

**Module Code**

**MHI222956**

**Big Data**

**2021 – 22**

**Semester A**

**BSC-ITMB**

---

**Module Team**

Mr Saleh alaraimi

---



**الجامعة الوطنية**  
**National University**

Science & Technology للعلوم والتكنولوجيا

**COLLEGE OF ENGINEERING**

## **COURSEWORK 2**

**BIG DATA**

**Students Number :** .....

**Students Names :** .....

Department of Electrical & Communications Engineering

## Abstract

In the age of technology, massive amounts of data are generated as much as 2.5 quintillion bytes daily. This enormous amount of data and technology has helped industries small and big to identify trends and insights with the help of analytics that helps in enhancing performance and decision making. One of the industries has taken sound advantage of big data in the healthcare sector. Big data collected from patients, intelligent devices, etc., have helped the healthcare sector make better strategic decisions, enhance patient experience, and identify trends and even cure for life-threatening diseases like cancer. One of the most common types of cancer that have taken the lives of many is breast cancer. Breast cancer can be reduced if identified earlier, thus predicting it can also help save a patient's life. The coursework highlights the need for big data in business and decision making and its use case in the healthcare sector. Big data issues have also been discussed with the solution. Furthermore, in the coursework, a breast cancer dataset is collected and then preprocessed and transformed to make predictions and analyses with the help of big data technology and analytics methodologies. Two machine learning algorithms are used and compared based on the evaluation metrics to identify which model performs better for the breast cancer prediction dataset.

# **Table of Contents**

<b>Abstract .....</b>	<b>2</b>
<b>Chapter 1: Introduction .....</b>	<b>5</b>
1.1. Aim .....	5
1.2. Objective.....	5
1.3. Big Data .....	5
1.4. Healthcare .....	6
1.5. Cancer.....	6
<b>Chapter 2: Case Study.....</b>	<b>7</b>
2.1. How big data impacts business operations .....	7
2.2. Big Data Use Cases .....	8
2.3. Issues with big data management and analytics .....	9
<b>Chapter 3: Big data Architecture .....</b>	<b>10</b>
3.1. Jupyter Notebook IDE .....	10
3.2. Python.....	10
3.2.1. Python Libraries .....	11
<b>Chapter 4: Data Collection .....</b>	<b>12</b>
4.1. Kaggle.....	12
4.2. The Dataset .....	12
4.2.1. The dataset attributes .....	13
<b>Chapter 5: Data Pre-processing .....</b>	<b>16</b>
5.1. Data Cleaning.....	16
5.2. Feature Selection .....	19
5.3. Balancing the dataset.....	23
5.4. Summary .....	25
<b>Chapter 6: Data Transformation .....</b>	<b>26</b>
6.1. Data Normalization .....	26
<b>Chapter 7: Data Visualization.....</b>	<b>27</b>
<b>Chapter 8: Data Analysis.....</b>	<b>32</b>
<b>Chapter 9: Machine Learning.....</b>	<b>33</b>
9.1. Data train and test split .....	33
9.2. Machine Learning Model .....	34
9.2.1. Decision Tree Classifier .....	34

9.2.2. Support Vector Machine (SVM) .....	34
9.3. Model Accuracy and Analysis .....	35
9.3.1. Decision Tree Classifier .....	35
9.3.2. Support Vector Machine.....	36
Chapter 10: Conclusion.....	39
Chapter 11: References .....	40

# **Chapter 1: Introduction**

The study is an individual assignment to examine a dataset using data mining techniques. For the coursework, customer personality analysis was utilized to examine and comprehend the relationship between the customer and the product sold by the business. The dataset analyses help supermarkets that employ customer personality analysis to understand their customers' requirements and personalities in order to stay up with demand.

## **1.1. Aim**

The aim of the coursework is to choose a real world information system and analyze the case by collecting data related to the information system and derive meaningful insights and knowledge to improve the business and develop a detailed report that supports all the steps that have been conducted.

Based on the knowledge gained from the module, students are expected to choose any real world information system for this case study. As a Big data professional, you are expected to analyse the case by collecting data related to the information system and analyse it to derive meaningful insights and knowledge to improve the business and develop a detailed report aligned with figures and snapshots that support all the steps that have been conducted.

## **1.2. Objective**

- Go Through real world information system.
- Select a suitable dataset.
- Develop a Big Data Application
  - Preprocess the data
  - Transform the data
  - Visualize the data
  - Give an analysis of the data
- Build two machine learning models
- Generate a Report based on the findings and the steps

## **1.3. Big Data**

**Big Data** is a massive collection of data that is rapidly expanding. It is a collection of data that is so massive and complicated that typical data management methods are incapable of effectively storing or analyzing it. Big data is a word that refers to data that is extremely large in size. The five pillars of big data are volume, velocity, variety, veracity, and value (Taylor,2002) Big data today is used in almost every industry or sector. But big data in its raw form is of no great use to make insights. Thus, big data analytics come into play in order to identify valuable insights and make predictions

**Big Data analytics** is the process of extracting valuable insights from large amounts of data, such as hidden patterns, undiscovered relationships, market trends, and consumer preferences. Big Data analytics offers several benefits, including improved decision making and the prevention of fraudulent actions (Simplilearn, 2020).

In the study, a breast cancer dataset from the healthcare sector is studied with the help of big data analytics using predictive analytics to determine if the patient's cancer tumor is malignant (cancerous) or benign (non-cancerous).

#### **1.4. Healthcare**

Healthcare is the prevention, treatment, and management of sickness, as well as the maintenance of mental and physical well-being, provided by the medical, nursing, and allied health professions (igi-global, n.d.). As big data is being used in almost all industry, it is in particular very important used in the healthcare sector to enhance and progress healthcare, hospitals, researchers, and pharmaceutical businesses are implementing big data solutions. With access to massive volumes of patient and population data, healthcare is improving treatments, doing more effective research on diseases such as cancer and Alzheimer's, producing new medications, and obtaining crucial insights into population health patterns (builtin.com, n.d.).

The healthcare area was chosen for the coursework because it is critical to improve and contribute to the well-being of human health, especially at a time when managing healthcare is critical.

#### **1.5. Cancer**

Cancer is a condition in which some cells in the body develop uncontrolled and spread to other regions of the body. Cancer may begin practically any place in the human body, which contains billions of cells. Human cells normally develop and multiply (a process known as cell division) to generate new cells when the body requires them. Cells die as they get old or injured, and new cells replace them (National Cancer Institute, 2021).

Cancerous tumors infiltrate neighboring tissues and can move to distant locations in the body to produce new tumors (a process called metastasis). Cancerous tumors are also known as malignant tumors. Benign tumors do not penetrate or spread into neighboring tissues. (National Cancer Institute, 2021).

**Breast cancer** is the most common kind of cancer, claiming the lives of thousands of people each year. Breast cancer is a cancer that begins in the breast. It might begin in either one or both breasts. Cancer develops when cells begin to proliferate uncontrollably (American Cancer Society, 2017).

Thus, the coursework determines whether the patient or tumor is malignant (cancerous) or benign (non-cancerous) based on the various attributes included in the dataset using big data analytics tools and techniques.

## **Chapter 2: Case Study**

### **2.1. How big data impacts business operations**

Big Data has been of significant importance for many years, but its scarcity has made it impossible to realize its full potential. However, with the advancement of technology in recent years, the relevance of technology has expanded and is now employed in practically all industries. The most recent technology has enabled the speed and efficiency required to analyze vast volumes of data that are growing by the day. Through the analysis of these massive data and technologies, one can gain surprising insights that can not only help organizations or sectors make better decisions but also assist and predict future forecasts, leading to innovations and well-informed decision making that further enhances the performance and growth of the organization (Ku, 2017).

Using and comprehending big data is a critical competitive edge for top organizations. To the degree that businesses can collect additional data from current infrastructure and clients, they will be able to unearth hidden insights that their competitors do not have access to, giving them the upper hand in the ever competing market place (Ku, 2017).

Businesses may aspire for higher quality with big data since its insights can help them target the same market they want and enhance choices and costs using analytical insights. In addition, big data gives business operations a more personalized and focused approach, which may save a company time and money while also increasing the efficiency of its business processes. Finally, big data may help you attain more excellent quality while simultaneously lowering expenses (InData Labs, 2019).

One of the benefits of big data is discovering patterns within an industry or a process using its numerous tools, such as visualization. These visualizations aid in transforming complex, structured, and unstructured data in enormous volumes into readily interpretable visuals. These advantages aid in the detection of hidden patterns, flaws, and possible opportunities, allowing companies to make more educated and weighted decisions (Sydorenko, 2021).

## **2.2. Big Data Use Cases**

Big data can be seen in almost all industries, healthcare, finance, or even governance. As a result, big data has a significant role to play. Thus has many use cases. This section discusses big data use cases in the healthcare industry.

The use of big data analytics in healthcare has several good and perhaps life-saving consequences. Big-style data, in essence, refers to the massive amounts of information generated by the digitalization that collects patient records and aid in the management of hospital performance, which would otherwise be too large and complicated for traditional technologies. This data is then aggregated and evaluated by a particular technology. When used in healthcare, it will utilize unique health data from a community (or an individual) to possibly help avoid epidemics, treat sickness, and reduce expenditures, and so on (Durcevic, 2020).

One such life-changing use of big data is in cancer prevention and cure. Medical researchers can utilize enormous data on cancer patient's treatment plans and recovery rates to identify trends and treatments with the best success rates in the real world. Researchers, for example, can evaluate tumor samples in biobanks connected to patient treatment data. Researchers may use this data to analyze how particular mutations and cancer proteins interact with different therapies and identify trends that will lead to better patient outcomes. This information can also lead to unexpected advantages, such as discovering that Desipramine, an antidepressant, can help treat some kinds of lung cancer (Durcevic, 2020).

Hospitals can make more informed strategic decisions with the health data available and get better insights into a patient's motives. Furthermore, through extensive data analytics, patient engagement and experience can be further enhanced as patients become more actively involved in the process and promote more data related to their health with the help of intelligent devices (Durcevic, 2020).



### **2.3. Issues with big data management and analytics**

There are various issues that one can face in big data management and analytics.

One of the enormous challenges or issues with big data management is the complexity of managing data quality. These issues are faced due to the data being from diverse sources. When integrating the data from different sources into one for analysis, it becomes problematic, especially when other data sources have different data formats. This makes integration of data difficult, making analyses with the big data complicated and also faulty—another issue when managing complex data is unreliable sources of data. Inconsistent data are not 100% correct or have duplicate or wrong data that can lead to alarming trends and analysis, leading to faulty decisions and predictions and taking up more IT infrastructure space leading to increased costs (Bekker, 2018).

Quality data is critical to make valuable insights. There are a plethora of ways available for data cleaning. But first and foremost. A good model is required for big data. It's essential to specify the data you need to collect. The data should be cleaned and prepared regularly, and as it comes in from various sources, organize and normalize it before sending it to any tool for analysis. The data can be separated for better analysis once uniform and cleaned (solvexia, 2019).

The most common trait of large data is its explosive growth. And one of the most critical difficulties of big data is just this. The design of your solution may be thought out and altered for upscaling with no additional work. But the fundamental issue isn't adding extra processing and storage capacity. The complexity of scaling up so that your system's performance does not degrade while remaining within budget (Bekker, 2018).

The first and most important precaution for such difficulties is a solid architecture for your big data solution. As long as your big data solution can brag about such a thing, there will be fewer complications later. Another critical step is to build your big data algorithms with future upscalability in mind. Aside from that, you must prepare for your system's maintenance and support so that any changes caused by data expansion are adequately addressed. Furthermore, conducting systematic performance audits can assist in identifying and addressing weak points promptly (Bekker, 2018).

