Comparison Of Random Forest and k-Nearest Neighbors For Breast Cancer Diagnosis

By Uthman Bello Maigari

CITY UNIVERSITY OF LONDON EST 1894

Motivation and Description of Data

My motivation for this study is. In the healthcare sector, knowledge is a critical component of problem solving and decision making. Due to the growing amount of data being generated every day, it has become increasingly crucial for healthcare institutions to efficiently manage and process their expertise in order to improve their services, achieve operational excellence and boost decision-making.

Using a diagnostic data set for breast cancer wisconsin with numeric variables, I have attempted to determine the likelihood that a growth is malignant or benign. Since the dataset is unique, high levels of accuracy and low incidences of false negatives are the most desirable results. Building two models for a classification task to use supervised learning approaches which are K nearest neighbour(KNN) and Random Forest (RF).

Exploratory Analysis

1) Dataset: Breast Cancer Wisconsin from Kaggle

2) This dataset contains 569 occurrences in the collection, and there are 32 numbers of attributes and it has no missing values.

3) The target class exhibited a slight mismatch, with a 62% to 37% split between malignant and benign tumors. Data must be split according to training and testing in order to account for this.

4) A predominant right-hand skew was evident in both classes of data in initial column summaries and histograms, demonstrating a natural correlation between the variables. In three columns, I also observed outliers that departed hugely from the mean, negatively affecting K-NN performance if left unchanged.

5) Keeping a fair balance between all these factors will depend on a feature selection approach. Feature and hyperparameter selection are critical due to the high variance between decision tree results.

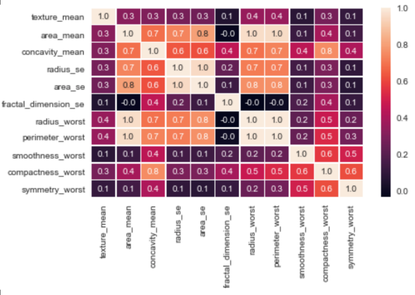

Figure 1: the correlation matrix of the best predictor variables

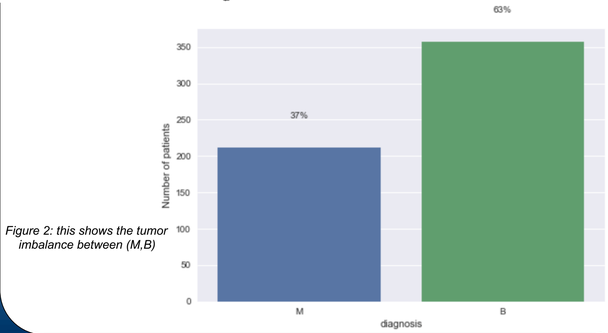

Figure 2: this shows the tumor imbalance between (M,B)

Model Comparison

RANDOM FOREST

1) Random forests, as the name suggests, consist of a large number of decision trees working together. A random forest produces class predictions from each tree, with the class with the most votes becoming the model's outcome
2) In addition to classification and regression, Random Forest is also a supervised learning method.
3) The Random Forest Algorithm makes it possible to deal with both continuous and categorical variables, so it can handle both regression and classification data sets. Classification problems are better solved with it.

Advantages
- Decision trees are less likely to overfit and more accurate when they are used.
- In the data, it automates the process of finding missing values.
- Using a rule-based approach, it does not require normalising data.

Disadvantages
- In order to combine the output of several trees, it requires a great deal of computational power and resources.
- Due to the combination of decision trees, it also requires a lot of time for training.

K Nearest Neighbour

1) Both regression and classification can be accomplished using K-nearest neighbors (KNN). Calculating the distance between the test data and all the training points allows KNN to predict the correct class for the test data.
2) According to the KNN algorithm, similar things exist near each other. As a result, similar things are close together.
3) It consists of identifying our nearest annotated data point, also called the nearest neighbor, in order to classify a given data point.

Advantages
- With K-NN, classification and regression problems can be solved simultaneously.
- K-NN is an algorithm that can be easily understood and implemented. Using K-NN algorithm, each new point is classified by its nearest neighbor K.

Disadvantages
- In spite of the ease of implementation of K-NN, the algorithm loses efficiency or speed very quickly as datasets grow.
- While the KNN algorithm performs well with few input variables, it struggles when the number of input variables increases.
- The K-NN algorithm always chooses the neighbors based on a distance criterion, so it is highly sensitive to outliers.

Hypothesis Statement and Methodology

Hypothesis

- Initially, random forest should outperform k-NN, but when scaling the features, k-NN should outperform random forest.
- A bayes optimization is expected in my case for hyperparameter tuning.
- K-NN is at significant risk of the "curse of dimensionality" because the data is high dimensional with low outputs. The distance between instances is calculated based on all attributes, as opposed to random forest where the distance is calculated based on a similarity metric.

Methodology

- To account for the imbalance in the classes, I normalised the data and performed CV sampling. To analyse this dataset, I have chosen to split it 70/30 and perform cross validation 10 times, which is in accordance with many approaches.
- To better reflect the real world application of the data (i.e. the risk that the disease may not be diagnosed), I have chosen predictive Bayes optimisation for my hyperparameter tuning. I will then compare the models on the basis of accuracy, F1 score and misclassified tumours. In Bayesian optimization, hyperparameters are predicted using the prior and posterior of a function and their uncertainty.

