ITRI626 ARTIFICIAL INTELLIGENCE II

Assignment 1

34292748
MNISI G

Contents

1 Introduction

Machine learning (ML) is a branch of computational science that studies the capacity of machines to learn things without being manually programmed [7]. Machine Learning is utilized where there are enormous datasets and humans are unable to analyze any information from the data output, which includes product recommendations, financial accuracy, predictive analysis, and other applications.

Big data (datasets that are complex and large) has made the field of Machine Learning to be relevant. One of the fundamental objectives of Machine Learning is to learn from data and generate predictions, which cannot be accomplished without the use of a variety of algorithms. This report will look at literature reviews on the chosen ML algorithms which are kNN, SVM and Random forest. Moreover classification dataset will be utilized and trained using each of the algorithms and be evaluated.

2 Literature Review

2.1 kNN

One of the most simple and effective approaches is the k-nearest-neighbors (kNN) which is a non-parametric classifier [5]. The nearest neighbor is determined using the k-value, which determines how many of the nearest neighbors are to be considered. This approach is biased because to apply it, one has to select an acceptable value for k and in most cases, the performance of the classification depends on this number [1].

A study by Uddin (2022), used different kNN algorithms to predict illness in the actual world. The algorithm was utilized because of the following reasons (1) how easy to comprehend and employ, (2) The algorithm's operations and computations are straightforward, (3) It can be modified in numerous ways to reduce its constraints and obstacles while increasing its reliability and relevance to a larger range of datasets [12]. The study compares several kNN algorithms (Generalised mean distance, Hassaant, Ensemble, Classic one) and evaluates them according to their recall, precision and accuracy. The average accuracy scores ranged between 64.22% and 83.62% with the Hassanaat kNN (83.62%) outperforming all the other kNN algorithms, following the Ensemble kNN algorithm with an average accuracy of 82.34%.

The study being evaluated shows that the kNN algorithm is good at predicting accurate results hence it is utilized. Moreover, kNN is good at classifying unknown data points to the closest neighbors that already exist (good classification method), has fast training time and it's easy to comprehend and employ. Lastly, when compared with other machine learning methods, it can be categorized at a high rate.

2.2 SVM

Support Vector Machine (SVM) is a supervised ML algorithm related to examining regression and classification datasets, so developing classifiers can be easily developed. It seeks to establish a choice threshold among two classes that allow label estimation using multiple feature vectors [6]. SVM is a linear classification, it can be extended to a non-linear classifier by employing the kernel technique [7]. In the kernel technique, the inputs are implicitly

mapped into feature spaces with high dimensions, this draws class distinctions so that the classification error decreases.

Biddle and Fallah (2021) conducted a study for classification for a new fault detection, recognition, and forecasting approach for automated cars utilizing SVM [2]. In this modern day, car sensors have rapidly increased because of technology, so they break down often and require maintenance now and then. The study was carried out in order for a user to know when there would be a defect in an automobile utilizing SVM methods. The study suggested an FDII (Fault detection, isolation, and identification) model which includes the present vehicle's condition as well as forecasting to estimate foreseeable vehicle health [2]

The model takes in a variety of independent sensor signals and applies a signal extraction of feature algorithms to them [2]. If there's a flaw found, the samples undergo processing further. If there are no flaws, the samples are sent to the module prediction to determine whether the signal from the sensor suggests decreased performance before the subsequent fault.

Utilizing an activation function, SVM converts the input information between the layers. The SVM algorithm yielded a prediction accuracy of 75.35% and a detection and identification accuracy of 94.94%. A study performed on the Car Evaluation Dataset by Singh also shows that the SVM is a better predictor alongside Artificial Neural Network (ANN)[10].

This algorithm was selected because it works perfectly with unstructured data and has the capacity to analyze high-dimensional data and to do effectively with tiny datasets.