## **Chapter 3: Big data Architecture**

Big data architecture refers to the conceptual and physical structure that defines how huge volumes of data are consumed, processed, stored, managed, and accessed (omnisci, n.d.).

The dataset was downloaded from the kaggle website and saved in the system, where it is being analyzed for prediction using machine learning. The dataset is then evaluated and preprocessed to remove errors from the data, making it easier to understand and providing improved accuracy when utilized with machine learning techniques. The data is then converted and used with the machine learning algorithm to determine if the patient's tumor is benign or malignant.

Jupyter integrated development environment (IDE) was utilized for the coursework to store, access, analyze, and visualize data. Jupyter Notebook was also utilized for machine learning algorithms and assessment measures. Python is a programming language for big data analytics and machine learning. To achieve the aim, many Python libraries were employed.

The following are the big data tools used to analyze the breast cancer dataset and make predictions based on the inputs:

### **3.1. Jupyter Notebook IDE**

Jupyter Notebook is an open-source internet program that provides an interactive computing environment. It creates papers (notebooks) by integrating inputs and outputs into a single file. It gives a single document including graphics, mathematical calculations, statistical modeling, and so on (Wickramasinghe, 2021).

All data intake, preprocessing, visualization, transformation, machine learning, and analysis were performed in the Jupyter Notebook IDE using Python 3.0 and associated modules. The data was saved in the system, and the dataset in.csv format was read, analyzed, and further processed using the Jupyter notebook.

### **3.2. Python**

Python is a free and open-source programming language with a very low learning curve. Python is a wonderful tool for businesses that want the software they use to be tailored to their specific needs, thanks to its ability as a general-purpose language and its large library of packages that aid in the development of a system for producing data models from scratch (Mayuresh, 2020).

Python libraries of many types aided in big data analytics and machine learning.

### 3.2.1. Python Libraries

- Pandas

Pandas is an analysis software package used for data visualization and created for the Python computer language. It helps manipulate numerical tables and time series in particular (javatpoint, n.d.).

Panda's library was used for data ingestion, reading, visualizing the dataset from its source. The library was also used to pre-process and analyze the dataset for any further changes.

- Numpy

NumPy is the foundational Python library for scientific computing. A Python library provides a multidimensional array of objects, derived objects, and various routines for fast array operations (NumPy, n.d.)

The NumPy library was used for further pre-processing the data, such as balancing the dataset for enhancing unbiased prediction

- Matplotlib.pyplot

Matplotlib is a data visualization and graphical plotting package for Python and its numerical extension NumPy cross-platform. As such, it provides an open-source alternative to MATLAB. Developers may also incorporate plots in GUI programs by using matplotlib's APIs (Application Programming Interfaces) (ActiveState, 2021).

Matplotlib. pyplot library was used for visualizing the dataset in graphs and further analyzing them.

- Seaborn

Seaborn is a free and open-source Python module based on matplotlib. It is used for exploratory data analysis and data visualization. Seaborn is simple to use with data frames and the Pandas library. The generated graphs can also be readily altered. (Katari, 2020).

Seaborn library was used to visualize the dataset in graphs and further analyze them.

- Sklearn

Sklearn is Python's most valuable and robust machine learning package. It offers a set of fast tools for machine learning and statistical modelings, such as classification, regression, clustering, and dimensionality reduction, via a Python interface (tutorialspoint, n.d.).

Sklearn library was used to split the data into train and split datasets and used for machine learning algorithms imported from the sklearn library separately.

## Chapter 4: Data Collection

The dataset for the coursework was collected from the kaggle

### 4.1. Kaggle

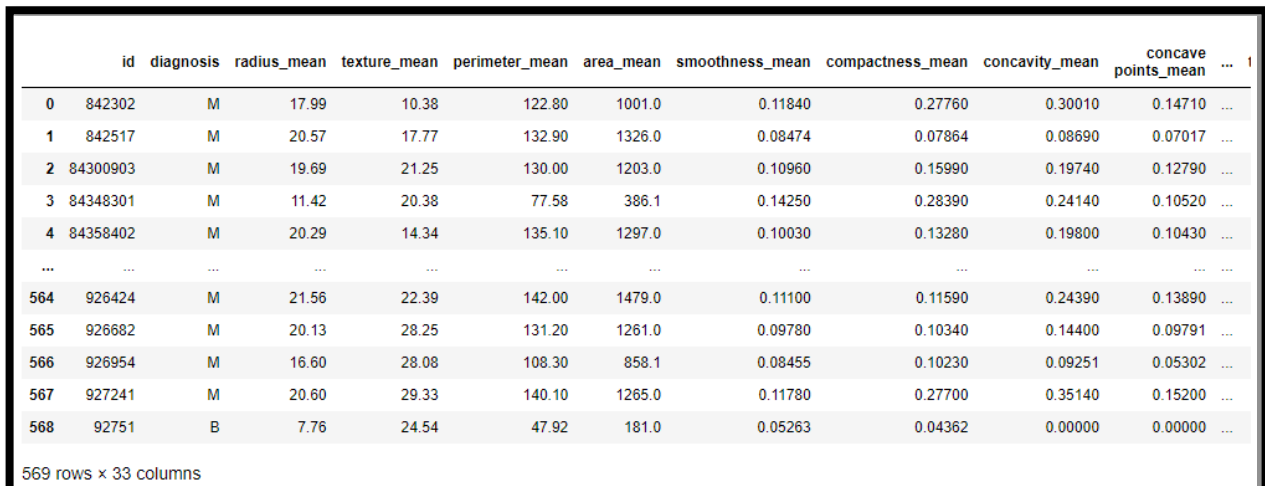
Kaggle is a subsidiary of Google LLC. It is an online community for data scientists and machine learning practitioners and a website that hosts machine learning competitions. Users use Kaggle to search and post data sets, study and construct models in a web-based data-science environment, collaborate with other data scientists and machine learning experts, and compete to solve data science challenges (Brownlee, 2017).

Some of the advantages of using Kaggle is that the data available are well defined and transparent and it is harder to deceive oneself with a poor test setting (Brownlee, 2017).

For the coursework, the breast cancer diagnostic dataset was collected from the kaggle site.

### 4.2. The Dataset

The Breast cancer diagnostic dataset UCI machine learning, under the license number of '[CC BY-NC-SA 4.0](#)'. The dataset is used to predict if the cancer is benign or malignant. The dataset features are computed from a digitized image of a Fine Needle Aspirate (FNA) of a breast mass and describe the characteristics of the cell nuclei present in the image (Dua et al., 2019).



	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	1
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 x 33 columns

Figure 0.1 The dataset snippet (self, 2021)

The dataset is free and accessible by all through the kaggle website or through UCI machine learning repository which is available in the UW CS ftp server. The dataset for the coursework is downloaded from the kaggle site in a .csv format and saved in the system and file where the analysis and machine learning algorithms is done (Dua et al., 2019)

The dataset downloaded has 569 rows and 33 columns or attributes

```
data.shape  
(569, 33)
```

Figure 0.2 Dataset shape (nos data points ,nos attributes)(self, 2021)

#### 4.2.1. The dataset attributes

There are a total of 33 columns or attributes present in the dataset .

```
data.columns  
  
Index(['id', 'diagnosis', 'radius_mean', 'texture_mean', 'perimeter_mean',  
      'area_mean', 'smoothness_mean', 'compactness_mean', 'concavity_mean',  
      'concave points_mean', 'symmetry_mean', 'fractal_dimension_mean',  
      'radius_se', 'texture_se', 'perimeter_se', 'area_se', 'smoothness_se',  
      'compactness_se', 'concavity_se', 'concave points_se', 'symmetry_se',  
      'fractal_dimension_se', 'radius_worst', 'texture_worst',  
      'perimeter_worst', 'area_worst', 'smoothness_worst',  
      'compactness_worst', 'concavity_worst', 'concave points_worst',  
      'symmetry_worst', 'fractal_dimension_worst', 'Unnamed: 32'],  
      dtype='object')
```

Figure 0.3 The dataset total columns( self,2021)

The following are the attributes present in the dataset (Dua et al., 2019)

1. ID number: The ID number is used for the unique identification of the data point and its data type is number or ID
2. Diagnosis (M = malignant, B = benign): it shows the diagnosis of the breast tissues, whether they are malignant or benign. This is the only string data type in the dataset

The dataset has been grouped by the ten real-valued features which are radius, texture, perimeter, area, smoothness, compactness, and concavity, concave points, symmetry, fractal dimension .These are computed for each cell nucleus. Furthermore, for each image, the mean, standard error, and "worst" or largest (mean of the three greatest values) of these characteristics were computed, yielding 30 features. All the following attributes are in a decimal data type. The following are the attributes

3. Radius mean: it is the mean of distances from centre to points on the perimeter
4. Texture mean: it is the standard deviation of gray-scale values

5. Perimeter mean
6. Area mean
7. Smoothness mean: it is the local variation in radius lengths
8. compactness mean: it is the formula  $(\text{perimeter}^2 / \text{area} - 1.0)$
9. concavity mean: it is severity of concave portions of the contour)
10. concave points mean (number of concave portions of the contour)
11. symmetry mean
12. fractal dimension mean: it is the formula ("coastline approximation" - 1)
13. Radius\_se : it is the standard error for the mean of distances from center to points on the perimeter
14. Texture\_se : it is the standard error for standard deviation of gray-scale values.
15. Perimeter\_se
16. Area\_se
17. Smoothness\_se: it is the standard error for local variation in radius lengths.
18. Compactness\_se: it is the standard error for  $\text{perimeter}^2 / \text{area} - 1.0$
19. Concavity\_se: it is the standard error for severity of concave portions of the contour
20. Concave points\_se: it is the standard error for number of concave portions of the contour
21. Symmetry\_se
22. Fractal\_dimension: it is the standard error for "coastline approximation" – 1
23. Radius\_worst: the "worst" or largest mean value for mean of distances from center to points on the perimeter.
24. Texture\_worst: the "worst" or largest mean value for standard deviation of gray-scale values.
25. Perimeter\_worst
26. Area\_worst
27. Smoothness\_worst: the "worst" or largest mean value for local variation in radius lengths.
28. Compactness\_worst: the "worst" or largest mean value for  $\text{perimeter}^2 / \text{area} - 1.0$
29. Concavity\_worst: the "worst" or largest mean value for severity of concave portions of the contour.