Experimental Results

1) Following a comparison of random forest and FSCMRMR prediction methods, 11 columns (23,20,2,29,11,28,25,27,4,8, and 14) were selected as the most significant predictors (over a specified threshold). Based on the analysis of the selections, the FSCMRMR were less correlated, so I chose to use them for our feature reduction run.
2) Bayes optimization with the full dataset showed KKN to be the most consistently reliable, with an F1 score of 96 percent. The emphasis has been on F1 scores since the balance between precision and recall provides a better representation of performance in the context of the domain. The performance of random forests at 95 percent was slightly worse.
3) KNN performed best with a feature-reduced dataset, outperforming random forest, which was again underperformed by Bayes optimization. In my F1 score, random forest performed under expectations (all figures below), whereas kNN performed near my target of 96 percent.
4) A large deviation was observed in feature selection and learning cycles for random forests. Leaf size remained relatively stable. In the maximum features/predictors sample, one was selected twice for Bayes optimisation (full/feature selected dataset)
5) In terms of hyperparameter tuning for kNN, city block was chosen as the preferred distance twice (once in each dataset), with Euclidean being chosen along with 10 neighbours for the first Bayes optimization run. The feature reduced dataset optimised at 16 neighbours under Bayes.

RF	*Model*	*KNN*
94.70	*Accuracy*	*95.88*
0.94	*Precision*	*0.98*
0.97	*Recall*	*0.94*
0.95	*F1_score*	*0.96*

Figure 3: KNN and Random Forest performance for the full dataset with Bayesian optimization

RF	*Model*	*KNN*
92.35	*Accuracy*	*95.29*
0.91	*Precision*	*0.99*
0.96	*Recall*	*0.93*
0.94	*F1_score*	*0.96*

Figure 4: KNN and Random Forest performance for the featured dataset with Bayesian optimization

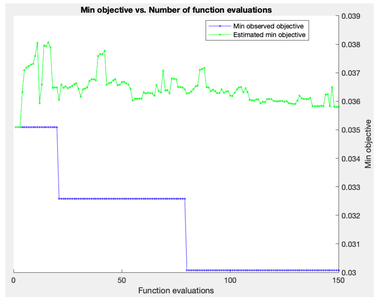

Figure 5: Random Forest feature selection, minimum objective v function for Bayes optimisation

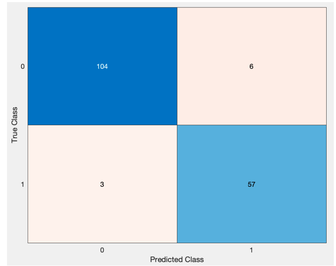

Figure 6: Random Forest Confusion Matrix, full dataset, Bayesian optimisation result

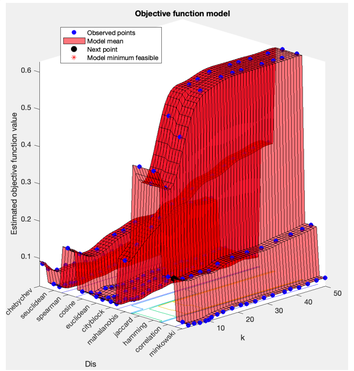

Figure 7: kNN full dataset Bayes optimisation objective function formulation

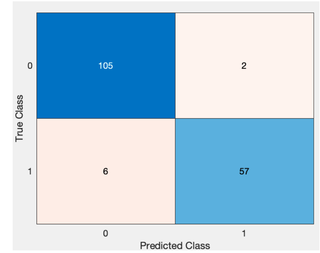

Figure 7: KNN Confusion Matrix, full dataset, Bayesian optimisation result

Future Work

- The most significant lesson learned from this project was that while the dataset initially achieved a high level of accuracy, tangibly improving it requires a highly structured and technical approach.
- Using recursive feature elimination (RFE), in which features are ranked and eliminated either before or after sampling, as well as simply selecting a small number of features would be instructive. For random forest and NKN, automated feature selection suggests 3 and 4 variables respectively.
- We would improve the model if we used data from more years.

Reference

- Dataset link: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data
- Md. Toukir Ahmed, Md. Niaz Imtiaz and Animesh Karmakar (2020). Analysis of Wisconsin Breast Cancer original dataset using machine learning algorithms for breast cancer prediction. https://www.journalbinet.com/uploads/2/1/0/0/21005390/67.02.09.2020_analysis_of_wisconsin_breast_cancer_original_dataset_using_data_mining_and_machine_learning_algorithms_for_breast_cancer_prediction.pdf
- Abien Fred M. Agarap (2019). On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset. https://arxiv.org/pdf/1711.07831.pdf
- Alam, M., Rahman, M. and Rahman, M. (2019). A Random Forest based predictor for medical data classification using feature ranking.
- Sasmitha Kularathe (2020). Prediction of breast cancer using KNN https://medium.com/analytics-vidhya/prediction-and-data-visualization-of-breast-cancer-using-k-nearest-neighbor-knn-classifier-df7adadc4872
- Genesis (2018). Pros and Cons of KNN https://www.fromthegenesis.com/pros-and-cons-of-k-nearest-neighbors/
- Yue, W., Wang, Z.,. and Liu, X. (2018). Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis.