Evaluating the performance of Nearest Neighbour and Probabilistic Classifiers on a Test Dataset

Conor Farrell
College Electronic Engineering
University of Galway
Galway, Ireland
c.farrell42@universityofgalway.ie

***Abstract*—This study compares K-Nearest Neighbours with probabilistic classifiers on the Wisconsin Breast Cancer Dataset. Evaluating KNN with different metrics and weight schemes against Naïve Bayes, Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). Testing with increasing dimensions form 10 top 30 features. KNN showed the best performance in most cases with LDA being marginally better in the full 30 feature case.**

I. Introduction

The selection of a machine learning algorithm for medical purposes must be done carefully to follow medical laws and to achieve high accuracy. The analysis uses the Wisconsin Breast Cancer Dataset, with tumour data from 569 patients. The main aims of the investigation:

- Evaluate KNN using different distance and weighting schemes.
- To compare KNN against probabilistic classifiers – Naïve Bayes, LDA, QDA.
- To determine the most reliable classification approach based on the data obtained.

II. Background

A. The Dataset

The Wisconsin Breast Cancer dataset (WBCD), provided by the University of Wisconsin is a dataset of 569 patients and 30 characteristics of the tumour, as well as the classification of the tumour either Benign or Malignant. The dataset has 3 sets of data: mean, standard error and maximum. For each set there are 10 attributes: radius, texture, perimeter, area, smoothness, compactness, concavity, symmetry and fractal dimensions. The dataset is complete, missing no values or rows in the dataset.

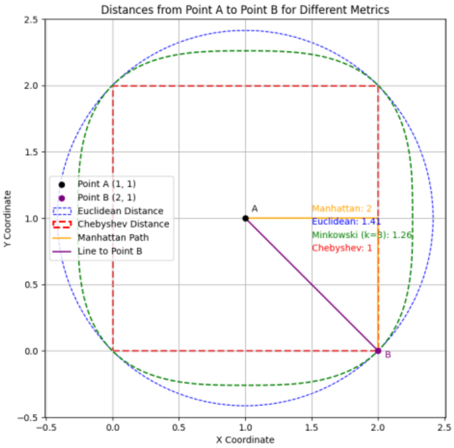
The dataset will need to be normalised due to the requirements of some of the classifiers. Some of the features 1e-3 (fractal_dimension2) up to 1e+3 (area3).

B. K-Nearest Neighbours

K-Nearest Neighbours (KNN) is a classifier which decides based on similarity. The classifier takes in a test point and finds the k closest neighbours from the training dataset. The distance is calculated using a specific distance metric. Then depending on the weighting, the majority target is chosen.

1) Distance Metrics

KNN can use different metrics to for determining the closest Neighbours, these can be Manhattan, Euclidean, Minkowski and Chebyshev. Where Minkowski is the general equation which can represent the other 3 for order 1, 2 and ∞ respectively. For the rest of the paper the Minkowski distance will refer to the k=3 case. Below is a diagram showing the differences between each case.



2) Weighting

KNN can have two different weighting metrics, uniform weighting and distance weighting. In uniform weighting, all points have an equal influence on the test data. In distance weighting, a point's influence is inversely proportional to the distance from the test data.

C. Probabilistic Classifiers

Probabilistic classifiers, make a prediction of the result based on the probability of an outcome with the inputs features. Probabilistic Classifiers can also return the likelihood of one result over another. This can be beneficial to show where more testing is required.

1) Naïve Bayes

Naïve Bayes is one of the simplest probabilistic classifiers, it works on the principle of Bayes' theorem:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Naïve Bayes assumes that each feature is independent of the other features, this is not true for the dataset being used, e.g. compactness is measured as perimeter^2/area. However, the simplification can still reveal interesting data.

2) Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) assumes that each class forms a Gaussian distribution and that each class has its own distribution which can be linearly separated. This is usually done by maximizing the distance between classes and minimizing the variance.

$$P(\,features \mid class\,) \sim N(\mu class, \Sigma)$$

LDA assumes that all classes share the same covariance matrix, and that data is normally distributed. This can be a limitation if either case is not true

3) Quadratic Discriminant Analysis

Quadratic Discriminant Analysis (QDA) operates similarly to LDA, with each class following a Gaussian distribution however each class has its own covariance matrix. This allows for the boundaries between classes to be more complex and improve performance. QDA is prone to overfitting for high-dimensional and small datasets.

