Classification with K-Nearest Neighbors Algorithm: Comparative Analysis between the Manual and Automatic Methods for K-Selection

Tsvetelina Mladenova[1], Irena Valova[2]
Department of Computer Systems and Technologies, University of Ruse, Ruse, 7017, Bulgaria

***Abstract*—Machine learning and the algorithms it uses have been the subject of many and varied studies with the development of artificial intelligence in recent years. One of the popular and widely used classification algorithms is the nearest neighbors' algorithm and in particular k nearest neighbors. This algorithm has three important steps: calculation of distances; selection of the number of neighbors; and the classification itself. The choice of the value for the k parameter determines the number of neighbors and is important and has a significant impact on the degree of efficiency of the created model. This article describes a study of the influence of the way the k parameter is chosen - manually or automatically. Data sets, used for the study, are selected to be as close as possible in their features to the data generated and used by small businesses - heterogeneous, unbalanced, with relatively small volumes and small training sets. From the obtained results, it can be concluded that the automatic determination of the value of k can give results close to the optimal ones. Deviations are observed in the accuracy rate and the behavior of well-known KNN modifications with increasing neighborhood size for some of the training data sets tested, but one cannot expect that the same model's parameter values (e.g. for k) will be optimally applicable on all data sets.**

***Keywords*—Machine Learning; KNN; classification**

I. INTRODUCTION

K-Nearest Neighbors (KNN) is a simple and easy-to-use supervised machine learning (ML) algorithm that can be used for solving classification problems in different domains - education, healthcare, livestock and crop production, administration, production, transport, etc. [1, 2, 3]. KNN is an extension of the idea of the nearest neighbor method introduced by E. Fix and J. Hodges in 1951 [4, 5, 6], which is based on the calculation of the distance between an unlabeled sample and the nearest sample - neighbor, from the training set. This extension was proposed by Cover and Hart [7] who added the parameter k representing the number of nearest neighbors to be considered for classification. For k=1, it can be said that the nearest neighbor algorithm, NN, is considered, and for k >1, KNN. The KNN algorithm is a relatively simple example of a non-parameterized classifier, easy to understand and implement [8, 9, 10]. The main stages of the KNN algorithm are three,(Fig. 1).

1) Calculate the distances from the unlabeled sample to each sample in the training set.

2) Processing the calculated distances and selecting the k neighbors [11] that will form the neighborhood.

3) Determining the class to which the unlabeled sample belongs - the classification stage.

The main challenges specific to the nearest-neighbor method can be divided into the following groups:

- Determination of the k parameter - the size of the neighborhood depends on it, which is decisive for the classification. It has been proven that the parameter is sensitive and at the same time important for the degree of efficiency of the model.
- Choosing the neighbors - the calculation of distances between samples is traditionally carried out using the Euclidean distance. Choosing the neighbors that will shape the neighborhood is no less important a step than choosing its size.
- Classification rule – the classification stage is the last stage in which the decision is made as to which class to classify the unlabeled sample. An inappropriate classification rule could mean an incorrect classification of the unlabeled sample.

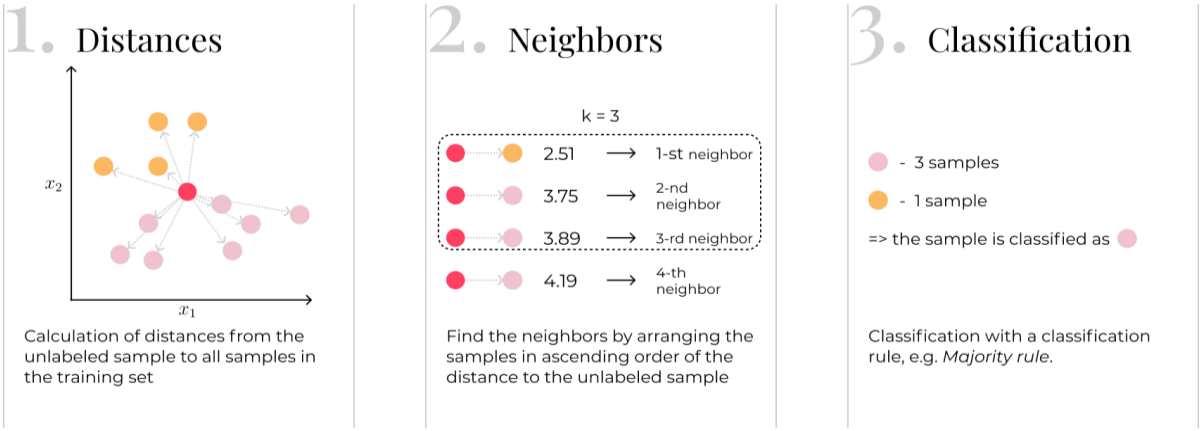

Fig. 1. Steps in classification with the method of nearest neighbors.

II. Methods for Calculating the k Value

Determining the value on which the neighborhood size will depend is a problem addressed by many scientific studies. The parameter is particularly sensitive, since at a value larger than necessary, the model's efficiency decreases. A smaller value, respectively, will make the model inaccurate. Additionally, a smaller value will mean that noise in the data will have a greater impact on the classification, and a larger value will mean that additional computing resources are required.

There are different approaches to choosing k. One approach uses a pre-definition of a single value, regardless of the particularities of the training set (type, size, subject area, etc.). The other approach is to determine the value of each set. It is used more, and with it, the value is most often found by the elbow method - the training set is trained with a series of k values and the one is found where the effectiveness of the model begins to decrease.

A third method is the *m-fold cross-validation* [12]. In this approach, the training set is divided into m disjoint sets. The cross-validation method is then applied to each set. Finally, the k value of the subset that gave the best results is taken as the k value of the entire training set.

A disadvantage of this approach is that the selection of the k value does not consider the distribution of the data [13]. Such a finding of the optimal value of k is called "manual search" and requires in-depth analysis by a person who understands and knows machine learning. In addition, such methods often require training the model several times, which means additional time for preparation and calculations.

One of the more automated methods for determining k is the dynamic one, where various approaches are applied to analyze the training set, feature distribution, class information, anomalies, etc., to find the best value.

It is important to note that no matter how the value for k is calculated, it does not guarantee that the model will have high efficiency or that this value will not change. Adding just one object or sample to the training set, or altering any step of the algorithm, may result in an inefficient neighborhood and an inapplicable value of k.

In [14] and [15] an approach is used where the value of k is determined automatically and can vary depending on the size of the set. According to the authors, this approach is suitable for reducing the set size, lowering the classification time, and is a suitable replacement for the traditional KNN for working with big data. The main idea on which the algorithm is built is to find local neighborhoods consisting of samples belonging to the same class. Thus, one can find the number of samples in each local neighborhood, as well as the similarities between the most distant sample in the neighborhood and its center.

By taking these samples as well as the centers of each neighborhood, the size of the training set is reduced while preserving its distribution. Each new, unlabeled sample is compared to the new, reduced set. In this way, there is no need to explicitly define a value for k. It should be noted that the authors propose the text classification algorithm where the classes from which the training set is composed are sufficient for the classification to be accurate.

In [16], [17], the number of neighbors on which the classification will depend is determined by the number of possible classes (1). Through experimental studies, it is proven that this determination of neighbors is appropriate when the boundaries between classes in the data are not clear enough and are face-to-face, so-called overlapping data.

$$k = c + 1 \quad (1)$$

Where:

c – the number of unique classes.

An approach with a dynamic selection of a value for k is also considered in [18], where the number of neighbors depends on the distribution of samples, relative to the classes in the training set. A Chinese text classification method is proposed, and it is proved through an experimental study that the modification achieves good results in classifying documents having classes with few samples.

Another approach used is equation (2).

$$k = \sqrt{N} \quad (2)$$

Where:

N – the number of samples in the training set.

A common practice of ML researchers is to apply the following rule – when the number of unique classes in the training set is even, the value of k is odd to avoid equality between neighboring classes. This is not a guaranteed approach, as samples from only one class or an equal number of samples from two or more classes can fall into the neighborhood, but it is an additional step to improve the model's performance.

The automated way of obtaining the neighborhood size has proven effective when the choice of a value for k needs to be made quickly and without additional steps before training the model. The results obtained with an automatically calculated value for k are comparable to the results obtained using some of the manual methods.

The main advantage of this approach is that it provides an additional step of automation and makes the nearest-neighbor method more accessible to users who do not understand ML. This is a solution that can be implemented in an algorithm aimed at automating learning processes and providing understandable results.

III. Neighborhood Size and Neighbors' Selection

The neighborhood size, in the nearest neighbor method, is determined by the chosen value for the k parameter. If the value is too small, noise in the data may have too much influence on the classification. Too large a value means more computation time and resources. The exact value of k is difficult to determine and depends on the characteristics of the training set. Taking into account the fact that the sought-after solution has to handle most small datasets, the possibility that the value of k can be further restricted should also be taken into account.

In Fig. 2, Fig. 3, and Fig. 4 can be seen how increasing the number of neighbors leads to a change in the accuracy of the model. The figures show the decision boundaries for different values of k.

At k=5, Fig. 2, the overfitting of the model is noticeable, i.e. the model tries to classify as many "single" cases as possible, making it unstable and unreliable.

The increase in the number of neighbors, for example, k=15, Fig. 3, results in a normalization of the classification, although areas where it can be said that there is overfitting are still observed.

Fig. 2, Fig. 3, and Fig. 4 visualize a case where the data in the training set is unbalanced and the boundaries between classes are not clearly defined. For the set in the figures, the traditional nearest neighbor method achieves a classification accuracy of 73% at k=5, 73% at k=15, and 76% at k=20. The increase in accuracy rate is a result of expanding the neighborhood. When the training set is large enough to allow this, this is not a problem, despite the additional computational resources required for the larger neighborhood. However, when the set is on the order of 25 samples, k=20 will mean that in practice the entire set will be used for training.