2.3 Random Forest

Random forest has been demonstrated in dealing with data differences in several classes particularly with enormous data sets [8]. Random forest is similar to kNN when it comes to being biased with values. When a dataset has multiple features, the accuracy of the algorithm also increases. RF is a predictive and classification approach that is built on the aggregate of multiple decision trees.

Utilizing random forest, a literature study was conducted by Simsekler et al (2021) with the aim to measure the satisfaction of patients by ranking the relevance of patient and provider-related characteristics in two frequent patient journey phases, check-in and consultation procedure [11].

The results of the study indicated that age features, a patient-related variable, are the most important predictor of patient satisfaction throughout both phases. The primary provider-related factors in each model constructed for the registration and consultation stages, respectively, are the total time spent for registration and the attentiveness and expertise of the doctor while listening to their inquiries [11]. A study was conducted by Ramya and Ganapathy to evaluate a Vehicle's Quality Performance using kNN and Random Forest. When the two algorithms were compared, the Random Forest performed better than the kNN by 7.05%[9].

In the study that was conducted, RF provided significant results in the drivers of a patient's satisfaction and the results can give healthcare facilities a better insight on what could be improved.

3 Materials and Methods

3.1 Materials

The dataset used is Car Evaluation Database from the University of California Irvine (UCI) Machine Learning Repository [3]. Bohanec created it to demonstrate DEX, an expert system for decision-making. This dataset was employed in machine learning to evaluate HINT (Hierarchy INduction Tool) and was shown to be capable of completely reconstructing the original hierarchical model. The dataset model bases its assessment of car acceptance on three concepts: cost, technical characteristics, and comfort.

The dataset comprises 6 variables (attributes), 1728 records (instances), and categorical variable type, and the target variable is CAR which implies car acceptability. Below is a table that explains the Car Evaluation dataset:

Data Set:	Multivariate	Number of Instances:	1728
Attributes:	Categorical	Number of Attributes	6
Associated Tasks:	Classification	Missing Values?	No

Table 1: Car Evaluation Dataset

The Car Evaluation Database contains the following four class values and six attributes. The class values are as follows:

Attributes	Denoted as
Acceptable	”acc”
Good	”good”
Unacceptable	”unacc”
Very good	”vgood”

Table 2: Class attributes

The six attributes are as follows: (1) buying: buying price (vhigh, high, med, low), (2) maint: the price of the maintenance (vhigh, high, med, low), (3) doors: number of doors (2, 3, 4, 5, more), (4) persons: capacity in terms of persons to carry (2,4, more) (5) luggage_boot: the size of luggage boot (small, med, big), and (6) safety: estimated safety of the car (low, med, high)

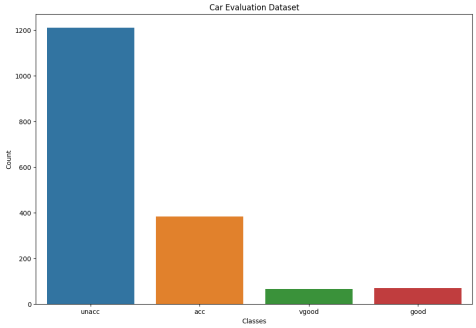


Figure 1: Frequency of the classes

From the figure above, 1728 cars were evaluated and the results show that 384 (22.28%) were acceptable, 69 (4.05%) were good, 1210 (69.85%) were unacceptable and 65 (3.82%) were very good. From this, it shows that more than half of the cars are unacceptable.

3.2 Methods

Data Preprocessing

Data preprocessing comprises five activities: data cleaning, reduction, scaling, partitioning, and transformation [4]. Data cleaning attempts to improve data quality through the imputation of missing values and outlier elimination. Data reduction is used to reduce data dimensions and thus the computing expenses associated with them. The goal of data scaling is to convert the original data into comparable ranges for predictive modeling.