30. Concave points\_worst: the "worst" or largest mean value for number of concave portions of the contour
31. Symmetry\_worst
32. Fractal\_dimension\_worst: the "worst" or largest mean value for "coastline approximation" – 1
33. Unnamed:32

data.dtypes	
id	int64
diagnosis	object
radius_mean	float64
texture_mean	float64
perimeter_mean	float64
area_mean	float64
smoothness_mean	float64
compactness_mean	float64
concavity_mean	float64
concave points_mean	float64
symmetry_mean	float64
fractal_dimension_mean	float64
radius_se	float64
texture_se	float64
perimeter_se	float64
area_se	float64
smoothness_se	float64
compactness_se	float64
concavity_se	float64
concave points_se	float64
symmetry_se	float64
fractal_dimension_se	float64
radius_worst	float64
texture_worst	float64
perimeter_worst	float64
area_worst	float64
smoothness_worst	float64
compactness_worst	float64
concavity_worst	float64
concave points_worst	float64
symmetry_worst	float64
fractal_dimension_worst	float64
Unnamed: 32	float64
dtype:	object

Figure 0.4 Dataset datatype (self,2021)

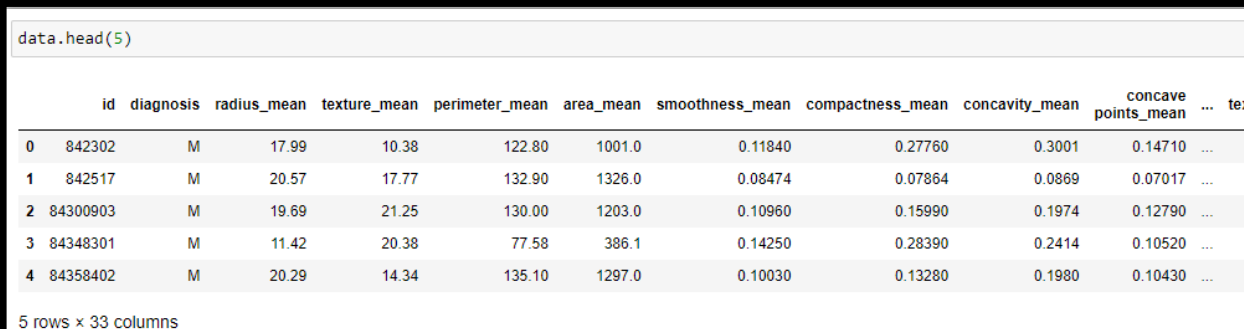
## Chapter 5: Data Pre-processing

Data preprocessing is converting raw data into well-formed data sets in order to use data mining methods. Raw data is frequently incomplete and formatted inconsistently. Data preprocessing is crucial in Machine Learning procedures to ensure that big datasets are prepared in such a manner that the data they contain can be processed and analyzed by learning algorithms. Data preprocessing is a very crucial part that can determine the success or failure of data analysis (Techopedia.com, 2019).

### 5.1. Data Cleaning

The act of preparing data for analysis by deleting or changing data that is incorrect, incomplete, irrelevant, redundant, or badly structured is known as data cleaning. Data cleaning is helpful as it can analyze the data that can provide any inaccurate results (Sisense, 2019).

- The first step is to analyze the dataset. Here the first 5 and last 5 dataset are analyzed as shown in figure 3.1.1. and 3.1.2.. Figure 3.2.3. shows the description of the dataset for analysis.

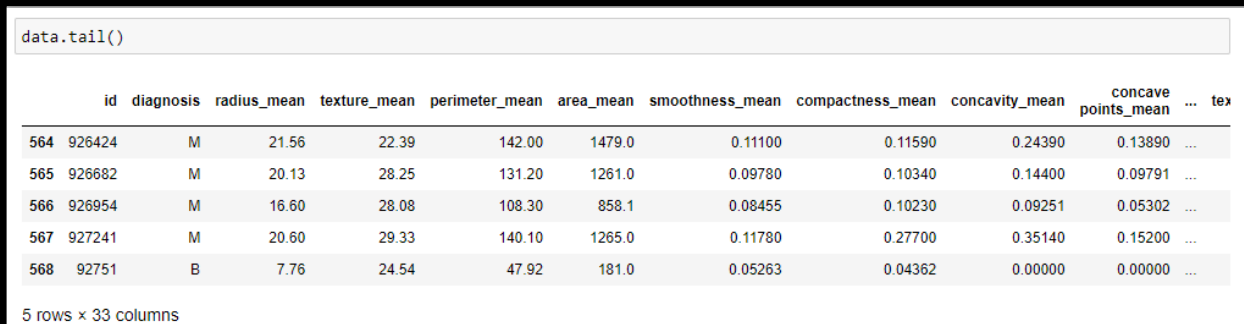


```
data.head(5)
```

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	tex
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...	...
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...	...
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...	...
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...	...
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...	...

5 rows x 33 columns

Figure 0.1 Analyzing the first 5 data points (self,2022)



```
data.tail()
```

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	tex
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	...	...

5 rows x 33 columns

Figure 0.2 Analyzing the last 5 datapoints (self,2022)



```
data.describe()
```

	id	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean
count	5.690000e+02	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000
mean	3.037183e+07	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341	0.088799	0.048919	0.038803
std	1.250206e+08	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813	0.079720	0.038803	0.038803
min	8.670000e+03	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.000000	0.000000	0.000000
25%	8.692180e+05	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920	0.029560	0.020310	0.020310
50%	9.060240e+05	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630	0.061540	0.033500	0.033500
75%	8.813129e+06	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.130700	0.074000	0.074000
max	9.113205e+08	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.426800	0.201200	0.201200

8 rows x 32 columns

Figure 0.3 Dataset Description(self,2022)

- To check for any duplicate values and analyze the unique values.

As seen below there are many unique values in the dataset and the diagnosis attribute has only two values which can be used as targets for the machine learning algorithms.

Moreover there are no duplicate values present in the dataset .

```
data.nunique()
```

id	569
diagnosis	2
radius_mean	456
texture_mean	479
perimeter_mean	522
area_mean	539
smoothness_mean	474
compactness_mean	537
concavity_mean	537
concave points_mean	542
symmetry_mean	432
fractal_dimension_mean	499
radius_se	540
texture_se	519
perimeter_se	533
area_se	528
smoothness_se	547
compactness_se	541
concavity_se	533
concave points_se	507
symmetry_se	498
fractal_dimension_se	545
radius_worst	457
texture_worst	511
perimeter_worst	514
area_worst	544
smoothness_worst	411
compactness_worst	529
concavity_worst	539
concave points_worst	492
symmetry_worst	500
fractal_dimension_worst	535
Unnamed: 32	0

dtype: int64

Figure 0.4 To check for unique values(self,2022)

```
data.duplicated().sum()
0
```

Figure 0.5 To check if there are any duplicate values(self,2022)

- To check any null values.

As seen in the figure below there are no null values present in the dataset , except in the Unnamed:32 columns which will be removed.

```
data.isnull().sum()
id                                0
diagnosis                        0
radius_mean                      0
texture_mean                     0
perimeter_mean                   0
area_mean                        0
smoothness_mean                  0
compactness_mean                 0
concavity_mean                   0
concave points_mean              0
symmetry_mean                    0
fractal_dimension_mean           0
radius_se                        0
texture_se                       0
perimeter_se                     0
area_se                          0
smoothness_se                    0
compactness_se                   0
concavity_se                     0
concave points_se                0
symmetry_se                      0
fractal_dimension_se             0
radius_worst                     0
texture_worst                    0
perimeter_worst                  0
area_worst                       0
smoothness_worst                 0
compactness_worst                0
concavity_worst                  0
concave points_worst             0
symmetry_worst                   0
fractal_dimension_worst          0
Unnamed: 32                      569
dtype: int64
```

Figure 0.6 To check for any null values(self,2022)

- To convert the diagnosis also the target of the dataset into numbers of '1' for malign tumor and '0' for benign tumor

```
data['diagnosis'] = data['diagnosis'].replace('M', 1)
data['diagnosis'] = data['diagnosis'].replace('B', 0)
```

Figure 0.7 Converting the diagnosis values to 0 and 1(self,2022)

- Removing the 'Unnamed:32' and 'id' column as they are not needed for machine learning.

```
data=data.drop(['Unnamed: 32', 'id'], axis=1)
```

Figure 0.8 Dropping columns(self,2022)

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean
0	1	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419
1	1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812
2	1	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069
3	1	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597
4	1	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809

5 rows × 31 columns

Figure 0.9 dataset after data cleaning(self,2022)

## 5.2. Feature Selection

Feature selection is the process of is the process of minimizing the number of input variables. Feature selection is preferable as is reduces modeling computational costs and increase the model performance (Brownlee, 2019).

The target of the dataset is the diagnosis attribute. The data will be preprocessed focusing on the target.

As there are a total of 31 attributes in the dataset, 30 of which are related to the diagnosis attribute which is the target . This is a large number of attributes to calculate. The relationship between diagnosis and other qualities is assessed using the correlation function, and those attributes that are less than the '0.5' threshold are removed, while those that are more are retained in the dataset for further preprocessing.

- Dataset correlation for Analysis

```
corr = data.corr()
data.corr()
```

	id	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean
id	1.000000	0.074626	0.099770	0.073159	0.098893	-0.012968	0.000098	0.050080	0.044
radius_mean	0.074626	1.000000	0.323782	0.997855	0.987357	0.170581	0.508124	0.676764	0.822
texture_mean	0.099770	0.323782	1.000000	0.329533	0.321088	-0.023389	0.236702	0.302418	0.293
perimeter_mean	0.073159	0.997855	0.329533	1.000000	0.988507	0.207278	0.556936	0.716136	0.850
area_mean	0.098893	0.987357	0.321088	0.988507	1.000000	0.177028	0.498502	0.685983	0.823
smoothness_mean	-0.012968	0.170581	-0.023389	0.207278	0.177028	1.000000	0.659123	0.521984	0.553
compactness_mean	0.000098	0.508124	0.236702	0.556936	0.498502	0.659123	1.000000	0.883121	0.831
concavity_mean	0.050080	0.676764	0.302418	0.716136	0.685983	0.521984	0.883121	1.000000	0.921
concave points_mean	0.044158	0.822529	0.293464	0.850977	0.823299	0.553995	0.831135	0.921391	1.000
symmetry_mean	-0.022114	0.147741	0.071401	0.183027	0.151293	0.557775	0.602641	0.500667	0.462
fractal_dimension_mean	-0.052511	-0.311631	-0.076437	-0.281477	-0.283110	0.584792	0.585389	0.336783	0.186
radius_se	0.143048	0.679090	0.275869	0.691765	0.732582	0.301467	0.497473	0.631925	0.698
texture_se	-0.007526	-0.097317	0.386358	-0.086761	-0.066280	0.068406	0.046205	0.076218	0.021
perimeter_se	0.137331	0.674172	0.281673	0.693135	0.726628	0.298092	0.548905	0.680391	0.710
area_se	0.177742	0.735864	0.259845	0.744983	0.800088	0.248552	0.455653	0.617427	0.690
smoothness_se	0.098781	-0.222600	0.006614	-0.202694	-0.166777	0.332375	0.135299	0.098564	0.027
compactness_se	0.033961	0.208000	0.191975	0.250744	0.212583	0.318943	0.738722	0.670279	0.490
concavity_se	0.055239	0.194204	0.143293	0.228082	0.207680	0.248396	0.570517	0.691270	0.439
concave points_se	0.078768	0.376169	0.163851	0.407217	0.372320	0.380676	0.642262	0.683260	0.615
symmetry_se	-0.017306	-0.104321	0.009127	-0.081629	-0.072497	0.200774	0.229977	0.178009	0.095
fractal_dimension_se	0.025725	-0.042641	0.054458	-0.005523	-0.019887	0.283807	0.507318	0.449301	0.257
radius_worst	0.082405	0.969539	0.352573	0.969476	0.962746	0.213120	0.535315	0.688236	0.830
texture_worst	0.064720	0.297008	0.912045	0.303038	0.287489	0.036072	0.248133	0.299879	0.292
perimeter_worst	0.079986	0.965137	0.358040	0.970387	0.959120	0.238853	0.590210	0.729565	0.855
area_worst	0.107187	0.941082	0.343546	0.941550	0.959213	0.206718	0.509604	0.675987	0.809
smoothness_worst	0.010338	0.119616	0.077503	0.150549	0.123523	0.805324	0.585541	0.448822	0.452
compactness_worst	-0.002968	0.413463	0.277830	0.455774	0.390410	0.472468	0.895809	0.754968	0.657
concavity_worst	0.023203	0.526911	0.301025	0.583879	0.512606	0.434926	0.816275	0.884103	0.752
concave points_worst	0.035174	0.744214	0.295316	0.771241	0.722017	0.503053	0.815573	0.881323	0.910
symmetry_worst	-0.044224	0.183953	0.105008	0.189115	0.143570	0.394309	0.510223	0.409464	0.375
fractal_dimension_worst	-0.029866	0.007066	0.119205	0.051019	0.003738	0.499316	0.687382	0.514930	0.368
Unnamed: 32	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

