Data Analysis Coursework 1

Breast Cancer Analysis

Chandrima Mukherjee(220990378)
ECS784P
School of Electronic Engineering and Computer Science
Queen Mary University of London Mile
End Road, London E1 4NS, UK
c.mukherjee@se22.qmul.ac.uk

Abstract

The report aims at implementing the Data Analysis techniques to analyse the two cancer types- malignant and benign. In this study, the University of California Breast Cancer Dataset was used to train and evaluate a two of the classification models Support Vector Machine, Linear Discriminant Analysis and K-Nearest Neighbors. Both the methods achieving precision and an F1 score above 90%. According to my analysis, attrition rates were greater for younger workers, those who worked overtime, had lower monthly incomes, and those who had been employed for a shorter amount of time.

1 Introduction and Background

Breast cancer is the second-leading source of cancer death for women and the most prevalent malignancy among them, accounting for almost one in three cancer diagnoses among women in the United States. Breast cancer develops when cells in the breast tissue, also known as a tumour, expand abnormally. The presence of a growth does not necessarily indicate the presence of cancer. Tumours can be benign, pre-malignant, or malignant (cancerous). In order to identify breast cancer, procedures like MRIs, mammograms, ultrasounds, and biopsies are frequently used.

2 Literature Review

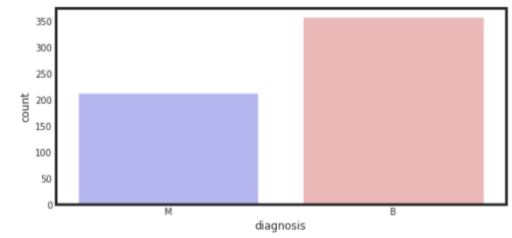
Given the findings of a breast fine-needle aspiration (FNA) test, which uses a tiny needle akin to one used for blood samples to remove some fluid or cells from a breast lesion or cyst (a lump, sore, or swelling). Because of this, I built a model that uses two training classes to categorise breast cancer tumours:
1 indicates that a cancer is present.
0 Means Negative (Not Cancerous)
The prediction falls into two groups because the labels in the data are discrete (i.e. Malignant or benign). This is a categorization issue in terms of artificial intelligence.
Therefore, the objective is to categorise whether a breast cancer is benign or malignant and to forecast the return and non-recurrence of malignant cases over time. To accomplish this, I have fitted two functions that can predict the discrete class of new input using machine learning classification techniques.
References to the papers that inspired this report have been added in the report section.

3 Dataset Review

The University of California, Irvine's machine learning library offers the Breast Cancer datasets. 569 samples of both malignant and benign tumour cells are included in the collection. The first two entries of the dataset contain, respectively, the samples' distinctive ID numbers and the associated diagnosis (M = malignant, B = benign). Columns 3-32 hold 30

real-value features that can be used to create a model to determine whether a tumour is benign or malignant. These features were computed from digital images of the cell nuclei.
A sample of a row is given below:

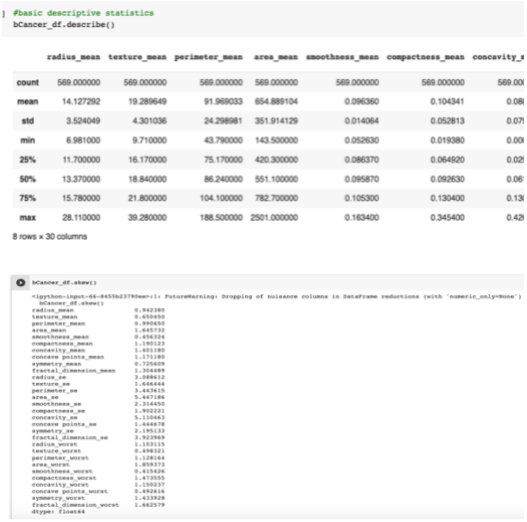
```
diagnosis                    M
radius_mean                20.57
texture_mean               17.77
perimeter_mean             132.9
area_mean                 1326.0
smoothness_mean          0.08474
compactness_mean         0.07864
concavity_mean            0.0869
concave points_mean       0.07017
symmetry_mean             0.1812
fractal_dimension_mean    0.05667
radius_se                  0.5435
texture_se                 0.7339
perimeter_se                3.398
area_se                     74.08
smoothness_se            0.005225
compactness_se            0.01308
concavity_se               0.0186
concave points_se          0.0134
symmetry_se               0.01389
fractal_dimension_se     0.003532
radius_worst                24.99
texture_worst               23.41
perimeter_worst             158.8
area_worst                 1956.0
smoothness_worst           0.1238
compactness_worst          0.1866
concavity_worst            0.2416
concave points_worst        0.186
symmetry_worst              0.275
fractal_dimension_worst    0.08902
Name: 1, dtype: object
```



