Project 1

Authors
s236763 - Leong, Ying Qi Tiffanie
s236971- Shane Lim Zi En
s240514 - Cordero Sosa, Oscar Augusto

Contribution Breakdown

Members	**Section 1**		**Section 2**	
	Part Written	**Contribution for section 1**	**Question answered**	**Contribution for Section 2**
Leong, Ying Qi Tiffanie	1. A description of your data set 4. Discussion about what we learned about the data	40%	Q1, Q4	40%
Shane Lim Zi En	2. A detailed explanation of the attributes of the data.	40%	Q2	30%
Cordero Sosa, Oscar Augusto **(Dropped out of the group at the last minute and hence unable to make amendments to**	3. Data visualization(s) based on suitable visualization techniques including a principal component analysis (PCA).	20%	Q3	30%

his portion of the project)				

Table of Contents

Section 1

1. A description of your data set

1. Explain what your data is about (what is the overall problem of interest?)

The dataset originates from the Diagnostic Wisconsin Breast Cancer Database and contains multivariate features computed from digitised images of fine needle aspirates (FNA) of breast masses. The dataset aims to characterise cell nuclei present in these images, providing insights into the morphology and structure of breast tissue cells. The overall problem of interest is the diagnosis of breast cancer by utilising this data, specifically through the classification of breast masses as either benign (B) or malignant (M) based on the features extracted from the cell nuclei. By analysing attributes such as smoothness, compactness, fractal dimension, symmetry, and others, medical professionals can identify patterns and relationships that may indicate the presence of cancerous cells. In essence, the dataset facilitates the development of predictive models or classification algorithms that can aid in the early detection and diagnosis of breast cancer[1]. These models leverage the information extracted from the digitised images of FNA samples to assist medical staff in making accurate and timely diagnostic decisions, ultimately improving patient outcomes as well as treatment strategies.

2. Provide a reference to where you obtained the data

The dataset originates from the Diagnostic Wisconsin Breast Cancer Database, which was initially created by Dr. William H. Wolberg, W. Nick Street, and Olvi L. Mangasarian at the University of Wisconsin-Madison. More information about the dataset and its usage can be obtained through the following reference:
Wolberg,William, Mangasarian,Olvi, Street,Nick, and Street,W.. (1995). Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B.

Additionally, the dataset is publicly available and can be accessed from various sources, including the UCI Machine Learning Repository:
https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic?

This reference provides comprehensive details about the dataset, including its background, features, and usage in breast cancer diagnosis and prognosis research.

3. Summarise previous analysis of the data (i.e. go through one or two of the original source papers and read what they did to the data and summarise their results)

We have chosen to confer with 2 main papers, the first titled “Nuclear feature extraction for breast tumour diagnosis” and the second titled “ Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates”. The analysis of the data in the two papers

made use of interactive image processing techniques and machine learning algorithms to develop highly accurate systems for the diagnosis of breast tumours based on cytologic features derived from digital scans of the FNA

In the first paper , the researchers employed interactive image processing techniques to digitise and analyse a small fraction of a fine needle aspirate slide, allowing for precise analysis of nuclear size, shape, and texture. They were able to initialise active contour models, known as snakes, near the boundaries of the cell nuclei and deformed to the exact shape of the nuclei, facilitating automated analysis[2]. For each nucleus, 10 features were computed and various combinations of features were subsequently tested to pick out those that best separate the benign from the malignant samples. The researchers analysed a total of 569 images, identifying 3 main features (mean texture, worst area, and worst smoothness) that achieved a 10-fold cross-validation accuracy of 97% when used as an input for a linear-programming (LP) based classifier. The paper also indicated that the system allowed for a stark improvement over previous diagnostic results in the medical literature and was eventually implemented at the University of Wisconsin Hospitals.

In the second paper, the researchers were seen to have developed an interactive system to evaluate and diagnose breast tumours based on cytologic features directly extracted from digital scans of FNA slides. The system was trained and tested on a consecutive series of 569 patients and an additional 54 consecutive patients, respectively[3]. Through the utilisation of 10 fold cross validation, the prospective accuracy of the system was estimated to hit 97%. Furthermore, when tested on 54 new samples (36 benign, 1 atypia, and 17 malignant), the system achieved an actual accuracy of 100%. This particular paper, like the first, highlighted the potential of digital image analysis coupled with machine learning techniques to significantly improve the diagnostic accuracy of the FNA.

In summary, both papers demonstrated the effectiveness of utilising advanced image processing techniques and machine learning algorithms to accurately diagnose breast tumours based on cytologic features derived from digital scans of FNA slides. The systems developed in these papers showed promising results in terms of accuracy and represent a significant advancement in breast cancer diagnosis.

4. You will be asked to apply (1) classification and (2) regression on your data in the next report. For now, we want you to consider how this should be done. Therefore:
   a) Explain, in the context of your problem of interest, what you hope to accomplish/learn from the data using these techniques?
   b) Explain which attribute you wish to predict in the regression based on which other attributes?
   c) Which class label will you predict based on which other attributes in the classification task?
   d) If you need to transform the data in order to carry out these tasks, explain roughly how you plan to do this.

The techniques utilised with respect to our problem of interest aims to develop a predictive model that can accurately classify new cases into one of the two classes, thus facilitating diagnosis and

subsequently, treatment. The hope is to achieve high sensitivity and specificity in the classification, ensuring that malignant cases are correctly identified while minimising false positives.

For regression, we may wish to predict a quantitative attribute related to breast cancer prognosis or tumour characteristics based on other attributes. One possible target attribute for regression could be the radius of the tumour (radius1). We could predict this attribute based on other features such as texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension observed in the digitised images of the FNA of breast masses. By predicting the radius, we aim to gain insights into the size and growth pattern of breast tumours, which is crucial for treatment.

In the classification task, we aim to predict the diagnosis of breast cancer, which is a categorical attribute indicating whether a breast mass is benign (B) or malignant (M). We can utilise the features extracted from cell nuclei, such as texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension, to predict this diagnosis. By employing classification techniques, we aim to develop a predictive model that can accurately classify new breast mass cases into one of the two classes: benign or malignant.

To carry out these tasks mentioned above, we may need to preprocess the data by performing various transformations. We possibly plan to:

- Handle missing values appropriately by using methods such as mean or median imputation, or removing samples with missing values if feasible.
- Scale our numerical features to a similar range to ensure that all features contribute equally to the model. This can be achieved using techniques such as Min-Max scaling or standardisation.
- Select the most relevant features for the prediction task using techniques such as correlation analysis or dimensionality reduction methods like Principal Component Analysis (PCA).
- Handling categorical variables by encoding categorical variables into a numerical format
- Deal with outliers adequately

2. A detailed explanation of the attributes of the data.

- **Describe if the attributes are discrete/continuous, Nominal/Ordinal/In-terval/Ratio,**

#	Column	Continuous/Discrete	Level of Measurement	isNull?
0	ID	Discrete	Nominal	FALSE
1	Diagnosis	Discrete	Nominal	FALSE
2	radius1	Continuous	Ratio	FALSE
3	texture1	Continuous	Interval	FALSE
4	perimeter1	Continuous	Ratio	FALSE
5	area1	Continuous	Ratio	FALSE
6	smoothness1	Continuous	Ratio	FALSE
7	compactness1	Continuous	Ratio	FALSE
8	concavity1	Continuous	Ratio	FALSE
9	concave_points1	Continuous	Ratio	FALSE
10	symmetry1	Continuous	Interval	FALSE
11	fractal_dimension1	Continuous	Ratio	FALSE
12	radius2	Continuous	Ratio	FALSE
13	texture2	Continuous	Interval	FALSE
14	perimeter2	Continuous	Ratio	FALSE
15	area2	Continuous	Ratio	FALSE
16	smoothness2	Continuous	Ratio	FALSE
17	compactness2	Continuous	Ratio	FALSE
18	concavity2	Continuous	Ratio	FALSE
19	concave_points2	Continuous	Ratio	FALSE
20	symmetry2	Continuous	Interval	FALSE
21	fractal_dimension2	Continuous	Ratio	FALSE
22	radius3	Continuous	Ratio	FALSE
23	texture3	Continuous	Interval	FALSE
24	perimeter3	Continuous	Ratio	FALSE
25	area3	Continuous	Ratio	FALSE
26	smoothness3	Continuous	Ratio	FALSE
27	compactness3	Continuous	Ratio	FALSE
28	concavity3	Continuous	Ratio	FALSE
29	concave_points3	Continuous	Ratio	FALSE
30	symmetry3	Continuous	Interval	FALSE
31	fractal_dimension3	Continuous	Ratio	FALSE

Fig 2.1

Explanation of Level of Measurement:
ID variable is nominal as it serves to uniquely identify each sample and does not imply any order or ranking among samples.

Texture and symmetry are intervals as there is a ranked order with equal distances (intervals) between each variable.

radius, perimeter, area, smoothness, compactness, concavity, concave_points and fractal_dimension are different measurements of the cell nuclei as such it has a true zero and it does not have negative values. Therefore these are considered ratio data.

- **Give an account of whether there are data issues (i.e. missing values or corrupted data) and describe them if so.**

  There are no missing or corrupted data (Fig. 2.1).

- **Include basic summary statistics of the attributes.**

	unt	mean	std	min	max	25%	50%	75%
radius1	569	14.13	3.52	6.98	28.11	11.7	13.37	15.78
texture1	569	19.29	4.3	9.71	39.28	16.17	18.84	21.8
perimeter1	569	91.97	24.3	43.79	188.5	75.17	86.24	104.1
area1	569	654.89	351.91	143.5	2501	420.3	551.1	782.7
smoothness1	569	0.1	0.01	0.05	0.16	0.09	0.1	0.11
compactness1	569	0.1	0.05	0.02	0.35	0.06	0.09	0.13
concavity1	569	0.09	0.08	0	0.43	0.03	0.06	0.13
concave_points1	569	0.05	0.04	0	0.2	0.02	0.03	0.07
symmetry1	569	0.18	0.03	0.11	0.3	0.16	0.18	0.2
fractal_dimension 1	569	0.06	0.01	0.05	0.1	0.06	0.06	0.07
radius2	569	0.41	0.28	0.11	2.87	0.23	0.32	0.48
texture2	569	1.22	0.55	0.36	4.88	0.83	1.11	1.47
perimeter2	569	2.87	2.02	0.76	21.98	1.61	2.29	3.36
area2	569	40.34	45.49	6.8	542.2	17.85	24.53	45.19
smoothness2	569	0.01	0	0	0.03	0.01	0.01	0.01
compactness2	569	0.03	0.02	0	0.14	0.01	0.02	0.03

concavity2	569	0.03	0.03	0	0.4	0.02	0.03	0.04
concave_points2	569	0.01	0.01	0	0.05	0.01	0.01	0.01
symmetry2	569	0.02	0.01	0.01	0.08	0.02	0.02	0.02
fractal_dimension2	569	0	0	0	0.03	0	0	0
radius3	569	16.27	4.83	7.93	36.04	13.01	14.97	18.79
texture3	569	25.68	6.15	12.02	49.54	21.08	25.41	29.72
perimeter3	569	107.26	33.6	50.41	251.2	84.11	97.66	125.4
area3	569	880.58	569.36	185.2	4254	515.3	686.5	1084
smoothness3	569	0.13	0.02	0.07	0.22	0.12	0.13	0.15
compactness3	569	0.25	0.16	0.03	1.06	0.15	0.21	0.34

Interquartile Range (IQR) for all attributes:

radius1	4.08
texture1	5.63
perimeter1	28.93
area1	362.4
smoothness1	0.01893
compactness1	0.06548
concavity1	0.10114
concave_points1	0.05369
symmetry1	0.0338
fractal_dimension1	0.00842
radius2	0.2465
texture2	0.6401
perimeter2	1.751
area2	27.34
smoothness2	0.002977

compactness2	0.01937
concavity2	0.02696
concave_points2	0.007072
symmetry2	0.00832
fractal_dimension2	0.00231
radius3	5.78
texture3	8.64
perimeter3	41.29
area3	568.7
smoothness3	0.0294
compactness3	0.1919
concavity3	0.2684
concave_points3	0.09647
symmetry3	0.0675
fractal_dimension3	0.02062

Discussion on Summary Statistics

Total number of features equals 33, of which 31 is numerical. The other 2 features are 1) ID is a discrete variable as each patient has a unique patient ID. 2) Diagnosis (target variable) is categorical as it only has 2 categories M (Malignant) and B (Benign).