32 rows x 32 columns

Figure 0.10 Dataset Correlation(self,2022)

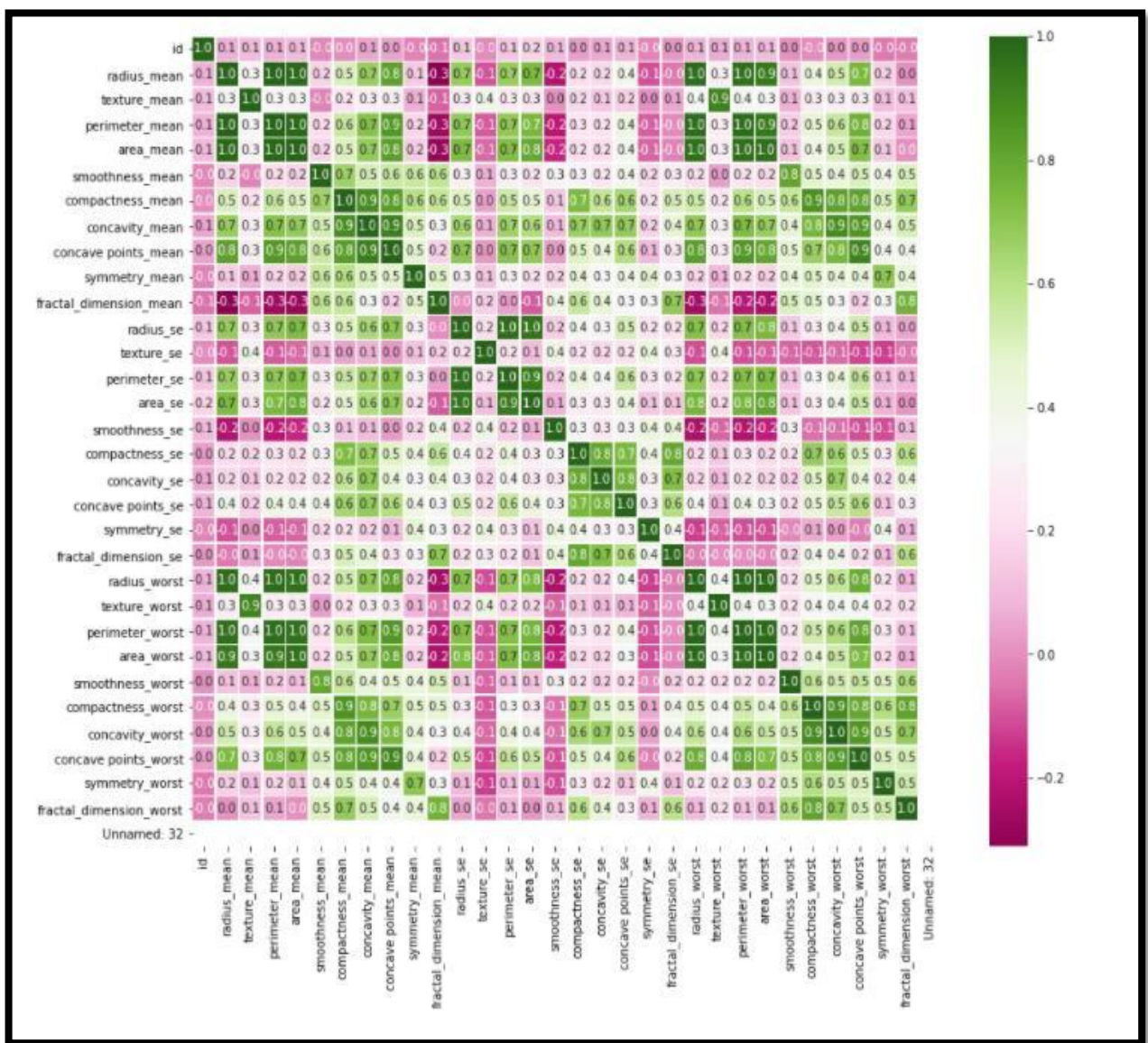


Figure 0.11 Correlation heatmap diagram(self,2022)

- To check the number of attributes and its names that are more than the 0.5 threshold

```
corr = data.corr()
cc=corr[abs(corr['diagnosis']) > 0.50].index
print('- Number of most correlated features = ', len(cc))
print('-----')
print('- Most correlated features is: \n ',cc)

- Number of most correlated features = 14
-----
- Most correlated features is:
Index(['diagnosis', 'radius_mean', 'perimeter_mean', 'area_mean',
      'compactness_mean', 'concavity_mean', 'concave points_mean',
      'radius_worst', 'perimeter_worst', 'area_worst', 'compactness_worst',
      'concavity_worst', 'concave points_worst', 'area_se'],
      dtype='object')
```

Figure 0.12 More than 0.5 threshold(self,2022)

- To check the number of attributes and its names that are less than the 0.5 threshold

```
cc2=corr[abs(corr['diagnosis']) <= 0.50].index
print('- Number of Least correlated features = ', len(cc2))
print('-----')
print('- Least correlated features is: \n ',cc2)

- Number of Least correlated features = 15
-----
- Least correlated features is:
Index(['texture_mean', 'smoothness_mean', 'symmetry_mean',
      'fractal_dimension_mean', 'texture_se', 'smoothness_se',
      'compactness_se', 'concavity_se', 'concave points_se', 'symmetry_se',
      'fractal_dimension_se', 'texture_worst', 'smoothness_worst',
      'symmetry_worst', 'fractal_dimension_worst'],
      dtype='object')
```

Figure 0.13 Less than 0.5 threshold(self,2022)

- Final dataset after analyzing the attributes

```
data = data[['diagnosis', 'radius_mean', 'perimeter_mean', 'area_mean',
            'compactness_mean', 'concavity_mean', 'concave points_mean',
            'radius_worst', 'perimeter_worst', 'area_worst', 'compactness_worst',
            'concavity_worst', 'concave points_worst', 'area_se']]
data
```

Figure 0.14 Changing the dataset to the reduced dataset(self,2022)



	diagnosis	radius_mean	perimeter_mean	area_mean	compactness_mean	concavity_mean	concave points_mean	radius_worst	perimeter_worst	area_worst	compactness_worst
0	1	17.99	122.80	1001.0	0.27760	0.30010	0.14710	25.380	184.60	2019.0	0.26010
1	1	20.57	132.90	1326.0	0.07864	0.08690	0.07017	24.990	158.80	1956.0	0.11825
2	1	19.69	130.00	1203.0	0.15990	0.19740	0.12790	23.570	152.50	1709.0	0.18710
3	1	11.42	77.58	386.1	0.28390	0.24140	0.10520	14.910	98.87	567.7	0.25940
4	1	20.29	135.10	1297.0	0.13280	0.19800	0.10430	22.540	152.20	1575.0	0.15710
...	...	...	...	...	...	...	...	...	...	...	...
564	1	21.56	142.00	1479.0	0.11590	0.24390	0.13890	25.450	166.10	2027.0	0.15450
565	1	20.13	131.20	1261.0	0.10340	0.14400	0.09791	23.690	155.00	1731.0	0.12530
566	1	16.60	108.30	858.1	0.10230	0.09251	0.05302	18.980	126.70	1124.0	0.10420
567	1	20.60	140.10	1265.0	0.27700	0.35140	0.15200	25.740	184.60	1821.0	0.28390
568	0	7.76	47.92	181.0	0.04362	0.00000	0.00000	9.456	59.16	268.6	0.05283

569 rows × 14 columns

Figure 0.15 Dataset after attribute reduction(self,2022)

### 5.3. Balancing the dataset

It is very important to have a balanced dataset as it help the model get higher and accurate model that have a balanced detection rate and balanced accuracy. This is especially important when using the dataset for a classification model. As the dataset is used for classification thus this section balances the dataset by comparing and deleting any extra data from the dataset (Amruthnath, 2020).

- There are 357 benign data and 212 Malignant data present in the dataset

```
data['diagnosis'].value_counts()
B      357
M      212
Name: diagnosis, dtype: int64
```

Figure 0.16 Diagnosis unique value count(self,2022)

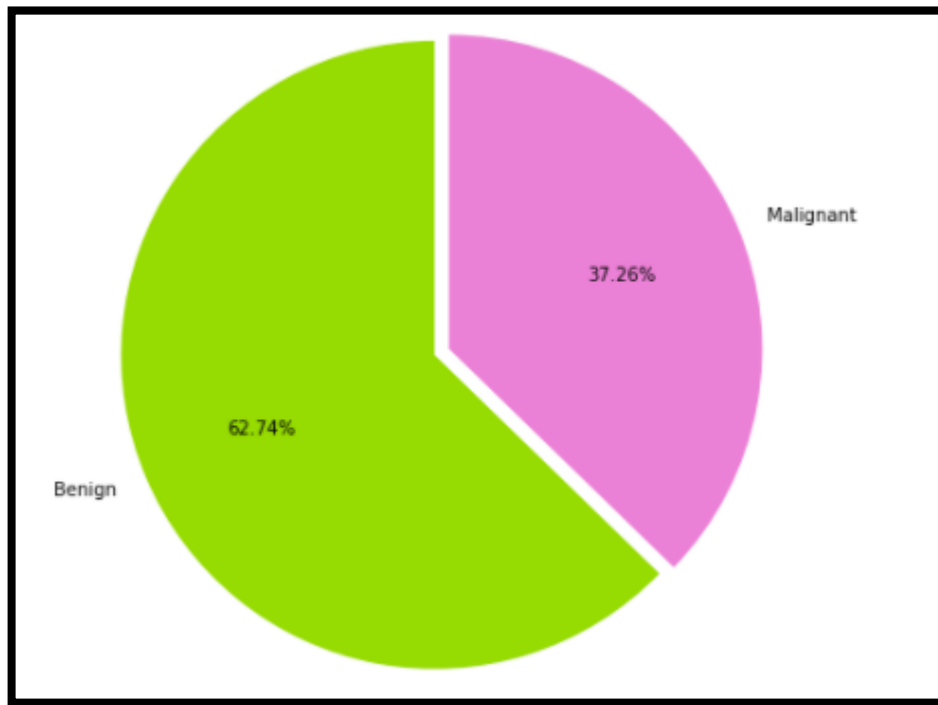


Figure 0.17 Diagnosis value graph(self,2022)

- The below code identifies any extra data

```
: #balancing the dataset
from sklearn import preprocessing
targets_all=data['diagnosis']
num_one_targets=int(np.sum(targets_all))
zero_targets_counter=0
indices_to_remove=[]
for i in range(targets_all.shape[0]):
    if targets_all[i]==0:
        zero_targets_counter+=1
        if zero_targets_counter>num_one_targets:
            indices_to_remove.append(i)

: unscaled_data=data.drop('diagnosis', axis=1)
```

Figure 0.18 Balancing the dataset(self,2022)

- The below is the number of data that were identified as extra

```
len(indices_to_remove)
145
```

Figure 0.19 Number of indices to remove from the dataset



- The below code is used to remove the extra data points and thus balanced

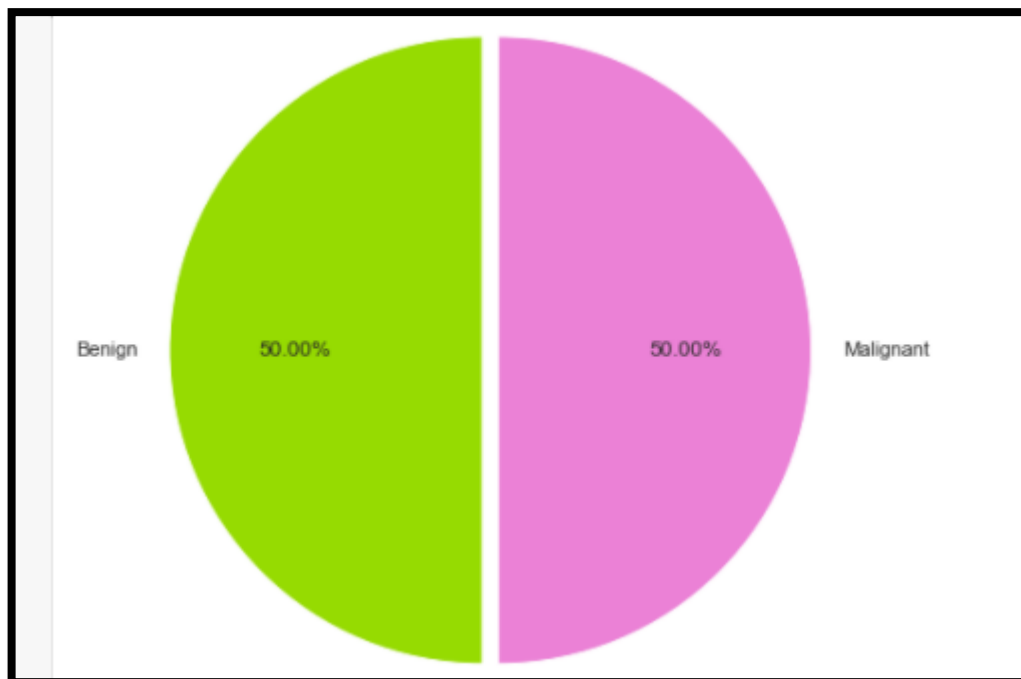
```
inputs = unscaled_data.drop(indices_to_remove,axis=0)
targets = targets_all.drop(indices_to_remove,axis=0)
```

*Figure 0.20 Removing the extra data from targets and data*

- The below proves that the data is balanced based on the diagnosis attribute

```
targets.value_counts()
1    212
0    212
Name: diagnosis, dtype: int64
```

*Figure 0.21 diagnosis value count*



*Figure 0.22 balanced diagnosis graph(self,2022)*

#### 5.4. Summary

Before preprocessing the data, there were 569 rows and 33 columns or attributes. After cleaning, feature selection and balancing the dataset, there are a total of 424 rows and 14 columns or attributes including the targets. While balancing the dataset the targets of the dataset was given which is the diagnosis attribute based on which the dataset was balanced and selected attributes.

## Chapter 6: Data Transformation

Data transformation refers to the process of changing data from one format to another. Data transformation is the method for dealing with the ever-increasing volume of data and effectively utilizing it for your organization. Because of the expansion in the number of data sources and devices, a massive amount of data is being created. Data transformation makes it simple for businesses to convert data (Sharma, 2020).

While preprocessing the dataset, it was divided into inputs and targets in order to classify the data for test and train in machine learning. As all the input and targets are numerical in value in the dataset. Thus, Normalization is done in order to transform the data.

### 6.1. Data Normalization

Data Normalization is one the most important strategies in data mining .Data normalization is a technique for converting source data into another format for efficient processing. The data is transformed here so that it can fall into a defined range .The basic goal of data normalization is to reduce or even eliminate redundant data. It has various advantages, such as making data mining algorithms more effective, allowing for speedier data extraction, and so on (Choudhury, 2021).

From the sklearn library the MinMaxScaler which is one of the most popular normalization method is imported and only the inputs (dataset without the targets or diagnosis columns) are normalized. The code is shown in figure 4.1. The transformation is shown in figure 4.2

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
data_scaled = scaler.fit_transform(inputs)
```

Figure 6.1 Normalization of inputs(self,2022)

```
data_scaled
array([[0.52103744, 0.54598853, 0.36373277, ..., 0.56861022, 0.91202749,
        0.27323299],
       [0.64314449, 0.61578329, 0.50159067, ..., 0.19297125, 0.63917526,
        0.12496355],
       [0.60149557, 0.59574321, 0.44941676, ..., 0.35974441, 0.83505155,
        0.16225522],
       ...,
       [0.62232003, 0.60403566, 0.47401909, ..., 0.25678914, 0.55945017,
        0.1716202 ],
       [0.45525108, 0.44578813, 0.30311771, ..., 0.27180511, 0.48728522,
        0.07724143],
       [0.64456434, 0.66553797, 0.4757158 , ..., 0.74976038, 0.91065292,
        0.14765633]])
```

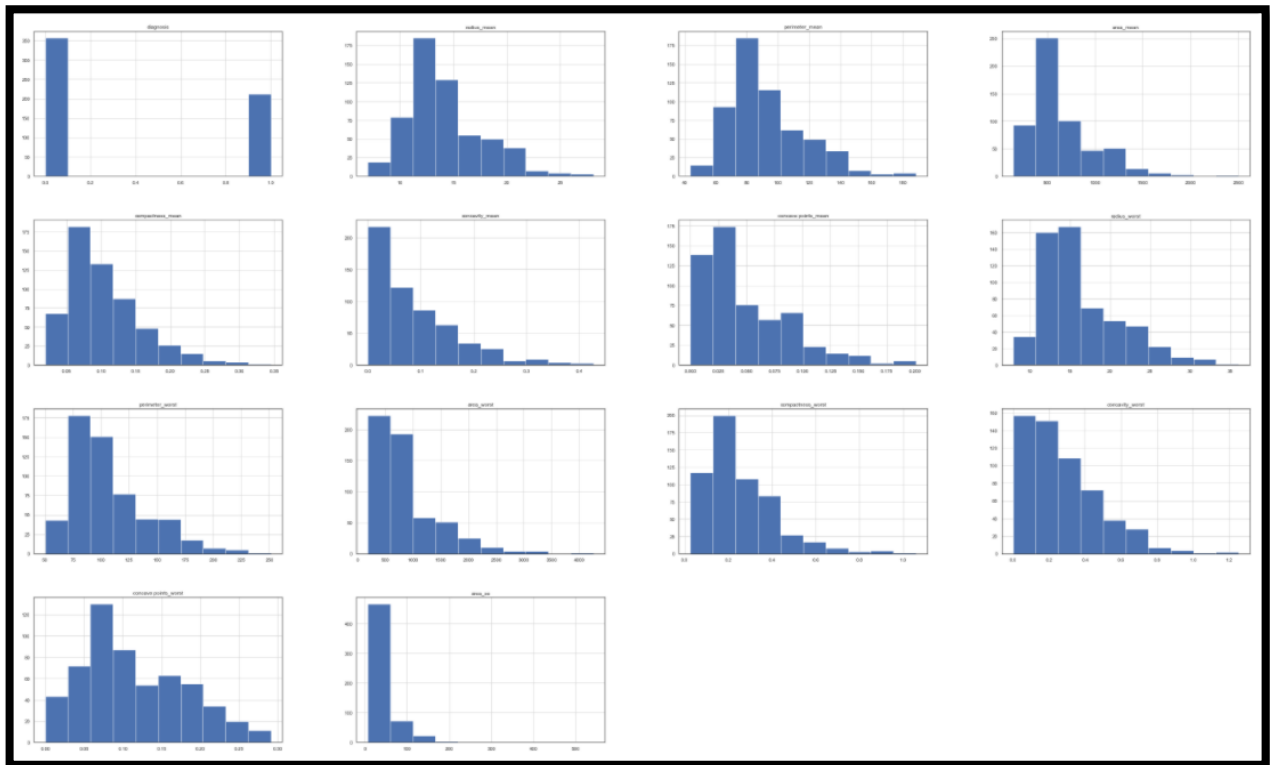
Figure 6.2 Normalized/ scaled dataset(self,2022)

## **Chapter 7: Data Visualization**

The graphical depiction of information and data is known as data visualization. Data visualization tools, which include visual components like as charts, graphs, and maps, give an easy method to observe and comprehend trends, outliers, and patterns in data (Tableau, 2018).

The following diagrams are the visualizations that are used for analysis and comparison of the data .

- The frequency distribution of the data before balancing the data



*Figure7.1 Frequency table of the columns of the dataset(self,2022)*

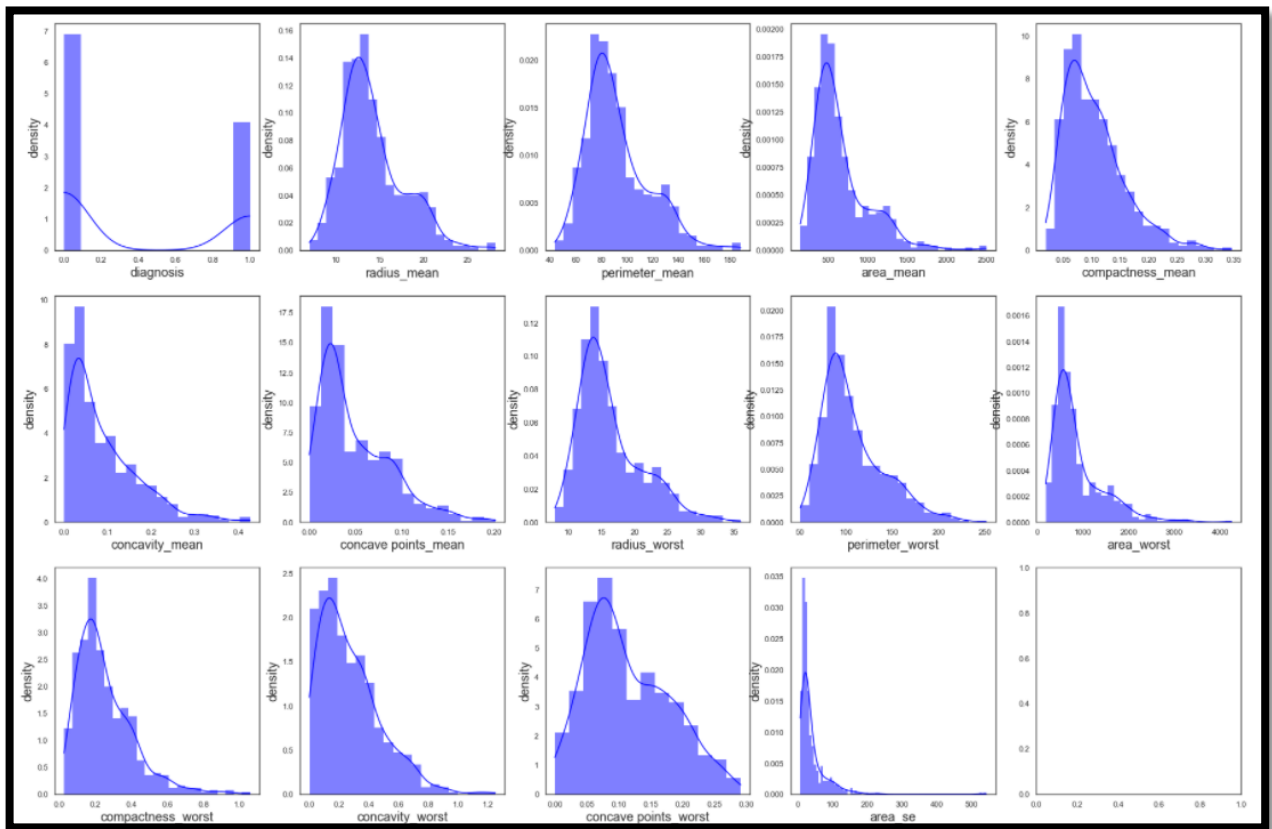


Figure7.2 Frequency distribution and graph design(self,2022)

- Frequency distribution of the diagnosis or targets after balancing and transformation.

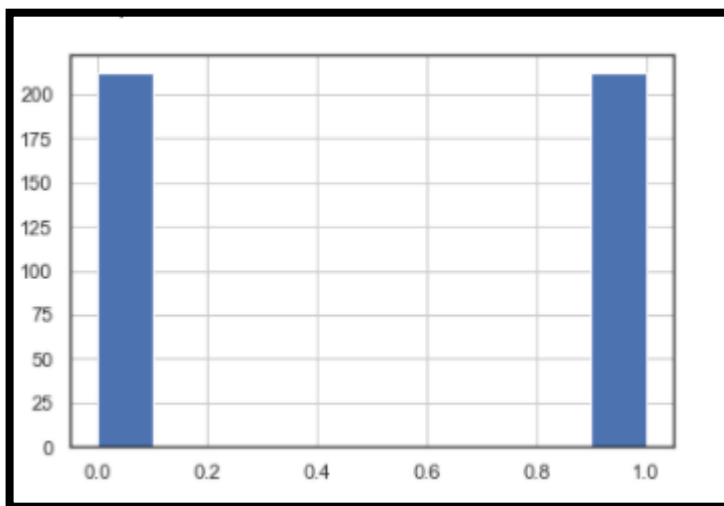


Figure7.3 frequency distribution of targets(self,2022)

- Frequency and graph distribution of the inputs or columns after preprocessing and transforming.

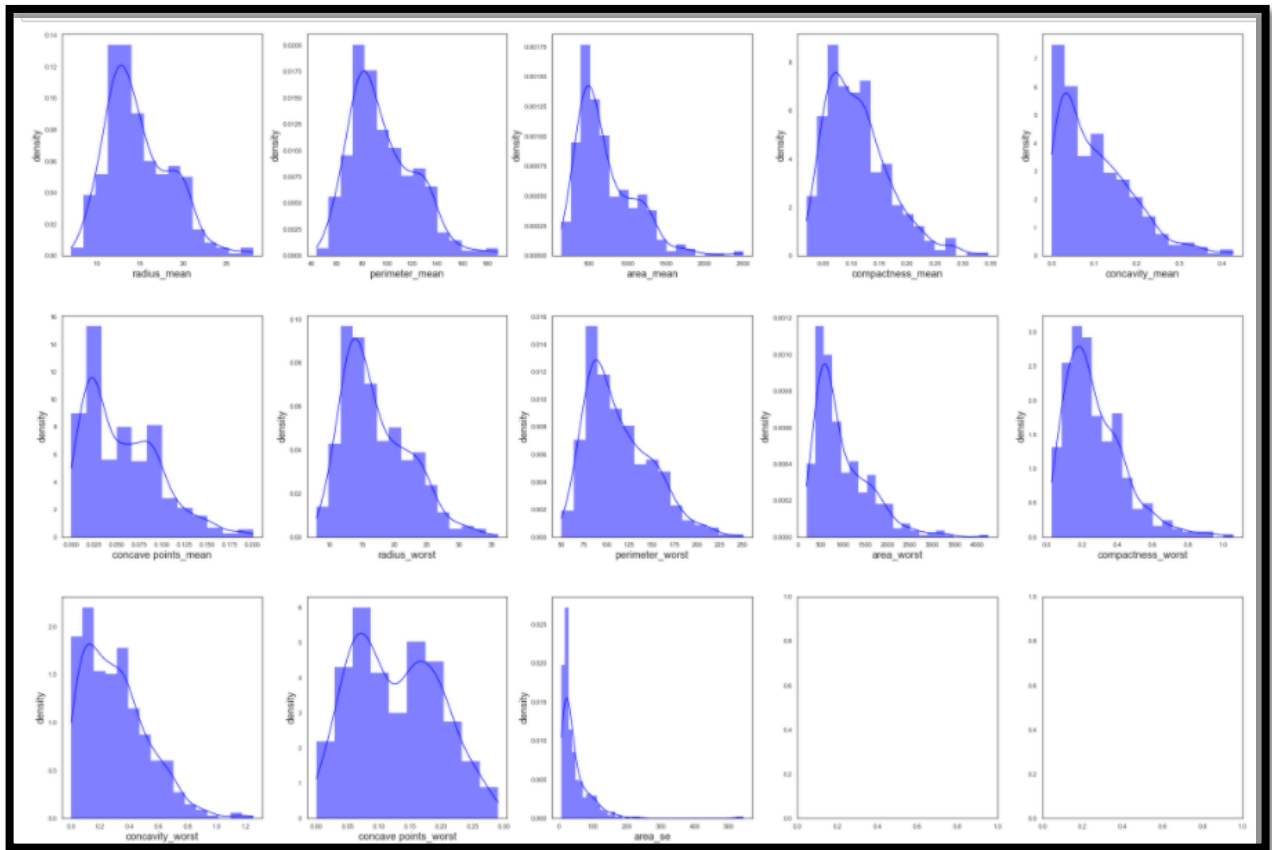


Figure7.4 Frequency distribution of inputs(self,2022)

- The correlation of the top 13 columns that have more than 0.5 correlation of significance with diagnosis or target column.

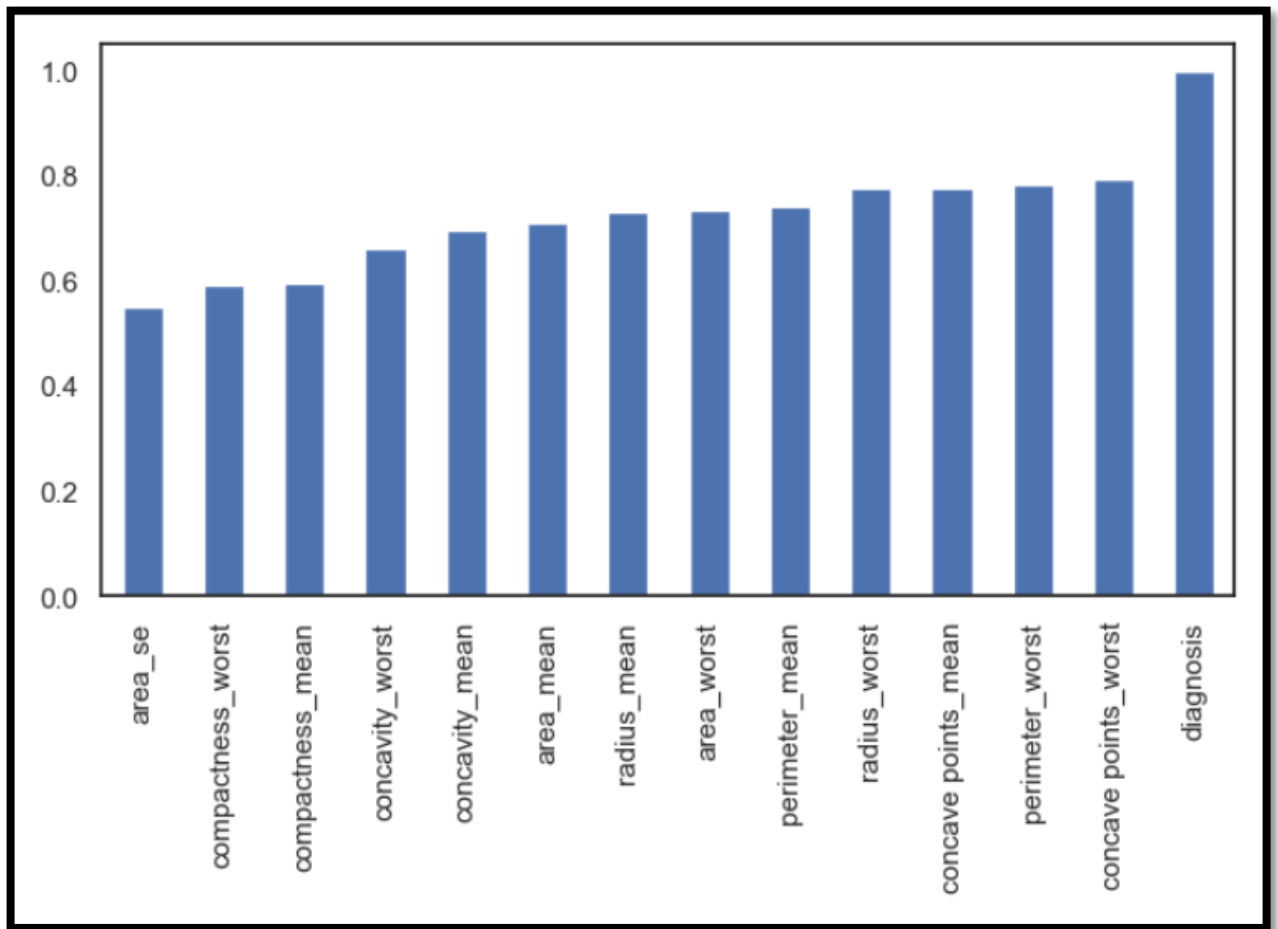


Figure7.5 Correlation graph to compare with diagnosis(self,2022)

- Visualization to understand the top four highly correlated columns to diagnosis and its relation to malign and benign tumor

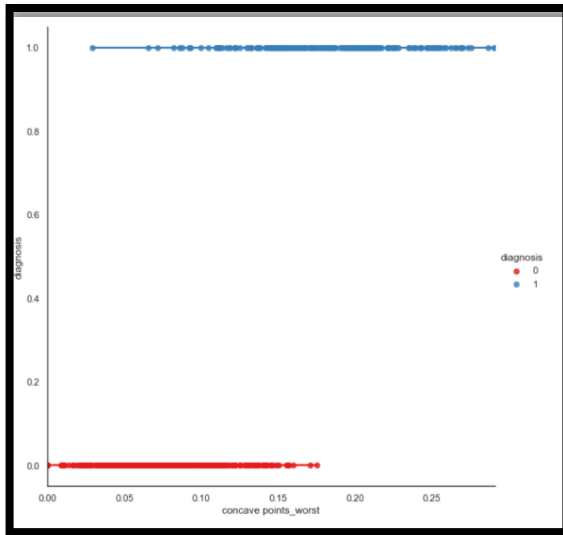


Figure7.7 diagnosis to concave points\_worst(self,2022)

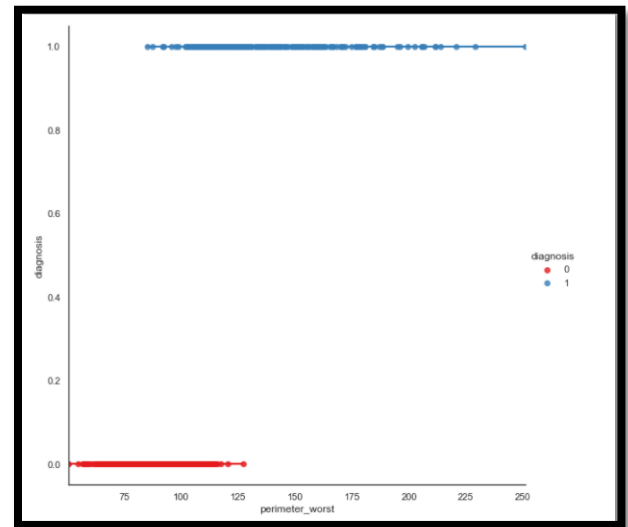


Figure 7.6 diagnosis to perimenter\_worst(self,2022)

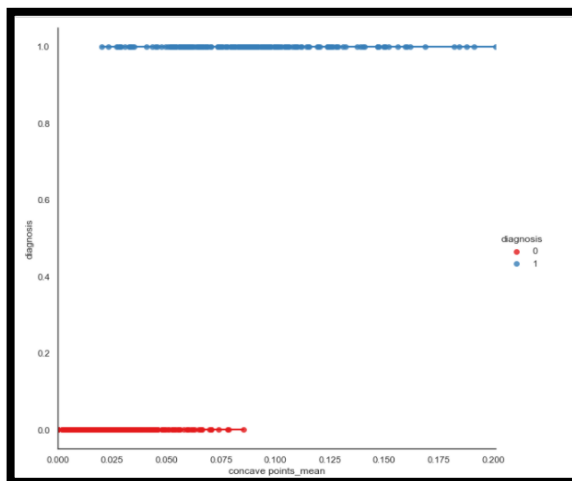


Figure 7.9 Diagnosis to concave\_point\_mean(self,2022)

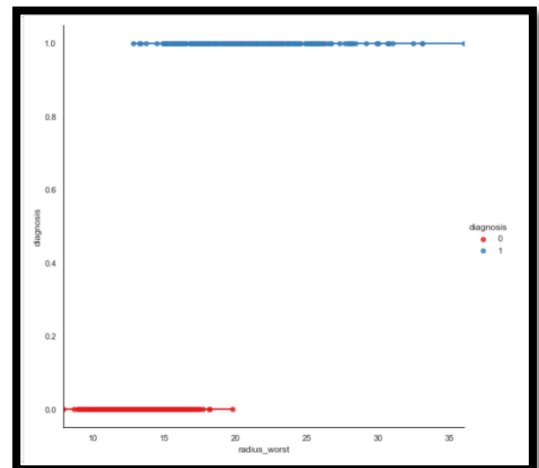


Figure7.8 diagnosis to radius\_worst(self,2022)

## **Chapter 8: Data Analysis**

The dataset is a raw dataset collected from kaggle that have the data of the measurements of the tumor that were derived from the images of breast cancer patients. The dataset before preprocessing had 569 rows and 33 attributes. The data was then preprocessed and transformed for analysis and machine learning. The data was processed using various visualization tools and correlation statistical analysis. Those attributes that had lower correlation of 0.5 with the diagnosis attribute were removed. Diagnosis is the target attribute while the 13 columns after preprocessing is the inputs attribute used to predict if the diagnosis of the patient is malign('1') or benign('0') . At this stage the dataset had only 424 rows and 14 attributes including the diagnosis attribute.

After preprocessing, transforming and visualizing the data, the following three insights were identified about the dataset:

- Before balancing the data, it was noticed that 357 cases were benign and 212 cases were malign. That is 62.74% of the recorded cases were benign and 37.25% of the recorded cases were malign. This may also indicate that from the recorded cases, the probability of the cases to be benign and non-cancerous is higher than the tumor being malign.
- The higher or bigger the mean and worst the higher are the chances for the cancer to be malignant or cancerous. In the graphs during analysis It was seen that concave points worst and mean and perimeter worst and radius worst all diagnosis were malignant as the number increased and were benign as he numbers decreased.
- From the correlation matrix it was noticed that all of the worst features or attributes had a significant correlation with the diagnosis attribute , while standard deviation or se of the attributes had the lowest significance with the diagnosis or target attribute . That is why most of the attributes after preprocessing are either worst or mean.



## **Chapter 9: Machine Learning**

Machine learning is a subset of artificial intelligence (AI) technology that allows computers to automatically learn and improve from experience without being explicitly designed. Machine learning is concerned with the creation of computer programs that can access data and utilize it to learn on their own. It predicts new values based on the past data as input. The basic goal is for computers to learn autonomously without human involvement or aid and then adapt their activities accordingly (Expert.ai Team, 2020).

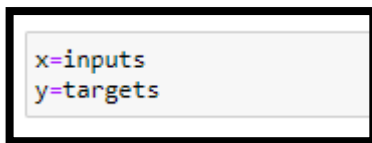
Machine learning are categorized into four- Supervised Machine learning, Unsupervised Machine learning , Semi-supervised machine learning and Reinforcement learning. For the coursework supervised machine learning is used to predict if the cancer is malign('1') or benign('0').Supervised machine learning can predict future occurrences that have been learned in the past using labeled data. After sufficient training, the system may offer objectives for any new input.

There are many types of supervised machine learning, such as Regression, classification etc. For the coursework two types of supervised machine algorithms are used to predict the outcome of the breast cancer tumor. The two algorithms are Decision Tree Classifier and Support Vector Machine.

### **9.1. Data train and test split**

The train-test split is a method of assessing the performance of a machine learning system. It can be used for any supervised learning technique and can be utilized for classification or regression tasks.The process entails splitting a dataset into two subgroups. The first subset, known as the training dataset, is utilized to fit the model. The second subset is not used to train the model; instead, the model is fed the dataset's input element, and predictions are generated and compared to expected values. The second dataset is known as the test dataset (Brownlee, 2020).

First the targets and inputs were initialized as x and y for splitting the data



