COMP551: Applied Machine Learning
Project 1

GROUP 54: Matthew Morvan, Alexander Tiou Fat, Anne-Sophie Fratzscher

January 2021

Abstract

The objective of this paper is to determine which classification method among K-Nearest Neighbors (KNN) and Decision Trees (DT) is most accurate. We made predictions using two benchmark datasets, one on breast cancer diagnosis and the other concerning hepatitis prognosis. First we discuss the cleaning of the datasets. Then, after implementing and optimizing both methods, we select the most highly predictive features, best cost functions, distance functions and hyper-parameters. Ultimately, we determined that KNN was better at predicting malignant cases of breast cancer, as well as predicting lethal cases of hepatitis, than DT. Additionally, we discuss the effects of imbalanced data on these methods.

1 Introduction

Machine learning (ML) has the potential to revolutionize many industries, especially healthcare [1, 2]. Two of the major applications of ML in healthcare are for diagnosis and for prognosis [3]. Many projects have sought to use ML on health data for diagnosis and prognosis, including for hepatitis [4] and breast cancer [5]. Popular ML techniques include the K-nearest-neighbour model (KNN) and decision tree model (DT). KNN is a non-parametric model that calculates distances between instances and predicts based on the majority label of the K most similar instances in space, whereas DT recursively splits data and predicts by traversing the tree from root to leaf [6]. Both algorithms have pros and cons: KNN is simple to implement but is sensitive to outliers and noisy features, whereas DT are not sensitive to outliers but can easily overfit data and are unstable [6].

The objective of this project was to implement, optimize and determine the most accurate machine learning models to make predictions on (1) breast cancer diagnosis and (2) hepatitis prognosis. The breast cancer dataset included categorical features extracted from breast mass images of benign or malignant tumours. The hepatitis dataset contained categorical and continuous features for patients that either survived or died after contracting hepatitis. In this project, we compared the KNN classifier and DT classifier for both datasets. To optimize accuracy, datasets were cleaned to remove outliers and best features were identified.

We found that the ideal number of neighbours for the KNN on both datasets was 5, giving an accuracy of 96.8% (for both datasets). Tuned DT also performed well for both datasets, with 95.2% accuracy using a tree of depth 5 for breast cancer data, and 88.9% accuracy using a tree of depth 2 for hepatitis data. Additionally, both classifiers performed better on the cancer dataset, likely due to the imbalanced nature of the hepatitis dataset.

2 Datasets

We analyzed two datasets for this project: the breast cancer dataset and the hepatitis dataset. The Diagnostic Wisconsin Breast Cancer Dataset was provided by the Clinical Sciences center of the University of Wisconsin while the Hepatitis Dataset was provided by the Jozef Stefan Institute of Carnegie Mellon University. Both were sourced from the UCI Machine Learning Repository and were downloaded as csv files from myCourses.

2.1 Breast Cancer Dataset

The breast cancer dataset initially had 699 instances with features that were computed from an image of an aspirate of a breast mass. Each instance was either labeled 2 for benign or 4 for malignant in the class column. Additionally, each instance had 10 features: ID, clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitoses. Aside from ID number, all the features where categorical, with a ranking of 1 to 10.

For pre-processing, the ID feature was removed, as the ID did not identify the sample class. We replaced the class values with 0 for benign and 1 for malignant. Additionally, 16 instances contained a missing attribute value, denoted by '?', were removed, leading to 683 instances. Furthermore, as KNN accuracy is negatively affected by outliers, any instance with a z-score above 3 (z-score > 3 is considered an outlier) for at least one feature was removed. This resulted in 632 remaining instances with 9 features in addition to the class label.

Analysis of the data showed that 458 instances were benign and 241 instances were malignant (34.5% malignant). For all the features, the distribution of benign instances was skewed right with most features having a rank below 5, whereas the malignant instance distribution was skewed left. Correlation analysis showed high correlation between uniformity of cell size and uniformity of cell shape. Mitoses had low correlation with all other features, whereas clump thickness had moderate correlation with all other features.

2.2 Hepatitis Dataset

The hepatitis dataset initially had 155 instances of patients infected with hepatitis. Each instance was either labeled 1 if the patient survived or 2 if the patient died. Additionally, each instance had 19 features: age, sex, given steroid, had fatigue, had malaise, had anorexia, had big liver, had firm liver, had palpable spleen, had spiders, had ascites, had varices, bilirubin level, alkaline phosphate level, SGOT level, albumin level, protime level, and had histology.

For pre-processing, we replaced the class values with 0 for survived and 1 for died. Additionally, 75 instances containing a missing attribute value were removed. Similarly to the breast cancer dataset, outliers were removed, resulting in 75 remaining instances with 19 features in addition to the class label.

Analysis of the data showed that 65 of the patients survived, whereas 10 died (13% died). Positive vs. negative distribution for categorical features and for continuous features varied greatly. Additionally, unlike for the breast cancer data, there was not one feature that had a correlation trend with all others. However, there was a strong correlation between ascites and albumin.

2.3 Ethical Concerns

There are ethical concerns when using ML, especially when working with biomedical data. As noted by Gerke et. al [7], not only are there the issues of informed consent to use and data privacy, but also algorithms can exhibit gender and racial biases. Additionally, there is the question of safety, as incorrect predictions in healthcare can in the best case delay diagnosis and in the worst case lead to unnecessary deaths.

3 Results

We implemented, optimized and compared two classification techniques (1) the K-Nearest-Neighbor model and (2) the Decision Tree model. We implemented the models based on examples from tutorial and collaborated through Google Colab. For both methods, datasets were split into training and test sets, where the test set was used to estimate the performance in all of the experiments after training the model with the training set. The final performance was evaluated using accuracy.

3.1 KNN Results

Before identifying an ideal K number of neighbors to compute distances on, it was important to identify the features that yielded the best results for each dataset, as well as the best distance function. For both datasets, we tried different combinations of highly correlated features and tested the models using K values ranging from 0 to 20 using both the Manhattan and Euclidean distance functions. The accuracy was then averaged for the runs and is shown below:

Table 1: Accuracy for Combinations of Features for Dataset 1

Features Tested	Euclidean	Manhattan
A	96	96.5
B	97	97.6
C	97.4	98

A = Testing over Cell Size and Cell Shape
B = Testing over Cell Size, Cell Shape, Bare Nuclei and Bland Chromatin
C = Testing over all features

Table 2: Accuracy for Combinations of Features for Dataset 2

Features Tested	Euclidean	Manhattan
A	96.1	83.2
B	97.3	90.1
C	97.2	83.7

A = Testing over Albumin and Ascites
B = Testing over Albumin, Ascites, Protime, Varices
C = Testing over all features

We see that using the Manhattan distance function on all features is more accurate for breast tumor diagnosis. However, the Euclidean distance function on albumin, ascites, protime and varices is better for predicting the hepatitis prognosis.

We then concentrated on finding the optimal K value. In order to increase certitude and determine our error rate, we performed 10-fold cross validation on both datasets using 50%, 80% and 90% of the data for training. Error rate was calculated as the accuracy subtracted from 1. For the breast cancer data, on average we obtained error rates of 3.5%, 3.2% and 3.2%, respectively, with K=5 for the 80-20 split having the best accuracy of 96.8% (3.2% error). For the hepatitis data, on average we obtained error rates of 3.3%, 5% and 3.2% for the splits respectively, with K=5 for the 90-10 split having the best accuracy of 96.8% (3.2% error). The error rate for different K for the datasets can be seen below:

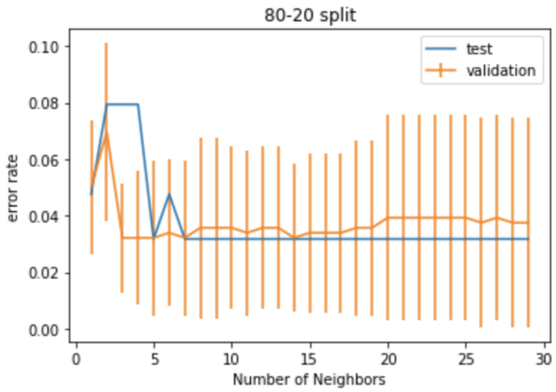


Figure 1: Breast cancer diagnosis error (80-20 split)

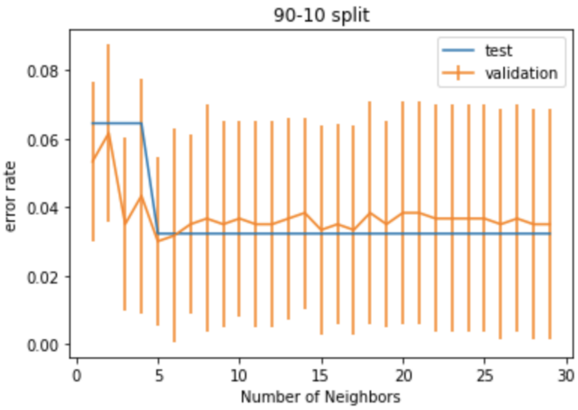


Figure 2: Hepatitis prognosis error (90-10 split)

Finally, we plotted the decision boundaries for both datasets. Two dimensions were picked for visualization. For the breast cancer dataset, we plotted the decision boundary for cell shape vs cell size. For hepatitis, we plotted the decision boundary for bilirubin vs albumin.

The decision boundary for breast cancer seems to indicate that tumors with cell size less than 5 and shape less than 5 tend to be malignant. The decision boundary for hepatitis was sparser, suggesting that patients with bilirubin levels between 1.0 and 2.0 and albumin levels below 3.5 are more likely to die.

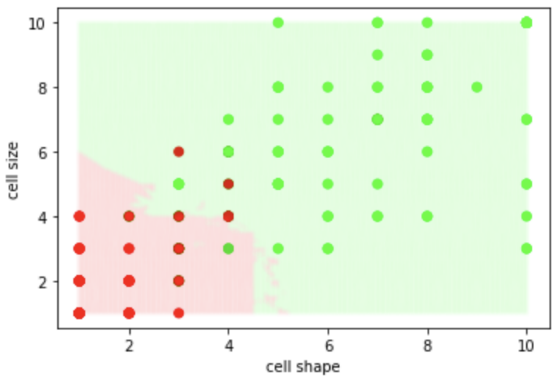


Figure 3: KNN decision boundary for breast cancer

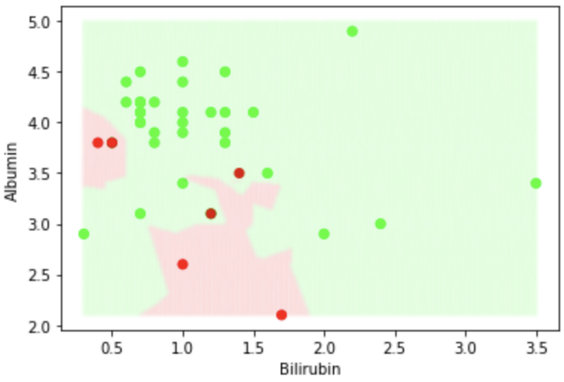
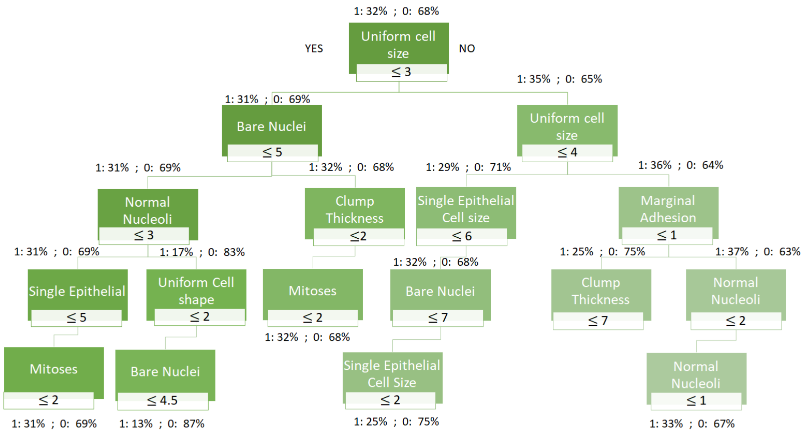