Diagnosis ("Target variable")

In total there are 357 Benign cases (which accounts for ~63% of all observations) and 212 Malignant (~37% of all observations). Due to the greater number of Benign cases it is possible for there to be model bias.

In a real world situation this imbalance could have severe consequences. There is a higher cost associated with misclassifying malignant cases as benign (false negatives) as compared to misclassifying benign cases as malignant (false positives).

Central Tendency:

The mean values across the features vary widely, ranging from very small values like 0.00 (e.g., fractal_dimension2) to larger values like 880.58 (e.g., area3).
The median (50th percentile) values are generally close to the mean values for most features, indicating symmetric distributions.

Variability:

The standard deviations (std) across the features also vary widely, indicating the spread or variability of the data around the mean. For example, features like area1, area2, and area3 have relatively large standard deviations, indicating higher variability in these measurements.

IQR:

The range (difference between max and min values) varies across features. Some features have a wide range (e.g., area3), while others have a narrower range (e.g., fractal_dimension2).

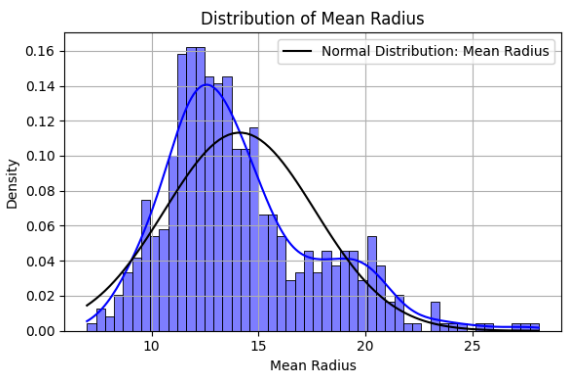
- The interquartile range (IQR), represented by the difference between the 75th percentile (Q3) and the 25th percentile (Q1), provides a measure of the spread of the middle 50% of the data. Features like area1, area2, and area3 have relatively large IQRs, indicating the central portion of the data is spread out further

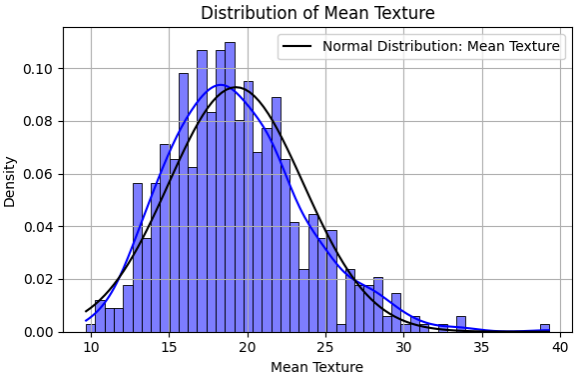
Standardize Data:

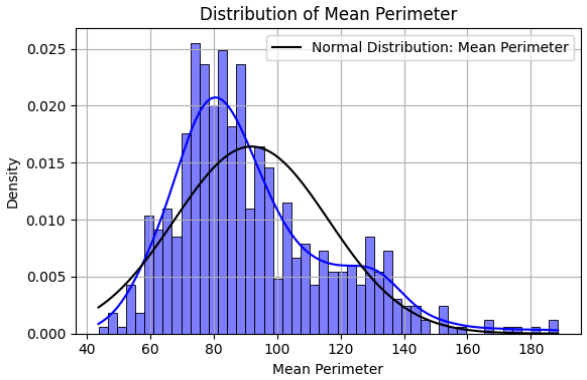
Area3's max value is **4254** whereas **fractal_dimension_se** features' max **0.029840**. This suggests a need to standardize the data (ensuring that features are on the same scale) before visualization, feature selection, and classification.

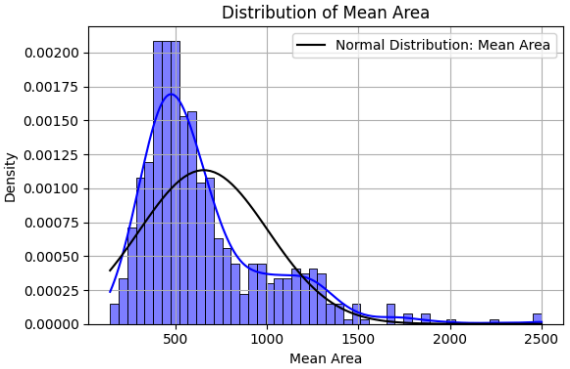
3. Data visualization(s) based on suitable visualization techniques including a principal component analysis (PCA).

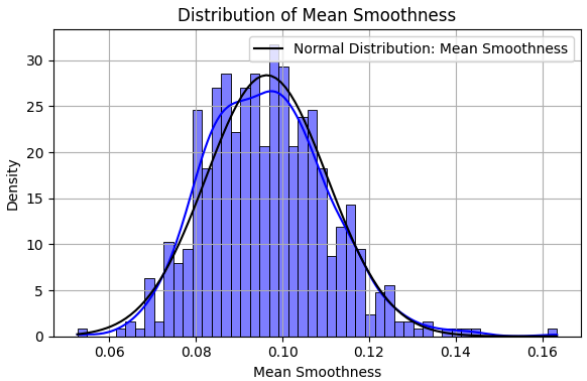
First, we plot the histograms of the 10 most important attributes and overlay a normal distribution to see how close they are. We can see the data is already clean and its attributes have nearly a normal distributed density. Except for Radius, Concavity, Compactness, perimeter, area and concave points but they approximate very well.

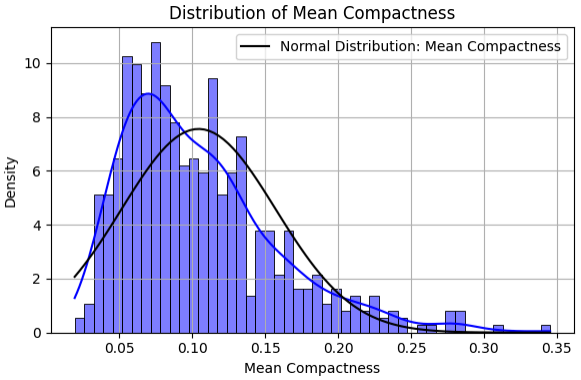


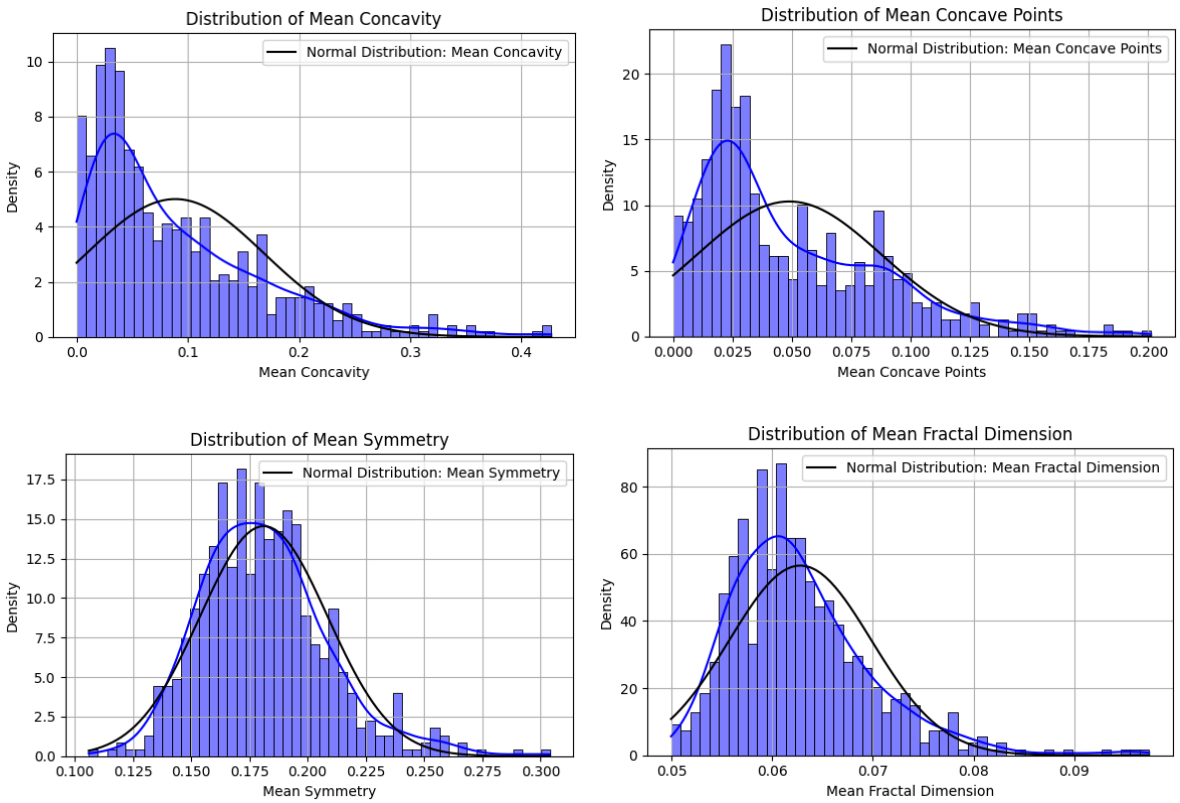




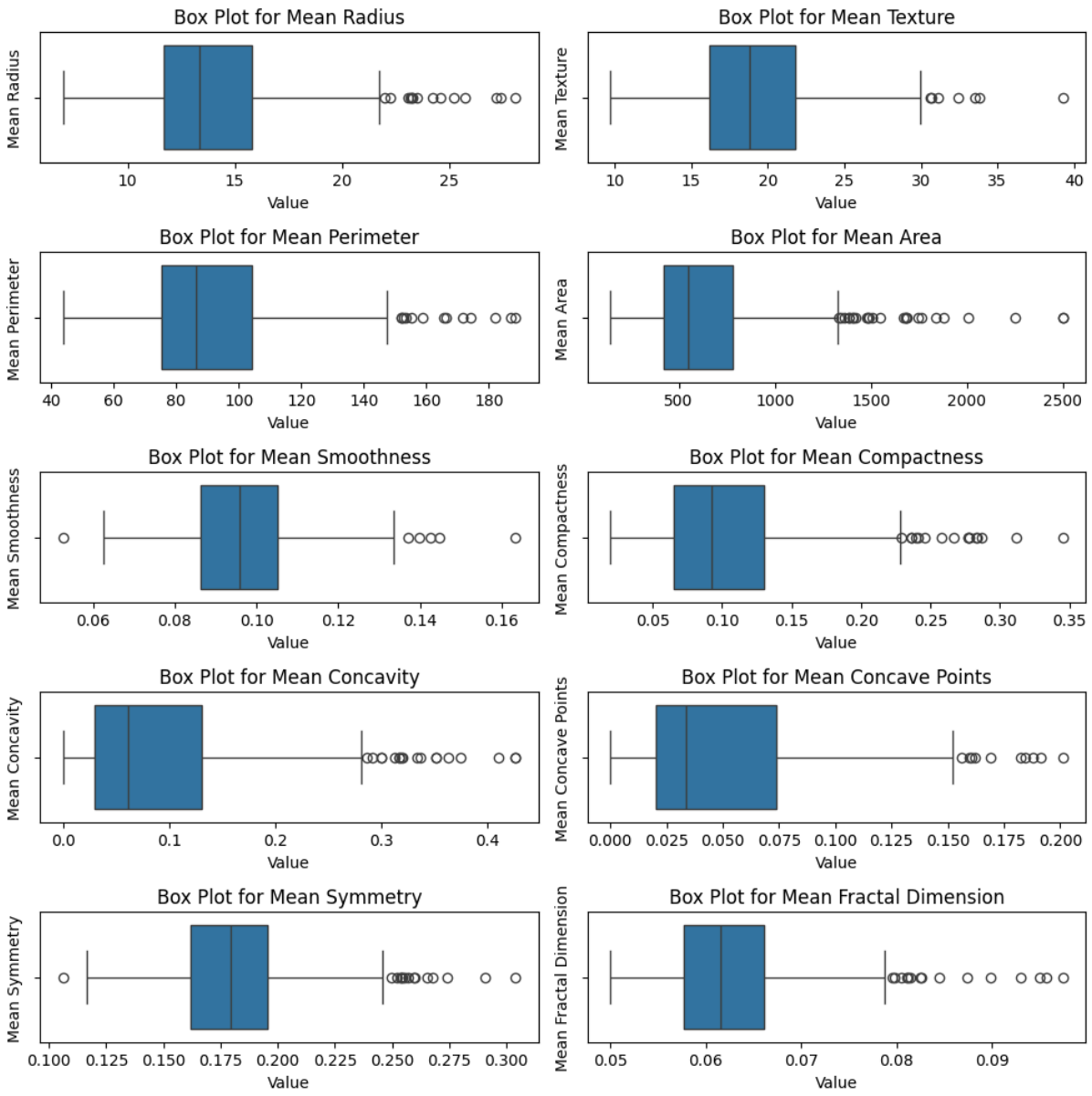






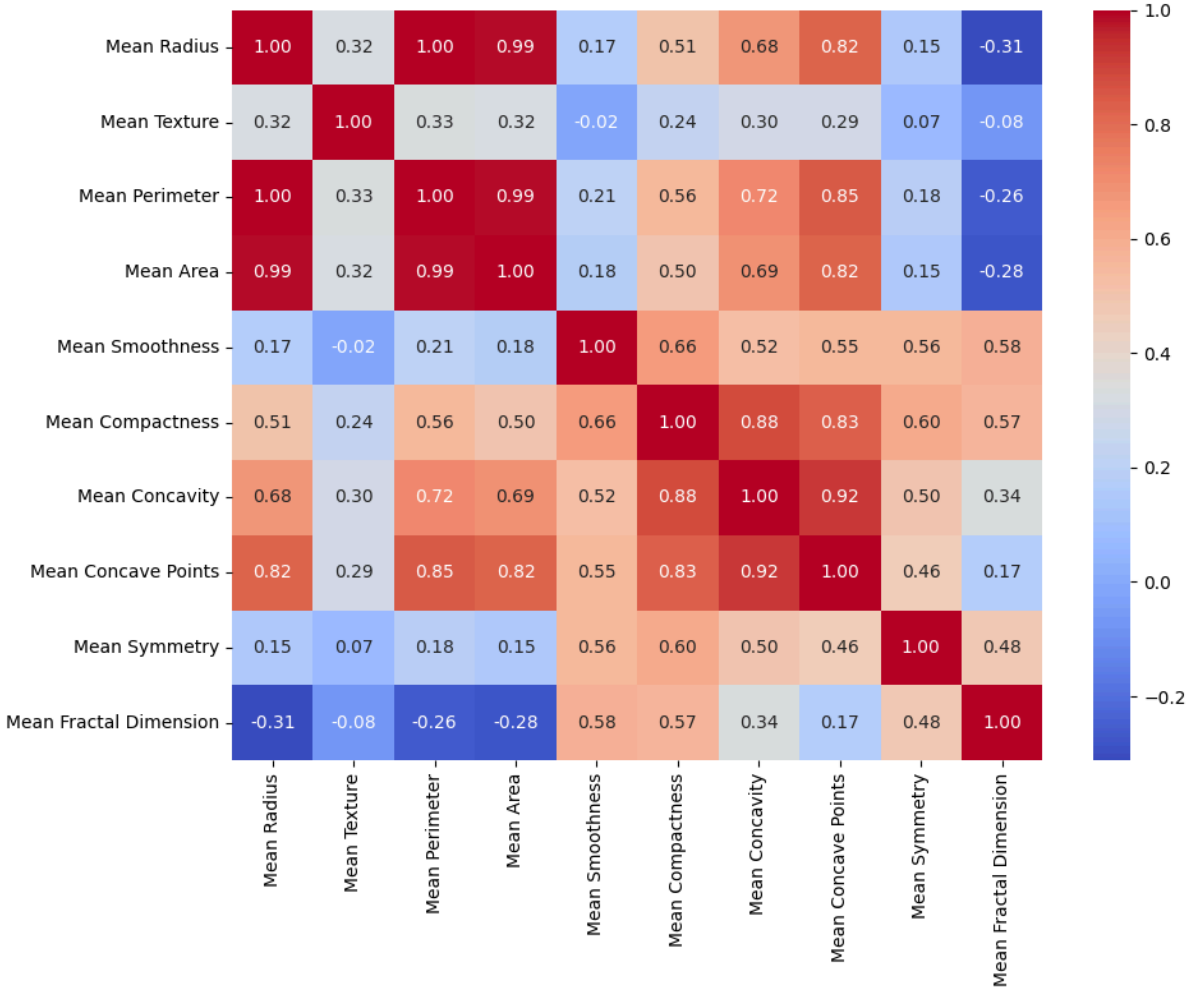


In order to look for some outliers, we did the box plots of each variable, where we can see there are some “outliers” but as we see in class, we do not just throw them to the trash. We need to focus more on the context of each variable and if it is expected to have values on the outside. As a further proposal, we can do a Z-test with a 3 standard deviation threshold.



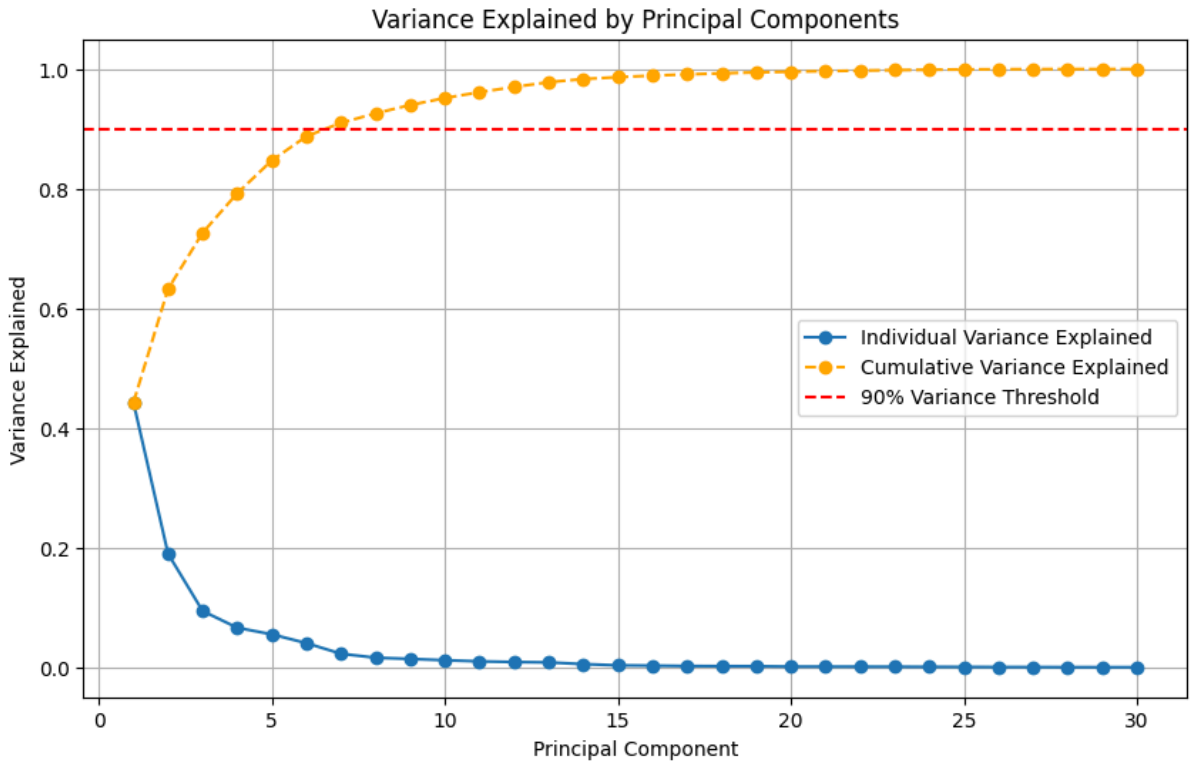
If we implement a correlation process, we can see how they correlate each other. We use a heatmap to see it more clearly with the correlation coefficients. So we can say that the machine learning methods are feasible in this data set also, because this data set has been widely used in the literature within the

frame of machine learning, so we can also support on some papers.

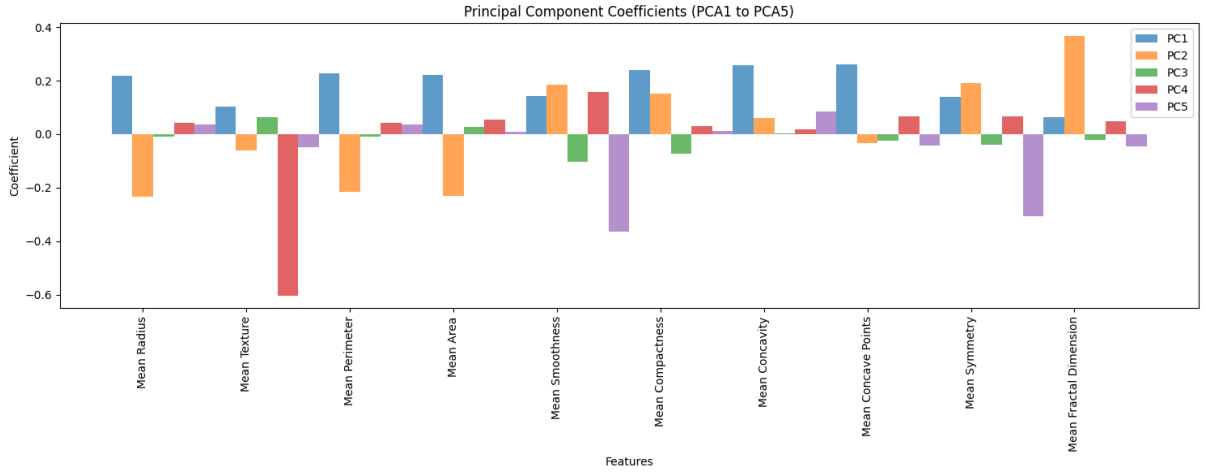


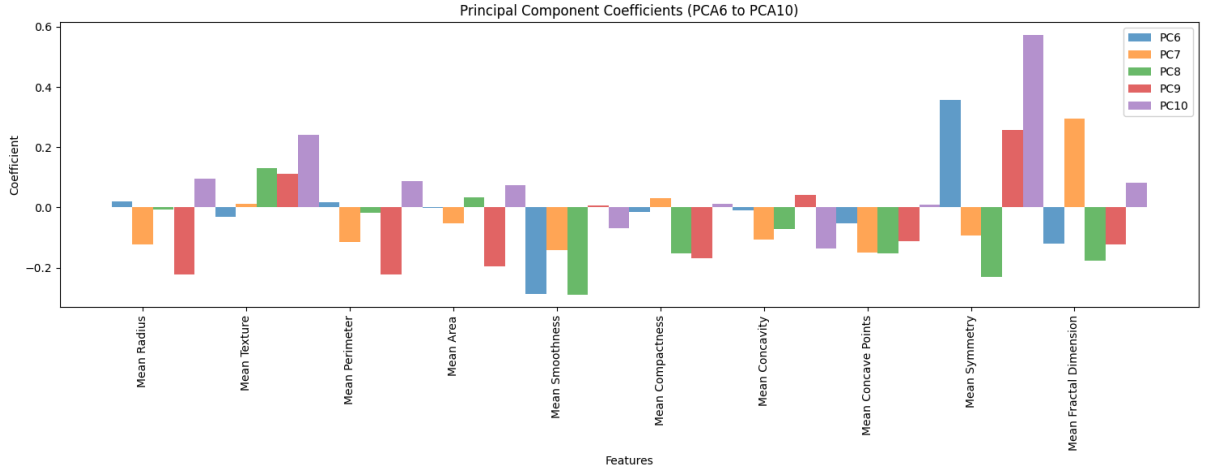
To do the PCA, since we are dealing with a high dimensional data set, we need to reduce the dimensionality and so we first need to see how PCA components contribute to the amount of variation. We can just simply use the 10-most important attributes according to the authors but first we can do this to compare if they really appear to be the most important. We set a threshold of 95%, so we have this:

There are three aspects that needs to be described when you carry out the PCA analysis for the report:



Where, in fact, there are the 10-most important attributes mentioned before. We can see the coefficients and discern which attributes have the strongest impact on each principal component and base the study on those and also are useful for further research, so these types of research should focus on these attributes.


