MLDM ASSIGNMENT

MBA-BABD

(2022-2024)

GROUP MEMBERS:

- ASHISH SAROCH 502204167
- ATINDERJIT SINGH 502204168
- HARJOT KAUR 502204169

INTRODUCTION

TOPIC - ANALYSIS OF BANKNOTE AUTHENTICATION SYSTEM USING DIFFERENT MACHINE LEARNING ALGORITHMS

Banknote authentication refers to the process of verifying the authenticity and validity of a banknote, ensuring that it is genuine and not counterfeit. Preserving the authenticity of higher denomination printed banknotes is one of the critical issues. It has a major role in the financial activities of every country. Therefore, our project is based on analysing BANKNOTE AUTHENTICATION i.e to distinguish between genuine and counterfeit banknotes. We have used various supervised machine learning algorithms such as Logistic Regression, Naïve Bayes, Decision Tree, Support Vector Machine (SVM) and Random Forest to differentiate genuine banknotes from counterfeit ones.

RELATED WORK

WE HAVE LISTED SUMMARY OF RELATED WORK :

REFERENCES	TECHNIQUES USED	PERFORMANCE MEASURES
Kumar *et al*., 2005	Probabilistic neural network, Multi-layer Perceptron, Radial Basis Function, Decision Tree, Naïve Base.	Sensitivity, Specificity, Accuracy
Shahani *et al*., 2018	Back Propagation Neural Network, Support Vector Machine	Accuracy, Sensitivity, Specificity, Precision
Nastoulis *et al*.,2006	Probabilistic Neural Network (PNN)	-
Omatu *et al*., 2007	Learning vector quantization	Reliability
Gillich *et al*., 2014	Wavelet Transform, Intaglio	RMSE (root mean squared error)

ABOUT THE DATASET

The dataset used to train the models is taken from UCI machine learning repository. Data was extracted from genuine and counterfeit banknote images. The dataset has 1372 observations. There are 5 attributes out of which 4 are the features and one is the target attribute. The dataset contains a balanced ratio of both classes which is 55:45(genuine: counterfeit). The target class contains two values: 0 and 1 where 1 represents genuine note and 0 represents fake note. Features of the dataset:

- ***Variance:*** Measure of how far a set of numbers is spread out.
- ***Skewness:*** Skewness is the measure of the lack of symmetry .
- ***Kurtosis:*** Kurtosis is a measure of whether the data are heavy- tailed or light-tailed relative to a normal distribution.
- ***Entropy:*** Image entropy is a quantity which is used to describe the amount of information which must be coded for, by a compression algorithm .
- ***Class :*** Class contains two values 0 representing genuine note and 1 representing fake note.

PROPOSED METHODOLOGY

For our project, our aim is to find the most accurate model for our problem (we have used cross validation to have more accurate result). Six supervised machine learning algorithms were used for the project:

- ***LOGISTIC REGRESSION:*** It is used for predicting the categorical dependent variable using a given set of independent variables. The outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.
- ***NAIVE BAYES:*** It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.It depends on the conditional probability.
- ***K-NN:*** K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

PROPOSED METHODOLOGY

- ***DECISION TREE:*** It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node.
- ***SUPPORT VECTOR MACHINE:*** In SVM, each data item is plotted on the graph and then classification is performed to find the hyper plane that differentiates the two classes.
- ***RANDOM FOREST:*** Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. It takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

EXPERIMENTAL ANALYSIS

FOLLOWING IS THE SUMMARY OF ANALYSIS (80% of the data was used for training and 20% for testing):

TECHNIQUES	ACCURACY	SPECIFICITY	SENSITIVITY	PRECISION
Logistic Regression	96.36%	95.54%	97.45%	94.26
Naïve Bayes	82.54%	83.43%	77.11%	81.25%
KNN	100%	100%	100%	100%
Decision Tree (Criterion = 'entropy')	99.27%	99.36%	99.15%	99.15%
Decision Tree (Criterion = 'gini')	98.90%	98.08%	100%	97.52%
SVM	100%	100%	100%	100%
Random Forest	99.63%	99.36%	100%	99.15%

CONCLUSION

“Performance metrics have been calculated to compare the performances of all the techniques. The result shows that KNN and SVM gives 100% success rate i.e highest accuracy. These techniques are an efficient way of solving the problem for all banking machines that accept all types of notes.”