Comparative Analysis of KNN and Decision Tree Algorithms for Classification in Information Mining

P. Sai Preethi

PG Scholar, Dept. of Computer Science Sri Venkateswara University, Tirupati

Abstract— *Information mining is a vital process in extracting valuable insights and patterns from diverse datasets. Classification, a key technique in data mining and machine learning, involves categorizing data into predefined classes or groups. Various classification algorithms, such as decision trees, Bayesian methods, case-based learning, artificial neural networks, and support vector machines, have been extensively studied and applied in this context. This paper focuses on two classification methods based on K-Nearest Neighbors (KNN) and Decision Tree. The study uses the Electricity_ board dataset, consisting of 45,781 examples, with five independent variables and one dependent variable, for analysis. The results demonstrate that KNN outperforms Decision Tree in terms of accuracy and precision for the classification task.*

I. INTRODUCTION

Data mining is a technology that involves discovering new relationships, hidden information, and meaningful patterns from datasets. It is part of the broader process known as "data discovery" or Knowledge Discovery in Databases (KDD). Data mining encompasses various techniques such as classification, clustering, association rule mining, and anomaly detection. The process of data mining aims to extract valuable information from datasets and present it in a comprehensible format. The success of data mining heavily relies on appropriate algorithms and a deep understanding of the datasets. This paper focuses on the classification process for the present study.

II. CLASSIFICATION

Classification plays a pivotal role in data mining and machine learning. Its primary objective is to construct a classifier that can accurately model and predict the class labels of unknown data instances. The performance of a classifier is measured by its classification accuracy. The supervised learning approach seeks to build a clear and unambiguous model that maps the indicator features to class labels. The classifiers are then utilized to assign class labels to testing instances, where the indicator feature values are known, but the class label is unknown. The classification of vast amounts of data can be time-consuming and computationally intensive, which may not be suitable for certain applications.

III. METHODOLOGY

Various kinds of characterization strategies have been proposed in writing that incorporates Decision Trees, Naïve Bayesian techniques, Artificial Neural Networks, Logistic Regression, SVM and KNN and so forth. In this paper, we assess the presentation of the KNN calculations on Electricity_ board informational index was utilized for the grouping contrasted and the Decision Tree.

3.1 Decision Tree

Decision tree learning is one of the most incredible procedures for regulated request learning. Decision trees are an essential recursive plan for conveying a continuous gathering process in which a case, portrayed by a lot of qualities, is given out to one of a disjoint plan of classes [3][5]. A decision tree is a tree structure which arranges a data test into one of its expected classes. Decision trees are used to isolate data by making decision principles from the gigantic proportion of available information. A decision tree classifier has an essential design which can be moderately taken care of and that successfully describes new data.

Decision trees include center points and leaves. Each center in the tree incorporates testing a particular property and each leaf of the tree implies a class. Regularly, the test differentiates a property assessment and a consistent. Leaf centers give a portrayal that applies to all events that show up at the leaf, or a lot of groupings, or a probability course over each possible game plan [7]. To portray a dark case, it is directed down the tree according to the potential gains of the properties attempted in moderate center points, and when a leaf is reached, the model is gathered by the class consigned to the leaf.

3.2 K-Nearest Neighbors (KNN)

The KNN is a non-parametric social event system, which is central regardless incredible all over [4]. The basic thought for k-NN depends resulting to choosing the distances between the attempted, and the status data tests to see its nearest neighbors. The attempted model is then moved to the class of its nearest neighbor [5].

The K-Nearest Neighbors (KNN) is a sensible regardless convincing technique for diagram. The KNN evaluation is a strategy for social event objects subject to closest orchestrating models in the part space. KNN is a kind of event based learning, or held perceiving where the end is simply approximated locally and all estimation is yielded until get-together [6]

For a data record D to be referenced, its K nearest neighbors is recuperated, and these improvements a neighborhood of D. Larger part extending a majority rule structure among the data records in the space is in general used to pick the requesting for D paying little mind to considered distance-based weighting. Regardless, to apply KNN we need to pick a reasonable convincing power for K, and the accomplishment of combination is a great deal of wards on this value. The fundamental drawbacks concerning KNN are (1) its low viability - being a languid learning methodology denies it in various applications, for instance, dynamic web burrowing for a gigantic vault, and (2) its dependence on the decision of an "unfathomable worth" for K.

IV. EXPERIMENTAL RESULTS

The investigations have been coordinated by using Python programming tongue. The Python Scikit-learn is a pack for data portrayal, gathering and portrayal. The Electricity_ board dataset used in this review was procured from the UCI ML storehouse dataset [7]. In this dataset there are 45781 examples and 5 highlights recorded and twenty class names, among which 20 examples have a place with the separately, Class wise details are shown in the table-1 and same shown in the figure-1. The standard dataset is divided two sets one for preparing (75%) and one more set for testing (25%).

