IRIS DATA ANALYSIS USING K NEAREST NEIGHBOUR (KNN) ALGORITHM

Salman Ahmad Khan
Department of Computer Science
Bahria University Islamabad, Pakistan
MS Data Science (2nd Semester)
Reg No: 01-249212-013
E-mail: Salmannahmed123@gmail.com

***Abstract* – Machine Learning is a data-driven approach using a different set of algorithms or models to predict the outcome in the form continuous or categorical (classification) form. Machine learning is further classified into three main categories, Supervised, Unsupervised, and Reinforcement learning. In this project, we have worked on supervised learning having labeled data set of IRIS. We have taken the iris dataset and used K-Nearest Neighbors (KNN) classification Algorithm. Our purpose is to build a model that is able to automatically recognize the iris species and the analysis of parameters and hyper-parameter tuning such as value of K, distance function, data splitting ratio and encoding methods in the case of the KNN algorithm using the IRIS dataset. Tools used for this project are VSCode using Python and different dependencies or libraries used such as Numpy, Pandas, Matplotlib, and machine learning library Scikit-learn.**

***Keywords* – Supervised learning, Classification technique, KNN, K values, Machine learning, Numpy, Pandas, Matplotlib, Scikit-learn, VScode, Jupyter, Anaconda.**

I. INTRODUCTION

Machine learning is used as a predictive approach to unseen data. In machine learning, the model or algorithm first learns how to perform the task by training the dataset, then testing [1]. In the Supervised learning approach, both the input and output data will be given to the model to learn from the data and predict continuously. for the prediction of continuous output problems, we used the Regression model while for categorical or classification problems, we use Classification algorithms [2]. In this project, we are using classification-based supervised learning to predict the label group of the IRIS data. KNN (instance base or lazy learner) is one of the simple algorithms that are used for classification based on different similarity measures such as Euclidean, Manhattan, etc. [3] The main objective of this work is to analyze the parameter and hyperparameter such as identifying the optimum number of K for the IRIS dataset, the splitting ratio deviation, the effect of using different Encoding types, and lastly the distance functions, which affects the model accuracy.

The Major steps involves in this project include of machine learning that are:

1. Data Collection
2. Data prepossessing and EDA
2. Choose algorithm(in this case KNN)
3. Creating object of the model
4. Train the model by training dataset (using different ratio)
5. Making prediction on unseen
6. Evaluation of the model

II. DATASET

In this project, the dataset is taken from the online source Kaggle, which is an open-source repository. The dataset contains 150 data points in which three different classes Setosa, Virginica, and Versicolor present. Each class has the same ratio of presence (50/150, 1/3). Total of 5 attributes in the dataset including petal length, Petal width, Sepal length, Sepal width, and Species. Out of these 5 attributes, 4 are in numerical attributes while the predictive output attribute "Species" are in categorical form. Figure.1 shows the exploration of the dataset and the relationship among all attributes and shows the dispersion of the sample point of the dataset. Fig.2 shows the overall data distributions.

Fig. 1. The pair-plot show the relation among the attributes with respect to the three classes Setosa, Virginica and Versicolor

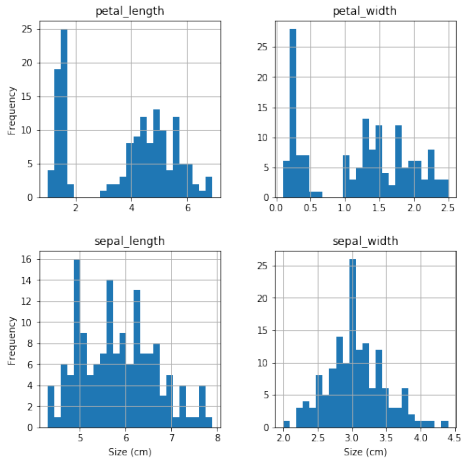

Fig. 2. Data Distributions

III. IMPLEMENTATION

We used Anaconda software(Visual Studio Code) for data manipulation, wrangling, and built a model using different libraries of Numpy, Pandas, Matplotlib, Seaborn, and Scikit Learn. we tried multiple different splitting ratios of training and testing data to obtain the best value of K, accuracy as well as the effect of these factors on the overall model and accuracy. The standard Scalar approach has been used for the normalization of the data.

IV. EVALUATION

The evaluation matrices depend on the chosen number of neighbor values for voting (K value) and the splitting . we get different values of accuracy in different K values, Similarity function, data splitting ratio, explained in below section.

V. DISCUSSION AND ANALYSIS

The main objective of this project is an in-depth analysis of the KNN model parameters and hyper-parameter in the case of the IRIS dataset, also from a general point of view. we built a KNN model discussed above using changes in the number of K values (Neighbour number for voting), the splitting ratio, different encoding techniques, and Distance functions. After in-depth literature review and implementation, we briefly explaining the factors one by one ;

1) Different values of K: The value of is actually the number of neighbors which u choose to be able to cast vote for your data, which have not been assigned any class yet. In this case of binary classification, the K value must be odd in because if we assigned K=2 or K=4, then the model equally include the neighbour data point one from each class in case of K=2 and two from each class in case of k=4, which will not perform accurately classification [4].
In IRIS dataset case, the accuracy directly decreases, when we increase the value of K. The cross-validation (10 fold) approach is used to get the misclassification or bad classification plot in Fig 3, as mentioned below accuracy decreases when we increase the values of K.

