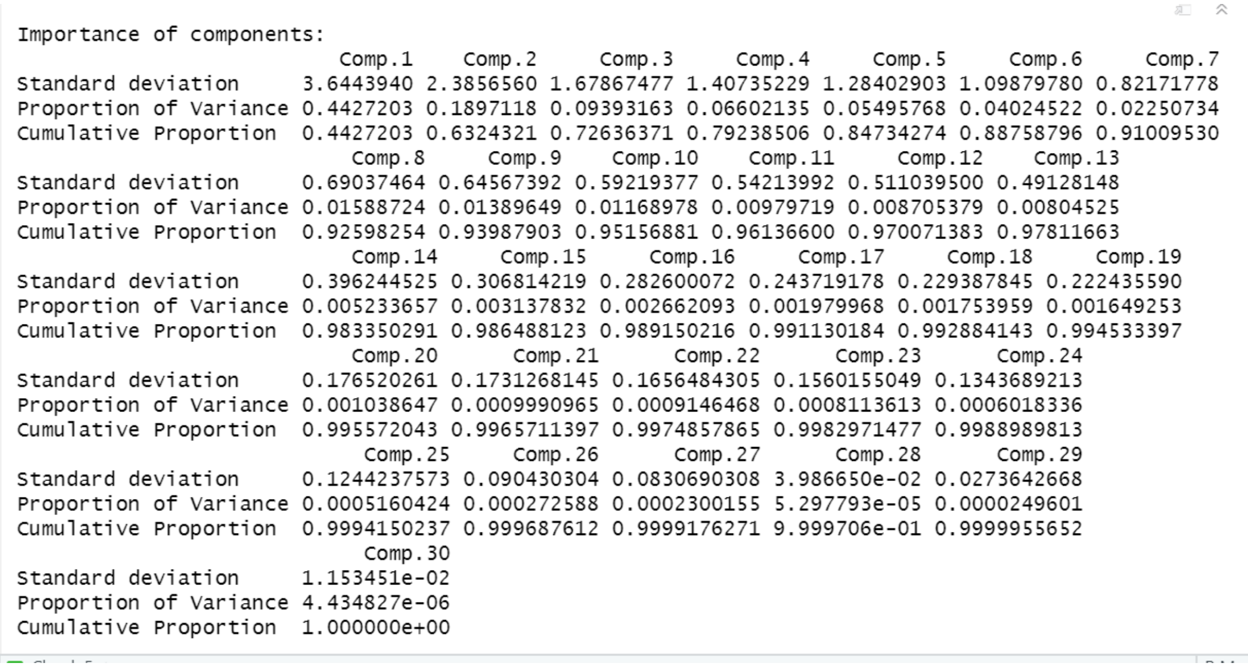
Figure 11: Correlation between the variables

Dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables. It can be divided into feature selection and feature extraction The aim of dimension reduction for this data set is to determine if there are any uncovered relationships between the proteins representing each feature and the generation of cancer, not only breast cancer but cancer in general. In this study, we do not have any a priori justification for specifying the particular structural model. Therefore, We proceed with exploratory methods. First, we will utilize the principal component analysis, determining the optimal number of principal components.

we will utilize the scree plots to visualize our results from the PCA. According to the interpreting techniques of the scree plot presented the values.

3.4.3 Principal component analysis

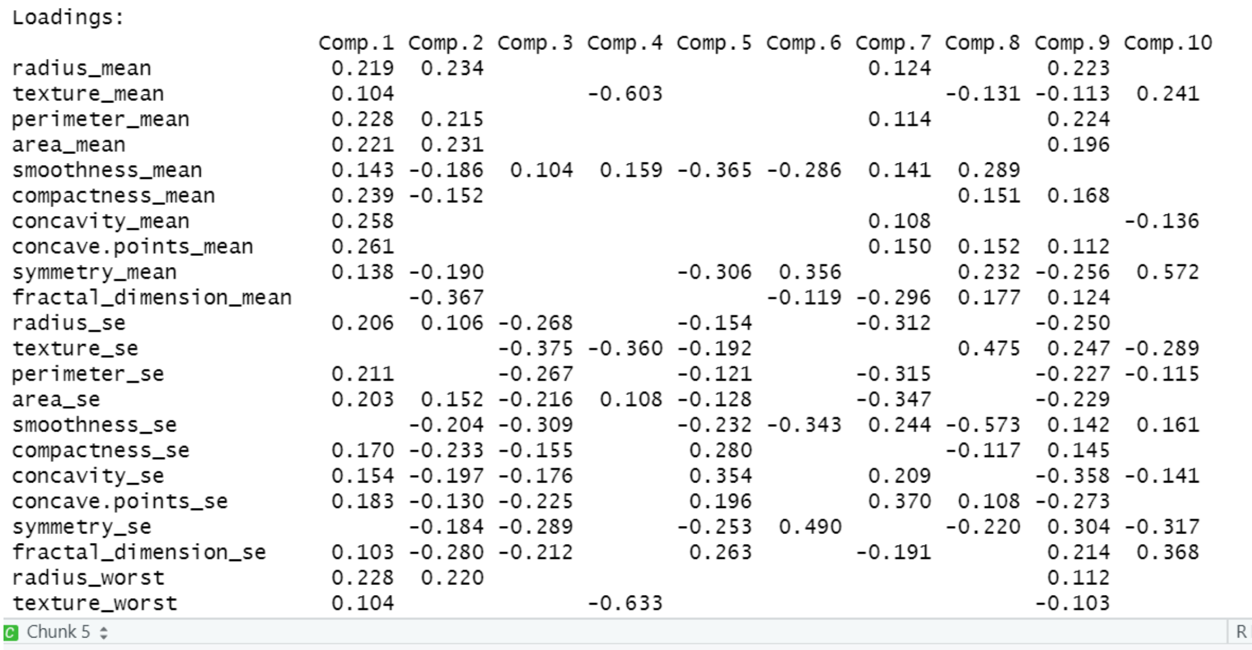
Principal Component Analysis (PCA) is a useful technique for exploratory data analysis, allowing better visualization of the variation present in a dataset with many variables. The principal components are orthogonal linear combinations that maximize the total variance. PCA is particularly helpful in the case of "wide" datasets, where you have many variables for each sample. When many variables are present, it is difficult visualize the data. PCA allowed us to see the overall shape of the data and identify which samples are similar and which are different. Variables that correlate with each other will contribute strongly to the same principal component. Each of the principal components will sum up to a certain percentage of the total variation in the dataset. If variables are strongly correlated with one another, we will be able to approximate the complexity of the dataset with just a few principal components.Principal component analysis can be broken down into five steps. The logical explanations of what PCA is doing and simplifying mathematical concepts such as standardization, covariance, eigenvectors and eigenvalues without focusing on how to compute them.



```
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5     Comp.6     Comp.7
Standard deviation     3.6443940 2.3856560 1.67867477 1.40735229 1.28402903 1.09879780 0.82171778
Proportion of Variance 0.4427203 0.1897118 0.09393163 0.06602135 0.05495768 0.04024522 0.02250734
Cumulative Proportion  0.4427203 0.6324321 0.72636371 0.79238506 0.84734274 0.88758796 0.91009530
                           Comp.8     Comp.9    Comp.10    Comp.11     Comp.12    Comp.13
Standard deviation     0.69037464 0.64567392 0.59219377 0.54213992 0.511039500 0.49128148
Proportion of Variance 0.01588724 0.01389649 0.01168978 0.00979719 0.008705379 0.00804525
Cumulative Proportion  0.92598254 0.93987903 0.95156881 0.96136600 0.970071383 0.97811663
                           Comp.14     Comp.15     Comp.16     Comp.17     Comp.18     Comp.19
Standard deviation     0.396244525 0.306814219 0.282600072 0.243719178 0.229387845 0.222435590
Proportion of Variance 0.005233657 0.003137832 0.002662093 0.001979968 0.001753959 0.001649253
Cumulative Proportion  0.983350291 0.986488123 0.989150216 0.991130184 0.992884143 0.994533397
                           Comp.20      Comp.21      Comp.22      Comp.23      Comp.24
Standard deviation     0.176520261 0.1731268145 0.1656484305 0.1560155049 0.1343689213
Proportion of Variance 0.001038647 0.0009990965 0.0009146468 0.0008113613 0.0006018336
Cumulative Proportion  0.995572043 0.9965711397 0.9974857865 0.9982971477 0.9988989813
                            Comp.25     Comp.26      Comp.27      Comp.28      Comp.29
Standard deviation     0.1244237573 0.090430304 0.0830690308 3.986650e-02 0.0273642668
Proportion of Variance 0.0005160424 0.000272588 0.0002300155 5.297793e-05 0.0000249601
Cumulative Proportion  0.9994150237 0.999687612 0.9999176271 9.999706e-01 0.9999955652
                            Comp.30
Standard deviation     1.153451e-02
Proportion of Variance 4.434827e-06
Cumulative Proportion  1.000000e+00
```

Figure 12: Summary of Principal component

The above visual Figure:12 shows the summary of the result object gives us the standard deviation, the proportion of variance explained by each principal component, and the cumulative proportion of variance explained. The first 5 components of variance are 84% of the variation in this data and 5 to 10 components are 97% of the variation in the data. you can see the standard deviation of size eigenvalues to determine the number of principal components. Retain the principal components with eigenvalues greater than one.



```
Loadings:
                        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10
radius_mean              0.219  0.234                              0.124         0.223        
texture_mean             0.104               -0.603                      -0.131 -0.113  0.241 
perimeter_mean           0.228  0.215                              0.114         0.224        
area_mean                0.221  0.231                                            0.196        
smoothness_mean          0.143 -0.186  0.104  0.159 -0.365 -0.286  0.141  0.289               
compactness_mean         0.239 -0.152                                     0.151  0.168        
concavity_mean           0.258                                     0.108               -0.136 
concave.points_mean      0.261                                     0.150  0.152  0.112        
symmetry_mean            0.138 -0.190               -0.306  0.356         0.232 -0.256  0.572 
fractal_dimension_mean        -0.367                      -0.119 -0.296  0.177  0.124        
radius_se                0.206  0.106 -0.268        -0.154        -0.312        -0.250        
texture_se                           -0.375 -0.360 -0.192                0.475  0.247 -0.289 
perimeter_se             0.211        -0.267        -0.121        -0.315        -0.227 -0.115 
area_se                  0.203  0.152 -0.216  0.108 -0.128        -0.347        -0.229        
smoothness_se                 -0.204 -0.309        -0.232 -0.343  0.244 -0.573  0.142  0.161 
compactness_se           0.170 -0.233 -0.155         0.280               -0.117  0.145        
concavity_se             0.154 -0.197 -0.176         0.354         0.209        -0.358 -0.141 
concave.points_se        0.183 -0.130 -0.225         0.196         0.370  0.108 -0.273        
symmetry_se                   -0.184 -0.289        -0.253  0.490        -0.220  0.304 -0.317 
fractal_dimension_se     0.103 -0.280 -0.212         0.263        -0.191         0.214  0.368 
radius_worst             0.228  0.220                                            0.112        
texture_worst            0.104               -0.633                             -0.103        
```

Figure 13: Eigen vectors Pc1 to pc10

Figure 13 : The above table shows these are the eigen vectors of pc1 to pc10.First we can see the Pc1 about loading smoothness mean ,symeetric mean,comapact se and concavity these are variable are effected because the values are so mal1 and other variable are big values.

3.4.4 Scree plot

A common method for determining the number of PCs to be retained is a graphical representation known as a scree plot. A Scree Plot is a simple line segment plot that shows the eigenvalues for each individual PC. It shows the eigenvalues on the y-axis and the number of factors on the x-axis. It always displays a downward curve. Most scree plots look broadly similar in shape, starting high on the left, falling rather quickly, and then flattening out at some point. This is because the first component usually explains much of the variability, the next few components explain a moderate amount, and the latter components only explain a small fraction of the overall variability. The scree plot criterion looks for the “elbow” in the curve and selects all components just before the line flattens out. (In the PCA literature, the plot is called a ‘Scree’ Plot because it often looks like a ‘scree’ slope, where rocks have fallen down and accumulated on the side of a mountain.)

Figure 14 :The scree plot visualizes from Pca the result shows in screen plot that we are able to explain the dataset with just a few components. The scree plot of the eigenvalues from the largest to the smallest. The ideal pattern is a steep curve followed by a bend and then a straight line using the components in the steep curve before the first 3 components that start the line trend. A biplot aims to represent both the observations and variables on the same plot. It shows the score (Yi) of each case on the first two PCs and the loading (wij) of each variable (investment type) on the first two PCs. Bottom axis=PC1 score; Left axis=PC2 score; Top axis= loadings on PC1; Right axis=loadings on PC2.PC1 is the difference between area+se and PC2 is made up almost analysis.

