MiniProject 1: Getting Started with Machine Learning

COMP 551, Winter 2022, McGill University

Manas Kale
261000867

Gauri Sharma
261026894

Avinash Bhat
260969537

Abstract

A large number of literature is evolving in machine learning that provides with an elaborate description of machine learning techniques that are being used to develop classifiers for disease diagnosis and prognosis. In order to learn more about this ever-evolving literature, through this project, we aimed to investigate the performance of two machine learning models (K-Nearest Neighbours and Decision Trees) on two benchmark datasets, Diabetic Retinopathy Debrecen dataset and Hepatitis dataset. The datasets were divided into training, testing and validation for each algorithm. We performed 3-folds cross validation to choose the best hyperparmaters and found that KNN gave the best overall testing accuracy for both the datasets - 69.57% for the Hepatitis dataset and 62.90% for the diabetes dataset. The decision tree algorithm gave a testing accuracy of 65.22% for the Hepatitis dataset and 62.47% for Diabetes dataset. We also observed a drasitc difference between validation and testing accuracies for the Hepatitis dataset, which implies the generalization error estimate was inaccurate, possibily because of class imbalance and small size of this dataset.

1 Introduction

Machine learning is the field of study that gives computers the ability to learn without explicitly being programmed (Samuel, 1959). It makes predictions and decisions based on the concepts of probability, statistics, calculus without explicitly programmed instructions. The main aim of this project was to implement two classification techniques: K-Nearest Neighbours and Decision Trees and compare the results obtained from these methods. The techniques mentioned above were to be implemented on two datasets: Diabetic Retinopathy Debrecen dataset and Hepatitis dataset. Both of these datasets were obtained from UCI's Machine Learning Repository (Dua and Graff, 2017).

The diabetic retinopathy dataset contains features extracted from the Messidor image set to predict whether an image contains signs of diabetic retinopathy or not (Antal and Hajdu, 2014). The hepatitis dataset was donated to the UCI machine learning repository by Gail Gong of Carnegie-Mellon University and is composed of medical records of patients. The hepatitis dataset contains a lot of of missing values. The existence of missing values in the dataset may affect the quality of the results analysis (Astuti et al., 2015). Therefore, these samples need to be dropped or imputed for handling the missing values.

After implementing the KNN and Decision Trees (DT) algorithm on the given datasets, we learnt that KNN outperformed the DT algorithm in case of the hepatitis dataset. For the diabetes dataset, the testing accuracies were very close for both algorithms. (Oladele T.O. and others, 2019) obtained an accuracy of 61.32% on the diabetes dataset which is similar to ours. The accuracy for KNN increased upto a point with an increase K values for both the datasets. The accuracies of different cost functions were similar mostly in case of decision trees.

2 Datasets

2.1 Hepatitis Dataset

The first dataset was the hepatitis dataset that contained the data of patients infected with hepatatis, including whether they survived or not. The task here was to predict whether a patient will survive or not, given other medical history data. This dataset contained information about 155 patients, out of which 32 died and 123 survived. There were 19 features, where some of them were binary (for instance steriod, antivirals, fatigue) whereas some were continuous (for instance bilirubin, albumin).

2.1.1 Statistical Analysis

We analysed the dataset for different statistics. From the fig 1, we can see we have a class imbalance problem, as class 2 has more datapoints in comparison to class 1. From the heatmap (as shown in fig 2), malaise and anorexia (0.599) followed by fatigue and malaise (0.595) are highly correlated.