Figure 4: KNN decision boundary for hepatitis

3.2 Decision Tree Results

To build the best Decision Trees for both datasets, we decided to focus on finding the ideal depth and cost function (misclassification, entropy, Gini index cost) for each tree. To tune these hyper-parameters, we performed a grid search with a 5-fold cross validation for the breast cancer dataset and a 10-fold cross validation for the hepatitis dataset.

For the breast cancer dataset, we deduced from our 5-fold cross validation that a depth of 5 using the Gini index cost function was the best hyper-parameters to build the decision tree, giving us an accuracy of 95.2%. The decision tree of depth 5 using the Gini index cost can be seen in Figure 5. For the hepatitis dataset, a decision tree of depth 2 using the entropy cost function gave us the lowest error rate based on our 10-fold cross validation. The test accuracy was 88.9%. The decision tree of depth 2 using the entropy cost function can be seen in Figure 6.



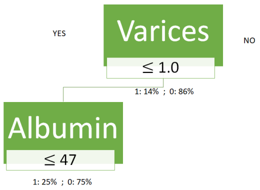


Figure 5: Best decision tree for breast cancer data

Figure 6: Best decision tree for hepatitis data

To better visualise and understand our classifications, we plotted the decision boundaries for both decision trees. Cell size and cell shape were the features used to plot the decision boundaries of the breast cancer dataset, while bilirubin and albumin were used for the hepatitis dataset. The decision boundaries can be seen below:

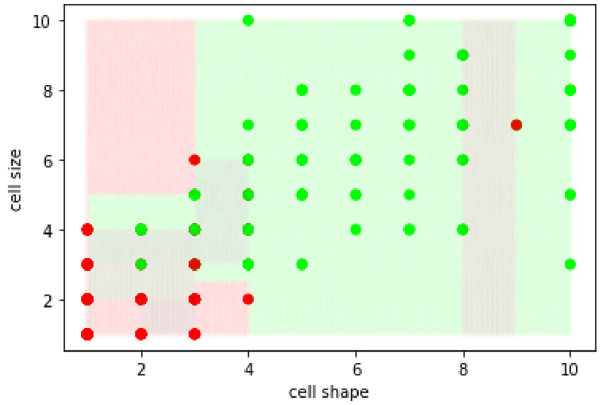


Figure 7: DT decision boundary for breast cancer

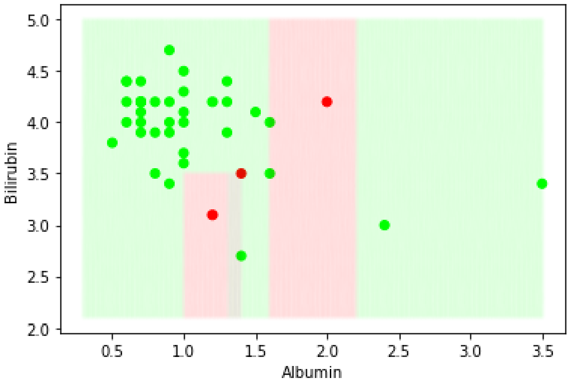


Figure 8: DT decision boundary for hepatitis

The decision boundary for breast cancer shows that a cell shape under 4 tends to be malignant, while others tend to be benign. For hepatitis, there are two distinct boundaries that suggest that patient have higher risk of death: one is when patients have albumin levels between 1.0 and 1.5 and bilirubin levels below 3.5, and the other is when a patient has an albumin level between 1.6 and 2.1.

