DETAILS OF ASSIGNMENT			
STUDENT NAME	Kartik Patel	ID NUMBER	101982976
EMAIL ADDRESS	101982976@student.swin.edu.au	PHONE CONTACT	0420690325
UNIT CODE * NAME	COS60008 – Introduction to Data Science		
ASSESSMENT TITLE	Assignment 2		
TUTOR'S NAME	Yu Sun	DATE OF SUBMISSION	29/05/2020

DECLARATION			
I declare that (the first four boxes must be completed for the assignment to be accepted): □ This assignment does not contain any material that has previously been submitted for assessment at this or any other university. This is an original piece of work and no part has been completed by any other student than signed below; □ I have read and understood the avoiding plagiarism guidelines and no part of this work has been copied or paraphrased **from any other source** except where this has been clearly acknowledged in the body of the assignment and included in the reference list; □ I have retained a copy of this assignment in the event of it becoming lost or damaged; □ (optional) I agree to a copy of the assignment being retained as an exemplar for future students (subject to identifying details being removed).			
Student acknowledgement (by typing your name you agree to the above):	I agree, Kartik Patel	**Date:**	29/05/2020

Executive Summary

The aim of this report is identification and classification problem-based data analysis and formulating the problem and determining the data needed to solve this problem. This report explains the problem formulation, data acquisition from UCI repository, data cleaning, and preparation for the exploration of data analysis. Also, at the end of the report included the data modelling part using two selected classification models and a comparison of both models using appropriate graphical visualisations.

Table of Contents

Introduction

As a data scientist, we need to work with a various data science project to solve data and business problems. The data science project has three phases to complete the project. The first phase is data preparation, need to require time and effort to gathering data for the project. The second phase is data building, using various mathematics and statics do the exploration of that data from planning to execution. The last stage is finishing, delivered the final project output to business and wrapping up the project. [1]

In this project, I have gone through all the three phases of the data science process. To complete this project, the first task is to find the appropriate data and gathering require Information, after collecting the data and information we are trying to clean that data and prepare data for the next step. The second task is exploratory data analysis on the clean dataset and explores the dataset all attributes with its relationship between this all dataset attributes. In addition to this, one research question is raised and try to get the answer to that question using statistics or graphical visualizations. The next part is data modelling on the data set, create train and test data set, and apply two classification model SVM and Gaussian Naïve Bayes on this data set and evaluate the target parameter to find accuracy and other parameters of evaluation and compare both model results using visualization graph. The last part of this project is a discussion about project work and the conclusion of the project outcomes.

Problem Formulation

Every year around more than 2 million women globally death by breast cancer diseases. To improve the advancements of the treatments and diagnostic ability of breast cancer, we need to build some early detection machines so physicians can treat disease aggressively and giving a better chance of survival from this disease.

This project aims to predict the recurrence of breast cancer using machine learning classification algorithms. To achieve this, we have used the dataset with 286 instances of breast cancer dataset Institute of Oncology University Medical Centre Ljubljana, Yugoslavia from UCI Machine Learning Repository.

Task 1 - Data Acquisition and Preparation

Task 1.1 Data Acquisition

Task 1.1.1 Data Set Selection

To meet the assignment requirement for the dataset to complete the data science project, the below link is used for the selection of the dataset.

https://archive.ics.uci.edu/ml/datasets.php?format=&task=cla&att=&area=&numAtt=less10&numIns=100to1000&type=&sort=nameUp&view=table

After carefully reviewing all the results of the classification data set, I have selected three dataset Blood Transfusion Service(https://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center), Hayes-Roth Dataset(https://archive.ics.uci.edu/ml/datasets/Hayes-Roth) and Blood cancer dataset(http://archive.ics.uci.edu/ml/datasets/Breast+Cancer). Once analyse given Information on these three datasets, I have chosen the breast cancer data set to complete this project.

Data Set Information:

This is one of the three domains and it provided by the oncology institute. This dataset includes 201 instances of one class and 85 instances of another class. There is a total of 10 attributes, with one class attribute of breast cancer recurrence or non-recurrence. Some attributes are linear, and some are nominal.

Attributes information:

1) Class: Breast cancer patient event is recurrence or non-recurrence.
2) Age: Breast cancer patient age at the time of diagnosis. (Range 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99)
3) Menopause: The patient is pre- or post-menopause at the time of diagnosis. (lt40, ge40, premeno)
4) Tumor size: Brest cancer patient tumor greatest diameter. (Range 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59)
5) Inv Node: Axillary lymph nodes. (Range 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39)
6) Node Caps: Breast cancer contains metastasize to a lymph node. (yes, no)
7) Degree of malignancy: Tumor histological grade (range 1-3).
8) Breast: Breast side. (left, right)
9) Breast quadrant: Breast four-quadrant, including the central part of the breast. (left-up, left-low, right-up, right-low, central)
10) Irradiation: Radiation therapy is given or not to the breast cancer patient. (yes, no)

Task 1.1.2 Data Loading

The breast cancer data set is given in .data format file, and its attributes value information is provided in .names file. First requires import python library for analysis and then after reading the data. The pandas library pd.read_csv method, the data of breast cancer .data file is read. Once data is loaded into the data frame, then after assign the column name into this dataset. The final prepared data frame is saved as a .csv format in the local drive on the Jupiter home directory path. The next step is to load data into panda's data frame to clean and prepare for analysis.

Task 1.2 Data Cleaning and Preparation

Task 1.2.1 Checking Null Values

The missing values are checking using panda method isnull().sum() or isnull().values.any(). There are no any missing value in breast cancer dataset.

Task 1.2.2 Checking Duplicate Values

The duplicate values are checking using the pandas' method duplicate (). In this data set, we found some rows are duplicate, and all the attribute values identical. To understand better ways about duplication, we need to sort the data set by columns age, class, and tumor size. To remove the duplicate values will use drop_duplicates () methods.