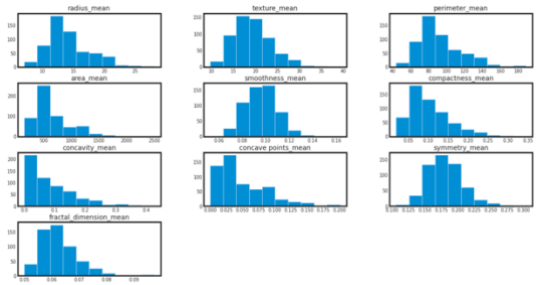
4 Data Exploration

Some initial exploratory data analysis (EDA) has been done to better comprehend data, use summary statistics and visualisations. Two approaches have been used to examine data:

- Descriptive statistics: - The process of distilling important traits from the data collection into straightforward numerical metrics. Metrics like mean, standard deviation, and association are frequently used.



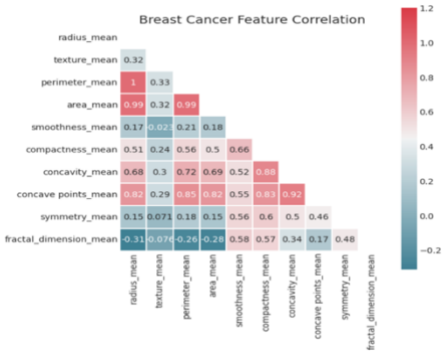
- Visualization:- It involves transforming the data, or portions of it, into abstract pictures or Cartesian space. I have done two types of data visualization.
  - Unimodal Visualization- Identifying the characteristics that are most useful in predicting either malignant or benign cancer is one of the primary objectives of visualising the data in this case. The other is to look for broad patterns that can help us choose models and hyperparameters. I have used histograms to comprehend each attribute of the dataset separately.



I can see that the concavity and concavity point characteristics might have an exponential distribution ( ). Additionally, I can see that the smoothness, texture, and symmetric properties may have

a Gaussian or nearly Gaussian distribution. This is intriguing because many machine learning methods rely on input variables having a Gaussian univariate distribution.

- Multimodal visualisation – I have used two multimodal techniques- Scatter Plot and Correlation Matrix.

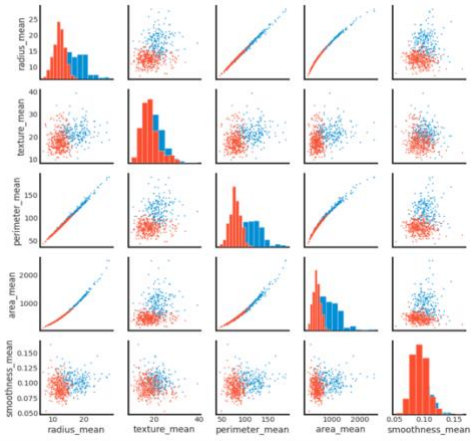


I can see that there is a significant positive correlation between mean values and parameters between 1 and 0.75.

A few parameters have moderately positive correlations (r between 0.5 and 0.75), such as concavity and area, concavity and perimeter, etc.

The mean area of the tissue nucleus has a strong positive association with the mean values of radius and parameter.

Similar to this, I observe a strong negative association between fractal dimension and the mean values of the radius, texture, and parameter.



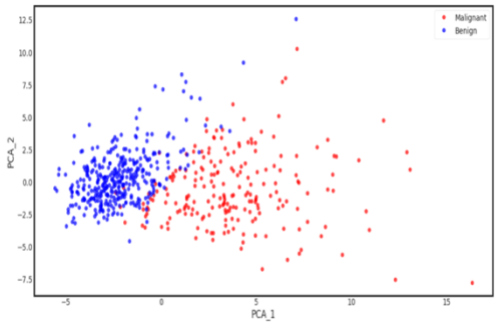
Cancer can be categorised using the average values of the cell's radius, circumference, area, compactness, concavity, and concave points. These parameters' larger levels frequently exhibit a correlation with malignant tumours. The averages for texture, smoothness, symmetry, and fractual dimension do not indicate a clear predilection for one diagnosis over another. There are no observable big outliers that call for additional cleanup in any of the histograms.