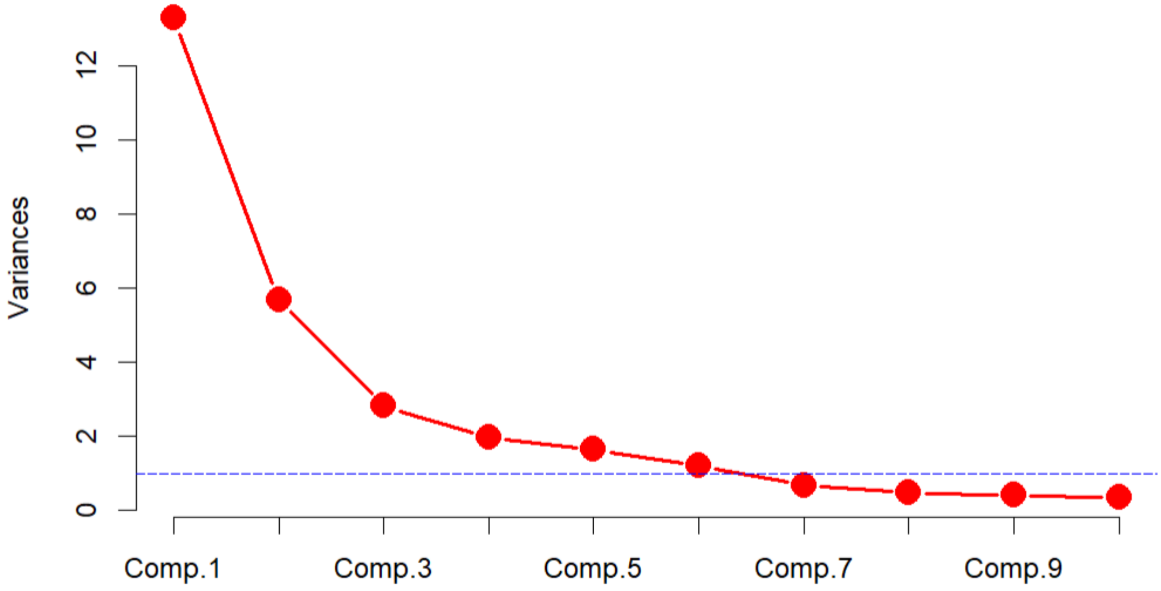


Figure 14: Scree plot of number of principal components

3.4.5 Biplot

Biplots are a type of exploratory graph used in statistics, a generalization of the simple two-variable scatterplot. A biplot overlays a score plot with a loading plot. A biplot allows information on both samples and variables of a data matrix to be displayed graphically.A biplot overlays a score plot and a loadings plot in a single graph. An example of X-axis and Y-axis points are the projected observations; vectors are the projected variables. If the data are well-approximated by the first two principal components, a biplot enables you to visualize high-dimensional data by using a two-dimensional graph

Figure 15 :A biplot aims to represent both the observations and variables on the same plot. It shows the score (Yi) of each case on the first two PCs and the loading (wij) of each variable on the first two PCs. Bottom axis=PC1 score; Left axis=PC2 score; Top axis= loadings on PC1; Right axis=loadings on PC2.PC1 is the difference between area+se and PC2 is made up almost analysis. The biplot shows angles between the vectors about the correlation between the variables. If they are 90 degrees they are not likely correlated like texture_se,smootheness_se and compactness_se. These two vectors are close concavity_se and concave point_se variables are close so these two variables are positively correlated. Vectors are close to 180 degrees are negatively correlated like freactal_demension_mean, radius mean, and area_se. There are so many variables fitted into two sides we cannot see the image properly.

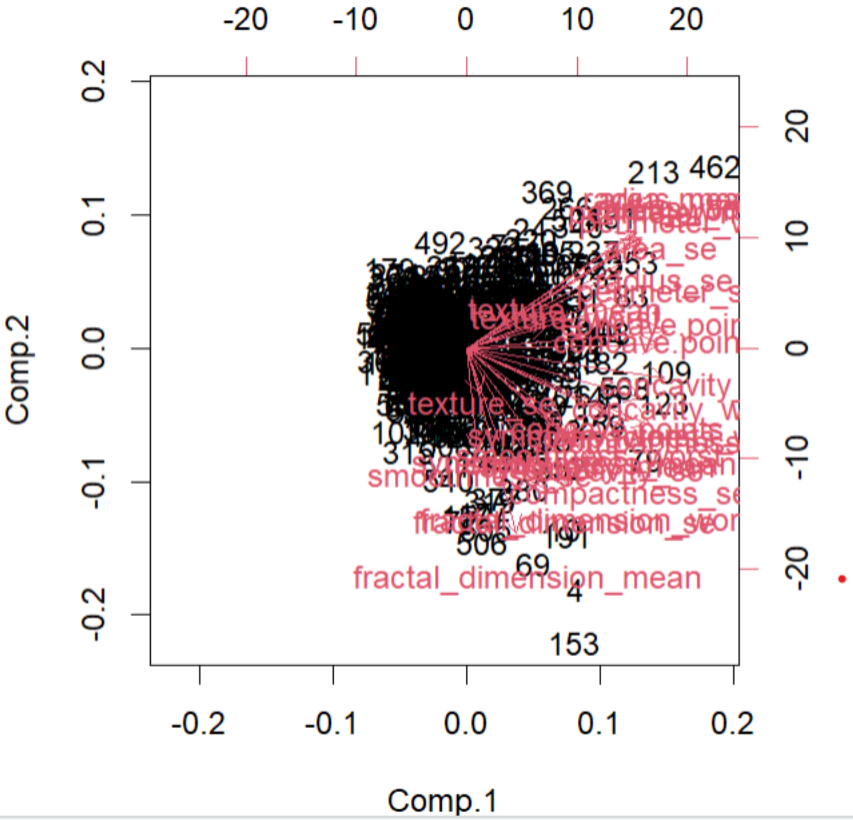
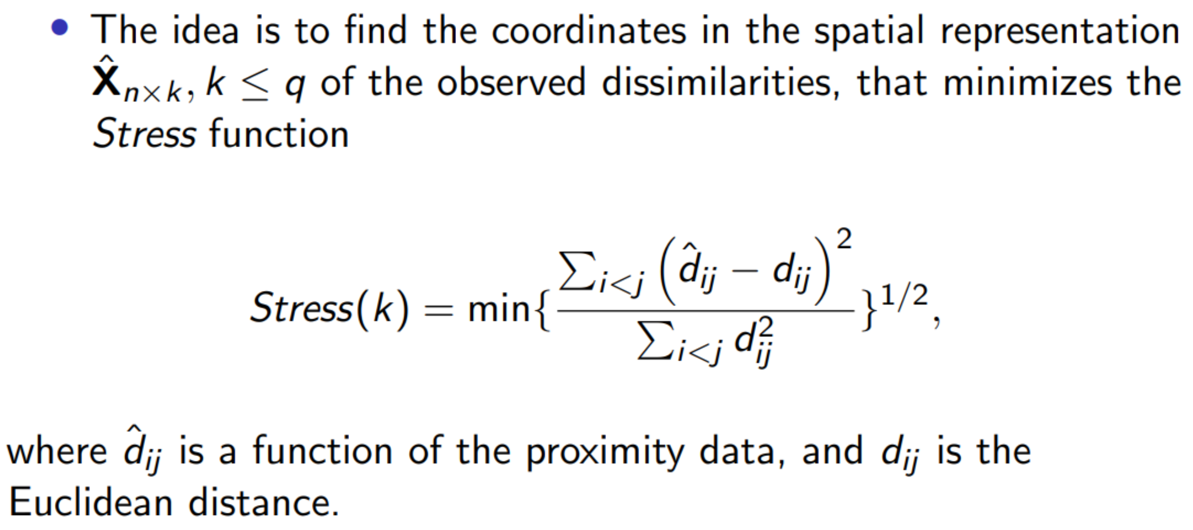


Figure 15: Biplot of number of principal components

3.4.6 Non metric MDS

Non-metric multidimensional scaling (NMDS) is an indirect gradient analysis approach which produces an ordination based on a distance or dissimilarity matrix. Unlike methods which attempt to maximise the variance or correspondence between objects in an ordination, NMDS attempts to represent, as closely as possible, the pairwise dissimilarity between objects in a low-dimensional space. Any dissimilarity coefficient or distance measure may be used to build the distance matrix used as input. As NMDS is an iterative algorithm, it can quickly become computationally demanding for large data sets. Further, multiple runs of the NMDS algorithm are needed to ensure a stable solution has been reached. If you do not have access to a platform that can effectively run NMDS, consider using principal coordinates analysis. The technique uses a „trial and error“ to find the best positioning in dimensional space. Goodness-of-fit is measured by „stress“ – a measure of rank-order disagreement between observed and fitted distances. Stress can be defined as a value representing the difference between distance in the reduced dimension compared to the complete multidimensional space. NMDS tries to optimize the stress as much as possible („pulling on all points a little bit so no single point is completely wrong, all points are a little off compared to distances)“.



- The idea is to find the coordinates in the spatial representation $\hat{\mathbf{X}}_{n\times k}$, $k \leq q$ of the observed dissimilarities, that minimizes the *Stress* function

$$Stress(k) = \min\{\frac{\sum_{i<j}\left(\hat{d}_{ij} - d_{ij}\right)^2}{\sum_{i<j} d_{ij}^2}\}^{1/2},$$

where $\hat{d}_{ij}$ is a function of the proximity data, and d_{ij} is the Euclidean distance.

3.4.7 stress plot

The Stress Plot allows you to plot stress results for static, nonlinear, dynamic, and drop test studies.To display this PropertyManager, run a static, nonlinear, dynamic study, or drop test study. Right-click Results and select Define Stress Plot.

Figure 16 : MDS finds a lower-dimension representation of the data that mimics the distances or minimizes the “stress” function. The stress plot shows an R2 value is 0.997. We can say that there is a good fit.

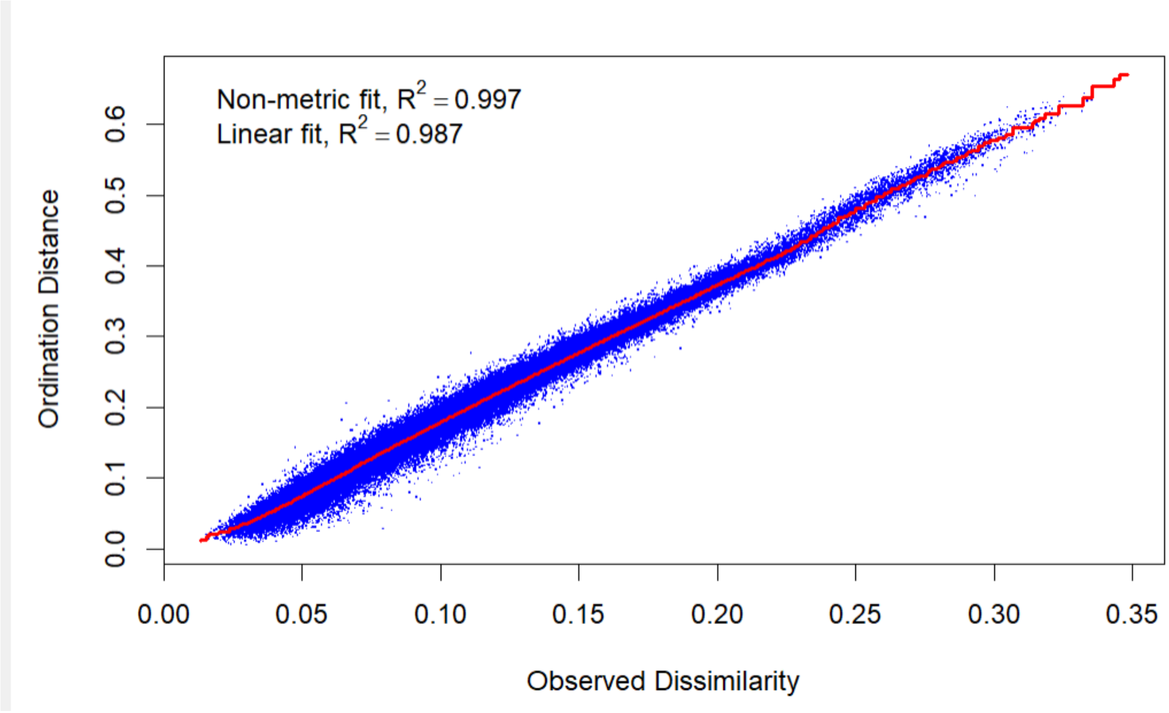
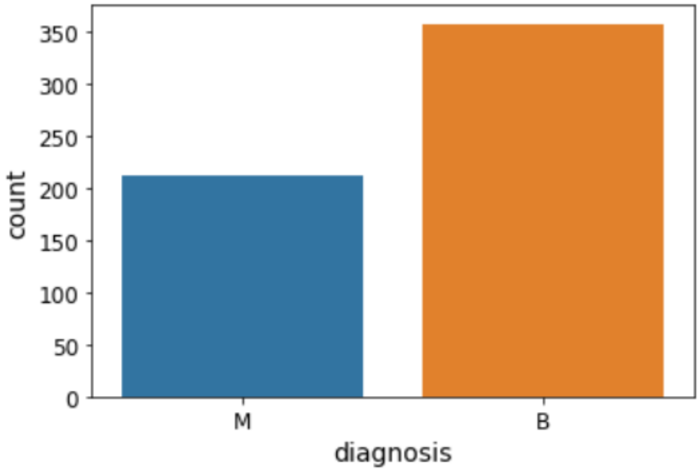


Figure 16: Stress plot

3.5 Classifications

we will visualize the diagnosis column in our dataset to see how many malignant and benign are present.