Task 1.2.3 Checking Missing Values

While analysing all attributes with a unique method to check the value of missing or some typo errors. We found two columns nodeCaps and breastQuad having the value '?' in records, replace this value '?' with 'missing.'

Task 2 - Data Exploration

Task 2.1 Exploring Dataset Attributes

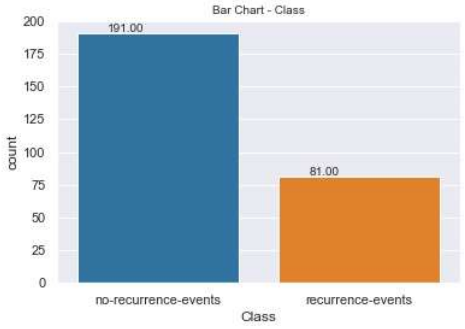
Task 2.1.1 Each Attributes Descriptive Analysis

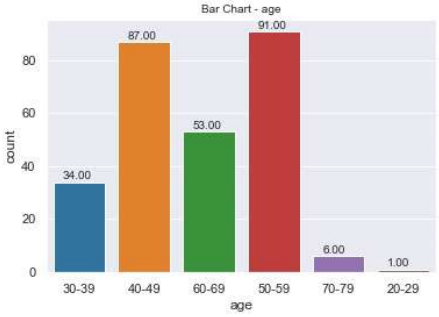
The total instance in the dataset is 272 after the cleaning process. Using pandas describe the method on all object we get the statistics result of the all attributes unique value, topmost value, and frequency of that value in the dataset.

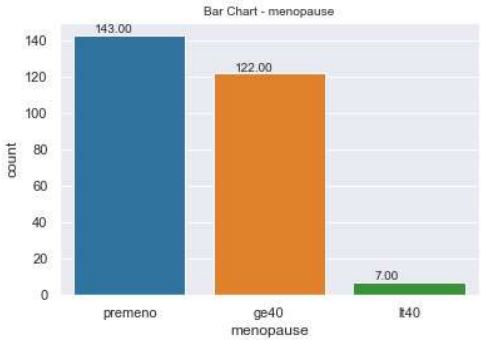
	Class	age	menopause	tumorSize	invNodes	nodeCaps	breast	breastQuad	irradiat
count	272	272	272	272	272	272	272	272	272
unique	2	6	3	11	7	3	2	6	2
top	no-recurrence-events	50-59	premeno	30-34	0-2	no	left	left_low	no
freq	191	91	143	60	200	209	143	103	205

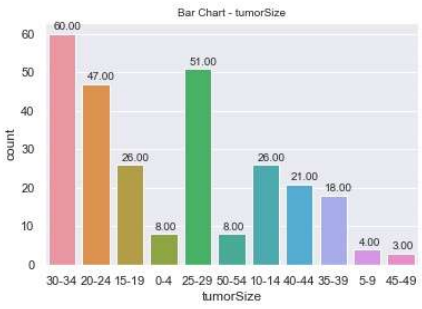
Task 2.1.2 Each Attributes Visualisation

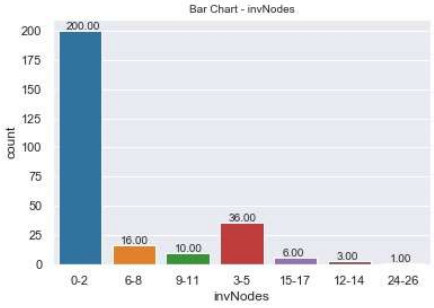
To Visualise the dataset, all attributes using the seaborn count plot because the major attributes are linear and nominal categorical variables with a string value.

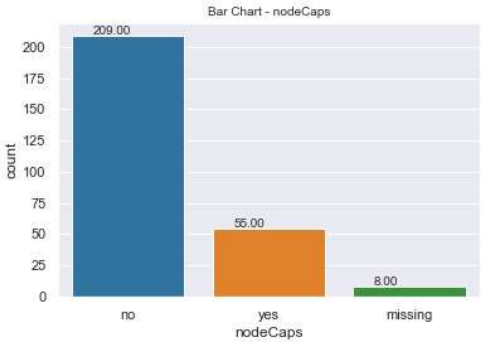


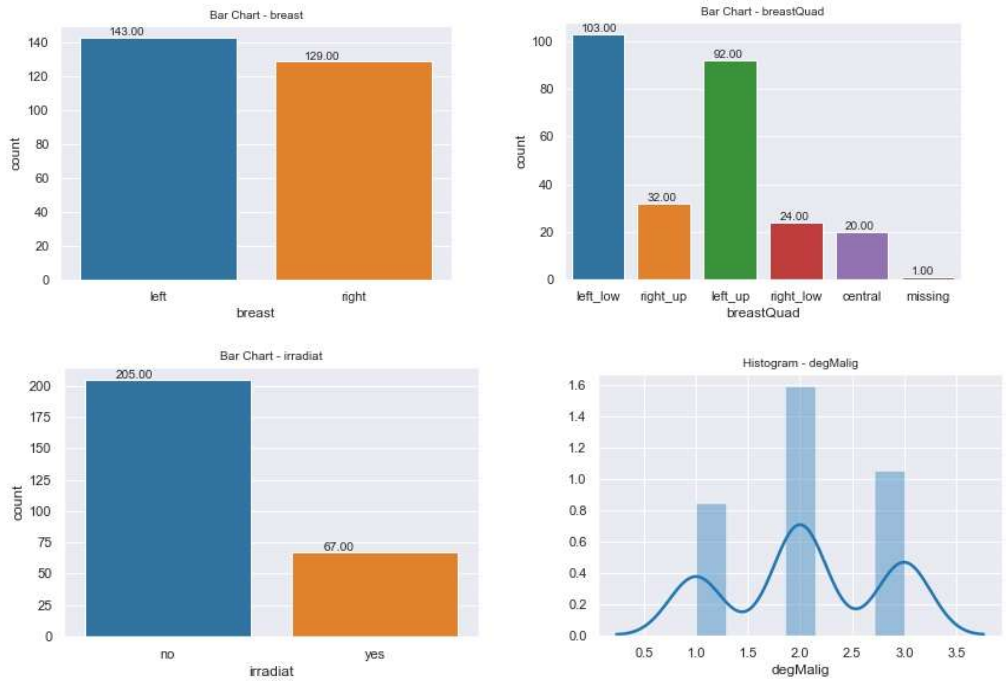












Below are some key observations based on the descriptive table and visualisations graph:

1) The recurrence event is shallow 30% only.
2) The age 50 -59 year is more risk for breast cancer in women
3) Breast cancer happens to women in pre-menopause.
4) Most of the women's breast cancer Tumor diameter range is 30-34.
5) Most of the Tumor is not converted to more aggressive disease.
6) The highest frequency for deg of malignancy is two which shown a higher risk of breast cancer in young 50-59 age of women.
7) Most women have breast cancer on the left side of the breast, and the quadrant is left side lower.
8) Most of the breast cancer women patients did not treat with radiation therapy.

Task 2.2 Relation between all Dataset Attributes

To the analysis of the relation between all attributes and for modelling, we need to convert the categorical string value into numeric value using the function. [4]

I. Age, Tumor Size, and Inv Node: These three columns are containing range value; for the modelling purpose, we convert this range value into its mean value of range.
   Age: '20-29':24.5,'30-39':34.5,'40-49':44.5,'50-59':54.5,'60-69':64.5,'70-79':74.5
   Tumor Size: '0-4':2,'5-9':7,'10-14':12,'15-19':17,'20-24':22,'25-29':27, '30-34':32,'35-39':37,'40-44':42,'45-49':47,'50-54':52
   Inv Node: '0-2':1,'3-5':4,'6-8':7,'9-11':10,'12-14':13,'15-17':16,'24-26':19

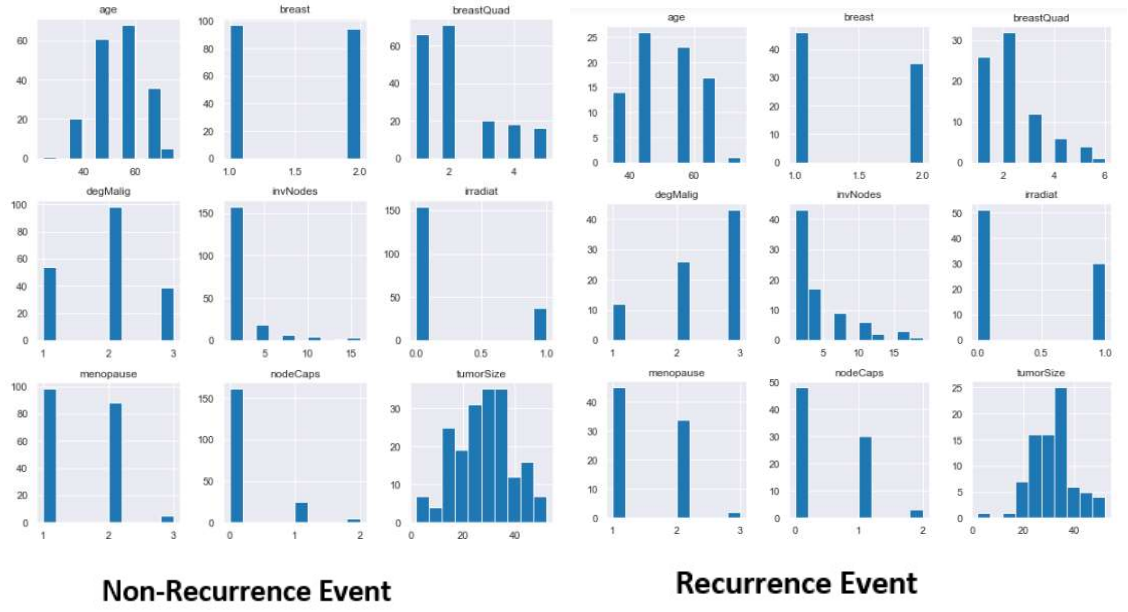
II. Class, Menopause, NodeCaps, Breast, BreastQuad, Irradiate: These columns categorical string value convert into distinct numerical value according to that categorical value.
   Class: 'no-recurrence-events':0 and 'recurrence-events':1
   Manopause: 'premeno':1,'ge40':2,'lt40':3
   NodeCaps: 'no':0,'yes':1,'missing':2
   Breast: 'left':1,'right':2
   Breast Quad: 'left_up': 1, 'left_low':2, 'right_up':3, 'right_low':4, 'central':5, 'missing':6

Irradiat: 'no':0,'yes':1

Also, we are using the outlier technique to identify any irrelevance of record in attributes of the dataset.

