CLASSIFICATION MODELS IN MACHINE LEARNING: KNEARESTCLASSIFIER AND DECISION TREE:

Fatai Olarinde JIMOH[1]
Universite de Bourgogne,
Centre de Universite, Condorcet
71200 Le Creusot.

Lateef Olalekan ADERINOYE[2]
Universite de Bourgogne,
Centre de Universite, Condorcet
71200 Le Creusot.

Abstract

Classification model is a data mining (machine learning) technique used to predict group membership for data instances. There are several classification techniques that can be used for classification purpose. In this paper, we present practical application of two classification techniques: decision tree classifier and k-nearest classifier with their strengths, weaknesses, potential applications and issues with their available solution. The goal of this study is to provide a practical review of these two classification techniques in machine learning.

Keywords: Machine learning, classification, classification model, classification issues.

1   Introduction

Machine Learning (ML) is a vast interdisciplinary field which builds upon concepts from computer science, statistics, cognitive science, engineering, optimization theory and many other disciplines of mathematics and science[1]. There are numerous applications for machine learning but data mining is most significant among all[2].Machine learning can mainly classified into two broad categories include supervised machine learning and unsupervised machine learning.

Unsupervised machine learning used to draw conclusions from datasets consisting of input data without labeled responses [3]. On the other hand, the Supervised machine learning techniques attempt to find out the relationship between input attributes (independent variables) and a target attribute (dependent variable)[4]. Supervised techniques can further classified into two main

categories; classification and regression.

Classification is a data mining (machine learning) approach that used to forecast group membership for data instances [6]. Although there are variety of available techniques for machine learning but classification is most widely used technique [7]. As students, we believed this is due its simplicity and easy application, especially, for the students, new researchers and/or data analysts. Classification is categorized as one of the supreme studied problems by researchers of the machine learning and data mining fields [1]. A general model supervised learning (classification techniques) is shown below in Figure 1.

This paper is organized as follows; in section 2 methodology of review is presented. Section 3 is divided into two subsections in which selected classification techniques has been discussed.

2 Methodology

A literature search was done from articles and datasets using databases such as Kaggle, google scholar, and some related web pages. The keywords used for literature search include; Machine learning, students performance dataset, classification, classification applications and classification algorithms. The articles that contains relevant data about classification techniques applications particularly, K-Nearest Classifier and Decision Tree Classifier were mainly used to assist our project. It is difficult to provide exhaustive application of the classification methods used on our data in order to simplify the project. Therefore, we focused only on its usefulness to retrieve results from our features and target..

3 The Classification Techniques used for prediction

3.1 Decision Tree Classifier

The core objective of decision tree is to produce a model that calculates the value of a required variable based on numerous input variables [6]. Usually all decision tree algorithms are constructed in two phases (i) tree growth; in which training set based on local optimal criteria is splitting recursively until most of the record belonging to the partition having same class label [14] (ii) tree pruning; in which size of tree is reduced making it easier to understand [2]. In this section we will focus on ID3 and C4.5 decision tree algorithm.

3.2 K-Nearest Neighbor Classifier

In K-nearest neighbor (KNN) technique, nearest neighbor is measured with respect to value of k, that define how many nearest neighbors need to be examine to describe class of a sample data point. One of the main advantage of KNN technique is that it is effective for large training data and robust to noisy training data.

4 Results and Analysis

We looked into classification of Students Performance using K-Nearest Neighbor Classifier and Decision Tree classification models. After fitting and application of the two model for predictions, we try to compare the predictions with the main dataset for better understanding and best model performance.

4.1 Score Accuracy for the two Models

```
d_y_pred = decision_tree.predict(X_test)

print('\n Classifier', c_y_pred[0:30]
        , '\n decision tree', d_y_pred[0:30])
```

```
Classifier [0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0]
decision tree [0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
```

And our two models proved to perform brilliantly as Knn Classifier model accuracy result is 0.9833333333333333 and Accuracy of Decision Tree model is 1.0.

```
print( '\n Accuracy of our Knn Classifier model is' , c_y_pred
        , '\n Accuracy of our decition tree model is' , d_y_pred)
```

```
Accuracy of our Knn Classifier model is 0.9833333333333333
Accuracy of our decition tree model is 1.0
```

```
[1]: from sklearn.model selection import cross val score
```

5   Conclusion

Our selection of classification models (KNearest Neighbor Classifier and Decision Tree Classifier) was based on interest and intention to exposed students like us to advantage and simplicity of these two models.

6   References:

[1] Ghahramani Z. "Unsupervised learning," in Advanced lectures on machine learning, ed: Springer, 2004; pp. 72- 112. https://doi.org/10.1007/978-3-540-28650-9$_5$

[2] Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: A review of classification techniques. ed, 2007.

[3] Zhang D, Nunamaker JF. Powering e-learning in the new millennium: an overview of e-learning and enabling technology. Information Systems Frontiers 2003; 5: 207-218. https://doi.org/10.1023/A:1022609809036

[4] Maimon O, Rokach L. Introduction to supervised methods, in Data Mining and Knowledge Discovery Handbook, ed: Springer, 2005 pp. 149-164.

[5] Ng A. "CS229 Lecture notes."

[6] Kesavaraj G, Sukumaran S. A study on classification techniques in data mining. in Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on, 2013; pp. 1-7.