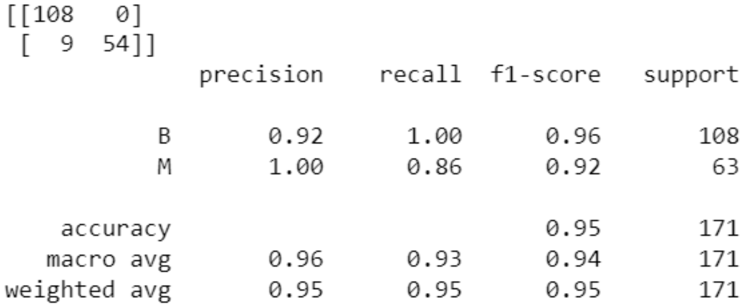


In this whole analyzing process we are going to convert our data and perform some data processing so that we can build a model which can classify the type of Breast cancer using this preprocessed data.

3.5.1 Knn classifier

In statistics, the k-nearest neighbors algorithm (k-NN) is a non-parametric supervised learning method first developed by Evelyn Fix and Joseph Hodges in 1951,[1] and later expanded by Thomas Cover.[2] It is used for classification and regression. In both cases, the input consists of the k closest training examples in a data set. The output depends on whether k-NN is used for classification or regression: In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor. In k-NN regression, the output is the property value for the object. This value is the average of the values of k nearest neighbors.

$$\hat{f}(x_q) <- \sum_{v \varepsilon V i=1}^{argmaxK} \delta(v, f(x_i))$$



```
[[108   0]
 [  9  54]]
              precision    recall  f1-score   support

           B       0.92      1.00      0.96       108
           M       1.00      0.86      0.92        63

    accuracy                           0.95       171
   macro avg       0.96      0.93      0.94       171
weighted avg       0.95      0.95      0.95       171
```

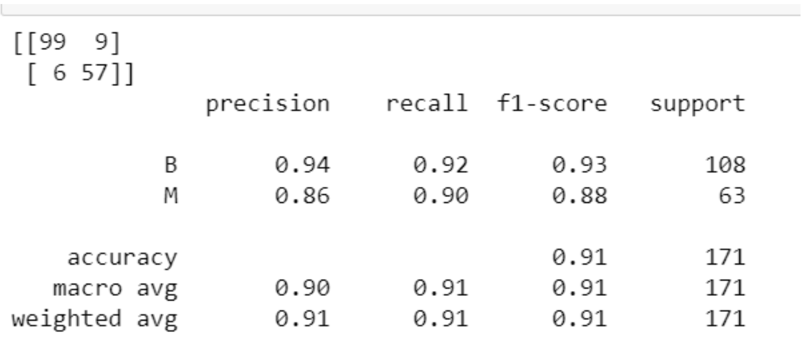
Figure 17: KNNClassifier

Figure 17 :The result shows that our KNN algorithm was able to classify 171 records in the test set with 95% accuracy. Which is excellent.Although the algorithm performed very well in this dataset. In the confusion matrix, we see that most patients have Cancers. Compared to Benign and Malignant f1 scores of 96 and 92, we have to predict that Benign is the most likely in Breast cancer. The confusion matrix shows that the true positive is 108 and the True negative is 54; the training data set is 70% in the actual data set.

3.5.2 Gaussian

Gaussian Processes are a generalization of the Gaussian probability distribution and can be used as the basis for sophisticated non-parametric machine learning algorithms for classification and regression. The Gaussian distribution, normal distribution, or bell curve, is a probability distribution which accurately models a large number of phenomena in the world. Intuitively, it is the mathematical representation of the general truth that many measurable quantities, when taking in aggregate tend to be of the similar values with only a few outliers which is to say that many phenomena follow the central limit theorem. Basically, it is the mathematical representation of how a large number of items follow the Central Limit Theory (CLT). The CLT says that , under mild conditions, the (normalized) sum of random values will tend to a gaussian distribution as the number of values in the sum increases. A Gaussian distribution can describe many examples of real-world data such as the ground state of a quantum harmonic oscillator or the distribution of demographic characteristics The Gaussian distribution occurs in many physical phenomena such as the probability density function of a ground state in a quantum harmonic oscillator. Any particle undergoing diffusion (such as in a mixed liquid) may have its location modeled accurately as a Gaussian distribution as a function of time. Even sepal width of irises have been found to follow a Gaussian distribution

$$p(Y=c|x) = \frac{p(x|Y=c)p(Y=c)}{\sum_{c'=1}^{c} p(x|Y=c')(p(Y=c'))}$$



```
[[99  9]
 [ 6 57]]
              precision    recall  f1-score   support

           B       0.94      0.92      0.93       108
           M       0.86      0.90      0.88        63

    accuracy                           0.91       171
   macro avg       0.90      0.91      0.91       171
weighted avg       0.91      0.91      0.91       171
```

Figure 18: Gaussian Classifier

Figure 18 :The result shows that our Gaussian algorithm of accuracy is 91%. Which is not excellent. In the above confusion matrix, the True positive is correctly classified as Benign cells, and the false positive is incorrectly classified as Benign cell type. The true negative is correctly classified as Malignant cells, and the false negative is incorrectly classified the malignant cells.

3.5.3 AUC-ROC Curve

AUC - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. The ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis. Defining terms used in AUC and ROC Curve: TPR (True Positive Rate) / Recall /Sensitivity

$$\text{TPR /Recall / Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{FPR} = 1 - \text{Specificity}$$
$$= \frac{FP}{TN + FP}$$

Although the comparison between two roc curves (Figure:19 and Figure:20), The y-axis shows the True Positive Rate, which is the same thing as Sensitivity, and the x-axis shows the False Negative Rate, which is the same thing as specificity. The comparison Figure:9 is the better-performing test since it covers a lot more area under the curve Figure:8, In summary, Figure:9 cover more of the areas its better-performing test.

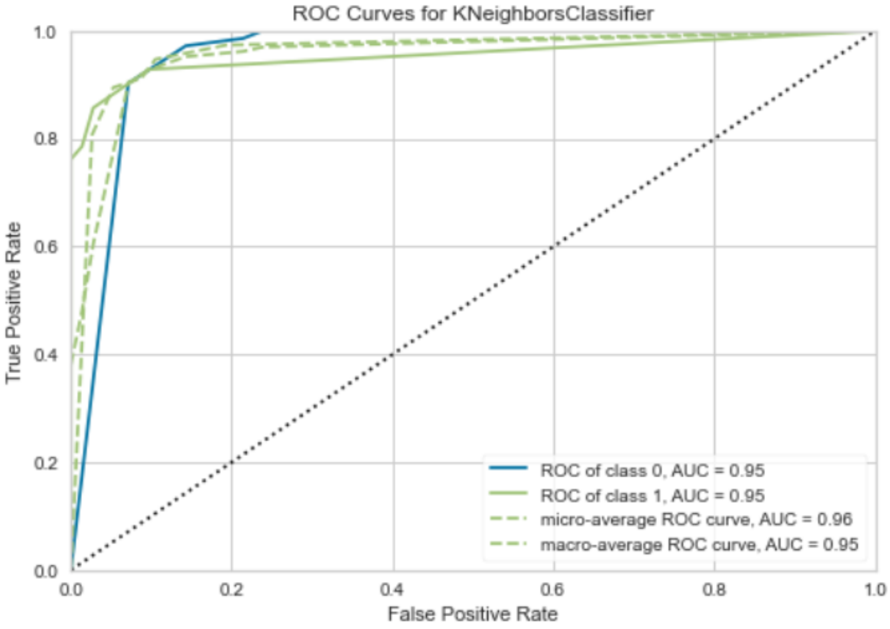


Figure 19: Knn_Roc curve

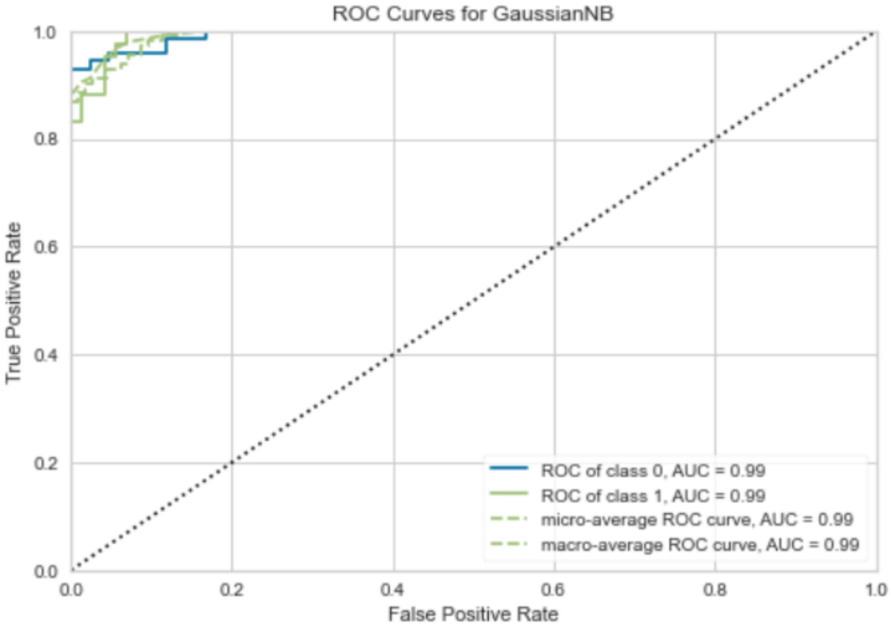


Figure 20: GB_Roc curve

3.5.4 Neural network

The basic idea behind a neural network is to simulate (copy in a simplified but reasonably faithful way) lots of densely interconnected brain cells inside a computer so you can get it to learn things, recognize patterns, and make decisions in a humanlike way. The amazing thing about a neural network is that you don't have to program it to learn explicitly: it learns all by itself, just like a brain! But it isn't a brain. It's important to note that neural networks are (generally) software simulations: they're made by programming very ordinary computers, working in a very traditional fashion with their ordinary transistors and serially connected logic gates, to behave as though they're built from billions of highly interconnected brain cells working in parallel. No-one has yet attempted to build a computer by wiring up transistors in a densely parallel structure exactly like the human brain. In other words, a neural network differs from a human brain in exactly the same way that a computer model of the weather differs from real clouds, snowflakes, or sunshine. Computer simulations are just collections of algebraic variables and mathematical equations linking them together (in other words, numbers stored in boxes whose values are constantly changing). They mean nothing whatsoever to the computers they run inside—only to the people who program them.

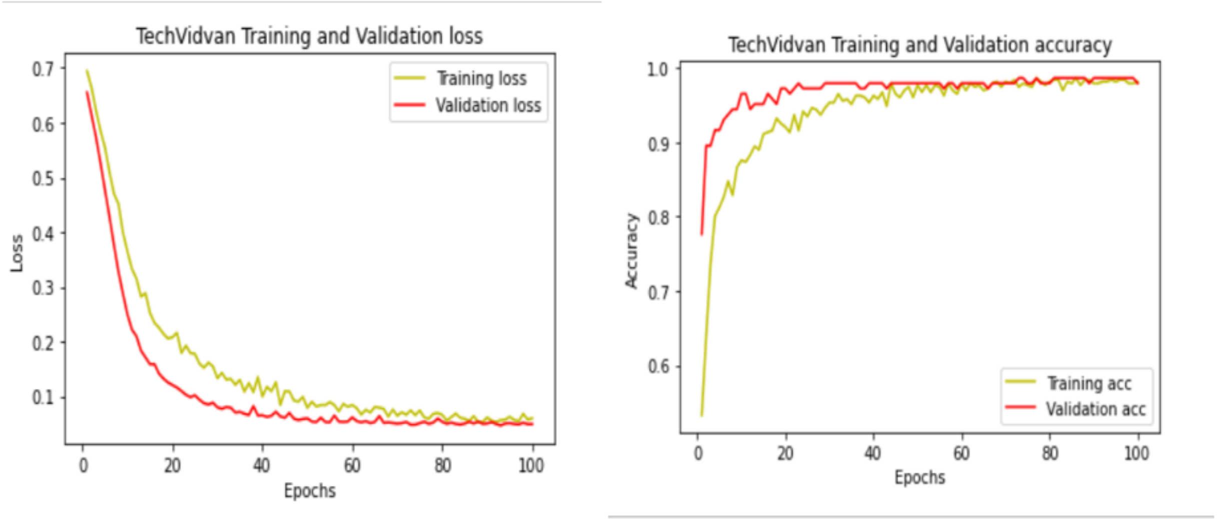


Figure 21: Neural network

Figure 21:In this we can see the our model is working very efficeny and accurately in the classifying whether the breast caner is of Malignant type or Benign type. The breast cancer data using Neural network model of accuracy is 98.8% on training data and 97.9% on accuracy validation data.As we have seen that our model is clasifying test data very efficiently and accurately.