5 Data Processing

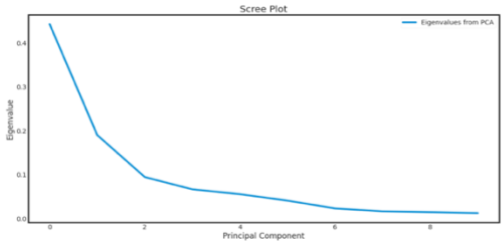
As in the EDA, I explored dataset and gained some insight on how the data is distributed and correlated to each other. Some interesting features have been identified. In the notebook feature selection has been used to reduce high-dimension data, feature-extraction and transformation has been used for dimension reduction.

- The goal was to improve the predictive capability of the analytics model, identify the data's most predictive characteristics and filter them.
- I have used Label encoding to transform class labels of diagnosis column from the original string representation of M and B into integer values.
- I have split the data set into 70:30.
- As seen in the exploratory analysis the raw data has different distributions that could effect our algorithms since most algorithms work well on features of same scale. So I have done standardization of the data using StandardScaler() from sklearn.preprocessing such that every attribute has a mean of 0 and SD(standard deviation) of 1.

- Many feature pairs divide the data beautifully as seen in the exploratory analysis. Thus it makes sense to employ one of the dimensionality reduction techniques to maintain as much information as possible while dealing with only two dimensions, So I have applied PCA(Pricipal component Analysis. After performing the linear PCA transformation, I now have a reduced dimensional subspace(in this instance 3D to 2D), where the samples are more evenly distributed along the new feature axes.



- I have made a scree plot to summarise the findings of PCA in order to determine how many principal components should be kept. At component 2, the “elbow” of the scree plot, there is the most pronounced shift in slope. Therefore it can be claimed that the first three components should be kept based on the framework of the scree plot.



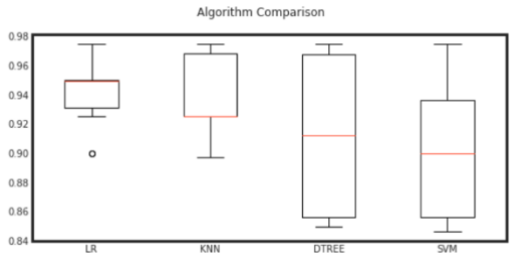
6 Learning Methods

I have used two learning methods – SVM and KNN. The reason I selected these two models is because I did a test on a number of different models like Logistic Regression, Decision Tree, KNN and SVM.

Logistic Regression	0.942244
K-nearest Neighbour	0.937179
Decision Tree	0.912308
Support Vector Machine	0.901987

Table A Exploring Algorithms without Standardization

The findings imply that KNN may merit further investigation. These accuracy numbers are merely averages. Examining the distribution of accuracy values computed across cross validation folds is always a good idea. Box and whisker plots allow us to represent that visually.

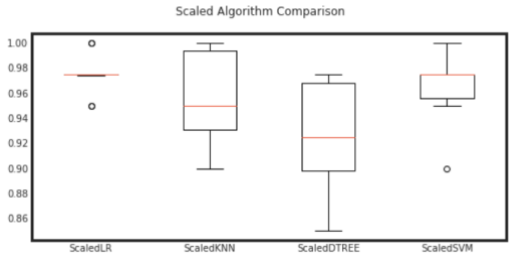


Except for SVM, all classifiers' results exhibit a similar tight distribution, which is encouraging as it indicates minimal variance. SVM's positive outcomes are acceptable.

It's conceivable that the diverse distribution of the attributes will affect how accurate algorithms like SVM are. I will conduct another spot-check using a standard version of the training dataset in the part that follows.

Logistic Regression	0.974936
K-nearest Neighbour	0.957372
Decision Tree	0.924679
Support Vector Machine	0.967436

Table B Exploring Algorithms with Standardization



The results show that data standardisation enhanced the performance of SVM, making it the most accurate technique to date.

The findings suggest further research into the SVM, and KNN algorithms. It is very likely that altering the present settings will result in models that are even more accurate.

- Support Vector Machine Algorithms (SVM) - The most popular classification algorithm. It is effective and works-well in high dimension spaces making it suitable for tasks that involve huge number of features. By using kernel functions to transform the input data into a higher-dimensional space where the data becomes separable, SVM can handle non linearly separable data successfully.
- K-Nearest Neighbour Algorithm(KNN) – KNN is a popular method used for both classification and regression tasks. It is straightforward yet effective. Because of its versatility it can be used in wide range of algorithms. Since KNN is a non-parametric algorithm, it makes no judgements about the distribution of the underlying data. This makes it a sensible option for issues where the data distribution is ambiguous or intricate.

