International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Comprehensive Review of K-Nearest Neighbor Classification in Supervised Learning

Ms. A. Kamatchi[1], Dr. V. Maniraj[2]

[1]Research Scholar, Department of Computer Science, A.V.V.M. Sri Pushpam College (Autonomous), Poondi,Thanjavur (Dt), Affiliated to Bharathidasan University, Thiruchirappalli,Tamilnadu
[2]Associate Professor, Research Supervisor, Head of the Department, Department of Computer Science, A.V.V.M.Sri Pushpam College (Autonomous), Poondi, Thanjavur(Dt), Affiliated to Bharathidasan University, Thiruchirappalli, Tamilnadu

Abstract
Machine learning, a key subset of artificial intelligence, has emerged as one of the most influential technologies in today's digital world. It powers many everyday applications—from search engines to video recommendations. For instance, platforms like Google and YouTube use machine learning algorithms to analyze user behavior and preferences, enabling personalized search results and content suggestions. At its core, machine learning systems operate by receiving input data, learning patterns from it, and producing relevant outputs. These systems are typically trained on historical data, allowing them to make predictions or decisions based on new inputs. This paper focuses on the application of the **K-Nearest Neighbor (KNN)** algorithm, one of the most straightforward and widely used classification methods in supervised learning. In supervised learning, both input features and their corresponding outputs (labels) are provided to the model during training. This labeled data enables the algorithm to learn from past examples and make accurate predictions when presented with new, unseen data. KNN operates by identifying the 'k' closest data points in the training set to a given input and assigning the most common class among them. This paper explores how KNN is applied to a model dataset and how it classifies new data points based on learned patterns.

Keywords: Artificial Intelligence, Machine Learning, Supervised Learning, K-Nearest Neighbor (KNN), Classification Algorithm, Labeled Data

1. Introduction

Today all human beings live in that era of technology where new things are making the work of human beings easier and faster every day. One of these topics is machine learning, whose name you have probably heard. With the help of machine learning, a machine completes tasks based on its learning, understanding, and experience. Nowadays, many programs online are being guided with the help of machine learning so that these machines recognize the habits, likes, and dislikes of humans. Now let us understand the fact with the help of an example. If you want to understand the camera of a smartphone that the object being captured in the camera is a book on multiple subjects in front of the Smartphone's camera, and in doing so, the Smartphone's algorithm will understand that the object of this texture, design, and shape is a book. In the future, if any book is placed in front of this Smartphone. In this case, the Smartphone will understand based on its experience that this object is the book. Still, on the contrary,

if a pencil is held in front of the camera, then the Smartphone algorithm will be unable to get the name or information of that object because the Smartphone has never received information about the pencil. Machine learning is currently used in other areas, including the financial sector, social media, robots, automation, gaming application, etc.

- In daily life, human beings use social media many times in which machine learning is used. Facebook and Google show relevant advertisements to users based on their past search activity and influence video results on YouTube.
- Machine learning techniques save time and produce better results even with limited resources.
- Many source programs help to increase the usefulness of algorithms through machine learning.
- It can handle multidimensional or multi-variety even when no dynamic and favourable conditions exist.

Computer experts first created the idea of creating human-like thinking and learning computers in the 1950s. In the effort, in 1950, the first computer game was developed that could beat the world champion player. Deep Blue Computer is one of the best examples of machine learning techniques, which defeated world champion Garry Kasparov (Chess Champion Player).

In the paper, the K-Nearest Neighbour classification algorithm has been described. K-NN algorithm is based on supervised learning. Supervised learning is a type of machine learning. Supervised learning can be understood in such a way that a supervisor in supervised learning can also call the teacher. Supervised is such learning in which train the machine from the supervisor data or labelled data. Then after the training, the machine can predict the correct output for any other input. In supervised learning, input and output data are already provided to the machine, called training data or labelled data.

The most important method of supervised learning is classification; classification is used to analyze data in supervised learning. The classification method is used for the prediction. The classification methods or algorithms divide the data sets into categories. Here the data sets are divided into classes based on different parameters. For example, G- mail divides its mail data into categories like Email, Spam, Advertisement, Promotion, etc. So if you want to output the data into categories, then the classification algorithm will be used using the K-NN algorithm and solve the classification type problem. K Nearest Neighbour (KNN) is a supervised learning algorithm that can perform classification tasks using neighbors' numbers (K). K- Nearest Neighbour is the simplest machine learning algorithm based on supervised learning. The KNN algorithm is mostly used in solving the classification problem.

2. Related Work

After studying the review papers of the last 10 years to understand the work done in the previous year related to the supervised machine learning algorithm. The study proposes Muhammad. I, Machine learning can be implemented in the form of association analysis. It will require supervised learning, unsupervised, and reinforcement learning, but a better understanding of the strengths and weaknesses of the supervised learning classification algorithms. The main objective of supervised learning is to create a concise model that gives class label distributions in predictor features. In supervised learning, train the model in such a way that trains the machine with labelled data. After the training, the machine can predict the correct output for any other input.

Several machine learning classifiers such as Decision Tree (DT), Naïve Bayes (NB), Support Vector Machine (SVM), and Random Forest (RF) algorithm have been studied to predict heart problems accurately. In the present scenario, machine learning is used in various fields in today's scenario.

Machine learning is used for handwriting recognition, medical diagnosis, and biometric recognition. Machine learning plays an important role in identifying diseases based on patient characteristics in the medical field. Doctors use machine learning software to diagnose various diseases like cancer, cardiac arrest, and many more .The K-NN algorithm is considered a well-known pattern recognition method and is one of the most prominent text classification algorithms. The K-NN machine learning algorithm is based on supervised learning, one of the classification's simplest machine learning algorithms. This paper explains the ideas related to the K-NN algorithm, principles, and implementation steps in detail .
Imandoust and Bolandraftar have analyzed the K-NN classification method, which is studied extensively for economic forecasts. The models for financial crisis forecasting have been one of the most attractive areas in financial research. In the study , expected in the present scenario, stock price prediction has become a challenging research topic; currently, the stock markets are considered a great trading area. The stock market prices keep on fluctuating. Therefore, there is no loss to the inventors in the stock market, so the K- NN algorithm has been applied to the data of the companies in the stock market so that the inventors have minimum losses. The machine learning algorithms are majorly used for various purposes like image processing, predictive analytics, and image processing. One of the main advantages of using a machine learning algorithm is what to do with the data; after that, it automatically performs its function very well . The study describes the survey paper and explains the essential definitions, principles, and methodologies of the supervised machine learning algorithm very well. Reddy, and Ravi Babu , explain in the paper that supervised machine learning algorithm is mainly used for various types of image classification, predictive modelling, data mining technique, etc. Bijalwan and Pascual analyzed the supervised machine learning classification techniques that are prominent in automatically classifying a set of text documents into different categories from a predetermined set. Soofi, Awan describes that data science uses data mining techniques to predict the group members, which come under a classification method. Many such classification techniques of machine learning are used for classification purposes. These include decision tree classification algorithms, K-NN, and support vector machines. The computer algorithm is studied in machine learning. This computer is trained from experience, and after learning from experience, the computer automatically improves itself. The paper uses machine learning techniques to build and compare the classifier models, including Bayes network, Logistic Regression, Decision Stump, J48, Random Forest, and Random Tree applied to agricultural data .
The main objective of the research is to help doctors and patients to predict diabetes as soon as possible through machine learning techniques. The prediction of the diagnosis of diabetes is based on these algorithms of machine learning. K Nearest Neighbours (KNN), Naive Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Logistic Regression (LR), and Random Forest (RF) are prominent in these algorithms .

3. Classification Methodology

The most important method of supervised learning is Classification; classification is used to analyze data in supervised learning. The classification method is used for the prediction. The classification methods or algorithms divide the data sets into categories . In supervised learning, input and output data are already provided to the machine, also called training data or labelled data [17]. According to the data, the machine gives its output, which completely depends on the quality of the training data. If the data quality is good, the machine's output will also improve. When a new input is given to the machine,

it will output only according to its previous experience and data. In this algorithm, the machine applies what is written in its past to the new data using the labelled example to predict future events.

Supervised learning can be understood with an example. With a supervised learning algorithm, a computer program or Machine learning model is given certain datasets (such as apple colour is red, weight 20 gram, round shape, and height 5 cm). Using these datasets, the computer program predicts the outputs. The two types of datasets are given to the Machine learning model. The first is feature data and the second is labelled data. An algorithm is used to train the ML model and determine what kind of output to predict. Suppose you have to train the ML model on how to recognize mangoes. So far, you have to use a supervised learning algorithm. Now you have to tell that ML model what the mango looks like; for this, you will give feature data to the ML Data(such as the mango colour is yellow, the shape is round, and the taste is sweet). When you have to give feature data to the ML model, using supervised first, it has been told that if something is an input appearance, colour is yellow, the shape is round, and taste is sweet. The output prediction will be mango because you used feature and label data to train the machine learning system. So the algorithm is called a supervised learning algorithm.

Table 2. Supervised Learning Dataset-II

Feature Data		**Label Data**
X1	**X2**	**Class**
10	100	Square
2	4	Root

If the value of X1 is 10 and X2 is 100 given to the above data set (Table 2), then the prediction of the output will be square.

3.1. K-Nearest Neighbor Classification Algorithm

In classification, the output variables are called labels

or categories because this output is in categorical for, means output gets divided into two or more categories such that an email can be categorized in two ways (spam or not spam), categorizing age group into three parts, children, young and old. K-Nearest Neighbour is the simplest machine learning algorithm based on supervised learning. The KNN algorithm is mostly used in solving the classification problem. The KNN classification algorithm need can be understood from this example. Suppose we have to be given two categories, i.e., category A and Category B, and have a new data point X1 and want to see which data point will lie in which of these categories. To solve this problem and need a K-NN algorithm.The category or class of any particular dataset can be easily solved with the help of the K-NN algorithm.

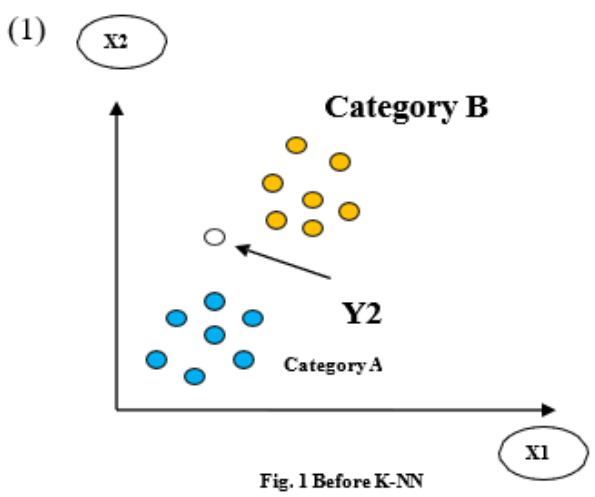
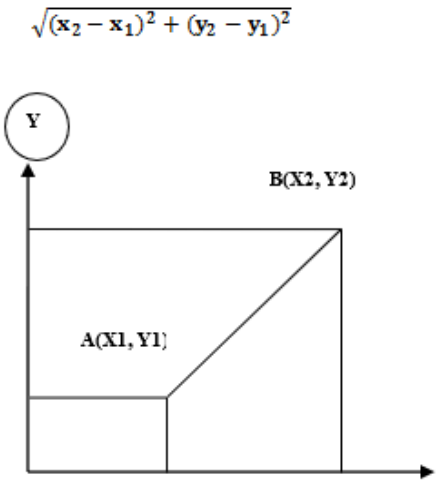


Fig. 1 Before K-NN



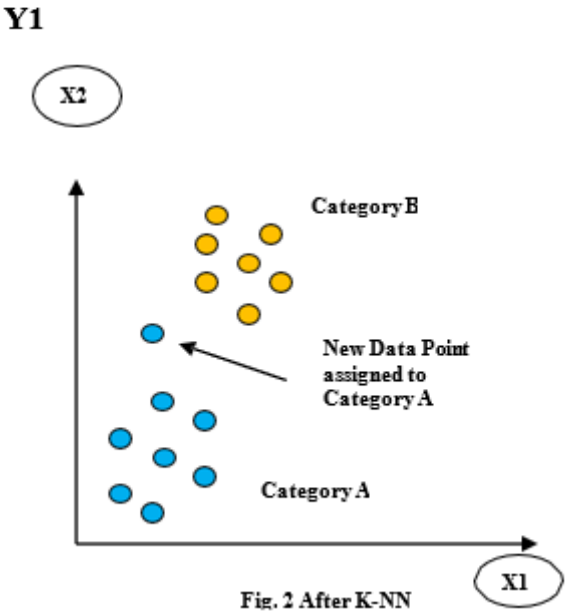


Fig. 2 After K-NN

4. Experiment and Evaluation

In the simplest terms, the result of the K-NN algorithm is as follows. The paper given a student dataset below in which maths and computer science marks are given, and the result based on marks is also associated.

The working steps of the K-NN algorithm can be understood as follows.

Step-1: First of all, selects the number k of the neighbours.

Step-2: Euclidean distance has to be calculated for the K number of neighbours.

Step-3*:* Take the K nearest neighbours per the calculated Euclidean distance.

Step-4*:* Among these K neighbours, count the number of data points in each category.

Step-5: New data points will be assigned to the category for which the number of neighbours is maximum.

Step-6: In this way, the K-NN classification model will be ready.

The formula for finding the Euclidian distance is as follows.

Euclidian distance between A1 and B1=

Now have a query that says student X that X student got 6 marks in maths and 8 marks in a computer science subject.

Query ⇨ Student (X) (Math=6, Computer Science=8)

From the above query, determine whether this X student passed or failed. To find this, I have to use the K-NN algorithm's help. Suppose taken the value of K as 3. It means that whatever the resulting value will come in this query should be very much to 3 neighbours. The paper has to find that the X query topples nearest to the neighbours of the tuple in the dataset. For this, take the help of the Euclidian distance formula.

D=(2) $\sqrt{(x_{02} - x_{A1})^2 + (y_{02} - y_{A1})^2}$

Where x_{02} is the first observed value, y_{02} is the second observed value of query tuple (new data point), and x_{A1} is the first actual value, y_{A1} is the second actual value of query tuple (new data point). From a distance calculated below and find the three nearest neighbours to X. For the given X 4, 5, and 7 are the nearest neighbours, so consider these three neighbours and not consider those neighbours (3, 6) who are distant neighbours.

$\sqrt{(6-4)^2 + (8-3)^2} = 5.38$ (3)

$\sqrt{(6-6)^2 + (8-7)^2} = 1$ (4)

$\sqrt{(6-7)^2 + (8-8)^2} = 1$ (5)

$\sqrt{(6-5)^2 + (8-5)^2} = 3.16$ (6)

$\sqrt{(6-8)^2 + (8-8)^2} = 2$ (7)

The result of the second (4) equation is passed, the third (5) equation is passed, and similarly, the result of the fifth (7) equation is also passed in the dataset (Table 3).

The result, which is in the dataset, shows two things pass and fail. Either student X could have passed or the failure. In the context, I got 3 (Three) pass values from the dataset (Table 3) based on the nearest neighbours value and got a 0 (Zero) fail value from the dataset.

5. Conclusion and Suggestions for Future Work

In the paper, supervised learning and the K-NN classification algorithm has been presented very easily. The problem faced in understanding the supervised machine learning and K-NN algorithm has been solved in this paper by easy example methods. The paper mentioned that supervised learning is that type of machine learning which works on labelled data. This type of machine is of a very basic type. In this, the data parameters have to be defined with utmost care. In supervised learning, a small piece of data is trained and applied to large-scale data. In this paper, a very small dataset has been used to explain the K-NN algorithm so that every user reading this paper can easily know the working of the K-NN classification algorithm. In this paper, the K-NN classification algorithm is applied to the student dataset in which the students' marks are given, and their result is linked with the labeled data based on the pass and fail. Now give the marks of a new student to the dataset. Then,

the K-NN algorithm will determine whether the new student will pass or fail. The evaluation and experimental results of the K-NN classification algorithm are explained with a small dataset and effective technique. For easy understanding of the K-NN algorithm, the K size is 3 (very small) in the K-NN classification algorithm. In the paper, the complete dataset is not given, which can show the prediction of the result. In the paper, given the query of only one student X in the dataset, predict the result in which their math and computer science marks are given. The paper implemented the K-NN classification algorithm on labelled data in supervised learning. In the future, work on the unsupervised learning and now work on the unlabeled data instead of working on the labelled data.

References

1. N. Kola, M. Kumar, "Supervised Learning Algorithms of Machine Learning: Prediction of Brand Loyalty," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 8, no. 11, pp. 3886-3889, 2019.
2. I. Muhammad, Z. Yan, "Machine Learning Approaches: A Survey," *ICTACT Journal on Soft Computing*, vol. 5, no. 3, pp. 946-952, 2015.
3. A. Juyal, C. Pande, "Performance Analysis of Supervised Machine Learning Algorithms on Medical Dataset," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, no. 6, pp. 1637-1642, 2020.
4. J. Sun, W. DU, "A Survey of KNN Algorithm," *Information Engineering and Applied Computing*, vol. 8, no. 11, pp. 1-10, 2018.
5. B. S. Imandoust, M. Bolandraftar, "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background," *Int. Journal of Engineering Research and Applications*, vol. 3, no. 5, pp. 605-610, 2013.
6. K. Alkhatib, K. Najadat, "Stock Price Prediction Using K-Nearest Neighbor (kNN) Algorithm," *International Journal of Business, Humanities and Technology,* vol. 3, no. 3, pp. 32-44, 2013.
7. B. Mahesh, "Machine Learning Algorithms- A Review," *International Journal of Science and Research (IJSR)*, vol. 9, no. 1, pp. 381- 386, 2018.
8. N. Burkart, F.M. Huber, "A Survey on the Explain ability of Supervised Machine Learning,' *Journal of Artificial Intelligence Research*, vol. 70, no. 6, pp. 245-317, 2019.
9. K.V. Reddy, R.U. Babu, "A Review on Classification Techniques in Machine Learning," *International Journal of Advance Research in Science and Engineering*, vol. 7, no. 3, pp. 40-47, 2018.
10. V. Bijalwan, V. Kumari, "Machine Learning Approach for Text and Document Mining," *Research Gate,* vol. 4, no. 5, pp. 1-9, 2019.
11. A. Soofi, A. Awan, "Classification Techniques in Machine Learning: Application and Issues," *International Journal of Basic & Applied Sciences*, vol. 4, no. 13, pp. 459-465, 2017.
12. A. Chaudhary, A. Kolhe, "Machine Learning Classification Techniques: A Company Study," *International Journal on Advanced Computer Theory and Engineering (IJACTE*), vol. 2, no. 4, pp. 21-25, 2013.
13. J. Sreedevi, J. Bai, "Newspaper Article Classification using Machine Learning Techniques," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 9, no. 5, pp. 872-877, 2019.

14. K. Karthikeya, K.H. Sudarshan, "Prediction of Agriculture Crops using KNN Algorithm," *International Journal of Innovative Science and Research Technology,* vol. 5, no. 5, pp. 1422-1424, 2019.
15. N. Krishnamoorthy, N. Umarani, "Diabetes Prediction in Healthcare Using KNN Algorithm," *International Journal of Multidisciplinary Educational Research,* vol. 10, no. 5, pp. 36-39, 2021.
16. M. Suyal, P. Goyal, "An Efficient Classifier Model for Opinion Mining to Analyze Drugs Satisfaction Among Patients," *Communications in Computer and Information Science (CCIS),* Springer Nature Switzerland AG, vol.1591, pp. 30-38, 2022. https://doi.org/10.1007/978-3- 031-07012-9_3
17. M. Suyal, P. Goyal, "A Two-Phase Classifier Model for Predicting the Drug Satisfaction of the Patients Based on Their Sentiments," *Communications in Computer and Information Science (CCIS),* Springer Nature Switzerland AG. vol. 1591, pp. 79-89, 2022. https://doi.org/10.1007/978-3-031-07012-9_7ss