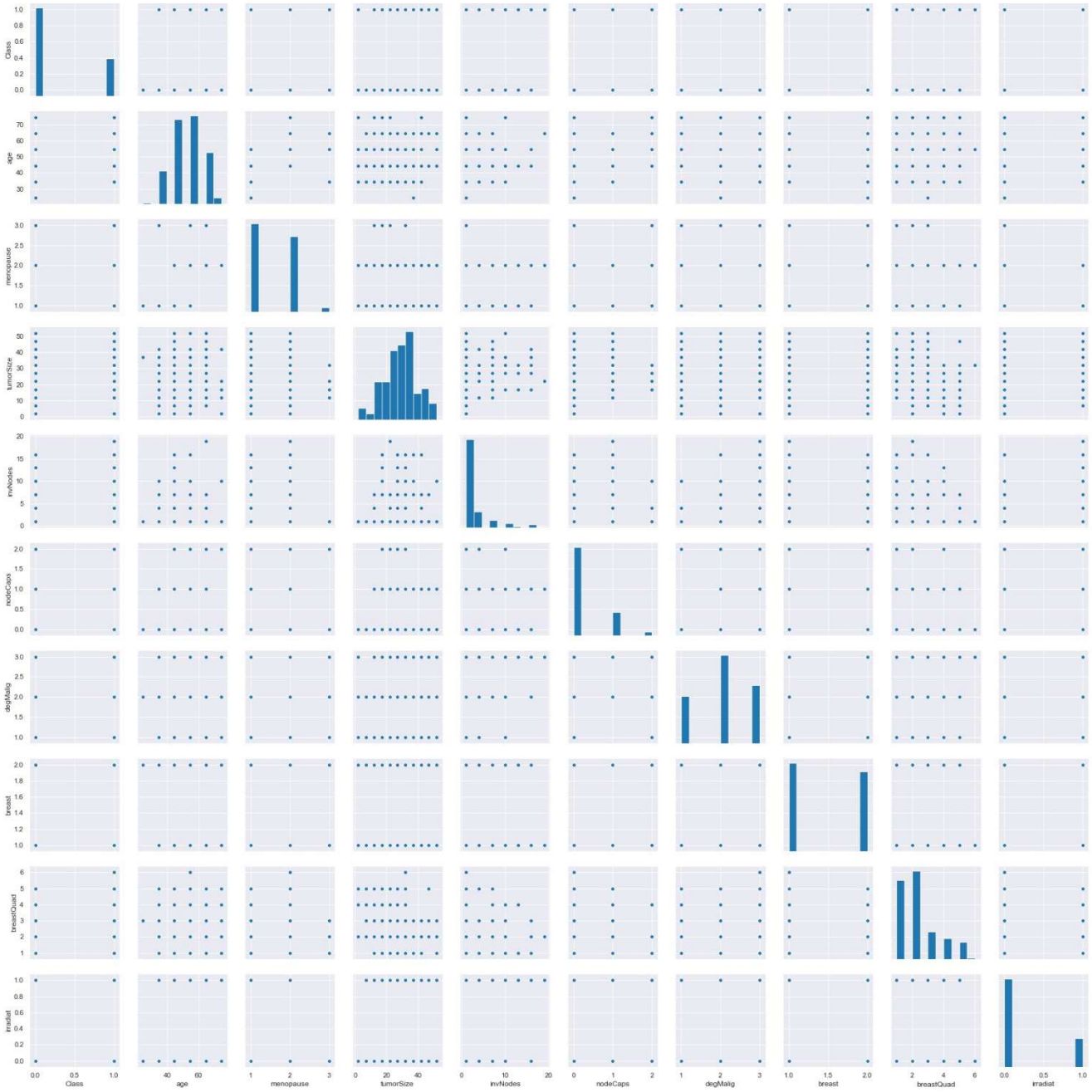
Task 2.2.1 Each attributes visualisation By Class

Each attributes of dataset are visualisation by class of recurrence and non-recurrence event.



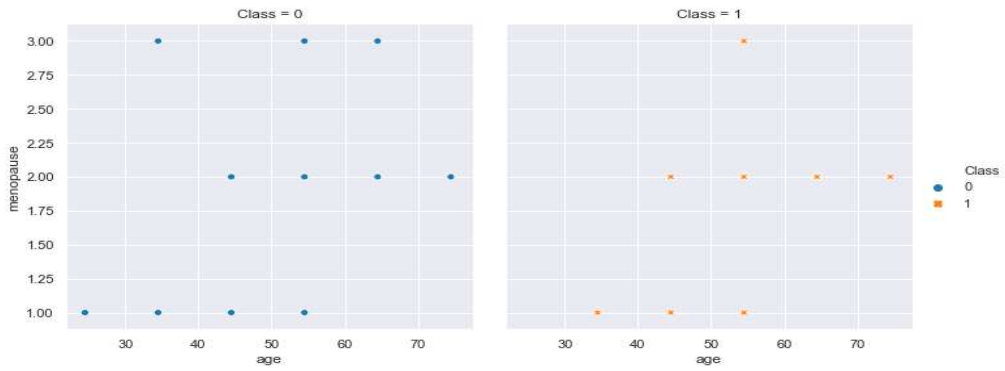
Task 2.2.2 Correlation Matrix

Task 2.2.3 Pair Scatter Plot

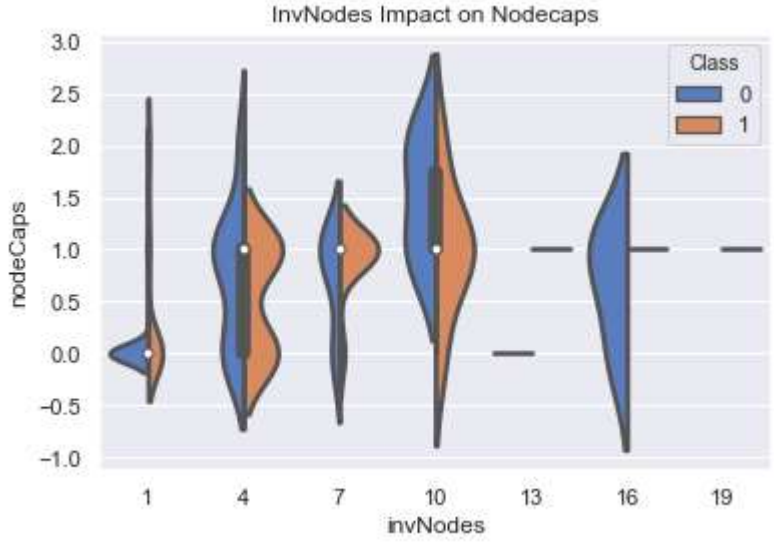


There is some crucial observation for each column relationship based on group by class each attributes visualisation, correlation matrix, and Scatter pair plots.