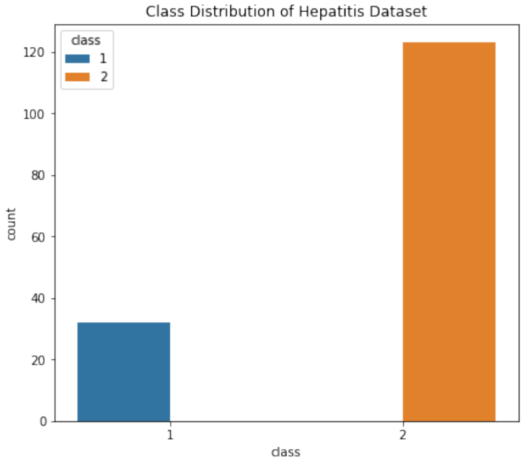


Figure 1: Class Imbalance Plot

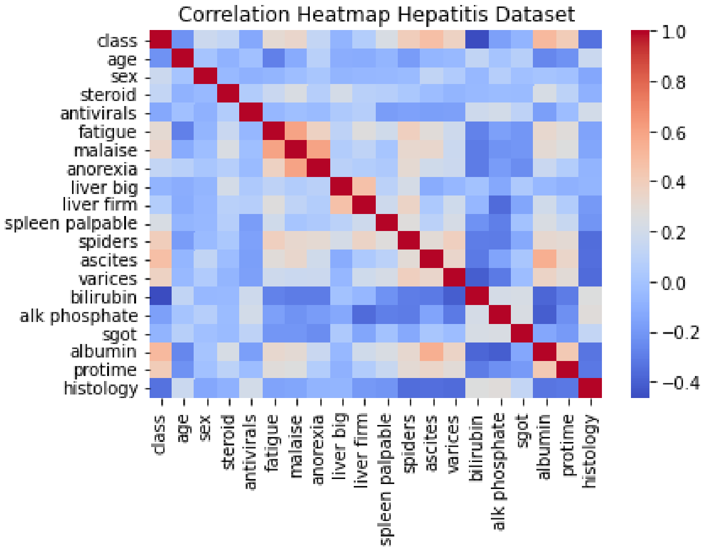


Figure 2: Heatmap for Hepatitis Dataset

2.1.2 Data Processing

We analysed the dataset and learnt that many features had missing values. First we tried augmenting the data by imputing the mean values for numerical features and mode values for the categorical variables. On comparing the original dataset with the augmented dataset, we found that statistially both of them were quite similar and therefore decided to use the original dataset. We also found that the ***protime*** feature had the most missing values and also does not correlate positively because of which we decided to drop the entire feature. We further dropped all the rows that had a missing value. This left us with 112 patient records.

2.1.3 Training, validation and testing sets

Due to the very small amount of patient records present in this dataset, we decided to use a ***80%:20% train:test split*** (i.e., 89 train and 23 test). The training set was further split into training and validation sets using our implementation of L-folds cross validation. Again due to the small size of the dataset, we used 3 folds cross-validation which resulted in a fold size of 29 records.

2.2 Diabetic Retinopathy Debrecen Dataset

The second dataset has substantially more records than the first. This dataset contains features extracted from the Messidor (Methods to Evaluate Segmentation and Indexing Techniques in the field of Retinal Ophthalmology) image set to predict whether an image contains signs of Diabetic Retinopathy (DR) or not. This is a type of diabetes that affects eyes. All features represent either a detected lesion, a descriptive feature of a anatomical part or an image-level descriptor of a patient's eyes.

There are a total of 1151 records with 18 useful features. Two attributes, *quality assesment* and *class label* are not features - the former is just the quality of the record and the latter indicates if DR is present or not. Features in this dataset are observations made on images of eyes such as diameter of optic disc, distance of the center of the macula and the center of the optic disc etc.

2.2.1 Statistical Analysis

We analysed the dataset for different statistics.The class distribution has been shown in fig 3 and it shows that it is balanced. From the heatmap that was obtained, MAs @ confidence level 0.5 and 0.6 (0.996) followed by MAs @ confidence level 0.6 and 0.7(0.994) are highly correlated.

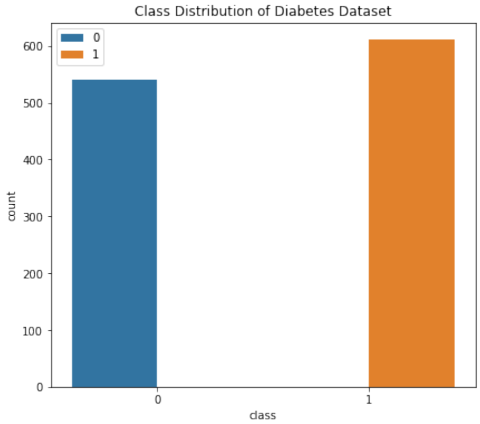


Figure 3: Class Distribution for Diabetes Dataset

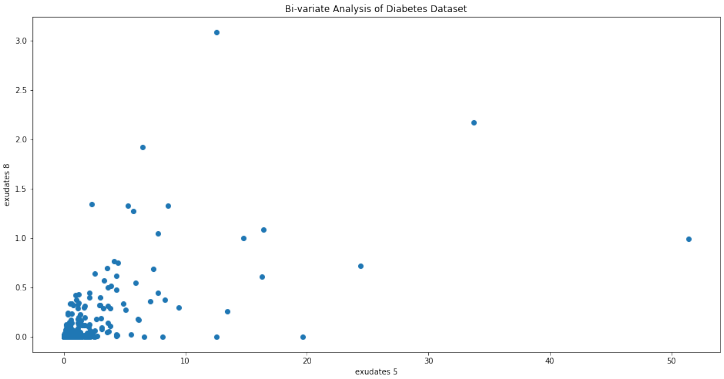


Figure 4: Bivariate analysis on diabetes dataset

2.2.2 Data Processing

Unlike the hepatitis dataset, this had no missing values. There were only 4 bad quality records, which were not significant enough to adversely affect any learning. We did not perform any pre-processing other than adding correct column names to features.

2.2.3 Training, validation and testing sets

We used a standard ***60%:40% train:test split*** (i.e., 690 train and 461 test). The training set was further split into training and validation sets using our implementation of L-folds cross validation. We used 3 folds, resulting in fold size of 230 records.

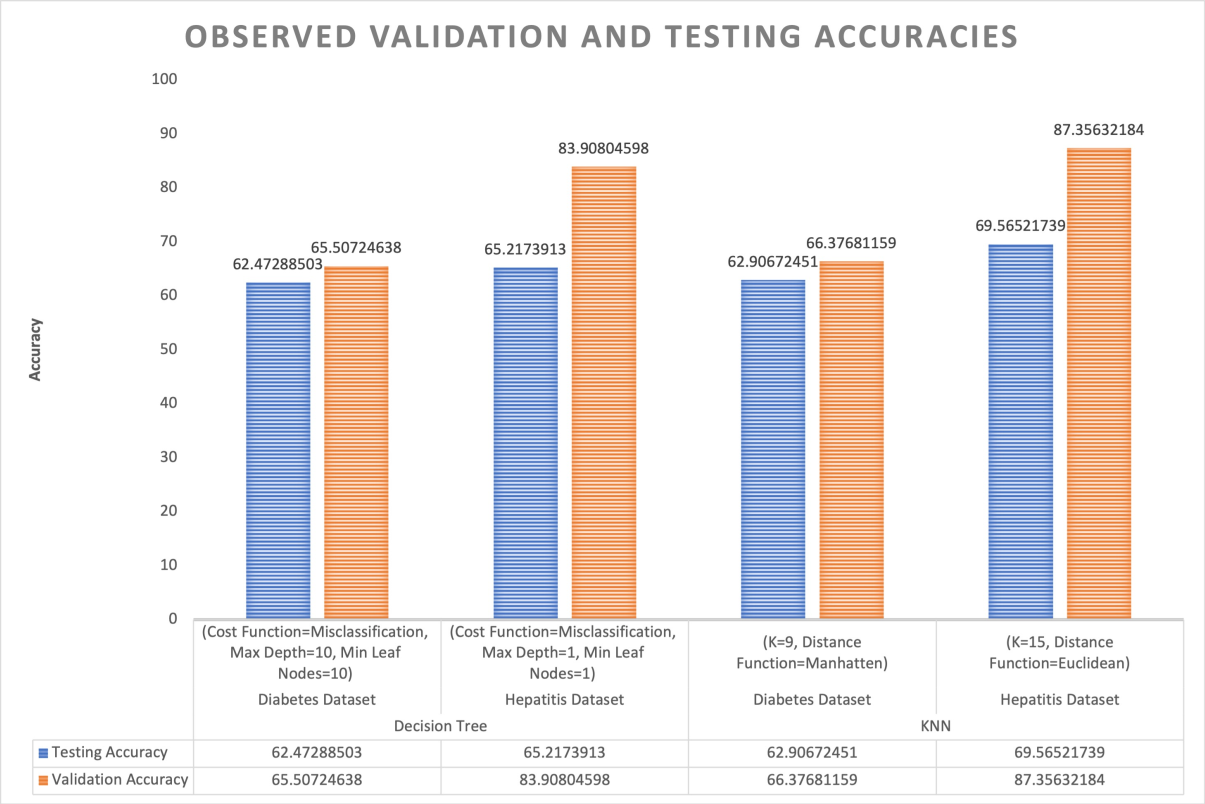
2.3 Ethical Concerns

These datasets contain medical records about patients and so there are various ethical issues associated with such kinds of healthcare data:

- ***Privacy*** - Patients' data could be being shared without their prior consent, unless otherwise informed and stated.
- ***Transparency*** - Patients' whose data is being shared are not aware about the purpose for which their data is being utilised.
- ***Fairness*** - The racial or the ethnic background was not considered, thus it might lead to incorrect results for different populations due to biases in training data.
- ***Accuracy*** - Accuracy of the model will be affected if it is biased or does not contain representative data (example: sex, race, ethnicity etc). The performance will be different for different use cases.

3 Results

We used the same pre-processing pipeline for both the algorithms in order to compare the results accurately. The only exception is that we normalize the features for KNN since this algorithm is sensitive to feature scaling. We performed grid search over various possible values to select the best hyperparameters for KNN and Decision Trees (DT). We also performed 3 folds cross-validation on each model and picked the one with the best validation accuracy for evaluating the test data. For the KNN model, we evaluated the performance with hyperparameters $K = \{1, 2, ...20\}$



	Decision Tree		KNN	
	Diabetes Dataset	Hepatitis Dataset	Diabetes Dataset	Hepatitis Dataset
	(Cost Function=Misclassification, Max Depth=10, Min Leaf Nodes=10)	(Cost Function=Misclassification, Max Depth=1, Min Leaf Nodes=1)	(K=9, Distance Function=Manhatten)	(K=15, Distance Function=Euclidean)
▤ Testing Accuracy	62.47288503	65.2173913	62.90672451	69.56521739
▤ Validation Accuracy	65.50724638	83.90804598	66.37681159	87.35632184

Figure 5: Accuracy metrics for both algorithms

and $distancefunction = \{euclidean, manhatten\}$. For the DT model, we evaluated performance with the hyperparameters $cost_function = \{misclassification, entropy, gini\}$, $maxdepth = \{1, 2, 5, 10, 50\}$, $minleafinstances = \{1, 2, 5, 10\}$.The best hyperparameters based on validation accuracies are given in fig. 5.

Figure 8 shows KNN decision boundary for the two most correlated features for Diabetes dataset. As expected from highly correlated features, we observed a clearly separable boundary in this feature space. For the Hepatitis dataset, we were unable to get a clean decision boundary because of the class imbalance. Nevertheless, the figure is still included in the code.

Figure 9 and 10 shows trend of different K values and max depth values for both datasets evaluated using cross validation. We observe accuracy for Hepatitis data remains the same after K=15, and reaches maximum at K=9 for Diabetes. Best accuracy is achieved for max depth=1 for Hepatitis and 10 for Diabetes.

4 Discussion and Conclusion

For the Hepatitis dataset (fig. 5), we observed that the KNN algorithm outperformed the DT algorithm for both validation and test sets by a narrow margin. This was an expected behaviour due to the very small size of this dataset and the extreme class imbalance as noted earlier in fig 1. Because of this, the generalization error estimate from the validation accuracy was also not a good estimate since the accuracy decreased for the test set in case of both the algorithms.
For the Diabetes dataset (fig. 5), although KNN had a better validation accuracy, its testing accuracy compared to DT was comparatively worse. This might have happened because the generalization estimate given by validation accuracy is not a good estimate.