III. Methodology

The Methodology will be broken down into 4 sections, Data Cleaning, KNN, Probabilistic Classifiers and Comparisons between the two.

A. *Data Cleaning*

The first step was to understand the dataset. The dataset is of 569 patients, with 30 characteristics for each and a target value of either benign or malignant. The next step was to check for missing data in the set.

The features had to be normalised due to the different orders of magnitude that each dataset scaled, with fractal_dimension2 being in the order of 1e-3 and area3 in the order of 1e+3. Different normalisation functions were compared, min max scaler, Z-score scaling, Max abs scaling, robust scaling and normalised scaling. Z-score scaling was chosen as it keeps all features to a range around 0 and preserves outliers which is useful in a medical context.

The data set was split 70/30 between the training set and test set with the proportion of malignant and benign targets within 10% of each other in both sets.

B. *K Nearest Neighbours*

The methodology for investigating KNN involved testing several parameters: number of neighbours, distance metric, weight scheme. Then the best configuration of parameters was used for the comparison between the other datasets. The best recall was used to minimize the number of false negatives. This was because a false negative would classify a malignant tumour as benign, which could mean vital cancer treatment is not given.

C. *Probabilistic Classifiers*

The different Probabilistic classifiers were Naïve Bayes, LDA and QDA. With the same test split for each classifier and a recall score calculated based on the test data the values were graphed.

D. *Comparison*

Once the best Nearest Neighbour value is determined, the recall is compared between the different classifiers, the decision boundaries (compressed from 10–30-dimensional space to 2-dimensional space). The confusion matrices are also determined. Additionally, an area under curve value is calculated, which measures the overall ability of a classifier to distinguish between classes.

E. *Different datasets*

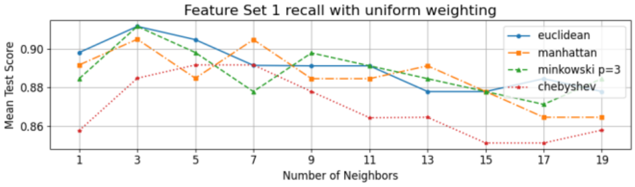
The dataset is divided into 3 main sections, mean, standard deviation and maximum. In testing an evaluation was done one only the mean values, on mean and standard deviation and on the complete dataset.

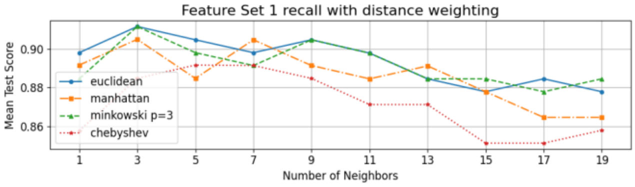
IV. Results

A. *Feature Set 1*

Feature Set 1 is made up of 10 dimensional vectors for each point in the dataset. Each value in the vector represents the mean of some measurement in the tumour.

The below graph shows how the mean test score varies for number of neighbours, distance metric and distance weighting.

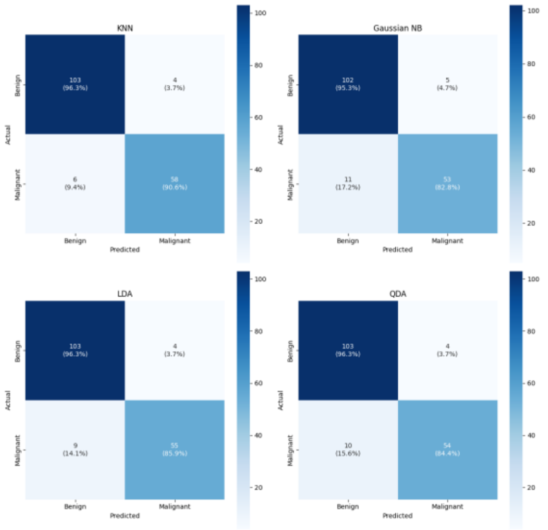




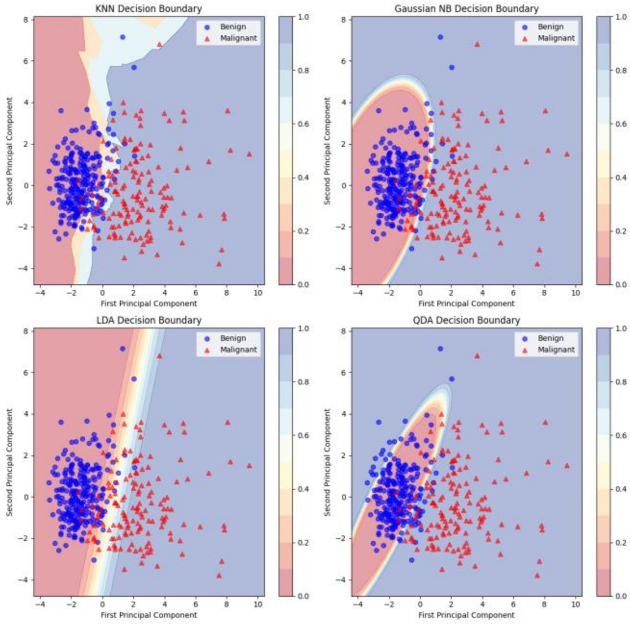
The above shows that changing the distance metric is a much more significant change then moving from uniform weighting to distance weighting

1) *Comparison with Probabilistic*

The confusion matrix for the different classifiers below with the KNN having the lowest number of false negatives.



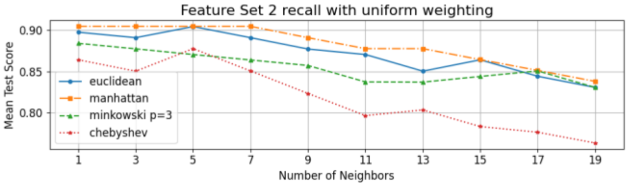
The boundary maps for each function are important to understand how a decision is reached and key positions where potential for error could occur. The graph shows how there may be issues with LDA as the boundary between the 2 regions isn't strictly linear and that some of the Malignant tumours are virtually indistinguishable from the benign tumours.

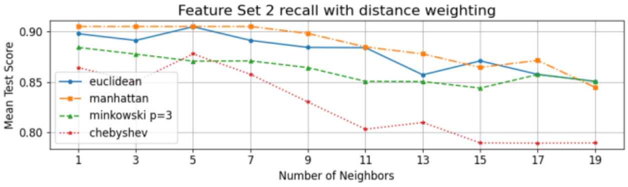


B. *Feature Set 2*

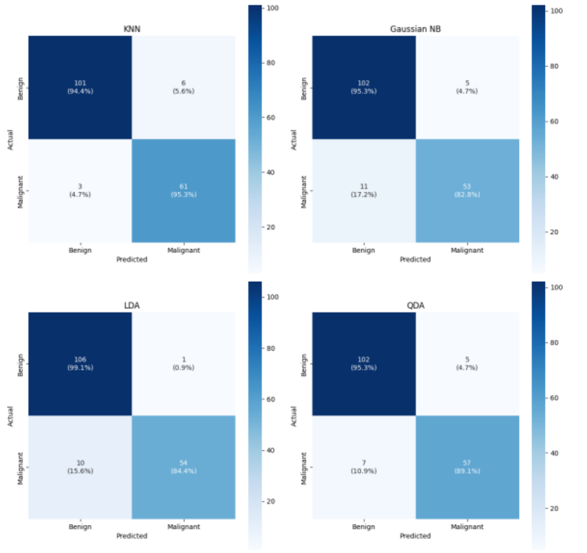
This set is made up of 20-dimensional vectors, with both means and standard deviations.

The data for feature set below will show results that are of a similar accuracy as feature set 1.