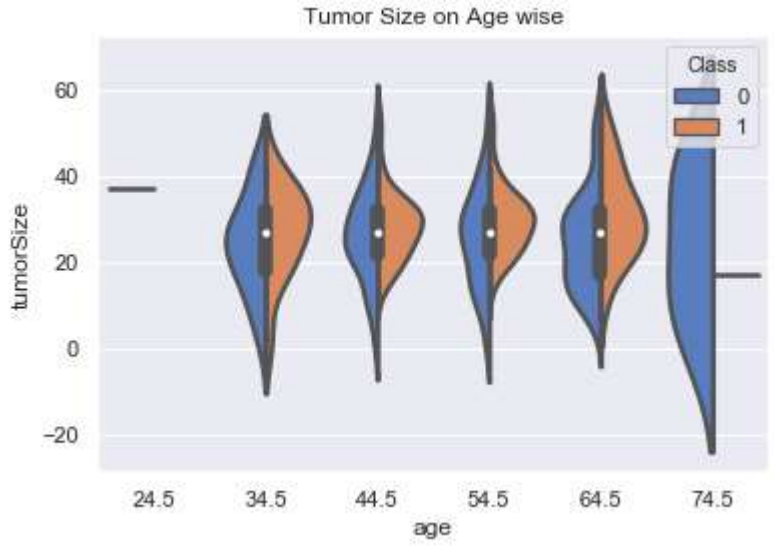
1) As per the Correlation matrix, age and menopause are highly or positively correlated to each other. Their correlation factor is 0.67. The relative plot shows that the age of women is increased, and menopause is going to stop, but the age range 50-59 is a higher risk of breast cancer after preventing menopause or post-menopause recurrence of breast cancer.



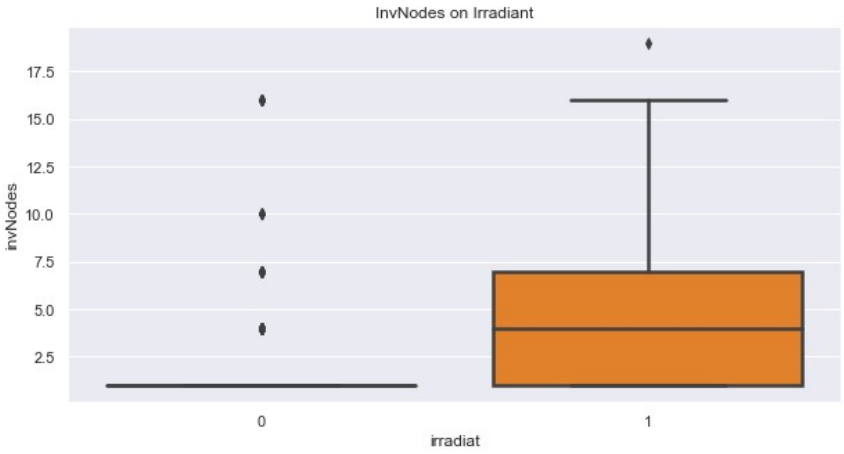
2) As per the Correlation matrix, the second pair is positively correlated is InvNodes and NodeCaps. Their correlation factor is 0.56. The violin plot clearly explains that most women breast cancer is a non-recurrence event with the match of historical examination. Also, over time the invnodes replace tumor into more severe in the disease of node caps and its high risk for women to survive from cancer.



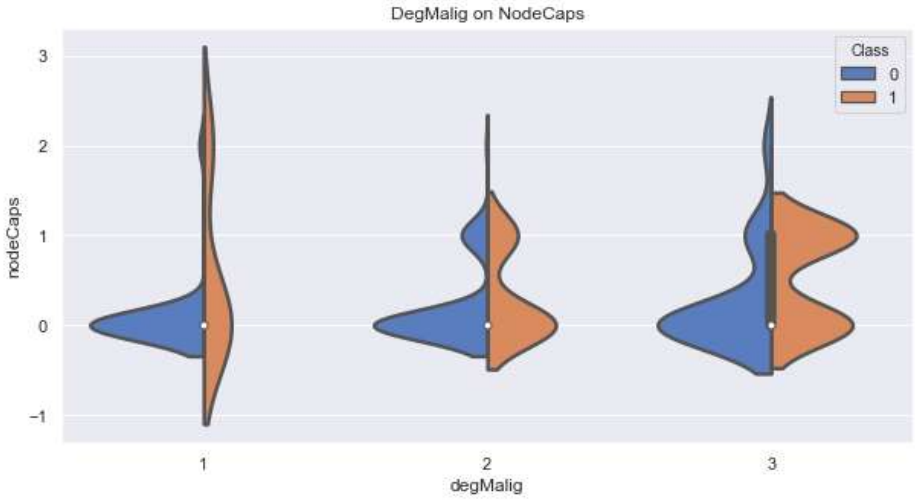
3) The risk of tumor size is high over the age. When age is increased, then the risk of breast cancer tumor size is also increased. For both class recursive and non-recursive event, the risk of tumor size increase in the proportion of age.



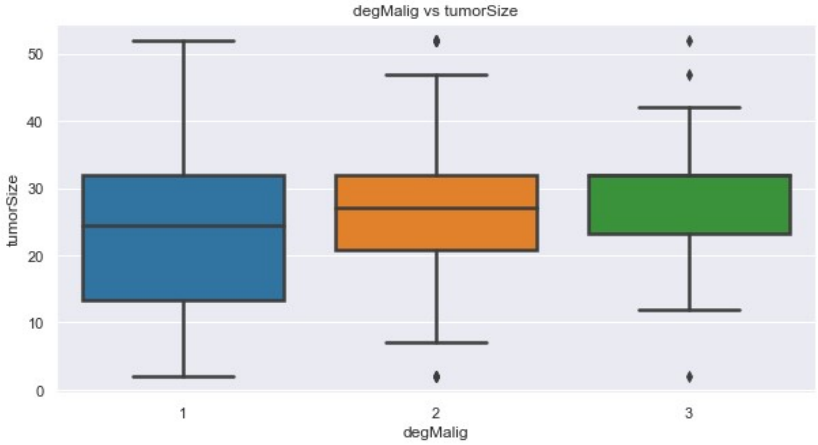
4) As per the below box plot, we can say that the invnodes or lymph node is creating a high risk for recurrence event breast cancer and that time radiation therapy given to that breast cancer affected women. For non-recurrence event inv node is less risk for radiation therapy.



5) As per the violin plot, it is clear that the degree of 3 is predominately cells of highly abnormal node caps for the type of classes. But in the given data set, more than 50% of the instance are degree 2, and that is equally divided of recurrence and non-recurrence breast cancer patients.



6) As per the box plot, it can be inferred that there is no relation between Degree of malignancy and tumor size because, for all types of degree malignancy, tumor size is almost the same. There is no change in tumor size based on the degree of malignancy.



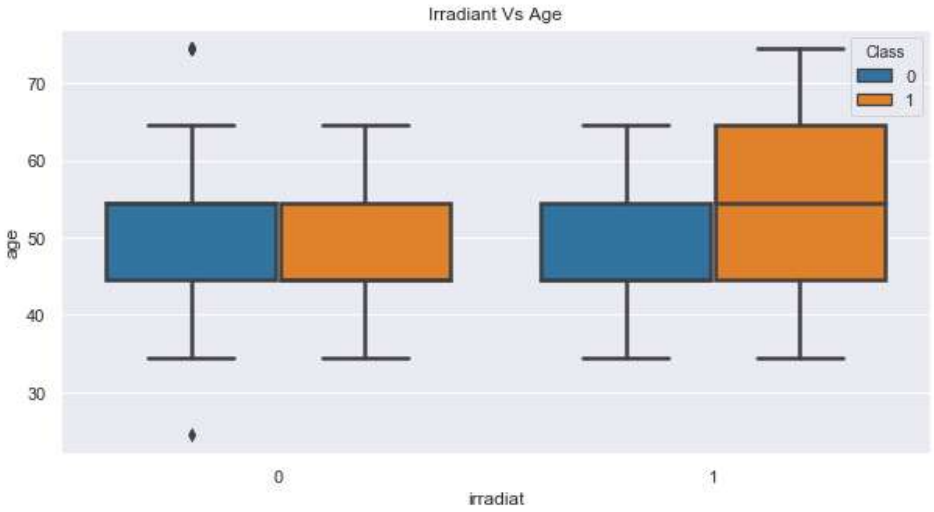
Task 2.3 Classification Question and Answer

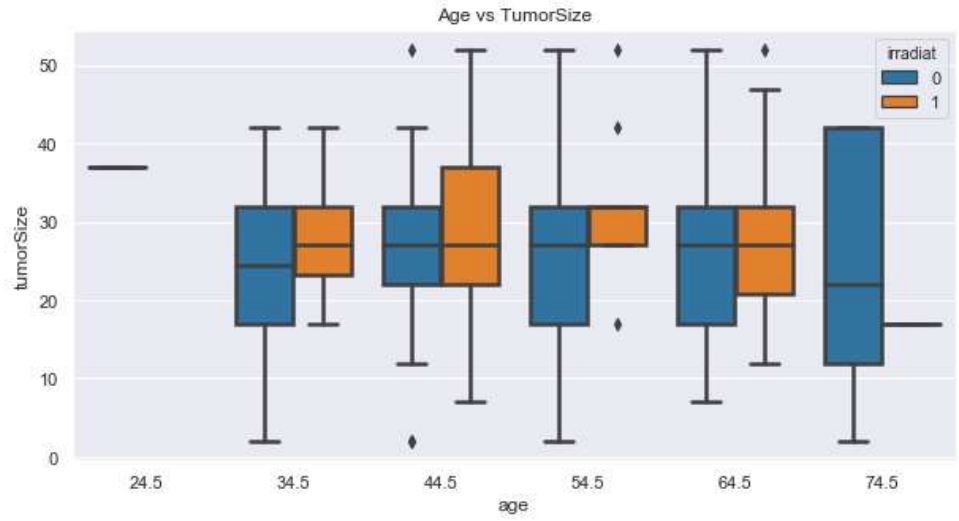
Which age group is a high risk of cancer recurrence with tumor size impacted into severe disease and maybe cure by radiation therapy?

To get the answer to this question, we need to use the four columns of the dataset class, age, tumor size, and irradiate. We need to plot the box plot age vs. radiation therapy over class variable, and the second box plot is age vs. tumor size over the class variable.

As shown in the first box plot, it is clear that the recurrence of breast cancer is increasing, and risky in higher age women need radiation therapy to recover from these diseases. The same way the second plot explains that the risk of tumor size is an increase over the age increase in women, but there is most of the breast cancer women are non-recurrence event.

If women age having more than 50 years, then the risk of breast cancer recurrence is very high, and the tumor size is an increase, which leads to serious condition into women. It may or not be cured by radiation therapy after the age of 75. The women age after 65 is challenging to recover from breast cancer, and if any women recover, then the chance of breast cancer recurrence is very high.





Task 3 – Data Modelling

Early detection of breast cancer is more useful to cure the women of disease rather than treatment after women seek and cancer symptoms are seen again. To predict the recurrence of breast cancer, we are using the class attributes as over target variables and apply the machine learning various classification algorithms to get the prediction high accuracy.

For this classification modelling, we are using class attributes as our target and age, menopause, tumor size, invNodes, nodeCaps, degMalig, irradiate attributes as our features. First, we create a feature dataset from this feature's columns for this modelling. To complete this data modelling part, we need to use the Scikit-Learn Library different modules. Also, create one empty data frame name of evaluation to store different classification algorithms results. This data frames contains column 'Train_Test_Suit','Model','Accuracy','F1 score','Recall','Precision'.