4 Discussion and Conclusion

Our objective was to find the ideal machine learning model and ideal hyper-parameters for predicting breast cancer diagnosis and for hepatitis prognosis. For both datasets, we saw how hyper-parameter tuning can greatly affect model accuracy. Additionally, training to test data ratio has an affect on accuracy, with an uneven split leading to better accuracy. This is likely due in part to the small size of our datasets, meaning that more data is needed during training to get accurate predictions. For the breast cancer dataset, accuracy was above 90% for both tuned models, with KNN outperforming DT. For the hepatitis dataset, the DT performed worse, with 88.9% accuracy, compared to KNN, which had 96.8% accuracy. However, KNN required more tuning than DT. KNN is sensitive to outliers and noise, so not only was it necessary to find optimal K and best distance function but also to find the best features. On the other hand, DT is much less sensitive to outliers and noise, so only ideal depth and cost function had to be optimized. However, as DT is based on a greedy splitting algorithm, DT was not very stable and can easily overfit data. Therefore, we would select our tuned KNN for predicting both breast cancer diagnosis and hepatitis prognosis.

Although our models were able to predict with greater than 90% accuracy, further optimization is necessary, as 90% accuracy in the medical field is not acceptable. Other distance functions, such as the Hamming distance function, could be tested for KNN due to the categorical nature of many features in both datasets. Additionally, tree pruning could be implemented to determine the most accurate tree depth. Most importantly, the imbalanced nature of the hepatitis dataset (13% death to 87% survival) likely impacts prediction. Techniques to overcome imbalanced data, such as the Synthetic Minority Oversampling Technique (SMOTE) [8] or Tomek links [9], may further improve accuracy. Other machine learning algorithms or ensemble learning methods may also further improve prediction [10].

5 Statement of Contributions

Matthew implemented KNN and tested the best distance functions, best K, and best features. Anne-Sophie pre-processed and ran statistics on the hepatitis dataset and implemented decision boundaries and cross validation. Alexander pre-processed the breast cancer dataset, implemented decision tree, and tested to find the best cost function and depth. Everyone contributed to writing the report.

References

[1] Ahuja A.S. The impact of artificial intelligence in medicine on the future role of the physician. *PeerJ*, 7:e7702, 2019.

[2] Song S Ding X Huang B Ge, Z. Data mining and analytics in the process industry: The role of machine learning. *IEEE Access*, 5:20590–20616, 2017.

[3] Heiser L.M. Gray J.W. Goecks J., Jalili V. How machine learning will transform biomedicine. *Cell*, 181:92–101, 2020.

[4] Kuhn L.A. Punch W.F. Raymer M.L., Doom T.E. Knowledge discovery in medical and biological datasets using a hybrid bayes classifier/evolutionary algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, page 3, 2003.

[5] Mangasarian O.L Wolberg W.H, Street W.N. Machine learning techniques to diagnose breast cancer from fine-needle aspirates. *Cancer Letters*, 77:163–171, 1994.

[6] Channe H.P. Jadhav, S.D. Comparative study of k-nn, naive bayes and decision tree classification techniques. *International Journal of Science and Research (IJSR)*, 5:1842–1845, 2016.

[7] Minssen T. Cohen G. Gerke, S. Ethical and legal challenges of artificial intelligence-driven healthcare. *Artificial Intelligence in Healthcare*, pages 295–336, 2020.

[8] Bowyer K.W. Hall L.O. Kegelmeyer W.P Chawla, N.V. Smote: Synthetic minority over-sampling technique. *Journal Of Artificial Intelligence Research*, 16:321–357, 2002.

[9] Prati R.C. Monard M.C. Batista, G.E. A study of the behavior of several methods for balancing machine learning training data. *Association for Computing Machinery*, 6:20–29, 2004.

[10] Zheng Z. Webb, G.I. Multistrategy ensemble learning: reducing error by combining ensemble learning techniques. *IEEE Transactions on Knowledge and Data Engineering*, 16:980–991, 2004.