The goal of data transformation is to rearrange what was originally collected data into formats that may be used by various data mining methods. It usually consists of two tasks: numerical data transformation and categorical data transformation. The goal of data partition is to split the entire data set into distinct groups based on its functional characteristics [4].

In terms of the Car Evaluation Dataset, it was imported and read the dataset as a CSV format. The header columns were renamed to the attribute's name given from the data and the four targets (classes) were changed to numerical so that the model can optimize efficiently. In this dataset, there were no missing values to be handled.

The dataset was split into three parts which include the dataset for training, validating, and training. The splitting ratio was 60-20-20 respectively. This is utilized to prevent overfitting. The training set is higher than the validating and testing data because the model needs to learn and observe from the training data and optimize the parameters.

The dataset was scaled (normalizing the features) so that the performance of each algorithm could improve and speed up the learning process. Each of the models was trained using the fit() function, which changes the model's parameters to increase the accuracy. The algorithms were validated after being trained, which is validating how the model works, this is done by using different that that was initially split. Lastly, each model is tested using the testing data, and the evaluation metrics are calculated to analyze how each model performs.

The following Python libraries were utilized numpy, pandas, matplot, and sklearn. NumPy includes support for huge arrays with multiple dimensions and matrices, as well as a wide set of high-level mathematics functions. Pandas are used when one is working with data, it is used for manipulating, cleaning, and analyzing data. Matplot is used for visualizing the results using graphs. Sklearn is used for various classifications, clustering methods, and many more.

Model Parameters

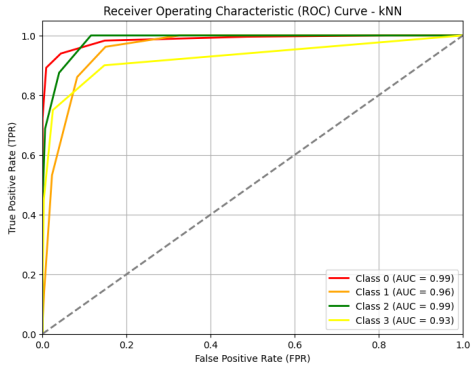
Each of the following ML algorithms uses different parameters, the kNN algorithm uses one parameter which is the number of neighbors. SVM uses the probability parameter which is used to calculate one of the metrics the kernel parameter and other parameters that are not specified manually. The Random Forest uses two parameters which include the number of trees and the depth of each tree.

4 Results

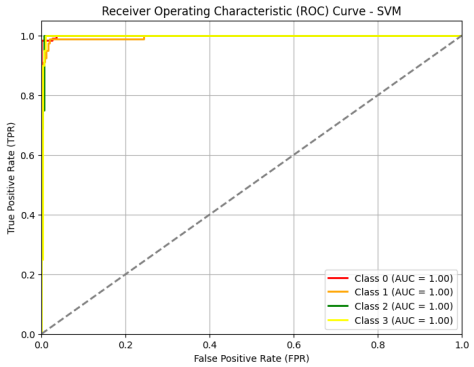
The following results were obtained when the three models were run using the test dataset

Algorithm	Class	precision	recall	f1-score	AUC
kNN	0	0.90	0.98	0.94	0.96351
	1	0.77	0.77	0.77	
	2	0.90	0.56	0.69	
	3	0.83	0.25	0.38	
	accuracy			0.87	
	macro avg	0.85	0.64	0.70	
	weighted avg	0.87	0.87	0.86	
SVM	0	1.00	0.97	0.98	0.99772
	1	0.88	0.99	0.93	
	2	0.82	0.88	0.85	
	3	0.93	0.70	0.80	
	accuracy			0.95	
	macro avg	0.91	0.88	0.89	
	weighted avg	0.96	0.95	0.95	
Random Forest	0	1.00	0.97	0.98	0.99380
	1	0.81	0.97	0.89	
	2	0.77	0.62	0.69	
	3	0.77	0.50	0.61	
	accuracy			0.93	
	macro avg	0.84	0.77	0.79	
	weighted avg	0.93	0.93	0.92	