Task 3.1 Splitting Dataset Train and Test

Before spitting the dataset into a train and test, we need to create one target variable as a Y recurrence cancer predictor and drop that variable from the feature dataset and create dataset X. The target variable Y needs to encoding due to categorical behaviour.

The dataset is divided into a training dataset and testing dataset; the training data set contains Information of prediction or output, and the model learns from this training data, and after on this model, we used the testing dataset to test our model prediction accuracy.

The dataset is splitting into three suits suit1 (50% train, 50% test), suit2 (60% train, 40% test) and suit3 (80% train, 20% test). This dataset contains high variation age, range, and nodes; for this, we need to bring all the features attributes of the dataset to the same levels, and this can be done by scaling the feature train and test dataset.

Task 3.2 Classification Models

This is a classification problem because the need to classify breast cancer is recurrence or non-recurrence; to achieve this, we will use the Gaussian Naïve Bayes and support vector machine, supervised classification model.

Gaussian Naïve Bayes: The naïve Bayes is a classification algorithm based on Bayes theorem. It is a simple supervised learning algorithm, fast, accurate, reliable, and best for classification problem to predict the result based on probability. It calculates the probability between the target(independent) variable with the feature variable. We are using this classification algorithm because we need to find out early detection of the probability of recurrence of breast cancer in women.

Support Vector Machine: The support vector machine is a supervised learning technique, which used in classification problems. It separates two classes and plots each data in n-dimensional space with the value of each feature being the value of a co-ordinate. We are using this classification algorithm because we create two classes of recurrence and non-recurrence events of breast cancer and separate both classes very well and find the best accuracy of early detection of breast cancer in women.

Below are results of classification for three different splitting suit datasets for this two-classification algorithm:

```
Suit 1 - GNB
Accuracy: 0.7720588235294118
F1 score: 0.754016656829956
Recall: 0.7720588235294118
Precision: 0.7596184419713832
Clasification report:
              precision    recall  f1-score   support

           0       0.79      0.92      0.85        96
           1       0.68      0.42      0.52        40

   micro avg       0.77      0.77      0.77       136
   macro avg       0.74      0.67      0.69       136
weighted avg       0.76      0.77      0.75       136

Confussion matrix:
[[88  8]
 [23 17]]
```

```
Suit 1 - SVC
Accuracy: 0.7058823529411765
F1 score: 0.5841784989858012
Recall: 0.7058823529411765
Precision: 0.49826989619377166
Clasification report:
              precision    recall  f1-score   support

           0       0.71      1.00      0.83        96
           1       0.00      0.00      0.00        40

   micro avg       0.71      0.71      0.71       136
   macro avg       0.35      0.50      0.41       136
weighted avg       0.50      0.71      0.58       136

Confussion matrix:
[[96  0]
 [40  0]]
```

```
Suit 2 - GNB
Accuracy: 0.7431192660550459
F1 score: 0.7351527116478922
Recall: 0.7431192660550459
Precision: 0.7321051192163298
Clasification report:
              precision    recall  f1-score   support

           0       0.79      0.86      0.82        76
           1       0.59      0.48      0.53        33

   micro avg       0.74      0.74      0.74       109
   macro avg       0.69      0.67      0.68       109
weighted avg       0.73      0.74      0.74       109

Confussion matrix:
[[65 11]
 [17 16]]
```

```
Suit 2 - SVC
Accuracy: 0.7155963302752294
F1 score: 0.638598344917857
Recall: 0.7155963302752294
Precision: 0.7027701077758974
Clasification report:
              precision    recall  f1-score   support

           0       0.72      0.97      0.83        76
           1       0.67      0.12      0.21        33

   micro avg       0.72      0.72      0.72       109
   macro avg       0.69      0.55      0.52       109
weighted avg       0.70      0.72      0.64       109

Confussion matrix:
[[74  2]
 [29  4]]
```

```
Suit 3 - GNB
Accuracy: 0.7818181818181819
F1 score: 0.7818181818181819
Recall: 0.7818181818181819
Precision: 0.7818181818181819
Clasification report:
              precision    recall  f1-score   support

           0       0.85      0.85      0.85        41
           1       0.57      0.57      0.57        14

   micro avg       0.78      0.78      0.78        55
   macro avg       0.71      0.71      0.71        55
weighted avg       0.78      0.78      0.78        55

Confussion matrix:
[[35  6]
 [ 6  8]]
```