7 Analysis, Testing and Results

Splitting the data into training and test sets is essential to prevent overfitting, as was described in the data pre processing part. This enables the generalisation of actual, unheard-of facts.

- Cross-validation - broadens the scope of this concept. I have define so-called folds to divide the data into comparable sized folds rather than having a single train/test split. All pleats are taken for training with the exception of one, known as heldout sample. After the training is complete I use the heldout sample to evaluate how well my trained model performed. A different fold is then pulled out to serve as the new heldout sample after the holdout sample has been tossed back with the other folds. With the remaining folds, training is done once more, and I have use the holdout sample to assess performance. Until every fold has had a chance to serve as a test or had a chance to serve as a test or holdout sample, this procedure is repeated. The cross-validation error also known as the predicted performance of the classifier, is then just the mean of the error rates calculated on each holdout sample.

Fold 1	0.9315
Fold 2	0.9526
Fold 3	0.9417

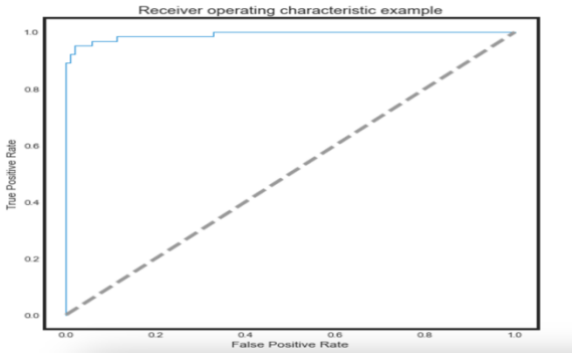
Table C 3-folds average scores

Average Score of the three folds being – 94.20%

- ROC Curve - In a ROC curve, you plot “True Positive Rate” on the Y-axis and “False Positive Rate” on the X-axis, where the values “true positive”, “false negative”, “false positive”, and “true negative” are events (or their probabilities) as described above. The rates are defined according to the following:

    1. True positive rate (or sensitivity)}: tpr = tp / (tp + fn)
    2. False positive rate: fpr = fp / (fp + tn)
    3. True negative rate (or specificity): tnr = tn / (fp + tn)

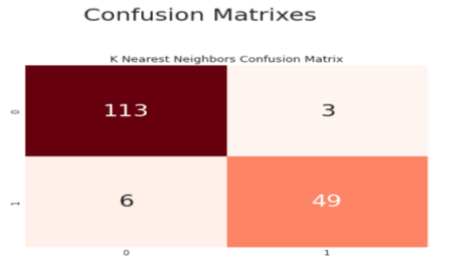
There are two possible predicted classes: "1" and "0". Malignant = 1 (indicates prescence of cancer cells) and Benign = 0 (indicates abscence).



The classifier made a total of 174 predictions (i.e 174 patients were being tested for the presence

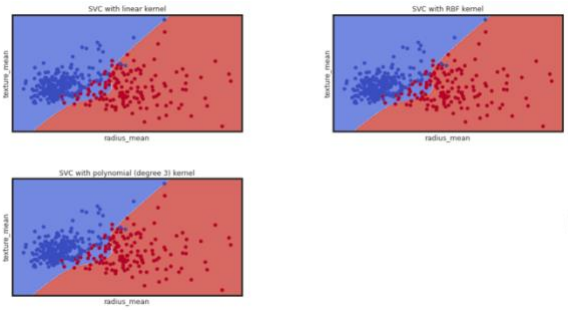
SVM Score is 94.74%

	precision	recall	f1-score	support
B	0.95	0.97	0.96	116
M	0.94	0.89	0.92	55
accuracy			0.95	171
macro avg	0.95	0.93	0.94	171
weighted avg	0.95	0.95	0.95	171



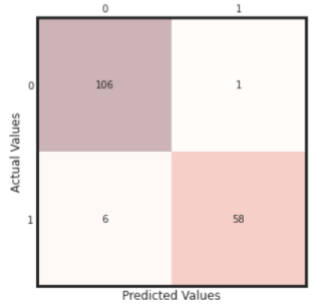
breast cancer). Out of those 174 cases, the classifier predicted "yes" 58 times, and "no" 113 times. In reality, 64 patients in the sample have the disease, and 107 patients do not.

- Optimising the SVM Classifier- In order to be tailored for a specific issue, machine learning models are parameterized. Models may contain a wide range of parameters, making it possible to formulate the best combination of parameters as a search issue. In this section, I have used scikit-learn to fine-tune the SVM Classification model's settings.