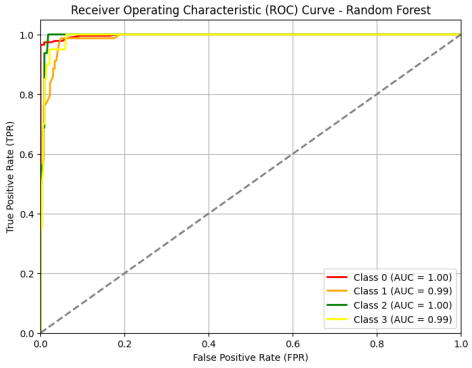
Table 3: Results of the three ML algorithms



(a) k-Nearest Neighbors (kNN)



(b) Support Vector Machine (SVM)



(c) Random Forest Model

From the following results, the kNN model has an overall accuracy of 0.85 and a ROC score of 0.96351, the SVM model has an overall accuracy of 0.95 and a ROC score of 0.99772, and the Random forest has an overall accuracy of 0.93 and the ROC score is 0.99380. From these metrics, SVM classification yields the best accuracy compared to the other two algorithms. This is because of its characteristics of handling unstructured data and can handle complex data. Lastly, SVM is capable of handling a classification using non-linear and linear data.

5 Conclusion

The Car Evaluation Database classification dataset was used in this assignment to evaluate car acceptability using three machine-learning algorithms. This assignment compared the three algorithms (kNN, SVM, Random Forest) and evaluated each algorithm using the metrics but it mainly focused on their overall accuracy and the ROC score. SVM was better having the highest accuracy of 95%, meaning that it predicts better than the other two algorithms.

References

[1] I. A. A. Amra and A. Y. Maghari. Students performance prediction using knn and naïve bayesian. In *2017 8th international conference on information technology (ICIT)*, pages 909–913. IEEE, 2017.

[2] L. Biddle and S. Fallah. A novel fault detection, identification and prediction approach for autonomous vehicle controllers using svm. *Automotive Innovation*, 4:301–314, 2021.

[3] M. Bohanec. Car Evaluation. UCI Machine Learning Repository, 1997. DOI: `https://doi.org/10.24432/C5JP48`.

[4] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang. A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. *Frontiers in Energy Research*, 9:652801, 2021.

[5] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer. Knn model-based approach in classification. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings*, pages 986–996. Springer, 2003.

[6] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu. Applications of support vector machine (svm) learning in cancer genomics. *Cancer genomics & proteomics*, 15(1):41–51, 2018.

[7] B. Mahesh. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 9(1):381–386, 2020.

[8] A. Paul, D. P. Mukherjee, P. Das, A. Gangopadhyay, A. R. Chintha, and S. Kundu. Improved random forest for classification. *IEEE Transactions on Image Processing*, 27(8):4012–4024, 2018.

[9] V. Ramya and K. Ganapathy. Evaluation of vehicle quality performance using random forest in comparison with knn to measure the accuracy, recall, and precision. In *2022 3rd International Conference on Intelligent Engineering and Management (ICIEM)*, pages 550–555. IEEE, 2022.

[10] Z. U. Rehman, H. Fayyaz, A. A. Shah, N. Aslam, M. Hanif, and S. Abbas. Performance evaluation of mlpnn and nb: a comparative study on car evaluation dataset. *International Journal of Computer Science and Network Security*, 18(9):144–147, 2018.

[11] M. C. E. Simsekler, N. H. Alhashmi, E. Azar, N. King, R. A. M. A. Luqman, and A. Al Mulla. Exploring drivers of patient satisfaction using a random forest algorithm. *BMC Medical Informatics and Decision Making*, 21(1):157, 2021.

[12] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide. Comparative performance analysis of k-nearest neighbour (knn) algorithm and its different variants for disease prediction. *Scientific Reports*, 12(1):6256, 2022.