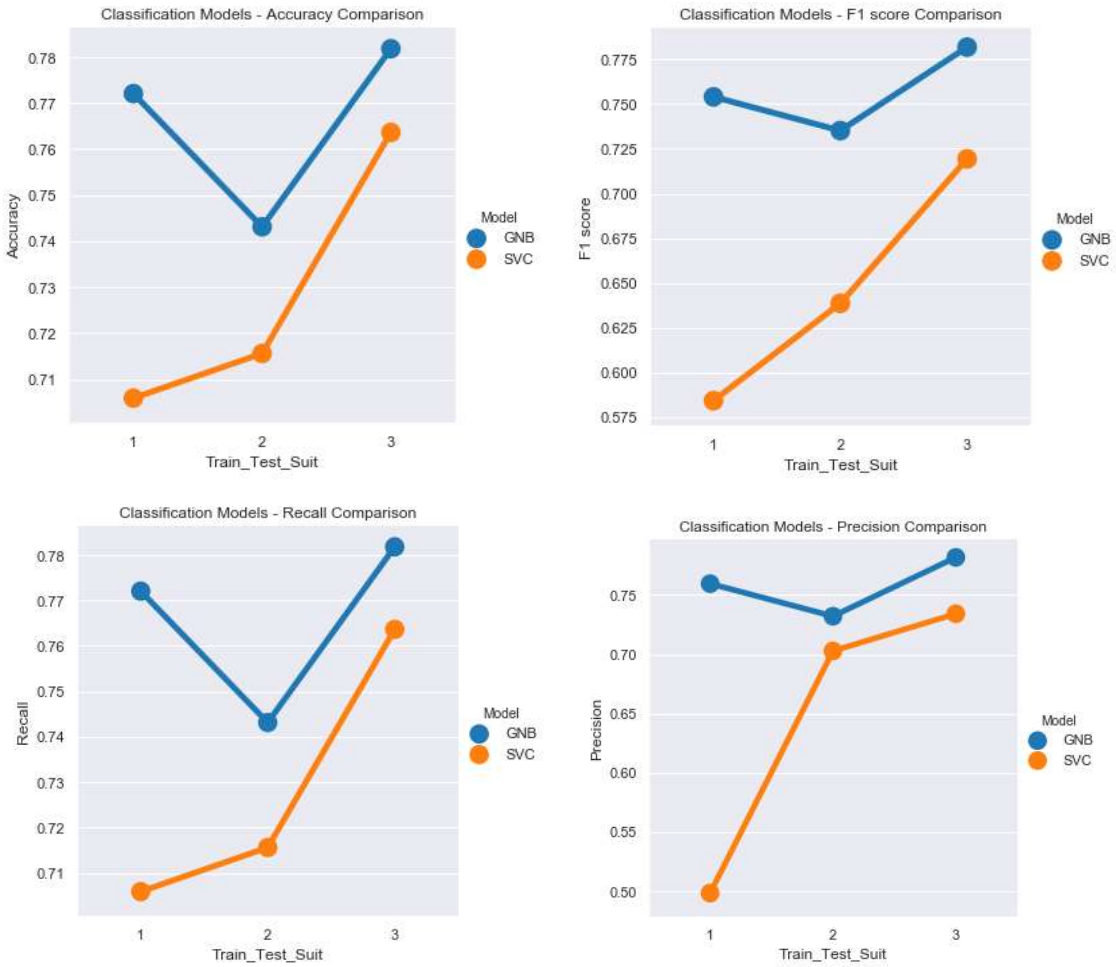
```
Suit 3 - SVC
Accuracy: 0.7636363636363637
F1 score: 0.7193438140806561
Recall: 0.7636363636363637
Precision: 0.7341818181818183
Clasification report:
              precision    recall  f1-score   support

           0       0.78      0.95      0.86        41
           1       0.60      0.21      0.32        14

   micro avg       0.76      0.76      0.76        55
   macro avg       0.69      0.58      0.59        55
weighted avg       0.73      0.76      0.72        55

Confussion matrix:
[[39  2]
 [11  3]]
```

This result is save in one data frame called evalution dataframe.

	Train_Test_Suit	Model	Accuracy	F1 score	Recall	Precision
0	1	GNB	0.772059	0.754017	0.772059	0.759618
1	2	GNB	0.743119	0.735153	0.743119	0.732105
2	3	GNB	0.781818	0.781818	0.781818	0.781818
3	1	SVC	0.705882	0.584178	0.705882	0.498270
4	2	SVC	0.715596	0.638598	0.715596	0.702770
5	3	SVC	0.763636	0.719344	0.763636	0.734182

Task 3.3 Comparison of Classification Models

The comparison of both classification models done using the seaborn line scatters plot by its model type.



Key Observations:

1) The Gaussian Naïve Bayes models having better accuracy, F1 score, precision, and recall compared to the Support Vector Machine model.
2) The Gaussian Naïve Bayes models give the highest accuracy of 78% for early detection of breast cancer in women.
3) The suit 3 splitting 80% train and 20% test data set, giving better accuracy, f1 score, precision, and recall value for both models.
4) The suit 1 splitting 50% train and 50% test data set giving lowest accuracy, f1 score, precision, and recall value for both models compare to the other two suits.

Conclusion

The classification models help out to cancer patients and medical treatment staff for early detection of breast cancer recurrence events. To develop this model, need analytical patient records & it cannot depend on actual breast cancer symptoms. Still, this model can be used to improve patient survival chances from breast cancer more straightforward and effective manner. We have investigated two classification models, and the Gaussian Naïve Bayes is giving high accuracy to predict the recurrence event of breast cancer in women. The maximum efficiency is 78% of this model, and we need more work to be done before giving to the clinical use of this model. This low accuracy is providing some help in the curing process of breast cancer women, but it is also given a false-positive result, which may not be detected. Also, if the data of age, tumor size, and Inv node are given in specific single value instead of range value, then the classification model prediction score is improving.

References

1) https://medium.com/cracking-the-data-science-interview/how-to-think-like-a-data-scientist-in-12-steps-157ea8ad5da8
2) https://www.datacamp.com/community/tutorials/naive-bayes-scikit-learn
3) https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/
4) https://chrisalbon.com/python/data_wrangling/convert_categorical_to_numeric/
5) https://towardsdatascience.com/building-a-simple-machine-learning-model-on-breast-cancer-data-eca4b3b99fa3
6) https://medium.com/@hannah.lgbhan/using-machine-learning-models-for-breast-cancer-detection-f95cf5414764
7) https://www.kaggle.com/kralmachine/seaborn-tutorial-for-beginners
8) https://www.cdc.gov/cancer/breast/basic_info/index.htm
9) https://ww5.komen.org/BreastCancer/ReturnofCancerafterTreatment.html
10) https://www.causeweb.org/usproc/sites/default/files/usclap/2018-1/Predictors_for_Breast_Cancer_Recurrence.pdf
11) http://archive.ics.uci.edu/ml/datasets/Breast+Cancer
12) Zwitter, M. & Soklic, M. Breast Cancer Data. (1988). Institute of Oncology, University Medical Centre Ljubljana, Yugoslavia.