3.6 Conclusion

In Conclusion, we can see that the dimension reduction was key for our analysis given our dataset had many features. We attempted PCA and Non-metric MDS analysis, both indicating cancer cells were contributing to the Breast cancer diagnosis. We move forward with three classification techniques KNN classifier, Gaussian, and Neural network, into the appropriate diagnosis. First, compare between theKNN and Gaussian in the result ROC curve, the Gaussian covered more area. All three classifiers' result indicating the accuracy neural network is the best classifier.

4 Brain stroke Prediction

4.1 Introduction

A stroke, sometimes called a brain attack, occurs when something blocks blood supply to part of the brain or when a blood vessel in the brain bursts. In either case, parts of the brain become damaged or die. A stroke can cause lasting brain damage, long-term disability, or even death. A stroke is an interruption of the blood supply to any part of the brain if blood flow was stopped for longer than a few seconds and the brain cannot get blood and oxygen, brain cells can die,and thr abilities controlled by that area of the brain and are lost.

Our goal is to predict if the patients had a stroke or not.This notebook aims at building a predictive model that helps in identifying whether a patient is likely to suffer from a brain stroke Our goal is to predict if they will experience a Brain Stroke or not. We are given attributes such as their Marital Status,Work Type,etc which can affect their mental state. We are also given attributes such as BMI,etc which describe their physical state of body.Machine learning algorithms will be able to detect the brain stroke effected or not effected. The prediction accuracy of the model will be analyzed by the end of this project. This project will provide insight of machine learning algorithms to perform classification on imbalanced dataset problems.

4.2 Literature Review

A stroke is a life-threatening condition that happens when part of your brain doesn't have enough blood flow. This most commonly happens because of a blocked artery or bleeding in your brain. Without a steady supply of blood, the brain cells in that area start to die from a lack of oxygen.

A stroke is a medical condition in which poor blood flow to the brain causes cell death. There are two main types of stroke: ischemic(his case), due to lack of blood flow, and hemorrhagic, due to bleeding. Both cause parts of the brain to stop functioning properly. Signs and symptoms of a stroke may include an inability to move or feel on one side of the body, problems understanding or speaking, dizziness, or loss of vision to one side. Signs and symptoms often appear soon after the stroke has occurred.If symptoms last less than one or two hours, the stroke is a transient ischemic attack (TIA), also called a mini-stroke. A hemorrhagic stroke may also be associated with a severe headache.The symptoms of a stroke can be permanent. The main risk factor for stroke is high blood pressure.Other risk factors include high blood cholesterol, tobacco smoking, obesity, diabetes mellitus, a previous TIA, end-stage kidney disease, and atrial fibrillation. An ischemic stroke is typically caused by blockage of a blood vessel, though there are also less common causes.A hemorrhagic stroke is caused by either bleeding directly into the brain or into the space between the brain's membranes. Bleeding may occur due to a ruptured brain aneurysm. Diagnosis is typically based on a physical exam and supported by medical imaging such as a CT scan or MRI scan. A CT scan can rule out bleeding, but may not necessarily rule out ischemia, which early on typically does not show up on a CT scan. Other tests such as an electrocardiogram (ECG) and blood tests are done to determine risk factors and rule out other possible causes. Low blood sugar may cause similar symptoms. The medications and treatments used vary depending on the type of stroke and how soon a person receives treatment after the stroke. There are also long-term treatments for stroke. These happen in the days and months after emergency treatment deals with a stroke's immediate threat.

Overall, your healthcare provider is the best person to tell you what kind of treatment(s) they recommend. They can tailor the information they provide to your specific case, including your medical history, personal circumstances and more. There are two mechanical devices approved by the FDA for use in this situation. They are the MERCI clot retrieval system, and the Penumbra Clot aspiration system.

Both devices are navigated into the blocked brain blood vessel by entering into the artery at the top of the thigh and using X-ray guidance. The MERCI device is a helical snare which grasps the clot and is then pulled out, and the Penumbra device uses aspiration to suction out the clot. In addition, an FDA off-label use of the rTPA to dissolve the blood clot is done by injecting the drug directly into the clot at a lower dose using the same catheterization techniques.

A healthcare provider can diagnose a stroke using a combination of a neurological examination, diagnostic imaging and other tests. During a neurological examination, a provider will have you do certain tasks or answer questions. As you perform these tasks or answer these questions, the provider will look for telltale signs that show a problem with how part of your brain works. Improve your lifestyle. Eating a healthy diet and adding exercise to your daily routine can improve your health. You should also make sure to get enough sleep (the recommended amount is seven to eight hours). There are many things you can do to reduce your risk of having a stroke. While this doesn't mean you can prevent a stroke, it can lower your risk. Actions you can take include: Avoid risky lifestyle choices or make changes to your behaviors. Smoking and tobacco use, including vaping, recreational drug use or prescription drug misuse, and alcohol misuse can all increase your risk of having a stroke. It's important to stop these or never start them. If you struggle with any of these, talking to your healthcare provider is important. Your provider can offer you guidance and resources that can help you change your lifestyle to avoid these behaviors. See your primary care provider for a checkup or wellness visit annually. Yearly wellness visits can detect health problems — especially ones that contribute to having a stroke — long before you feel any symptoms.

4.3 Project Goals

The Goal of the analysis of brain stroke the patients suffered or did not suffer, The analysis of gender, age, residence type, ever-married, work type, and smoking status suffered or not suffered, and visualize the distribution of the columns. Dealing with balance set. Predict the machine learning algorithm of KNeighbors, Random forest, Logistic regression, and decision tree classifications.

4.4 Problem statement

- Apply Machine learning models to verify that the model is a good fit
- Finalize a model based on it's metrics
- Predict if the person will have a stroke or not using the final model.

4.5 Project Descriptions

4.5.1 Data loading

The project covers the entire data analytics work flow other than data collection as the data was collected from available kaggle dataset.

4.5.2 Data Collection

As mentioned earlier the data is collected from the kaggle dataset.The data set entries are 4982 entries and 11 features.Stroke feature are defined by 1 if patient had stroke or 0 if not.The other 5 features are categorical variable and 4 features are binary values.

“Column Information”

1) gender: “Male”, “Female” or “Other”
2) age: age of the patient
3) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
4) heartdisease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
5) evermarried: “No” or “Yes”
6) worktype: “children”, “Govtjov”, “Neverworked”, “Private” or “Self-employed” 7) Residencetype: “Rural” or “Urban”
7) avgglucoselevel: average glucose level in blood

8) bmi: body mass index
9) smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown" Note: "Unknown" in smoking_status means that the information is unavailable for this patient
10) stroke: 1 if the patient had a stroke or 0 if not

Stroke feature are defined by 1 if patient had stroke or 0 if not.The other 5 are labelled columns and 4 are numerical columns.

4.5.3 Data cleaning

Data cleaning is the process of removing incorrect, duplicate, or otherwise erroneous data from a dataset. These errors can include incorrectly formatted data, redundant entries, mislabeled data, and other issues; they often arise when two or more datasets are combined together. Data cleaning improves the quality of your data as well as any business decisions that you draw based on the data.

There is no one right way to clean a dataset, as every set is different and presents its own unique slate of errors that need to be corrected. Many data cleaning techniques can now be automated with the help of dedicated software, but some portion of the work must be done manually to ensure the greatest accuracy.

"Check missing values"

Figure 22: shows no "Null" values in this data set.

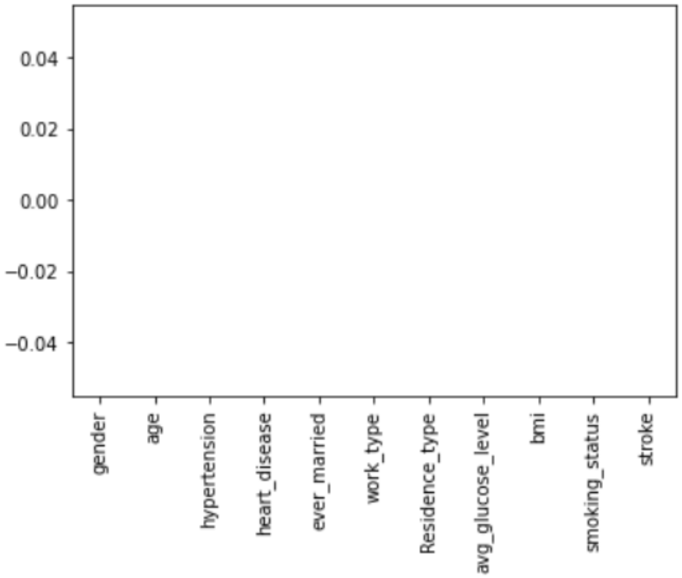


Figure 22: No Nulls in the Data

4.5.4 Exploratory analysis

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate. Originally developed by American mathematician John Tukey in the 1970s, EDA techniques continue to be a widely used method in the data discovery process

“Verifying the correlation between our variables”

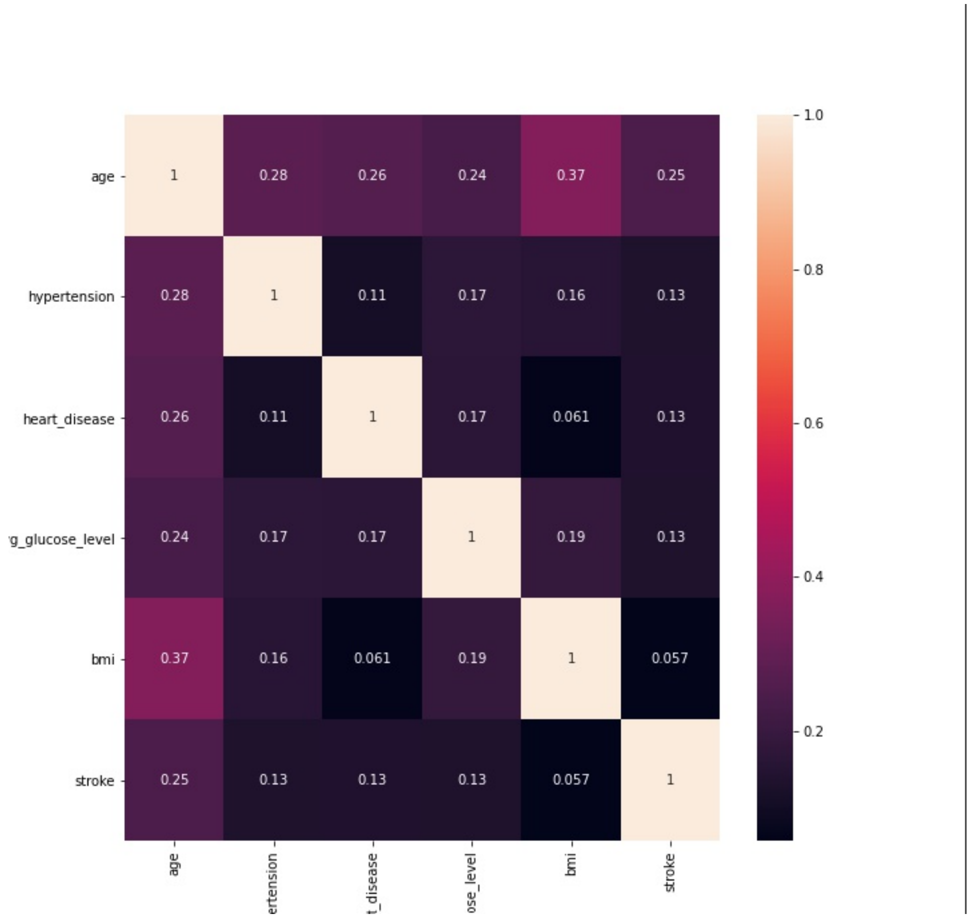


Figure 23: Correlation Between the variables

Figure 23 :The above visuals we can look at the diagonal value 1 that is correlated that matching with the same data points.The pearsons coefficient of r values are so small .These are not highly corelation between the variables.we can see the plot of other variables there is no higher corrlation between the variables so as expected the bmi and the heart diseases are little bit high,Which make sense as well stroke and bmi.

“Categorical variables comparing with target variables”

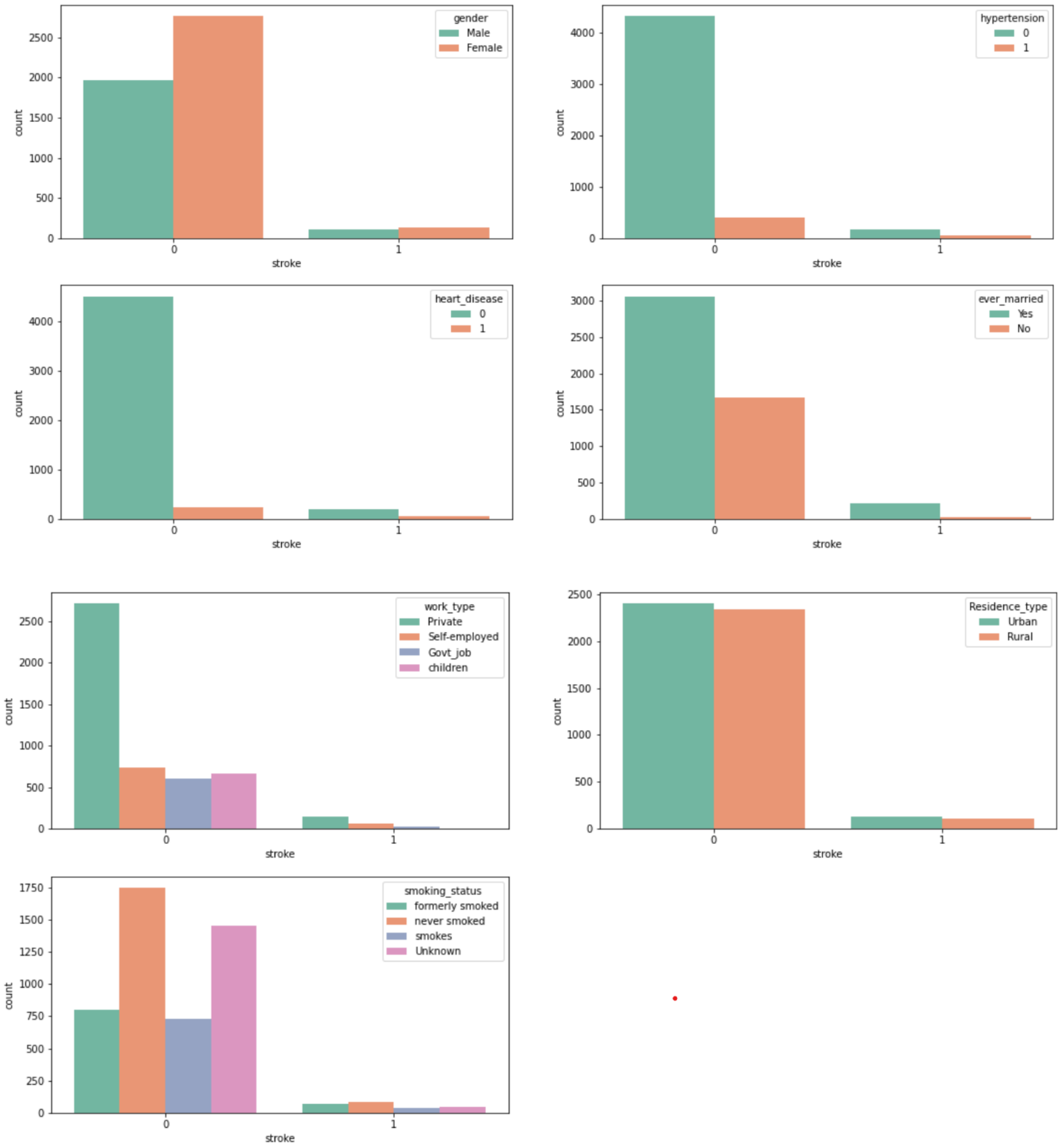


Figure 24: Categorical variables vs Stroke

Figure 24:We can see that the above visuals show we can come to the conclusion that we have more women than men in our base,Womens are more likey to have a stroke. There are only a few people who have hypertension and a few people with heart disease. The work type visual show most private job people are most suffered a stroke. When compared with every married visual most never-married people are suffering. Similarly work type visual says the private sector of people have a stroke, self-employed likely have a stroke, and govt job sector people got a stroke. The residence type of rural and urban patients of the base is balanced. A large portion of patients have never smoked in their lifetime, the majority being women. As compared to males many female patients do not smoke. The probable reason is that the data captured have

a majority of female records. It is observed that patients who formerly smoked are a little higher than those who currently continue smoking.

“Continuous variables”

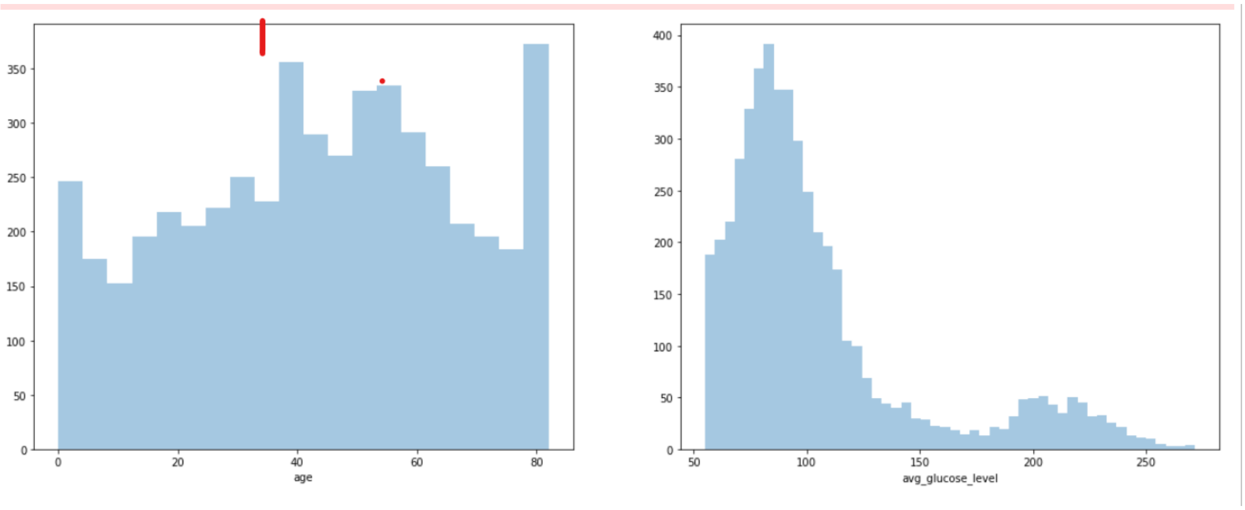


Figure 25: continuous variable distribution

Figure 25 & 26:Continuous variable we can see that age is well balanced.We have a people of all ages and we have a good concentration in the average of 80 years when we look at gulcose level We can see that the most people do not have level very high. When we can look at the BMI its practically a normal distribution then we can say that the data is well distributed.

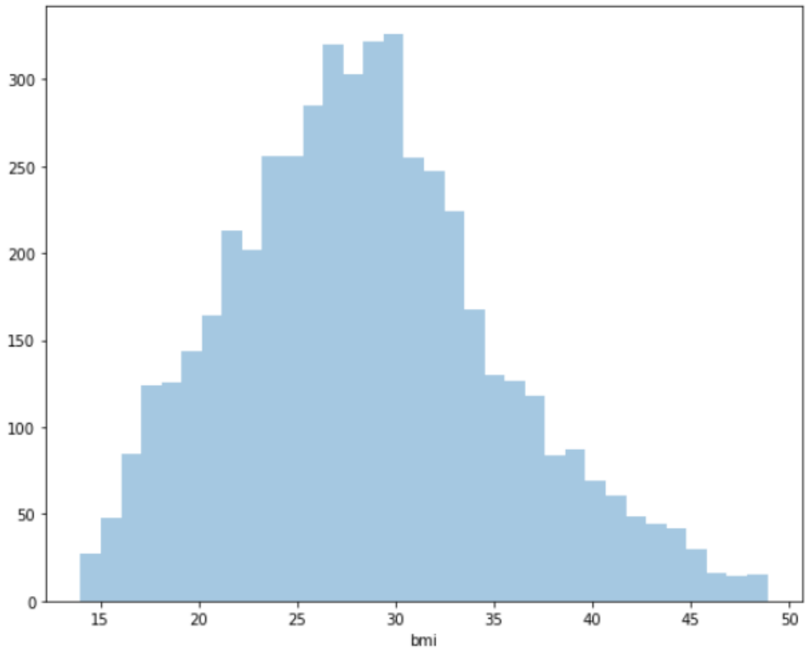
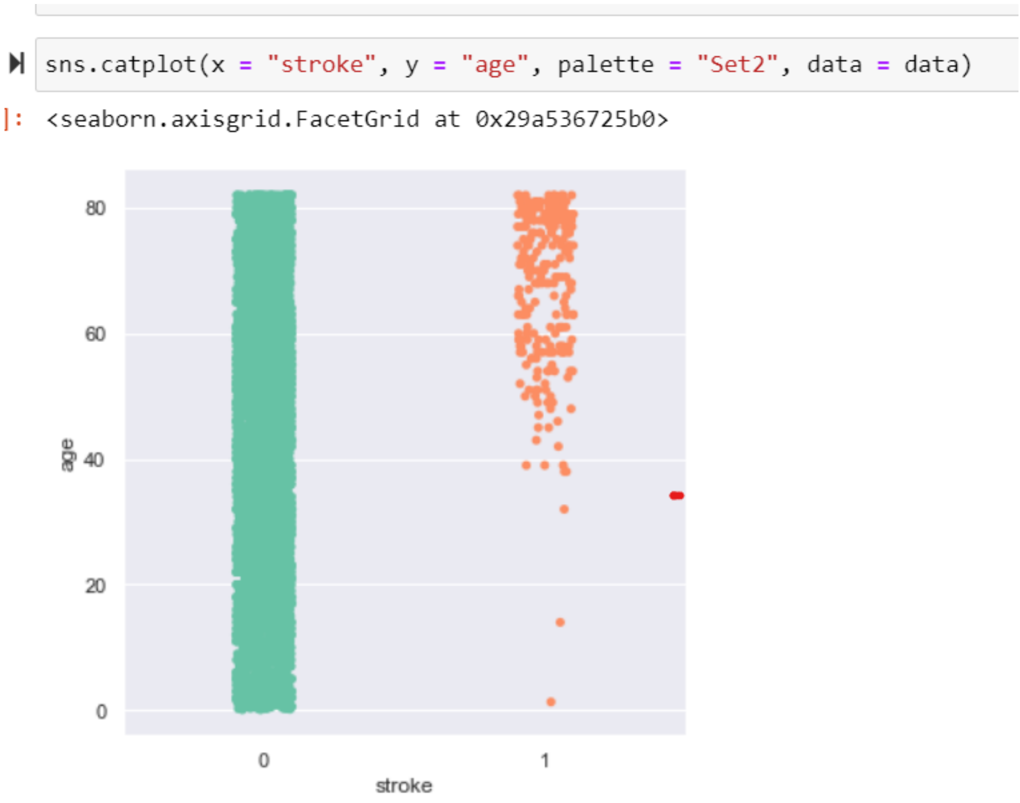


Figure 26: Continuous variable distribution

“Continuous variable comparing with our target variable”

Figure 27:The below visual shows we look at the age variable. You can see that only people above 40 years old start to have a stroke and then the below age is very rare, and see the 0 to 20 years below age very low have a stroke. The trend is like the older the person, the more likely is to have a stroke.