Table-1
Dataset Label Type Information

S. No	Label	Count
1	Bank	1410
2	Automobile Industry	1403
3	Bpo Industry	1397
4	Cement Industry	1403
5	Farmers1	1418
6	Farmers1	1405
7	Health Care Resources	1400
8	Textile Industry	1405
9	Poultry Industry	2888
10	Residential (individual)	2867
11	Residential (individual)	2884
12	Food Industry	2905
13	Chemical Industry	2830
14	Handlooms	2887
15	Fertilizer Industry	2876
16	Hostel	2857
17	Hospital	2906
18	Supermarket	2874
19	Theatre	2870
20	University	2896

We survey our two models using assorted execution estimations like Accuracy, Precision and Recall, the Experimental results are showed up in the table-2 and same showed up in the Figure-2.

Table-2
Performance of classifiers

Algorithm	Accuracy	Precision	Recall
Decision Tree	89.56	89.5	89.6
KNN	91.85	92	92

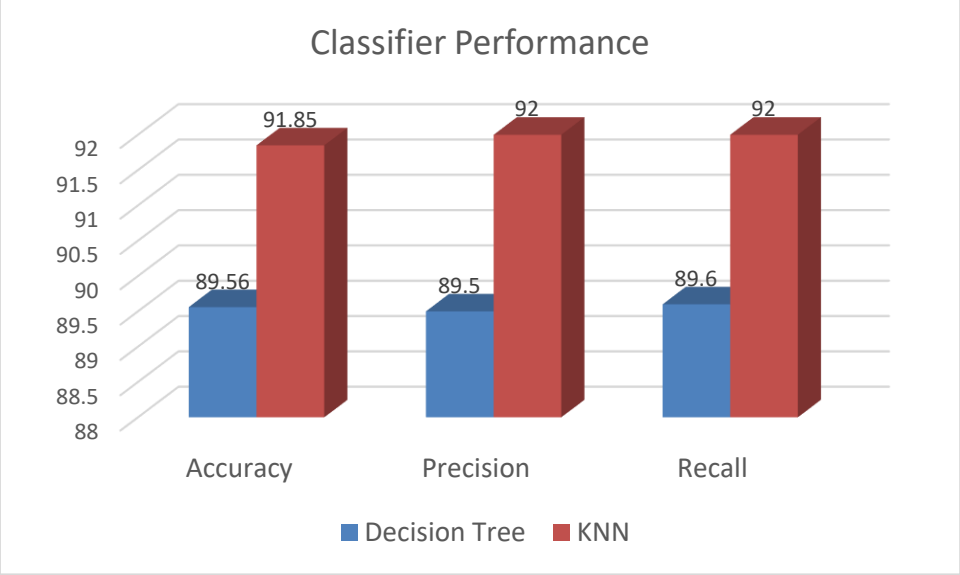

Figure-2: Experimental Results

We find in the Figure-2, the introduction of the KNN estimation has accomplished 91.85% precision and Decision Tree has achieved 89.56% As the result from assessment among the two computations, we find that most vital precision of Classification model is KNN (91.85%). So, the KNN algorithm have got highest accuracy, with a 2.29% difference when compared to Decision Tree algorithm.

V. CONCLUSION

The two machine learning algorithms (Decision Tree and KNN) are presented in the study in order to train the model and estimate the electricity usage of the Electricity_ board. By contrasting and evaluating the performance of the underlying Decision Tree classifier, the study has made progress toward KNN. In order to construct a system to predict the Electricity_ board dataset, this study identifies the applicability of these two classifiers, along with its drawbacks and advantages. In this investigation, the Decision Tree method had an accuracy value of 89.56% while the KNN approach had a high accuracy value of 91.85%. As a result, the KNN algorithm was utilised to predict electricity use with greater accuracy.

REFERENCES

[1] D. Hand, H. Mannila, P. Smyth.: Principles of Data Mining. The MIT Press. (2001)

[2] G Ravi Kumar, K Venkata Sheshanna and G Anjan Babu, "Sentiment analysis for airline tweets utilizing machine learning techniques", International Conference on Mobile Computing and Sustainable Informatics, PP:791-799, Publisher:Springer, Cham, 2020

[3] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques.2nd ed. San Francisco: Morgan Kaufmann, 2005.

[4] J. Han and M. Kamber," Data Mining concepts and Techniques", the Morgan Kaufmann series in Data Management Systems, 2nd ed. San Mateo, CA; Morgan Kaufmann, 2006.

[5] N. Michael, "Artificial Intelligence - A Guide to Intelligent Systems",2nd edition, Addison Wesley, 2005.

[6] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Reading, MA: Addison-Wesley, 2005.

[7] UCI machine learning repository. http://archive.ics.uci. edu/ml/.