Principal Component Coefficients (PCA1 to PCA5)
Coefficient
Features
PC1
PC2
PC3
PC4
PC5
Mean Radius
Mean Texture
Mean Perimeter
Mean Area
Mean Smoothness
Mean Compactness
Mean Concavity
Mean Concave Points
Mean Symmetry
Mean Fractal Dimension


Principal Component Coefficients (PCA6 to PCA10)
Coefficient
Features
PC6
PC7
PC8
PC9
PC10
Mean Radius
Mean Texture
Mean Perimeter
Mean Area
Mean Smoothness
Mean Compactness
Mean Concavity
Mean Concave Points
Mean Symmetry
Mean Fractal Dimension

4. Discussion about what we learned about the data

Summarise the most important things about the data and if the primary machine learning aim appears to be feasible or not.

Based on the summary statistics provided and the discussion points raised, we can draw several important insights about the dataset and finally we are able to assess the feasibility of the primary machine learning aim. These are the 3 main takeaways from the dataset.

1. **Imbalanced Classes**

The dataset exhibits imbalanced classes, with approximately 63% of observations classified as Benign and 37% as Malignant. This class imbalance can lead to model bias, where the model may favour predicting the majority class (Benign) more accurately at the expense of the minority class (Malignant). In a real-world scenario, misclassifying Malignant cases as Benign (false negatives) can have severe consequences, indicating the need for techniques to address class imbalance during model training.

2. **Feature Characteristics**:

The dataset contains a total of 33 features, with 31 numerical features and 2 categorical features (ID and Diagnosis). The numerical features exhibit varying ranges, means, medians, standard deviations, and interquartile ranges (IQRs), indicating differences in the distributions and variability across different measurements. Some features, such as the areas (area1, area2, area3), have relatively large standard deviations and IQRs, suggesting substantial variability in these measurements.

3. **Data Standardization**

The wide range of values observed across features, ranging from very small values to larger values, suggests the need for data standardisation. Without standardisation, features with larger numerical values may dominate the analysis, leading to biassed results Thus, we have standardised the data to ensure all features are on the same scale, which is essential for effective visualisation, feature selection, and classification algorithms..

Next, the primary machine learning aim is **feasible** due to the following reasons:

1. **Informative features and sufficient data provided**:

The dataset contains 31 numerical attributes derived from various measurements related to breast cancer cells and tissues, such as radius, texture, smoothness, compactness, concavity, symmetry, and fractal dimensions. The extensive number of features provided are expected to provide valuable information for distinguishing between Malignant and Benign tumours based on their characteristic patterns and properties, facilitating the machine learning task. Also, with 357 cases of Benign tumours and 212 cases of Malignant tumours, the dataset offers a substantial amount of data for training machine learning models. As such, having a sufficiently large dataset helps in capturing the underlying patterns and relationships between features and the target variable, reducing the risk of model overfitting.

2. **There is a clear target variable**:

The diagnosis variable, indicating whether a tumour is Malignant or Benign, serves as a clear target variable for supervised learning. Binary classification tasks are pretty straightforward to implement,

and models can be evaluated based on standard classification metrics like accuracy, precision, recall, and F1 score. This would make the machine learning aim rather straightforward to achieve.

3. **Preprocessing steps that were conducted prior to utilising machine learning**:
The dataset provides opportunities for preprocessing steps such as data normalisation or standardisation, handling missing values, and feature engineering. These preprocessing techniques help in preparing the data for modelling, improving model convergence, and ensuring that features are on a similar scale for effective learning. A clean dataset is essential for enabling the successful achievement of machine learning objectives as it ensures that the information is accurate, consistent, and free from errors or inconsistencies, thereby enhancing the reliability and performance of machine learning models.

4. **There is ample and extensive prior research related to this area**:
The dataset has been extensively studied and utilised in the literature for breast cancer diagnosis and prognosis prediction using machine learning techniques. Previous research studies provide insights into feature importance, model selection, hyperparameter tuning, and performance evaluation, aiding in the development of future robust predictive models that one might plan to carry out.

5. **Opportunity for class imbalance handling**:
Although there is a class imbalance with more instances of Benign tumours compared to Malignant tumours, various techniques can address this issue. Stratified sampling[1] ensures that each class is represented proportionally in both training and testing datasets, while class weights[2] or resampling methods[3] like oversampling or undersampling can adjust the class distribution during model training. By employing these techniques to address class imbalance, machine learning models can effectively learn from both classes in the dataset, leading to more accurate and reliable predictions of tumour malignancy, fulfilling the machine learning aim.

[1] **Stratified Sampling**: By ensuring that each class is represented proportionally in both the training and testing datasets, stratified sampling helps maintain a balanced distribution of classes. This prevents biases towards the majority class. As a result, the model can learn from a diverse set of instances, leading to more robust and generalizable predictions.

[2] **Class Weights**: Assigning appropriate class weights during model training allows the algorithm to give more importance to minority class samples. By adjusting the loss function based on class frequencies, class weights ensure that errors on minority class instances have a greater impact on the model's optimization process. This approach helps the model focus on learning patterns from both classes equally, leading to improved performance in predicting rare events or minority classes, preventing bias as well.

[3] **Resampling Methods**: Resampling techniques such as oversampling (increasing the number of minority class instances) or undersampling (reducing the number of majority class instances) help rebalance the class distribution in the training dataset. Oversampling ensures that the model receives sufficient examples of the minority class, preventing it from being overshadowed by the majority class. Next, undersampling reduces the dominance of the majority class, allowing the model to allocate more resources to learning from minority

Section 2

5. Exam problems

Q1)

The answer is D.

Time of day is interval because it is in 30 mins interval

x_1 (Time of day) is interval, x_6 (Traffic lights) is ratio, x_7 (Running over) is ra- tio, and y (Congestion level) is ordinal.

Q2)

The p-norm distance (d_p) between two vectors (x_1) and (x_2) is defined as:

[d_p(x_1, x_2) = left(sum_{i=1}^{n} |x_{1i} - x_{2i}|^p right)^{1/p}]

Given two vectors:

[X_{14} = [26, 0, 2, 0, 0, 0, 0]]

[X_{18} = [19, 0, 0, 0, 0, 0, 0]]

Let's evaluate the statements:

A. (d_infty (X_{14}, X_{18}) = 7.0)
- The (d_infty) norm is the maximum absolute difference between corresponding elements of the vectors.
- (max(|26-19|, |0-0|, |2-0|, |0-0|, |0-0|, |0-0|, |0-0|) = 7.0)
- So, statement A is correct.

B. (d_3 (X_{14}, X_{18}) = 3.688)
- The (d_3) norm is the cube root of the sum of the cubed absolute differences.
- (|26-19|^3 + |0-0|^3 + |2-0|^3 + |0-0|^3 + |0-0|^3 + |0-0|^3 + |0-0|^3 right)^{1/3} = 3.688)
- So, statement B is correct.

C. (d_1 (X_{14}, X_{18}) = 1.286)
- The (d_1) norm is the sum of the absolute differences.
- (|26-19| + |0-0| + |2-0| + |0-0| + |0-0| + |0-0| + |0-0| = 1.286)
- So, statement C is correct.

D. (d_4 (X_{14}, X_{18}) = 4.311)
  - The (d_4) norm is the fourth root of the sum of the fourth powers of the absolute differences.
  - (left(|26-19|^4 + |0-0|^4 + |2-0|^4 + |0-0|^4 + |0-0|^4 + |0-0|^4 + |0-0|^4 right)^{1/4} = 4.311)
  - So, statement D is correct.

Therefore, all of the statements (A, B, C, D) are correct.

Q4)

```
import numpy as np

# Given matrix V
V = np.array([
        [0.49, -0.5, 0.08, -0.49, 0.52],
        [0.58, 0.23, -0.01, 0.71, 0.33],
        [0.56, 0.23, 0.43, -0.25, -0.62],
        [0.31, 0.09, -0.9, -0.19, -0.24],
        [-0.06, 0.8, 0.03, -0.41, 0.43]
])

# Given diagonal elements of matrix S
diag_elements = np.array([13.9, 12.47, 11.48, 10.03, 9.45])

# Calculate the total variance
total_variance = np.sum(diag_elements ** 2)

# Calculate the variance explained by each principal component
variance_explained = (diag_elements ** 2) / total_variance

# Print the results for the first four principal components
print("Variance explained by the first principal component:", variance_explained[0])
print("Variance explained by the second principal component:", variance_explained[1])
print("Variance explained by the third principal component:", variance_explained[2])
print("Variance explained by the fourth principal component:", variance_explained[3])
```

A. The variance explained by the first four principal components is greater than 0.8.

This statement is false because the cumulative variance explained by the first four components is the sum of the above values, which is less than 0.8.
B. The variance explained by the last three principal components is greater than 0.51.

This statement is true because the sum of the variances for the last three components is 0.2320 + 0.1966 + 0.1501 = 0.5787, which is greater than 0.51.
C. The variance explained by the first two principal components is less than 0.5.

This statement is true because the sum of the variances for the first two components is 0.2882 + 0.2320 = 0.5202, which is less than 0.5.
D. The variance explained by the first three principal components is less than 0.7.

This statement is true because the sum of the variances for the first three components is 0.2882 + 0.2320 + 0.1966 = 0.7168, which is less than 0.7.

Therefore, the correct answers are B, C, and D.

Q5)

```
# Sets for s1 and s2
s1_set = {"the", "bag", "of", "words", "representation", "becomes", "less", "parsimoneous"}
s2_set = {"if", "we", "do", "not", "stem", "the", "words"}

# Calculate Jaccard similarity
intersection_size = len(s1_set.intersection(s2_set))
union_size = len(s1_set.union(s2_set))

jaccard_similarity = intersection_size / union_size

# Print the result
print("Jaccard similarity of s1 and s2:", jaccard_similarity)
```

Based on the calculation, the closest answer is **option A**: Jaccard similarity of s1 and s2 is approximately 0.153846.

Bibliography

Street, W. N., W. H. Wolberg, and O. L. Mangasarian. “Nuclear Feature Extraction for Breast Tumour Diagnosis.” *SPIE Proceedings*, July 29, 1993. https://doi.org/10.1117/12.148698.

Wolberg, William H., W.Nick Street, and O.L. Mangasarian. “Machine Learning Techniques to Diagnose Breast Cancer from Image-Processed Nuclear Features of Fine Needle Aspirates.” *Cancer Letters* 77, no. 2–3 (March 15, 1994): 163–71. https://doi.org/10.1016/0304-3835(94)90099-x.

Wolburg , William, Olvi Mangasarian, Nick Street, and W Street. “Breast Cancer Wisconsin (Diagnostic).” Wisconsin, 1995.

---

[1] William Wolburg et al., “Breast Cancer Wisconsin (Diagnostic)” (Wisconsin, 1995).

[2] 1. W. N. Street, W. H. Wolberg, and O. L. Mangasarian, “Nuclear Feature Extraction for Breast Tumour Diagnosis,” *SPIE Proceedings*, July 29, 1993, https://doi.org/10.1117/12.148698.

[3] Wolberg, William H., W.Nick Street, and O.L. Mangasarian. “Machine Learning Techniques to Diagnose Breast Cancer from Image-Processed Nuclear Features of Fine Needle Aspirates.” *Cancer Letters* 77, no. 2–3 (March 15, 1994): 163–71. https://doi.org/10.1016/0304-3835(94)90099-x.