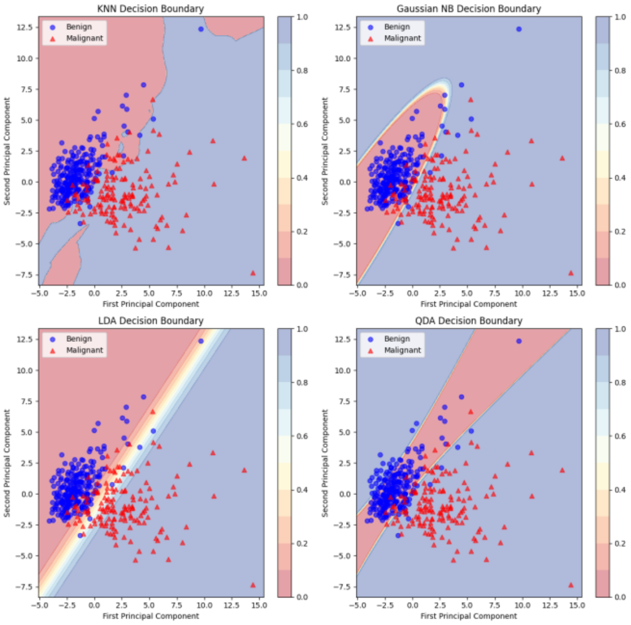




The exact best parameters for the K nearest neighbour changes but the best recall value remains the same.



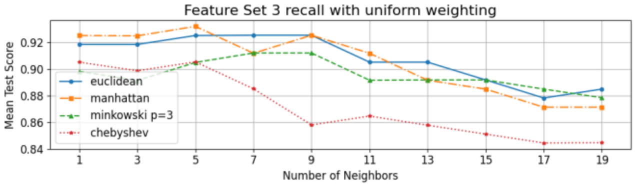
The 4 confusion matrices are virtually identical to the 4 confusion matrices found in Feature Set 1.

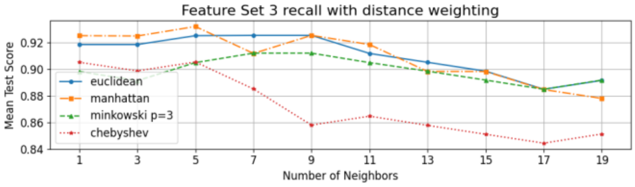


The above decision boundary graph implies that the model is significantly worse than the accuracy and recall values suggest, this is most likely a simplification due to the feature reduction from 20-dimensions to 2-dimensions, QDA is the worst example this with the confusion matrix stating only 13 total points were mislabelled but the decision boundary graph showing at least double that.

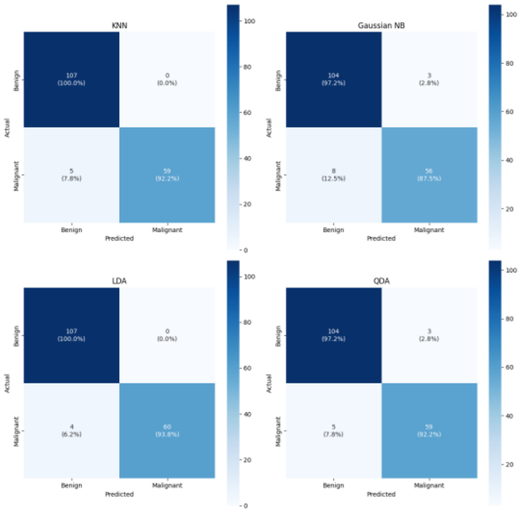
C. *Feature Set 3*

This is the full data set, a 30-dimensional vector for each entry. This is made up of means, standard deviations and maximum values of each measurement.