The classifier accuracy score is 0.96

	precision	recall	f1-score	support
0	0.95	0.99	0.97	107
1	0.98	0.91	0.94	64
accuracy			0.96	171
macro avg	0.96	0.95	0.96	171
weighted avg	0.96	0.96	0.96	171



The SVM algorithm has two key parameters that can be tuned:

- The value of C (how much the margin should be relaxed)
- Kernel type and size.

SVM (the SVC class) uses the Radial Basis Function (RBF) kernel with a C value of 1. The grid search will be performed using a standardized copy of the training dataset, as with KNN. A number of simple kernel types and C values (less than and more than 1.0) will be tested.

Using Python scikit-learn, you can tune parameters for algorithms in two simple ways:

- Grid Search Parameter Tuning.
- Random Search Parameter Tuning.
- Decision Boundaries for different classifiers – Let's see the decision boundaries produced by the linear, Gaussian and polynomial classifiers.
- Algorithm Tuning - In this part, I looked into fine-tuning the parameters for two algorithms: KNN, and SVM. These algorithms showed promise in the spot-checking in the previous section.

Tuning for SVC

```
→ Model Training Accuracy: 0.940 +/- 0.034
→ Tuned Parameters Best Score:  0.9446794871794871
→ Best Parameters:
 {'clf__C': 1.0, 'clf__kernel': 'linear'}
```

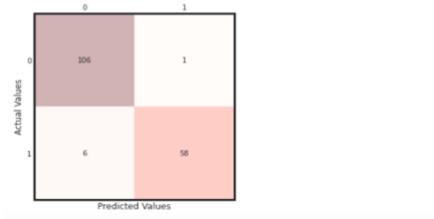
Tuning for KNN

```
→ Model Training Accuracy: 0.915 +/- 0.047
→ Tuned Parameters Best Score:  0.9396153846153847
→ Best Parameters:
 {'clf__n_neighbors': 19}
```

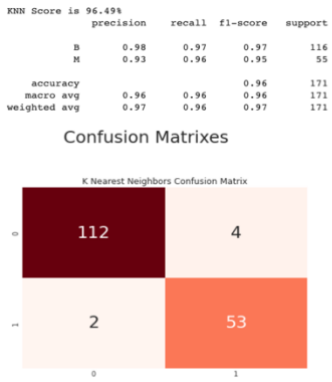
After tuning –
SVC matrix and score –

SVM Score is 97.08%

	precision	recall	f1-score	support
0	0.96	1.00	0.98	107
1	1.00	0.92	0.96	64
accuracy			0.97	171
macro avg	0.98	0.96	0.97	171
weighted avg	0.97	0.97	0.97	171



KNN score and matrix –



8 Concluding Remarks

I used Python to complete a classification predictive modelling issue. The stages covered in detail were:

- Issue Specification (Breast Cancer data).
- the dataset being loaded.
- Analyze data with different distributions but the same size.
    - Review algorithms (KNN looked good).
    - Examine algorithms using standardised criteria (KNN and SVM looked good).
- Algorithm Tuning (K=19 for KNN was fine, SVM with an RBF kernel and C=100 was best)..
- Finish the model (use all training data and confirm using validation dataset)

9 References

- F. S. Ahadi, M. R. Desai, C. Lei, Y. Li and R. Jia, "Feature-Based classification and diagnosis of breast cancer using fuzzy inference system," *2017 IEEE International Conference on Information and Automation (ICIA)*, Macao, China, 2017, pp. 517-522, doi: 10.1109/ICInfA.2017.8078962.
- S. Ghosh, S. Mondal and B. Ghosh, "A comparative study of breast cancer detection based on SVM and MLP BPN classifier," *2014 First International Conference on Automation, Control, Energy and Systems (ACES)*, Adisaptagram, India, 2014, pp. 1-4, doi: 10.1109/ACES.2014.6808002.
- Umesh D R and B. Ramachandra, "Association rule mining based predicting breast cancer recurrence on SEER breast cancer data," *2015 International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT)*, Mandya, India, 2015, pp. 376-380, doi: 10.1109/ERECT.2015.7499044.
- S. Mallick, R. Dash, R. Dash and R. Rautray, "Breast Cancer Data Analysis using Machine Learning Approaches," *2021 International Conference in Advances in Power, Signal, and Information Technology (APSIT)*, Bhubaneswar, India, 2021, pp. 1-4, doi: 10.1109/APSIT52773.2021.9641294.
- Ş. Niţică, G. Czibula and V. -I. Tomescu, "A comparative study on using unsupervised learning based data analysis techniques for breast cancer detection," *2020 IEEE 14th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, Timisoara, Romania, 2020, pp. 000099-000104, doi: 10.1109/SACI49304.2020.9118