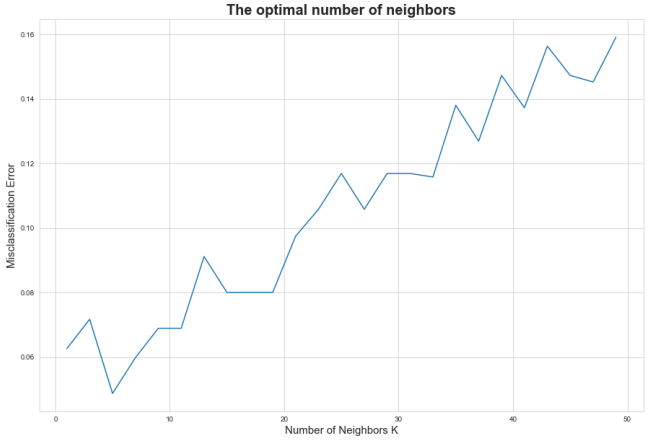

Fig. 3. shows the Misclassification graph

Based on this result, we clearly choose the k values is 3 based on the accuracy.

2) Different training and test split ratios: We used 4 different training testing ratios applied to the iris dataset and check out the effect on the model based on accuracy as well as confusion matrices.

TABLE I
Different Training and Testing Ratio

Training Ratio	Testing Ratio	Accuracy	Optimum Value of K
80	20	1.0	1
70	30	0.97777	3
60	40	0.9833	3
50	50	0.9866	8

Based on the above result, we conclude that the data splitting ratio will affect the accuracy of the model with an optimum value of K. The Value of K is automatically changed with respect to the splitting ratio as mentioned above. In Iris's case, we choose the K value based on accuracy.
In Fig 4, the confusion matrices for each split ratio and other classification reports of the Iris dataset in the case of KNN.

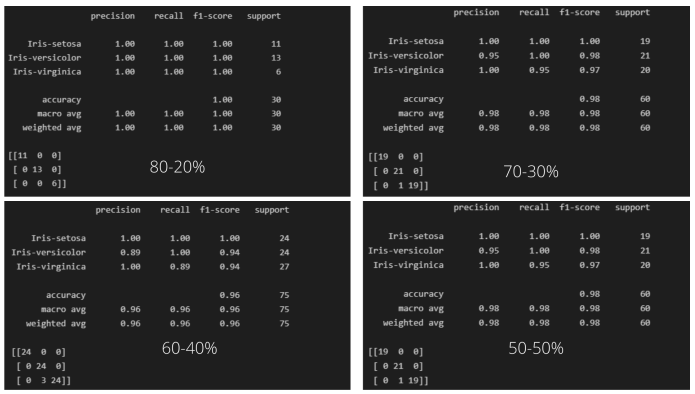

Fig. 4. Confusion Matrices and Classification Report for Each Ratio

3) Different distance functions: In this project of Iris data, the Minkowski function is used as a similarity distance

function, which is built-in of the model shown in fig.5

```
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
                     metric_params=None, n_jobs=None, n_neighbors=3, p=2,
                     weights='uniform')
```

Fig. 5. Minkowski Distance Function

4) Different label encoding techniques: Llabel encoding is a technique in which the categorical or label values are converted into a numerical value (starting from 0) to machine-readable form [5]. In case of this project, there is no need for encoding in this project the attributes were already in numerical form.

VI. CONCLUSIONS

After completing this project by implementing the as well as via literature review, we conclude that, in the K Nearest Neighbour model, the value of K plays a vital role and impacts the model performance and accuracy. choosing the optimal value of k is based on the nature of the data as well as the data splitting ratio.

In the case of this project, we get different k values using different data split ratios and get different accuracy and classification rate. For K= 3, and the data split ratio is 70-30, we get the accuracy of 0.97777. For K=1, the 80-20 split ratio has an accuracy rate of 1.0, which is the highest accuracy in the remaining values.

ACKNOWLEDGEMENTS

We get help in implementation from different literature sources and also borrow some code from online sources Kaggle [4], Towadsdatascience, medium and geek for geeks [5] [6] [2].

REFERENCES

[1] I. El Naqa, M. J. Murphy, "What is machine learning?", *in machine learning in radiation oncology*, pp. 3–11, Springer, 2015.

[2] R. Sathya, A. Abraham, *et al.*, "Comparison of supervised and unsupervised learning algorithms for pattern classification", *International Journal of Advanced Research in Artificial Intelligence*, vol. 2, no. 2, pp. 34–38, 2013.

[3] D. Cheng, S. Zhang, Z. Deng, Y. Zhu, M. Zong, "kNN algorithm with data-driven k value", *in International Conference on Advanced Data Mining and Applications*, pp. 499–512, Springer, 2014.

[4] Yousefnami, "Why does increasing K decrease variance in KNN?", , Nov 2021, URL: https://towardsdatascience.com/why-does-increasin

[5] "ML: Label encoding of datasets in Python", , Sep 2021, URL: https://www.geeksforgeeks.org/ml-label-encoding-o

[6] Lnbalon, "Iris dataset EDA and classification analysis", , Aug 2017, URL: https://www.kaggle.com/code/lnbalon/iris-dataset-