```
x=inputs
y=targets
```

*Figure 9.1 Renaming the inputs and targets(self,2022)*

The below figure splits the dataset into train and test. 80% of the dataset is train dataset and 20% is test dataset.

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)
```

*Figure9.2 splitting the dataset into test and train(self,2022)*

## 9.2. Machine Learning Model

### 9.2.1. Decision Tree Classifier

Decision Trees (DTs) are a type of non-parametric supervised learning approach that is commonly used for regression and classification problems. The objective is to build a model that predicts the value of a target variable using basic decision rules derived from data attributes. A tree is an example of a piecewise constant approximation (scikit learn, 2009).

```
from sklearn.tree import DecisionTreeClassifier
tree = DecisionTreeClassifier()
tree.fit(x_train, y_train)

DecisionTreeClassifier()
```

*Figure9.3 Fitting the data into the ML algorithm(self,2022)*

### 9.2.2. Support Vector Machine (SVM)

The "Support Vector Machine" (SVM) is a supervised machine learning technique that can be used for classification and regression tasks. It is, however, largely employed in categorization difficulties. Each data item is plotted as a point in n-dimensional space (where n is the number of features you have), with the value of each feature being the value of a certain coordinate in the SVM algorithm. Then, we accomplish classification by locating the hyper plane that best distinguishes the two classes (look at the below snapshot) (Ray, 2019).

```
from sklearn.svm import SVC
svm = SVC(kernel='linear')
svm.fit(x_train,y_train)

SVC(kernel='linear')
```

*Figure 9.4 Fitting the data into the ML algorithm(self,2022)*

### 9.3. Model Accuracy and Analysis

#### 9.3.1. Decision Tree Classifier

The decision tree classifier has a training accuracy of 100 percent and a test accuracy of around 93 percent. The training score exceeds the test score. To corroborate the model's correctness, assessment measures such as the confusion matrix, f1 score, and ROC AUC curve were utilized to describe the model's performance. The evaluation measures are capable of distinguishing between model outcomes. According to the confusion matrix in figure 7.7, 36 are true negative, 43 are true positive, 4 are false positive, and 2 are false negative. The f1 score accuracy using the confusion matrix is 93 percent, which is the same as the model's testing accuracy. The ROC AUC curve is also 93% highlighting that the accuracy of the decision tree classifier is 93%.

```
tree.score(x_train,y_train)
1.0
```

Figure 9.5 DTC train accuracy(self,2022)

```
tree.score(x_test,y_test)
0.9294117647058824
```

Figure 9.6 DTC test accuracy(self,2022)

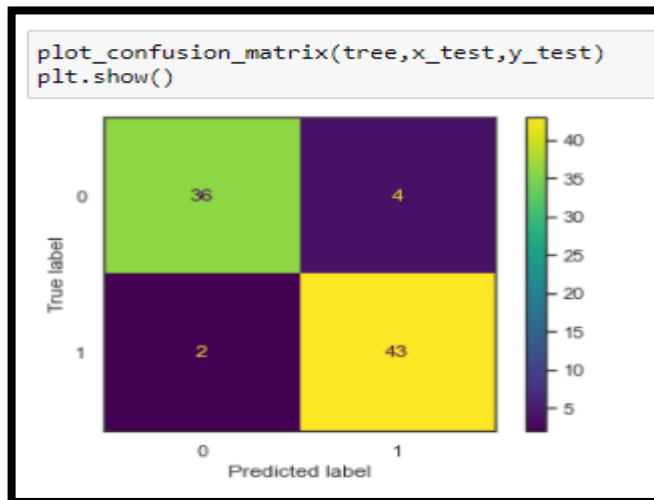


Figure 9.7 DTC Confusion matrix(self,2022)

```
prediction_dt = tree.predict(x_test)
print(classification_report(y_test, prediction_dt))
metrics.plot_roc_curve(tree, x_test, y_test)
```

	precision	recall	f1-score	support
0	0.95	0.90	0.92	40
1	0.91	0.96	0.93	45
accuracy			0.93	85
macro avg	0.93	0.93	0.93	85
weighted avg	0.93	0.93	0.93	85

Figure 9.8 DTC metrics(self,2022)

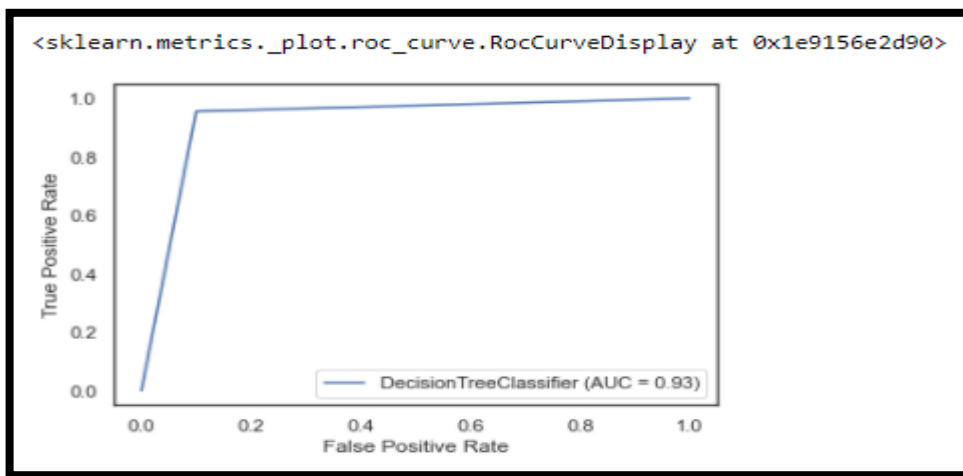


Figure 9.9 DTC ROC AUC curve(self,2022)

### 9.3.2. Support Vector Machine

The support vector machine has a training accuracy of 94.1 percent and a test accuracy of around 96.4 percent. The training score exceeds the test score. To confirm the model's correctness, assessment measures such as the confusion matrix, f1 score, and ROC AUC curve were utilized to describe the model's performance. The evaluation measures are capable of distinguishing between model outcomes. According to the confusion matrix in figure 7.12, 37 are true negative, 44 are true positive, 3 are false positive, and 1 is false negative. The f1 score accuracy is 96 percent when using the confusion matrix, which is the same as the model's testing accuracy. The ROC AUC curve is also 99% highlighting that the accuracy of the decision tree classifier is 96%

```
svm.score(x_train,y_train)  
0.9410029498525073
```