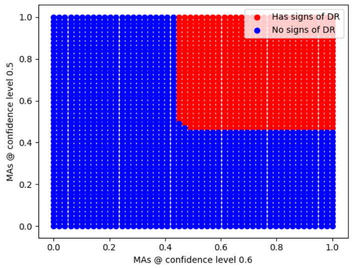


Figure 6: KNN decision boundary

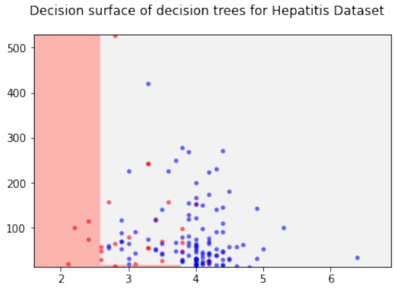


Figure 7: Hepatitis dataset decision boundary

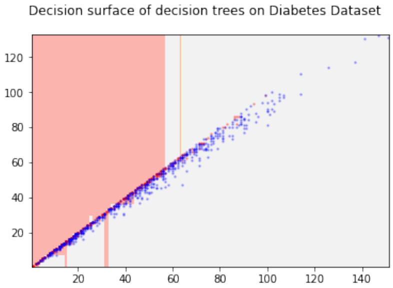


Figure 8: Diabetes dataset decision boundary

The decision tree (DT) is more generalizable. For the KNN algorithm, we observed a better accuracy on both the datasets with an increase in the K value. In fact, for the hepatitis dataset we observed that the accuracy did not change once it reached the K value of 5 for both the distance metrics. For both the datasets, both the distance metrics offer a comparable accuracy. When we consider the hyperparameters for decision trees,we see that the accuracies of the cost functions are often very similar and do not have a distinct trend across datasets. Higher max depth values also lead to decrease in accuracy, possibly due to overfitting.

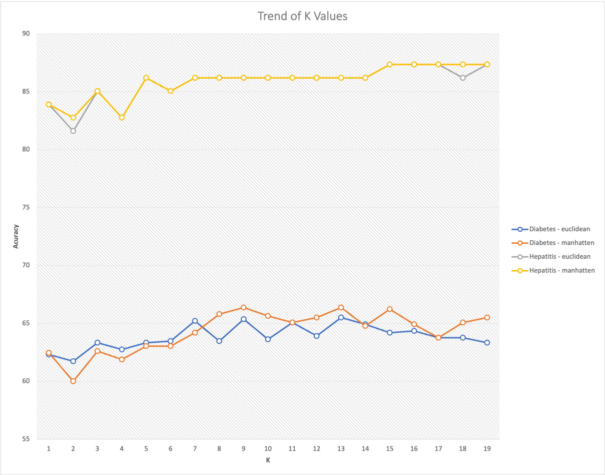


Figure 9: Trend of K values

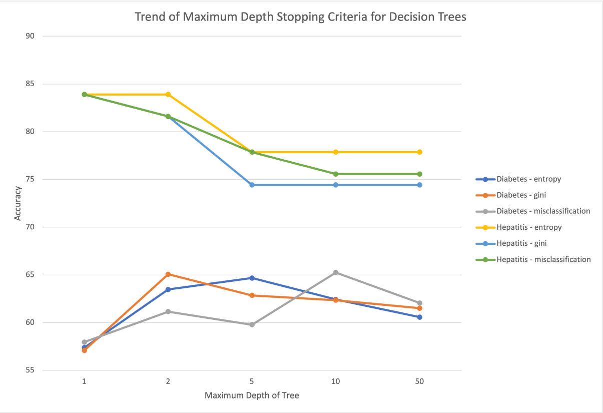


Figure 10: Trend of max depths

For further work, we aim to see if these results can be improved by generating more data to balance class labels for Hepatitis dataset and one-hot encoding binary data.

5 Statement of Contributions

Avinash implemented decision trees, augmentation, decision boundary and plots. *Gauri* did data cleaning and processing, statistical analysis, data visualization, report writing. *Manas* implemented KNN, cross validation, data preprocessing, report writing.

References

Bálint Antal and András Hajdu. 2014. An ensemble-based system for automatic screening of diabetic retinopathy. *Knowledge-Based Systems*, 60:20–27, Apr.

Tri Astuti, Hanung Adi Nugroho, and Teguh Bharata Adji. 2015. The impact of different fold for cross validation of missing values imputation method on hepatitis dataset. In *2015 International Conference on Quality in Research (QiR)*, pages 51–55.

Dheeru Dua and Casey Graff. 2017. UCI machine learning repository.

Ogundokun R.O. Oladele T.O. et al. 2019. Application of data mining algorithms for feature selection and prediction of diabetic retinopathy. *ICCSA, Lecture Notes in Computer Science*, 11623.

A. L. Samuel. 1959. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):210–229.