In Figure 28: The conclusion when we look at the BMI we can see that its more common people to have a stroke when the bmi is on average ,not too high and not too low we can see that our gulcose variable we can see that it is well distributed.



```
sns.catplot(x = "stroke", y = "age", palette = "Set2", data = data)
```

```
]: <seaborn.axisgrid.FacetGrid at 0x29a536725b0>
```

Figure 27: Continuous variable vs target variable

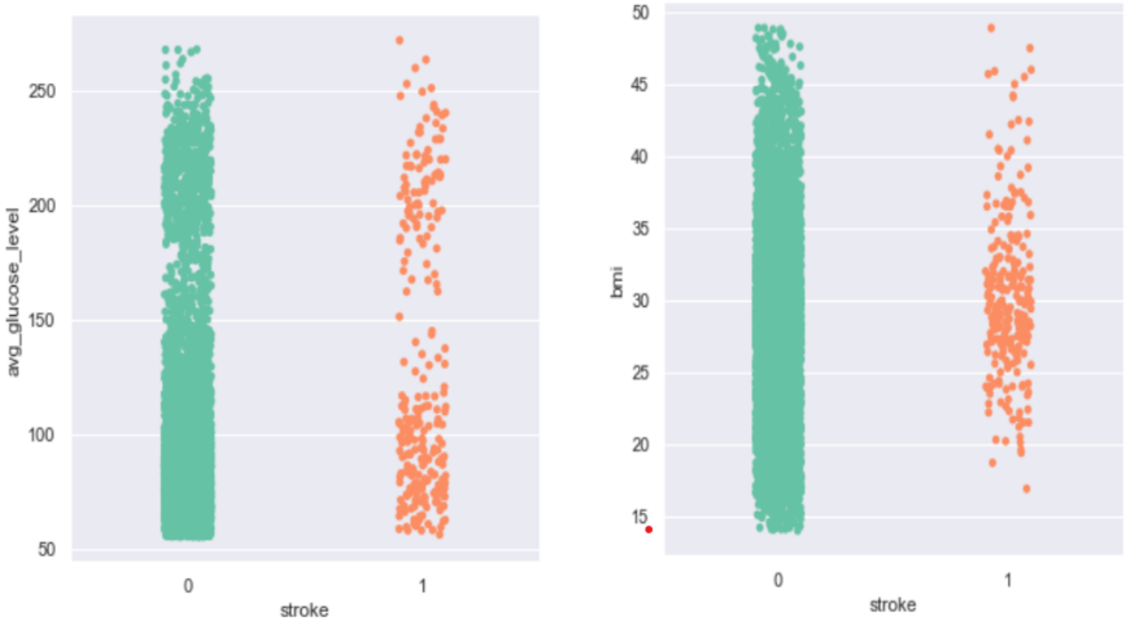


Figure 28: Continuous variable vs target variable

Figure 29 image shows the distribution of strokes visually. Most of the strokes affected (5%) and not involved people (95%) in this data.If the data set is used as the base dataset for the predictive models. We can say that the data is imbalanced. In our data set, the target variable is a stroke.

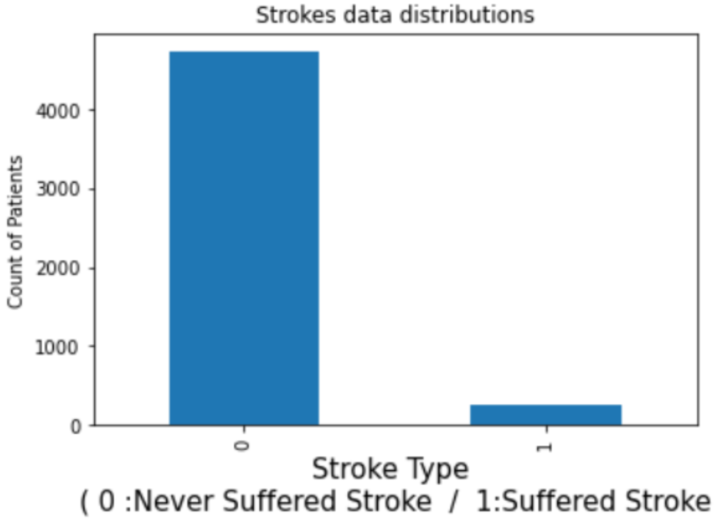


Figure 29: Data is imbalanced

4.5.5 Dealing with imbalanced data

Imbalanced classification involves developing predictive models on classification datasets that have a severe class imbalance.The challenge of working with imbalanced datasets is that most machine learning techniques will ignore, and in turn have poor performance on, the minority class, although typically it is performance on the minority class that is most important

Resampling data is one of the most commonly preferred approaches to deal with an imbalanced dataset. There are broadly two types of methods for this i) Undersampling ii) Oversampling. In most cases, oversampling is preferred over undersampling techniques. The reason being, in undersampling we tend to remove instances from data that may be carrying some important information. In this article, I am specifically covering some special data augmentation oversampling techniques: SMOTE and its related counterparts.

4.5.6 Smote

SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line. Specifically, a random example from the minority class is first chosen. Then k of the nearest neighbors for that example are found (typically k=5). A randomly selected neighbor is chosen and a synthetic example is created at a randomly selected point between the two examples in feature space.

SMOTE is an oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together.

4.5.7 Balancing class

Analyzing the data we can see that we have a lot more data with Non Stroke, so the models will learning more about this data than when the person doesn't has Stroke and can't learn about when this person will has Brain Stroke, then we need to balance the classes and see what we can do with this data.

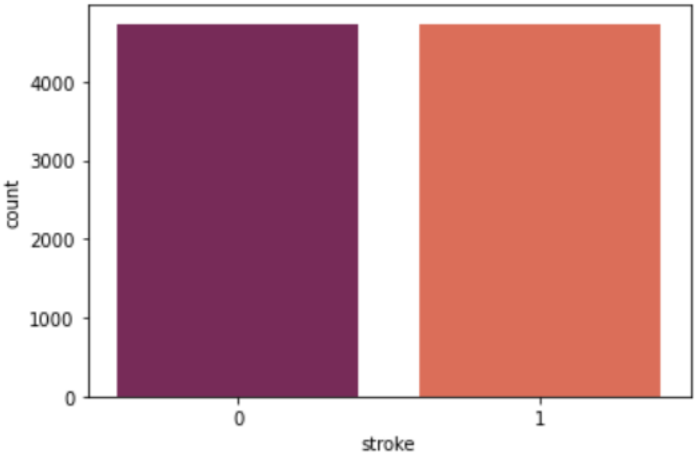


Figure 30:They never suffered a stroke and suffered stroke both the sample proportions are the same so will start the classification models with balance data.

4.6 Data Preprocessing

4.6.1 Preparing data

As you see our data, some of categorical columns have only two unique values; like 'Gender' or 'Ever Married' columns. We can simply use 'get_dummies' function to deal with them but in this case, the number of features and the complexity of the model will increase.

So, to avoid of this, I convert the values of 'gender', 'ever married' and 'residence type' work_type columns to binary values

4.6.2 Dataset splitting

One of the first decisions to make when starting a modeling project is how to utilize the existing data. One common technique is to split the data into two groups typically referred to as the training and testing sets23. The training set is used to develop models and feature sets; they are the substrate for estimating parameters, comparing models, and all of the other activities required to reach a final model. The test set is used only at the conclusion of these activities for estimating a final, unbiased assessment of the model's performance. It is critical that the test set not be used prior to this point. Looking at the test sets results would bias the outcomes since the testing data will have become part of the model development process. I have determined how many instances are considered suffered [Suffered ="1"] The exciting part I have been waiting for all this time: is training machine learning algorithms. First performed a 70/30 train-test split to test our algorithms' performance, splitting our balanced data set into two pieces.

4.6.3 Data scaling

In practice, different types of variables are often encountered in the same data set. A significant issue is that the range of the variables may differ a lot. Using the original scale may put more weight on the variables with a large range. In order to deal with this problem, the technique of features rescaling need to be applied to independent variables or features of data in the step of data preprocessing. The terms “normalization” and “standardization” are sometimes used interchangeably, but they usually refer to different things.

The goal of applying feature scaling is to make sure features are on almost the same scale so that each feature is equally important and make it easier to process by most machine-learning algorithms.

This means that you’re transforming your data so that it fits within a specific scale, like 0-100 or 0-1. You want to scale data when you’re using methods based on measures of how far apart data points, like models, or KNN. With these algorithms, a change of “1” in any numeric feature is given the same importance.

$$X_n = \frac{x - \mu}{\sigma} = \frac{x - Mean(x)}{StdDiv(x)}$$

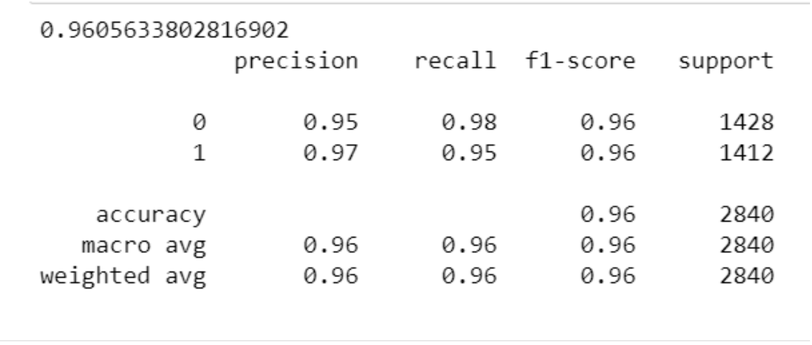
4.7 Model classifier

In this section I will train five types of classifiers and decide which classifier will be more effective in detecting strokes identifier. As we have already split the data into training and testing sets and separated the features from the labels I just had to pass the data to the model classifiers.To get a better algorithem which one is the best on our data

4.7.1 K-nearest neighbors

K nearest neighbor is based on supervised learning technique.k-nearest neighbor is one of the simplest algorithem techniques and Knn algorithm assumes the similarity between the e set of data (training data) consists of a set of input data and correct responses corresponding to every piece of data.Based on this training data, the algorithm must generalize such that it is able to correctly (or with a low margin of error) respond to all possible inputs. The algorithm should produce sensible outputs for inputs that were not encountered during training.As our data we can get best accuracy.Though it would be a better to check the accuracy from all classifiers as we havethe ability of processing the data. The KNN algorithm is also part of a family of “lazy learning” models, meaning that it only stores a training dataset versus undergoing a training stage. This also means that all the computation occurs when a classification or prediction is being made. Since it heavily relies on memory to store all its training data, it is also referred to as an instance-based or memory-based learning method.

$$\hat{f}(x_q) <- \sum_{v \varepsilon V i=1}^{argmaxK} \delta(v, f(x_i))$$



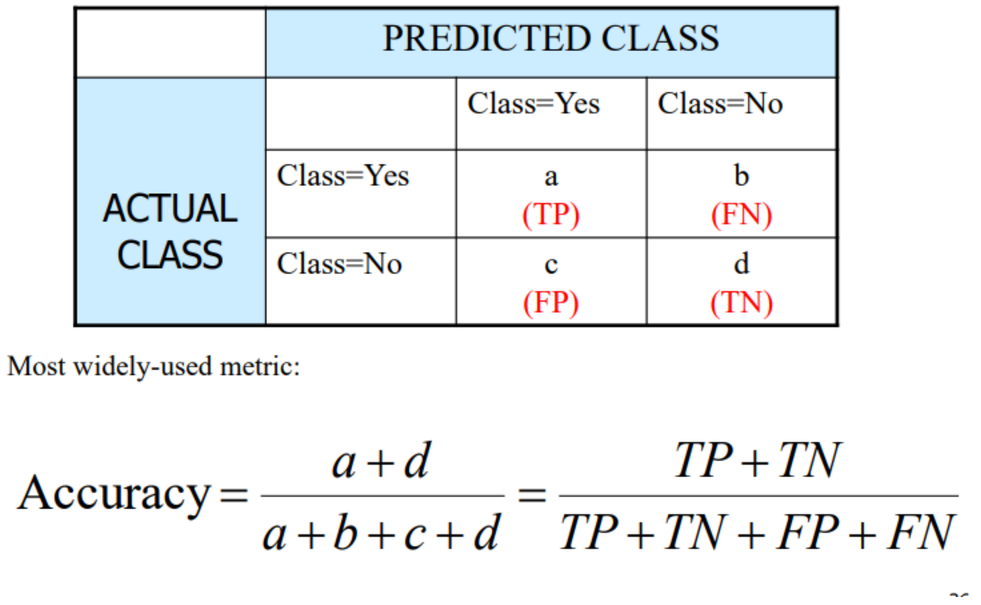
```
0.9605633802816902
              precision    recall  f1-score   support

           0       0.95      0.98      0.96      1428
           1       0.97      0.95      0.96      1412

    accuracy                           0.96      2840
   macro avg       0.96      0.96      0.96      2840
weighted avg       0.96      0.96      0.96      2840
```

Figure 30: KNeighbors classifier

Figure 30 : The result shows the Knn algorithm was able to classify all the 2840 records. In the test set, 96% accuracy is excellent. Algorithms performed very well with the data set. The precision of target variable 0 is 96% and ! is 97%, which is high and compared to the f1-score is 96%, and 96% is balance. The recall 0 is 97% and 1 is 96%.

4.7.2 Confusion matrix

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing Confusion matrix is a very popular measure used while solving classification problems. It can be applied to binary classification as well as for multiclass classification problems.The diagonal elements show the number of correct classifications for each class The off-diagonal elements provides the misclassifications.Confusion matrix helps in identifying the accuracy of the model classifier by classifying True Poistive,True Negative, False Positive and False Negative values. The left most part gives the result for the training dataset the 70% of the actual dataset. The middle part gives the result for the testing dataset. The right most part is the prediction for the actual dataset.



		PREDICTED CLASS	
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

Most widely-used metric:

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

Figure 31: The confusion matrix of the KNN classifier the true positive cases and True negative cases are correctly classified as Low-Risk and High-Risk 98% and 94%.The true negative cases in which we predicted the brain strokes High_Risk.The True positive-negative cases which we predicted don't have a stroke. The False negative cases and False positive cases are incorrectly classified as Low-Risk and High_Risk is 6% and 2%. False positive cases which we predicted have High-risk of stroke also known as "Type 1 error" and False negative cases which predicted Have Low_Risk of stroke also known as "Type 11 error.

4.7.3 Logistic regression

This type of statistical model (also known as logit model) is often used for classification and predictive analytics. Logistic regression estimates the probability of an event occurring, such as voted or didn't vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In logistic regression, a logit transformation is applied on the odds—that is, the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds, and this logistic function is represented by the following formulas:

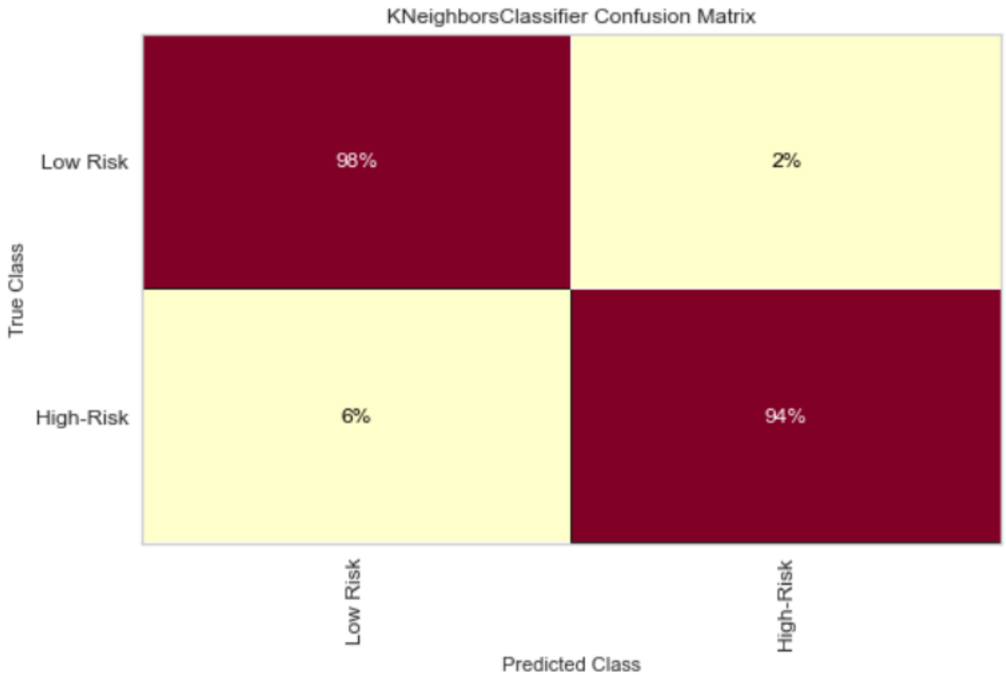
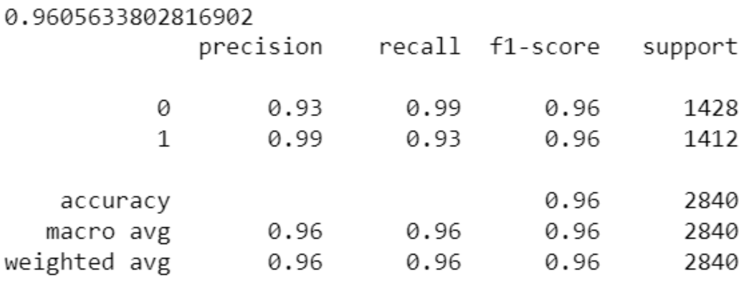


Figure 31: KNeighbors classifier of the confusion matrix

Logit(pi) = 1/(1+ exp(-pi))

ln(pi/(1-pi)) = Beta_0 + Beta_1X_1 + . . . + B_kK_k

In this logistic regression equation, logit(pi) is the dependent or response variable and x is the independent variable. The beta parameter, or coefficient, in this model, is commonly estimated via maximum likelihood estimation (MLE). This method tests different values of beta through multiple iterations to optimize for the best fit of log odds. All of these iterations produce the log-likelihood function, and logistic regression seeks to maximize this function to find the best parameter estimate. Once the optimal coefficient (or coefficients if there is more than one independent variable) is found, the conditional probabilities for each observation can be calculated, logged, and summed together to yield a predicted probability. For binary classification, a probability less than .5 will predict 0 while a probability greater than 0 will predict 1. After the model has been computed, it's best practice to evaluate how well the model predicts the dependent variable, which is called goodness of fit.



```
0.9605633802816902
              precision    recall  f1-score   support

           0       0.93      0.99      0.96      1428
           1       0.99      0.93      0.96      1412

    accuracy                           0.96      2840
   macro avg       0.96      0.96      0.96      2840
weighted avg       0.96      0.96      0.96      2840
```

Figure 32: Logistic Regression Classifier

Figure 32: The result shows the Logistic regression classifier of accuracy is 96% is excellent. The algorithm

performed well in the dataset. The precision of target variable 0 is 93% and ! is 1, which is high compared to the f1-score is 96%, and 96% is balanced. The recall 0 is 93% and 1 is 1

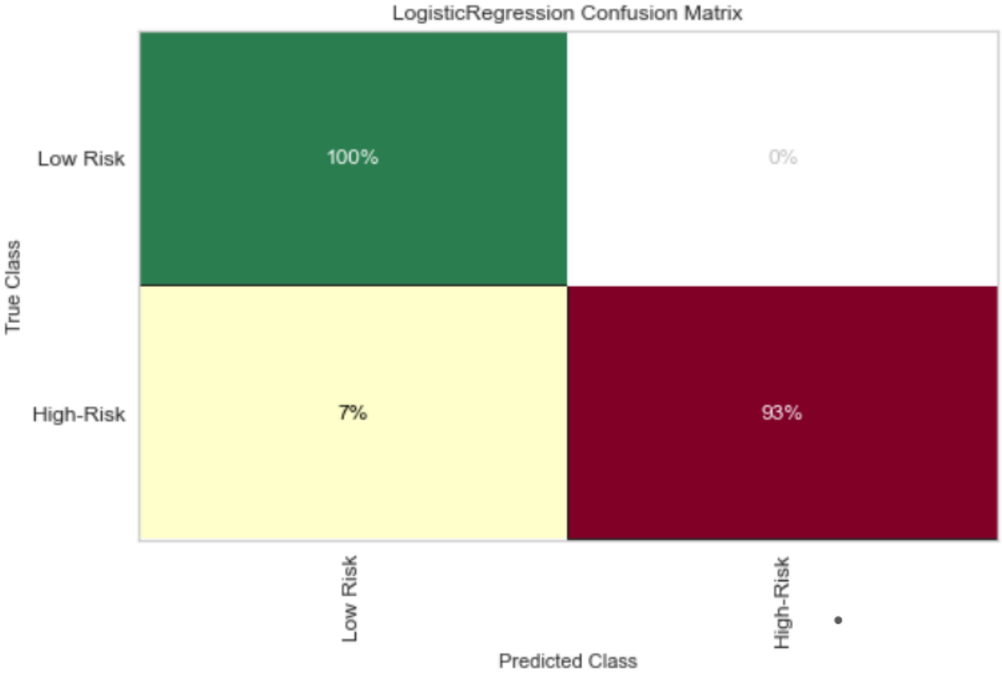


Figure 33: Logistic Regression of Confusion matrix

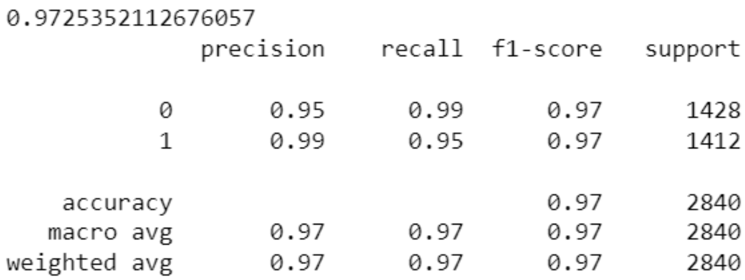
Figure 33: The confusion matrix of the Logistic Regression classifier the true positive cases and True negative cased are correctly classified as Low-Risk and High-Risk 100% and 93%.The true negative cases in which we predicted the brain strokes High_Risk.The True positive-negative cases which we predicted don't have a stroke. The False negative cases and False positive cases are incorrectly classified as Low-Risk and High_Risk is 7% and 0%. False positive cases which we predicted have High-risk of stroke also known as "Type 1 error" and False negative cases which predicted Have Low_Risk of stroke also known as "Type 11 error.

4.7.4 Random forest classifier

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training Random forests generally outperform decision trees, but their accuracy is lower than gradient-boosted trees.[citation needed] However, data characteristics can affect their performance. Random forest is a supervised learning algorithm. The "forest" it builds is an ensemble of decision trees, usually trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result.

Figure 34: The result shows the Random Forest classifier of accuracy is 97% is excellent. The prediction of algorithm performed very well in the dataset. The precision of target variable 0 is 95% and 1 is 99%, which is high compared to the f1-score is 97%, and 97% is balanced. The recall 0 is 99% and 1 is 95%. The Random Forest classifier is the best model

Figure 35: The confusion matrix of the Random Forest classifier the true positive cases and True negative cased are correctly classified as Low-Risk and High-Risk 99% and 95%.The true negative cases in which we predicted the brain strokes High_Risk.The True positive-negative cases which we predicted don't have a



```
0.9725352112676057
              precision    recall  f1-score   support

           0       0.95      0.99      0.97      1428
           1       0.99      0.95      0.97      1412

    accuracy                           0.97      2840
   macro avg       0.97      0.97      0.97      2840
weighted avg       0.97      0.97      0.97      2840
```

Figure 34: Random Forest Classifier

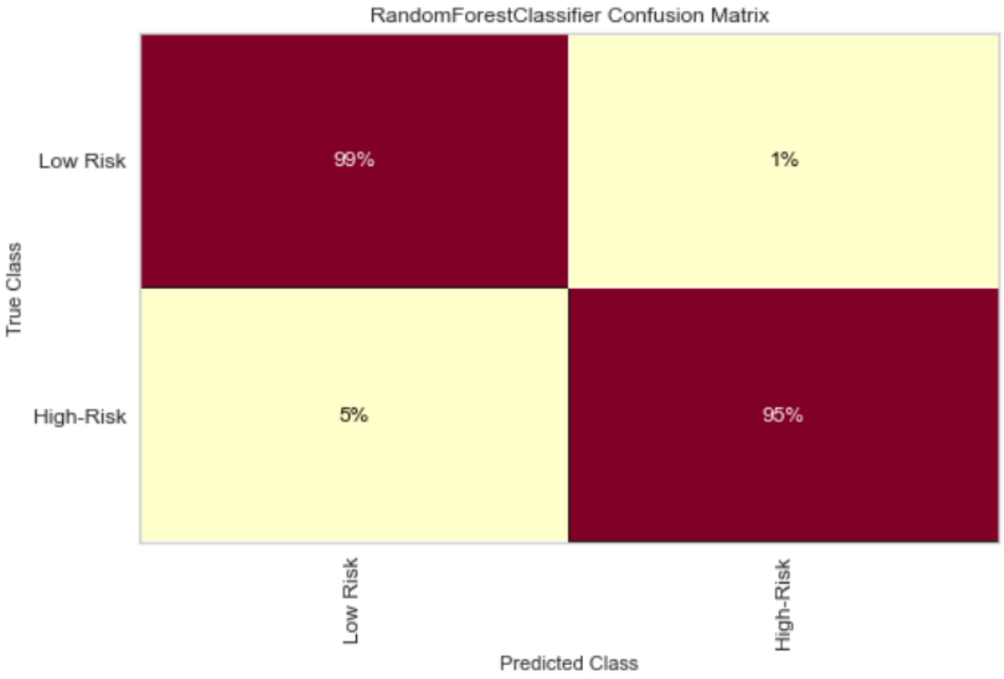
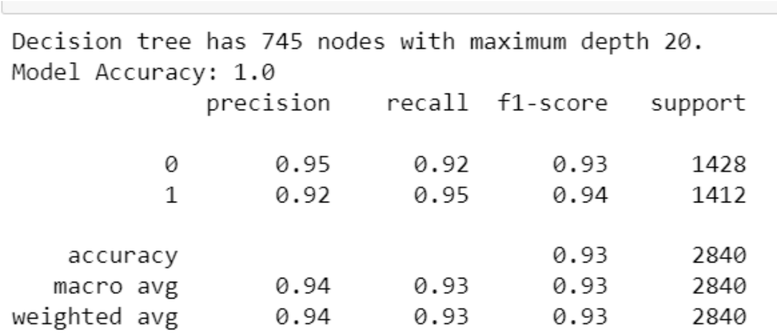


Figure 35: Random Forest of the Confusion matrix

stroke. The False negative cases and False positive cases are incorrectly classified as Low-Risk and High_Risk is 5% and 1%. False positive cases which we predicted have High-risk of stroke also known as “Type 1 error” and False negative cases which predicted Have Low_Risk of stroke also known as “Type 11 error.

4.7.5 Decision tree

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.A decision tree is a map of the possible outcomes of a series of related choices. It allows an individual or organization to weigh possible actions against one another based on their costs, probabilities, and benefits. They can can be used either to drive informal discussion or to map out an algorithm that predicts the best choice mathematically. For instance, in the example below, decision trees learn from data to approximate a sine curve with a set of if-then-else decision rules. The deeper the tree, the more complex the decision rules and the fitter the model.



```
Decision tree has 745 nodes with maximum depth 20.
Model Accuracy: 1.0
              precision    recall  f1-score   support

           0       0.95      0.92      0.93      1428
           1       0.92      0.95      0.94      1412

    accuracy                           0.93      2840
   macro avg       0.94      0.93      0.93      2840
weighted avg       0.94      0.93      0.93      2840
```

Figure 36: Decision Tree Classifier

Figure 36: The result shows the Decision Tree classifier of accuracy is 94% is good. The decision tree has 745 nodes with a maximum depth of 20 and the model accuracy is 1. The algorithm performed well in the dataset. The precision of target variable 0 is 95% and 1 is 92%, which is not high compared to the f1-score is 93%, and 97% is balanced. The recall 0 is 92% and 1 is 95%.

Figure 37: The confusion matrix of the Decision Tree classifier the true positive cases and True negative cased are correctly classified as Low-Risk and High-Risk 93% and 94%.The true negative cases in which we predicted the brain strokes High_Risk.The True positive-negative cases which we predicted don’t have a stroke. The False negative cases and False positive cases are incorrectly classified as Low-Risk and High_Risk is 6% and 7%. False positive cases which we predicted have High-risk of stroke also known as “Type 1 error” and False negative cases which predicted Have Low_Risk of stroke also known as “Type 11 error.

“Compare between Four Algorithms”

The final results of four machine learning models we have very good models and some not-so-good models. In our KNN algorithm and Logistic regression of accuracy is 96%is we had good accuracy. when compared to the Random Forest classifier excellent accuracy of 97% is the best model, The decision tree classifier of accuracy is 94%. They predict the person will have a stroke in this model. Finally, our Random forest classifier is the best model because the accuracy is high and the misclassification of high risk(actual) and low risk(prediction) is lower at 5%. Random forest classifiers predict that 99% of the time the person is at low risk while he was actually at a lower risk.so it predicts that 95% of the time the person is at high risk when one was actually at high risk and only 5% of the time it misclassifies to Low Risk when the person is actually at High Risk. The value is the lowest among all the models. The Matters the most because you don’t want to misclassify when the person will be at High Risk. Only 1% of the time it misclassifies to High Risk when the person was actually at Low Risk.

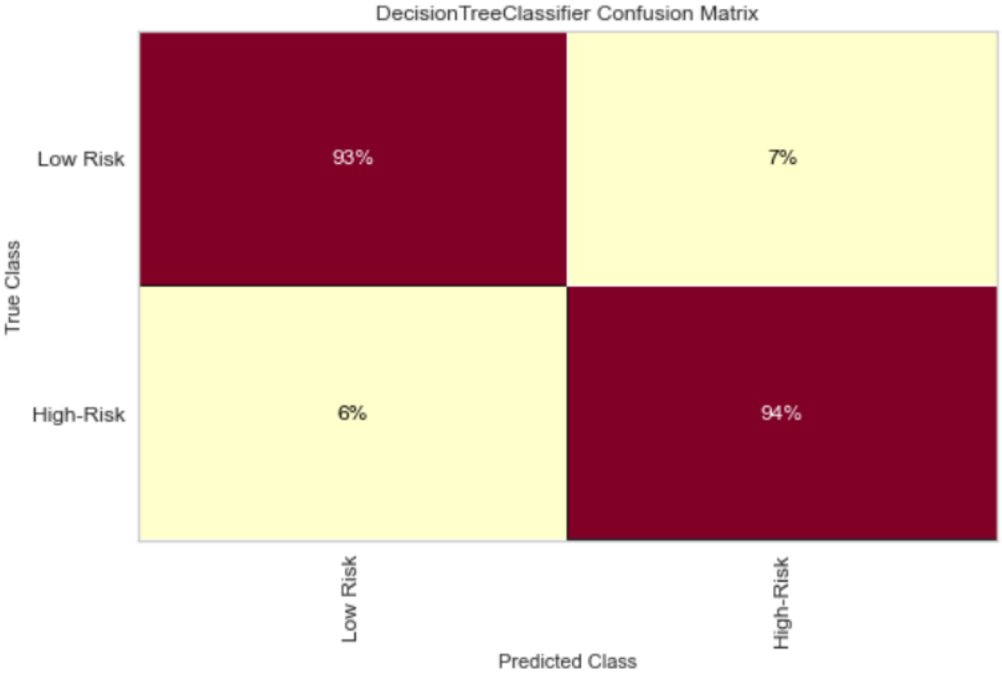


Figure 37: Decision Tree Classifier

4.8 Conclusion

In summary, having a highly imbalanced data set. we can see that numerical variable and categorical variables in our data visualization get a good gain. In our data analysis, we can say that older people and women are more likely to have a stroke. A smoker patient got a stroke in our data and private sector job people got a stroke. A. Few people got strokes with hypertension and heart disease. In our data, we used smote technique with an imbalance to balanced data. Apply the machine learning models the KNN classifier accuracy and Logistic regression accuracy both are the same. The decision tree classifier of the accuracy is small when compared to Knn and logistics. The Random forest classifier had great accuracy and the model is a good fit. We predict the person will have a stroke. Random forest classifier gives the best model

5 Conclusion

In conclusion the comparative study of the actual expenses vs budget and the two vacation destinations,The expenses in India are less when compare to Dubai with a minor deviation from the budget.In the breast cancer analysis, the dimension reduction of principal component analysis and nonmetric MDS both indicated to cells were contributors to Breast cancer diagnosis. To predict Breast cancer of three classifications the results indicate the neural network is the best model. In the brain stroke data, the Random forest classifier is a good fit to predict that most the person is at High Risk.

6 Biblogrphy

"A New Way of Working." Monday.com, https://monday.com/.

"Stress Plot Propertymanager." Stress Plot PropertyManager - 2021 - SOLIDWORKS Help, https://help.solidworks.com/2021/english/SolidWorks/cworks/HIDD_HELP_STRESSPLOTTYPE.htm.

Learning, UCI Machine. “Breast Cancer Wisconsin (Diagnostic) Data Set.” Kaggle, 25 Sept. 2016, https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data.

Breast Cancer Genetics :: The National Breast Cancer Foundation. (2019). Retrieved from https://www.nationalbreastcancer.org/breast-cancer-genetics

Team, TechVidvan. “Breast Cancer Classification Using Machine Learning.” TechVidvan, 11 Aug. 2021, https://techvidvan.com/tutorials/breast-cancer-classification/.

Chris Woodford. Last updated: August 30. “How Neural Networks Work - a Simple Introduction.” Explain That Stuff, 30 Aug. 2021, https://www.explainthatstuff.com/introduction-to-neural-networks.html.

The Literature Review ::Retrieved From https://repositorio-aberto.up.pt/bitstream/10216/70529/2/30766.pdf.

DeepAI. “Gaussian Distribution.” DeepAI, DeepAI, 17 May 2019, https://deepai.org/machine-learning-glossary-and-terms/gaussian-distribution.

Brownlee, Jason. “Gaussian Processes for Classification with Python.” MachineLearningMastery.com, 2 Aug. 2020, https://machinelearningmastery.com/gaussian-processes-for-classification-with-python/.

Narkhede, Sarang. “Understanding AUC - Roc Curve.” Medium, Towards Data Science, 15 June 2021, https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.

“Non-Metric Multidimensional Scaling - Gusta Me.” Google Sites: Sign-In, https://sites.google.com/site/mb3gustame/dissimilarity-based-methods/nmds.

Joshuaebner. “Non-Metric Multidimensional Scaling (NMDS): What? How?” Archetypal Ecology, 18 Feb. 2018, https://archetypalecology.wordpress.com/2018/02/18/non-metric-multidimensional-scaling-nmds-what-how/.

Mangale, Sanchita. “Scree Plot.” Medium, Medium, 28 Aug. 2020, https://sanchitamangale12.medium.com/scree-plot-733ed72c8608.

Rick Wicklin on The DO Loop. “What Are Biplots?” The DO Loop, 6 Nov. 2019, https://blogs.sas.com/content/iml/2019/11/06/what-are-biplots.html.

“Different Kinds of Breast Lumps.” Stony Brook Cancer Center, https://cancer.stonybrookmedicine.edu/breast-cancer-team/patients/bse/breastlumps.

“Breast Tumors.” National Breast Cancer Foundation, 18 May 2022, https://www.nationalbreastcancer.org/breast-tumors/.

“Introduction to Dimensionality Reduction.” GeeksforGeeks, 27 Sept. 2022, https://www.geeksforgeeks.org/dimensionality-reduction/.

Akbasli, Izzet Turkalp. “Brain Stroke Prediction Dataset.” Kaggle, 16 July 2022, https://www.kaggle.com/datasets/zzettrkalpakbal/full-filled-brain-stroke-dataset.

Italosimoes. “HEART ATTACK - Analysis and Predictions.” Kaggle, Kaggle, 14 May 2021, https://www.kaggle.com/code/italosimoes/heart-attack-analysis-and-predictions.

“What Is the K-Nearest Neighbors Algorithm?” IBM, https://www.ibm.com/topics/knn.

Harrison, Onel. “Machine Learning Basics with the K-Nearest Neighbors Algorithm.” Medium, Towards Data Science, 14 July 2019, https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761.

“What Is Logistic Regression?” IBM, https://www.ibm.com/topics/logistic-regression.

Italosimoes. “Brain_stroke_analysis_and_predictions.” Kaggle, Kaggle, 12 Aug. 2022, https://www.kaggle.com/code/italosimoes/brain-stroke-analysis-and-predictions.

“Stroke: What It Is, Causes, Symptoms, Treatment & Types.” Cleveland Clinic, https://my.clevelandclinic.org/health/diseases/5601-stroke.

“Random Forest.” Wikipedia, Wikimedia Foundation, 29 Nov. 2022, https://en.wikipedia.org/wiki/Random_forest.

“What Is a Decision Tree Diagram.” Lucidchart, https://www.lucidchart.com/pages/decision-tree.

“1.10. Decision Trees.” Scikit, https://scikit-learn.org/stable/modules/tree.html.

“Stroke: What It Is, Causes, Symptoms, Treatment & Types.” Cleveland Clinic, https://my.clevelandclinic.org/health/diseases/5601-stroke.

abdallahhassan22. “EDA + Stroke Prediction.” Kaggle, Kaggle, 15 Aug. 2022, https://www.kaggle.com/code/abdallahhassan22/eda-stroke-prediction.

Raphaelmarconato. “Brain Stroke - Eda, Balancing and ML.” Kaggle, Kaggle, 11 Aug. 2022, https://www.kaggle.com/code/raphaelmarconato/brain-stroke-eda-balancing-and-ml.

Satpathy, Swastik. “Smote: Overcoming Class Imbalance Problem Using Smote.” Analytics Vidhya, 6 Jan. 2021, https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/.

“About Stroke.” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 2 Nov. 2022, https://www.cdc.gov/stroke/about.htm.

Markham, Kevin. “Simple Guide to Confusion Matrix Terminology.” Data School, Data School, 3 Feb. 2020, https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/.

KDnuggets. “Data Transformation: Standardization vs. Normalization.” JPT, JPT, 4 Apr. 2022, https://jpt.spe.org/data-transformation-standardization-vs-normalization?gclid=CjwKCAiAp7GcBhA0EiwA9U0mtiqFanx9lmfLnQJxq
2ZEIAxiZ8X.

“Data Splitting”:: Retrived From https://bookdown.org/max/FES/data-splitting.html

Brownlee, Jason. “Smote for Imbalanced Classification with Python.” MachineLearningMastery.com, 16 Mar. 2021, https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/.

By:IBM Cloud Education. “What Is Exploratory Data Analysis?” IBM, https://www.ibm.com/cloud/learn/exploratory-data-analysis.

“Stroke (Brain Attack).” UCLA Health System, https://www.uclahealth.org/medical-services/radiology/clinical-services/stroke-brain-attack.

Sherrer, Kara. “Data Cleaning: Definition, Methods & Steps.” TechnologyAdvice, 30 June 2022, https://technologyadvice.com/blog/information-technology/data-cleaning/.