Figure 10.10 SVM Train Dataset Accuracy(self,2022)

```
svm.score(x_test,y_test)  
0.9647058823529412
```

Figure 10.11 SVM Test Dataset accuracy(self,2022)

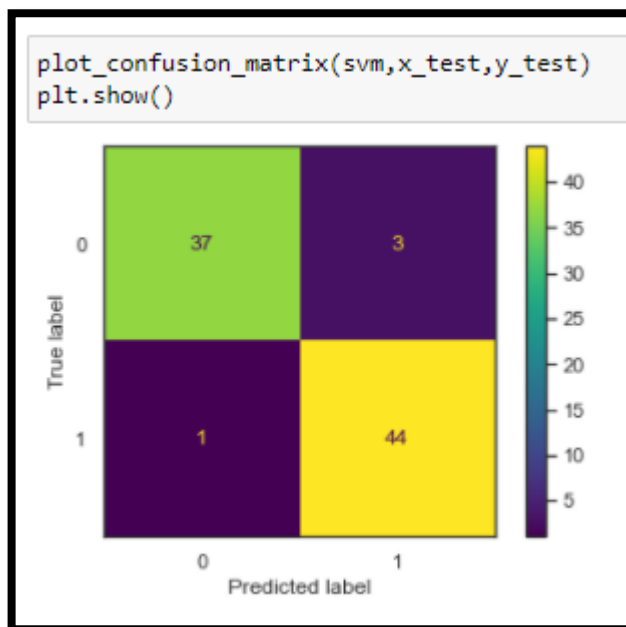


Figure 10.12 SVM Confusion matrix(self,2022)

```
prediction_lr = svm.predict(x_test)
print(classification_report(y_test,prediction_lr))
metrics.plot_roc_curve(svm, x_test, y_test)
```

	precision	recall	f1-score	support
0	0.97	0.95	0.96	40
1	0.96	0.98	0.97	45
accuracy			0.96	85
macro avg	0.97	0.96	0.96	85
weighted avg	0.96	0.96	0.96	85

Figure 10.13 SVM metrics(self,2022)

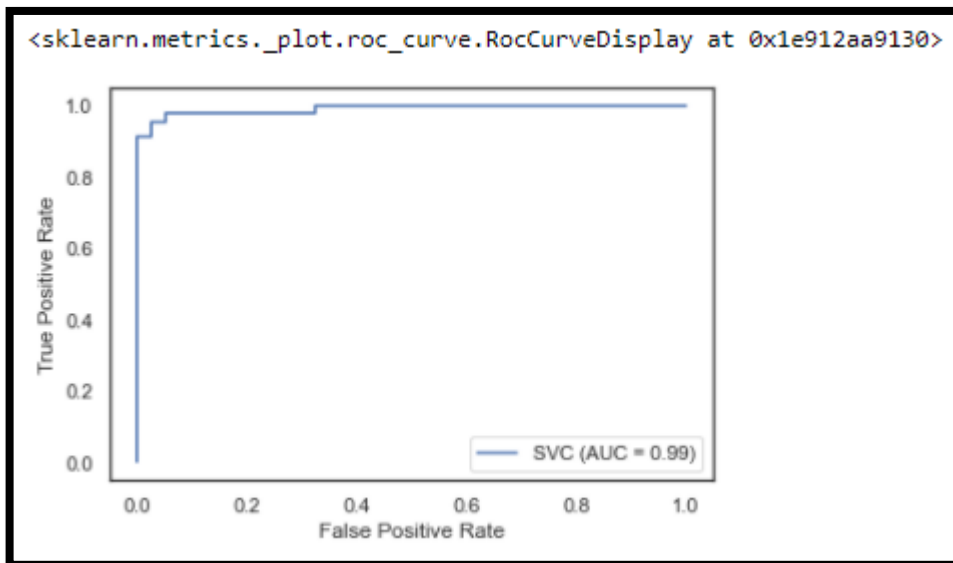


Figure 10.14 SVM ROC AUC score(self,2022)

When comparing the two machine learning algorithms it was noticed that the SVC machine learning algorithms performed better than the decision tree classifier. The svc algorithm had an accuracy of 96% while the decision tree classifier had an accuracy of 93% that is svc performed 3% better than decision tree classifier.

## **Chapter 10: Conclusion**

Finally, the goal of the course was realized by using big data analytics techniques and technology to preprocess the data, convert, display, and analyze it, as well as highlight three significant discoveries. For the coursework, the healthcare area was investigated, and a breast cancer detection dataset was chosen—the coursework stresses big data and its usefulness, particularly in the healthcare sector. The problems and solutions associated with big data were also explored. The raw data from Kaggle was obtained and preprocessed. The data was also adjusted to avoid data bias. Normalization was used to transform the data, which was then viewed and compared. Valuable insights were discovered as a result of the visualization. The data was then fed into two machine learning algorithms, both of which performed well. Although the support vector machine gave better performance when compared to the decision tree classifier. Finally, all the information was then detailed in the coursework.

## **Chapter 11: References**

- ActiveState (2021). *What Is Matplotlib In Python? How to use it for plotting?* [online] ActiveState. Available at: <https://www.activestate.com/resources/quick-reads/what-is-matplotlib-in-python-how-to-use-it-for-plotting/> [Accessed 4 Jan. 2022].
- American Cancer Society (2017). *What is Breast Cancer?* [online] Cancer.org. Available at: <https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html> [Accessed 4 Jan. 2022].
- Amruthnath, N. (2020). *Why balancing your data set is important?* [online] R-bloggers. Available at: <https://www.r-bloggers.com/2020/06/why-balancing-your-data-set-is-importa> [Accessed 2 Jan. 2022].
- Bekker, A. (2018). *The “Scary” Seven: big data challenges and ways to solve them.* [online] Scnsoft.com. Available at: <https://www.scnsoft.com/blog/big-data-challenges-and-their-solutions> [Accessed 4 Jan. 2022].
- Brownlee, J. (2017). *How to Get Started with Kaggle.* [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/get-started-with-kaggle/> [Accessed 2 Jan. 2022].
- Brownlee, J. (2019). *How to Choose a Feature Selection Method For Machine Learning.* [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/> [Accessed 2 Jan. 2022].
- Brownlee, J. (2020). *Train-Test Split for Evaluating Machine Learning Algorithms.* [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/> [Accessed 3 Jan. 2022].
- Choudhury, A. (2021). *Top 8 Data Transformation Methods.* [online] Analytics India Magazine. Available at: <https://analyticsindiamag.com/top-8-data-transformation-methods/> [Accessed 3 Jan. 2022].
- Dua, Dheeru, Graff and Casey (2019). *UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set.* [online] Uci.edu. Available at: <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29> [Accessed 2 Jan. 2022].
- Durcevic, S. (2020). *18 Examples of Big Data In Healthcare That Can Save People.* [online] BI Blog | Data Visualization & Analytics Blog | datapine. Available at: <https://www.datapine.com/blog/big-data-examples-in-healthcare/#how-to> [Accessed 4 Jan. 2022].



Expert.ai Team (2020). *What is Machine Learning? A definition - Expert System*. [online] Expert.ai. Available at: <https://www.expert.ai/blog/machine-learning-definition/> [Accessed 3 Jan. 2022].

InData Labs (2019). *Impact of Big Data on Business – InData Labs*. [online] InData Labs. Available at: <https://indatalabs.com/blog/impact-of-big-data-on-business> [Accessed 4 Jan. 2022].

javatpoint (n.d.). *Python Pandas | Python Pandas Tutorial - javatpoint*. [online] www.javatpoint.com. Available at: <https://www.javatpoint.com/python-pandas> [Accessed 4 Jan. 2022].

Katari, K. (2020). *Seaborn: Python*. [online] Medium. Available at: <https://towardsdatascience.com/seaborn-python-8563c3d0ad41#:~:text=Seaborn%20is%20a%20data%20visualization> [Accessed 4 Jan. 2022].

Ku, L. (2017). *The Impact of Big Data in Business*. [online] PlugandPlay. Available at: <https://www.plugandplaytechcenter.com/resources/impact-big-data-business/> [Accessed 4 Jan. 2022].

National Cancer Institute (2021). *What Is Cancer?* [online] National Cancer Institute. Available at: <https://www.cancer.gov/about-cancer/understanding/what-is-cancer> [Accessed 3 Jan. 2022].

numpy (n.d.). *What is NumPy? — NumPy v1.19 Manual*. [online] numpy.org. Available at: <https://numpy.org/doc/stable/user/whatisnumpy.html> [Accessed 4 Jan. 2022].

omnisci (n.d.). *What is Big Data Architecture? Definition and FAQs | OmniSci*. [online] www.omnisci.com. Available at: <https://www.omnisci.com/technical-glossary/big-data-architecture> [Accessed 5 Jan. 2022].

Ray, S. (2019). *Understanding Support Vector Machine algorithm from examples (along with code)*. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/> [Accessed 3 Jan. 2022].

scikit learn (2009). *1.10. Decision Trees — scikit-learn 0.22 documentation*. [online] Scikit-learn.org. Available at: <https://scikit-learn.org/stable/modules/tree.html> [Accessed 3 Jan. 2022].

Sharma, R. (2020). *6 Methods of Data Transformation in Data Mining*. [online] upGrad blog. Available at: <https://www.upgrad.com/blog/methods-of-data-transformation-in-data-mining/> [Accessed 3 Jan. 2022].

Simplilearn (2020). *What is Big Data Analytics and How It's Being Used*. [online] Simplilearn.com. Available at: <https://www.simplilearn.com/what-is-big-data-analytics-article> [Accessed 4 Jan. 2022].

Sisense (2019). *What is Data Cleaning? | Sisense*. [online] Sisense. Available at: <https://www.sisense.com/glossary/data-cleaning/>. [Accessed 2 Jan. 2022].

solvexia (2019). *15 Big Data Problems You Need to Solve*. [online] [www.solvexia.com](http://www.solvexia.com). Available at: <https://www.solvexia.com/blog/15-big-data-problems-you-need-to-solve#complex-systems> [Accessed 4 Jan. 2022].

Sydorenko, I. (2021). *Big Data and Its Business Impacts*. [online] [labelyourdata.com](http://labelyourdata.com). Available at: <https://labelyourdata.com/articles/big-data-business-impact> [Accessed 4 Jan. 2022].

Tableau (2018). *Data visualization beginner's guide: a definition, examples, and learning resources*. [online] Tableau Software. Available at: <https://www.tableau.com/learn/articles/data-visualization> [Accessed 5 Jan. 2022].

Taylor, D. (2021). *What is BIG DATA? Introduction, Types, Characteristics, Example*. [online] [www.guru99.com](http://www.guru99.com). Available at: <https://www.guru99.com/what-is-big-data.html#6> [Accessed 3 Jan. 2022].

Techopedia.com. (2019). *What is Data Preprocessing? - Definition from Techopedia*. [online] Available at: <https://www.techopedia.com/definition/14650/data-preprocessing> [Accessed 2 Jan. 2022].

tutorialspoint (n.d.). *Scikit Learn - Introduction - Tutorialspoint*. [online] [www.tutorialspoint.com](http://www.tutorialspoint.com). Available at: [https://www.tutorialspoint.com/scikit\\_learn/scikit\\_learn\\_introduction.htm](https://www.tutorialspoint.com/scikit_learn/scikit_learn_introduction.htm) [Accessed 4 Jan. 2022].