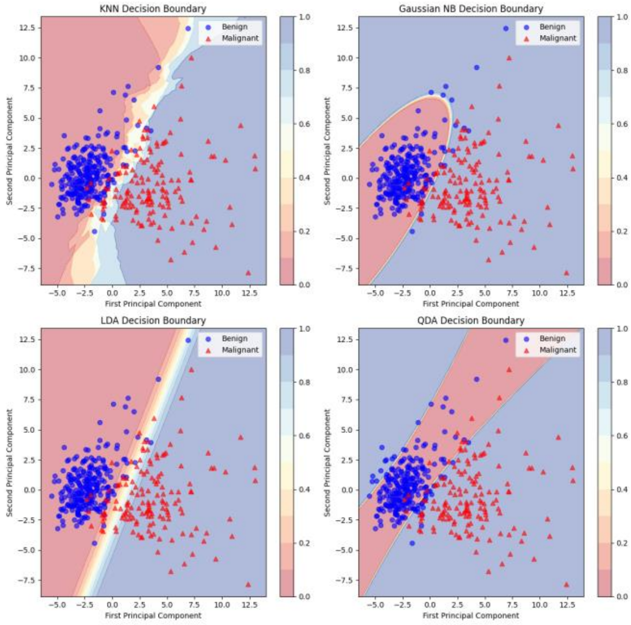




The KNN graph shows a marginal improvement on the previous 2 feature sets.



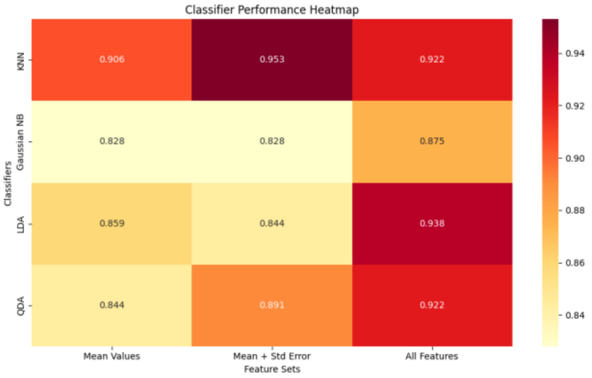
The confusion matrix is very similar to the previous two.



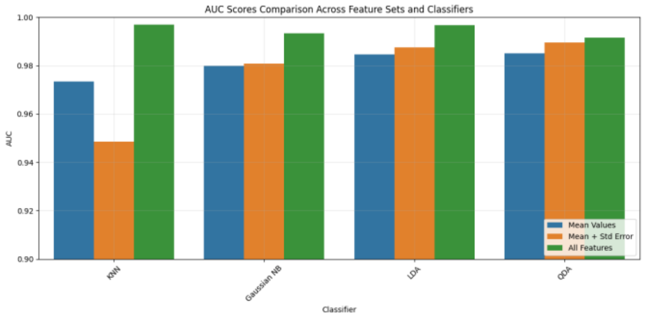
The decision boundaries are like the ones in feature set 2. Not entirely accurate due to the reduction in dimensions.

D. *Overall Performance:*

The performance of all the classifiers across the different sets is the most important as that chooses which model should picked for actual diagnoses and tests.



The above shows clearly that K-nearest neighbours is the best overall classifier. The only time a classifier did better is with LDA in the full feature set.



The above compares the AUC, the higher the value indicates a better classifier performance. The results of the above graph seem at odds with the classifier heatmap. With the best classifier for recall being the worst in AUC.

V. CONCLUSIONS

The report has detailed a thorough analysis of classifications of tumours for breast cancer. The main result is the overall high performance of KNN compared to the probabilistic classifiers. The improvements between the restricted dataset and the full dataset were minimal at best suggesting that mean values are sufficient for classification purposes.

When comparing between AUC and recall the results showed that LDA had the best discriminative ability. Where KNN minimises false negatives. In a medical context the recall is likely more important due to the risk of misdiagnosis. Another reason the KNN might struggle with the AUC is that the parameters were optimised for recall rather than all around classification like with accuracy or f1.

KNN was not the greatest classification method in the case of the full dataset with LDA having the highest recall. This result was still lower than the absolute best result of KNN in the middle feature set of mean and standard deviation.

The analysis also showed that for this application the difference between weighting scheme had minimal if any impact on the result and quality of the classifier. However, the metric for determining the distance between the points could have a significant impact on the accuracy of the classifier. Manhattan and Euclidean were typically the best for any given scenario and Chebyshev was significantly worse in almost all cases especially as the number of neighbours was increased.

Further research can determine if deep learning methods could improve on the results found in the above classifiers, however that is outside of the scope of this assignment.