Tec Empresarial
P-ISSN: 1659-2395; E-ISSN: 1659-3359

A CLASSIFICATION RESEARCH USING K-NEAREST NEIGHBOUR PIMA DIABETES DATASET

Dr. Monika Saxena (Corresponding Author)

Associate Professor, Bennett University, monika.saxena@bennett.edu.in

Keshav Goyal

MBA pursuing, m22mbag0056@bennett.edu.in

Prashant Tiwari

MBA Pursuing, m22mbag0022@bennett.edu.in

Abstract:

The procedure of extracting information and knowledge from enormous volumes of data is called data digging. Data analysis is the primary use case for data mining. Numerous techniques, including joining mining, regression, forecast, classification, grouping, etc., are used in data mining. In order to use a model to detect unknown objects or patterns whose class designations are unclear, sorting is expressed as an operation of exploring a cluster of models that justify and changes data-classes and concept. One type of supervised learning problem is classification. This means that the problem of classifying data patterns in machine learning is to separate out individual arrangements from a group of arrangements based on their features and determine which patterns belong to which class. Different classifiers can be used to classify patterns. A classifier is a computer programme that takes in a arrangement or data point's feature vector and delegates it to one of a number of predetermined classes. K-Nearest Neighbour is employed to classify patterns. This study presents the certainty of k-nn using 3-datasets from UCI ML cohort, with a focus on the data mining classification technique. This paper's primary objective is to present a classification of the diabetes dataset and compare between the different nearest neighbours sets for data mining. A straightforward yet effective method for assortment in research is the k-nn classifier.

Keywords: KNN, Classification, Diabetes, Supervised learning, Data mining

Introduction:

Diabetes has emerged as a significant global health concern, affecting number of individuals and posing substantial challenges to healthcare systems. Prompt identification and efficient treatment of diabetes are vital in reducing its complications and enhancing patient outcomes. Machine learning algorithms have made significant progress in healthcare, particularly in the identification of those who are at risk of acquiring diabetes. Out of these models, the K-Nearest Neighbors (KNN) classifier is notable for being a straightforward yet effective method for predictiveanalytics.

The purpose of this introduction is to give a broad understanding of how the KNN classifier might be used for predicting diabetes. We will explore the basic principles of diabetes, the importance of predictive analytics in its treatment, and the underlying principles of the KNN algorithm. In addition, we will examine the utilization of KNN in the prediction of diabetes, its benefits, constraints, and possible directions for future investigation.

Healthcare professionals can utilize the KNN classifier to effectively utilize data-driven analysis in order to swiftly identify patients who are at risk of diabetes. This allows for preventive treatments and the implementation of individualized healthcare programs. In conclusion, incorporating KNN-based prediction models into clinical practice shows potential for increasing preventive healthcare measures and improve the quality of life for persons impacted by diabetes.

All that data mining is an information repository that allows enormous amounts of data to be stored, reborn, and stored in databases and data depositary. Due to its incorporation of various integration methods from diverse fields such as high-performance computing, database technology, ML, pattern recognition, statistics, neural networks, information retrieval, and data visualization, it is also referred to as Knowledge Discovery in Databases (KDD). The data mining system architecture can be enhanced by incorporating key components such as a database and a data warehouse. Here, somehow server is responsible for fetching the user-data based on the knowledge base for the user's data-mining inquiry.

In ML, the first thing to be evaluated is the prediction certainty of differentiating algorithms that are extracted from experimental data, or examples. However, interpretability or transparency of a classifier is often important in practice. The exactness of k-nn classifiers in categorizing a database is examined in this study. Because of their significant role in exploratory pattern analysis, many amazing and varied classification learning algorithms, including Support Vector Machine (SVM), k-nn, and Naive Bayes Classifier which are classified as project associated algorithms.

The various categories of data digging strategies are as follows:

Classification:

This trail involves classifying the given data occurrence into one of the previously determined target classes. A case in point is classifying a customer in a credit-card databset as either a reliable or a outlier based on a variety of geographic and past expense criteria.

Clustering:

The purpose of k-NN is to determine how many closest neighbors should be examined when considering a value by computing the nearest neighbor on different computations of k. The basic idea behind a k-NN is to calculate a distance matrix to determine the separations between data points. The algorithm then looks through the dataset to find the K nearest answers. Here are a few instances of distance metrics: The distance in space between two points is called the Euclidean distance. Finding all the "radius" between every data point and reference point in the data-set is the first step in the k-NN algorithm. These distances are sorted in the second step, after which the k closest objects are selected to carry out the third and last step of categorization. Lastly, k-nn finds the k closest points among N points to a data-point from the 2-dimensional function space. The number of neighbors from a dataset that are considered is k.

Objective:

In the space of data diging, most classification experiments has been observed with respect to different datasets. The purpose of this paper is to analyze the PIMA diabetes dataset over a range of k-values while taking data normalization into consideration, and to write about different data classification techniques related to kNN classification.

Literature Review:

This section presents a sense of the literature review, containing technical paper reviews on the k-nn classification techniques as applied to various functions. It also summarizes the current data-mining research being undertaken on various functions. Following that, the given conclusions and disadvantages were revealed from another research:

The writer concentrates on the k-nn technique for differences in this work [1]. The author tested this model with various matrix and the three criteria for classification (random, consensus, and majority) in light of parameter “k”. According to the results, the k-nn technique is applied with each form of distances ie Euclidean distance and Manhattan. These distances help with performance and categorization, but they consume time. As an outcome, they devise 2 forms of stretch that produce the best results (98.70 & 98.70; thus).

In this paper [2], the researcher noticed and focused on the choice of k values, and the experiment results show that the suggested approach consistently outperforms other classifiers over a wide range of k, and its efficacy has been demonstrated with good performance.

The KNN classifier is one of the most utilized neighborhood classifiers in this pattern recognition work [3]. It does, however, have substantial downsides, such an immense computational intricacy, complete dependence on training data, and uniformity of weight over classes. This project suggests a novel strategy for increasing k-nn classification performance considering Genetic Algorithms (GA) to address the indicated issue.

Methodology:

Any data-mining procedure cannot be performed directly on the source dataset. A data set is required to prepare for the procedure. Unexpected numbers, missing values, and data dimensions that are excessively large with undesirable qualities or attributes are examples of impurities in data collected from various sources [4].

Before the data can be used, these impurities should be removed, and data should be pre-processed. Some of the preprocessing approaches discussed in this section were used in this project.

A. Normalization

Some attributes in a dataset may have values in the upper numeric range, while others may have values in the lower numeric range [5]. Some classification methods, such as neural networks & its variations, distance valuation, requiring that the values of all characteristics be few in a range [6]. For neural networks or k-NN, for example, input values can be -1, 0, or +1.

B. k-nearest neighbor Classification

The distance between data-points in a training dataset can be used to classify them, which is a simple yet successful strategy. We can use a variety of measures to compute the distance, which will be explained further below.

Dataset used : PIMA Data set

Table II shows how the pima Indian Diabetes database from the UCI ML warehouse is used to create a real valued forcasting between 0 & 1.

This was transformed to a binary choice using a threshold of 0.448. Class value 1 means "Tested positive for diabetes." It was founded by the National Institute of Diabetes and Digestive and Kidney Diseases.

Name of data-set	Instances	Attributes	Class Levels
PIMA	786	9	2

Table -1: Dataset Information

Experimental Results:

The experimental data collection includes some P.i.m.a. data at the 2-class level. Performance either differs or stays the same depending on the data. Here, both nominal and numerical data are used in the collection and evaluation of data of various sizes and sorts. k-nn techniques are used to check the exactness of data-sets. Using the Python, the implementation is finished. This approach uses the Pima dataset and the k-nn model to investigate, how this methodology helps anticipate foreign class values & calculates prediction closeness using the confusion-matrix.

Dataset Name / k=	3	5	7
p.i.m.a.	77.60	76.04	77.08

Table-2: KNN performance with various k-values expressed as a percentage

When the k-value rises, efficiency rate of KNN first falls and then rises. This is because class borders become less pronounced with higher values of “k”, which also lessen the impact of fault on classification. The max certainty is attained instantly while “k” is still a minute numbers, and it is subsequently diminishing; however, for Pima, the maximum accuracy is obtained gradually for various k-values.

Conclusion & Future work:

The K-NN classifier is a popular neighborhood classifier in data digging and pattern identification. It does, however, have some downsides, including high computing intricacy, complete dependance on the training dataset, and no-load fluctuation among collection. The accuracy in various k selections is the emphasis of this strategy to improve classification performance.

However, when using alternative k values Each dataset's accuracy increases, decreases, and then improves again when the numbers are odd. The implementation makes k-NN a very effective classifier. It produces positive outcomes given the magnitude of the data collection rises. In order to enhance the effectiveness of the classifier model, future research might combine feature selection techniques with a variety of datasets.

References:

1. Suguna, N., & Thanushkodi, K. (2018). An Improved k-Nearest Neighbor Classification Using Genetic Algorithm. IJCSI International Journal of Computer Science Issues, 7(4), 2.

2. Gou, J., Du, L., Zhang, Y., & Xiong, T. (2012). A New Distance-Weighted k-Nearest Neighbor Classifier. Journal of Information and Computational Science, 9(6), 1429-1436.

3. Bagui, S. C., Bagui, S., & Pal, K. (2003). Breast Cancer Detection using Nearest Neighbor Classification Rules. Pattern Recognition, 36, 25-34.

4. Mohapatra, D., Tripathy, J., Mohanty, K. K., & Nayak, D. S. K. (2021). Interpretation of Optimized Hyper Parameters in Associative Rule Learning using Eclat and Apriori. In 2021 5th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 879-882). IEEE. DOI: 10.1109/ICCMC51019.2021.9418049.

5. Mohapatra, D., Tripathy, J., & Patra, T. K. (2021). Rice Disease Detection and Monitoring Using CNN and Naive Bayes Classification. In S. Borah, R. Pradhan, N. Dey, & P. Gupta (Eds.), Soft Computing Techniques and Applications (Advances in Intelligent Systems and Computing, vol 1248). Springer, Singapore.

6. Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical Pattern Recognition: A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(1), 4-37.