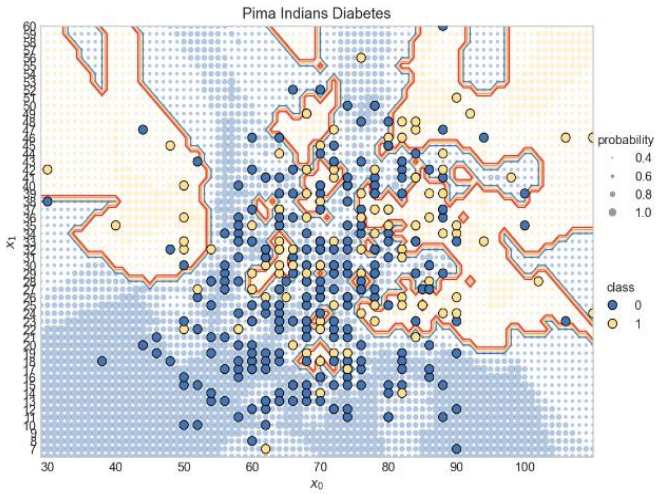

Fig. 2. Decision boundaries for k=5.

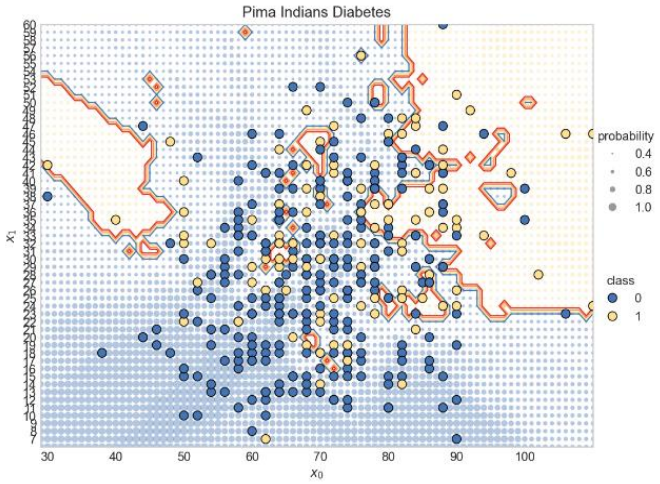

Fig. 3. Decision boundaries for k=15.

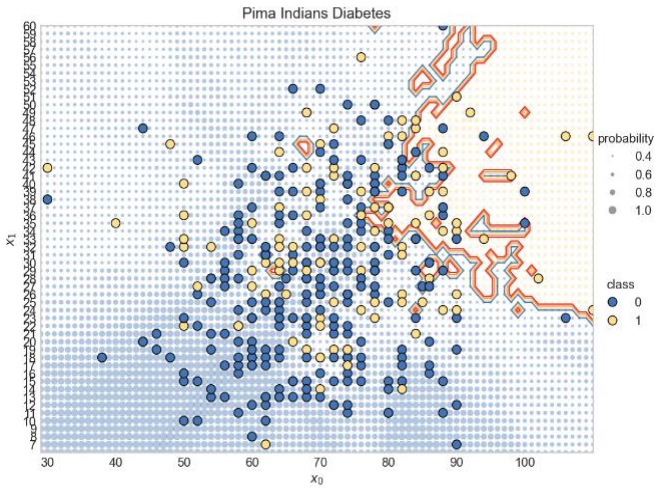

Fig. 4. Decision boundaries for k=20.

IV. Methods for Calculating the Distances between Samples and Forming the Neighborhood

The calculation of the nearest neighbors consists of using a function to calculate the distance between the unlabeled sample and all other samples from the training set. There is no single metric for distance measurement that is applicable in all cases [19], although attempts for finding such a metric are not lacking [20, 21, 22].

According to [20], the neighborhood must meet two criteria: 1) the neighbors are close to the unlabeled sample and 2) the neighbors are symmetrically located around it. The idea of nearest neighbors takes into account only the first criterion [23]. Therefore, the neighborhood may not be symmetrically located if the data in the set of neighbors is inhomogeneous.

A. Nearest Centroid Neighborhood (NCN)

In [23] a new definition of the term "neighborhood" is proposed, which does not require user-defined parameters. The proposed algorithm is called Nearest Centroids or Nearest Centroid Neighborhood (NCN). The basic idea can be described as follows: let be a point whose neighbors are to be found and added to the set of points. The neighbors are such that they fulfill the following conditions: 1) they are as close as possible to and 2) the center between the neighbor and the point is as close as possible [24].

B. Distance Weighted K-Nearest Neighbor (WKNN)

In 1975, S. Dudani proposed a modification of the nearest neighbor algorithm [25]. The author argues that it is logical that the distances between individual neighbors should be proved in the form of a "weight" that varies depending on the distance from the unlabeled sample to its neighbors.

In [26], a modification of S. Dudani's algorithm (DWKNN) is proposed for calculating neighbor weights. The proposed algorithm reduces the weights of the nearest neighbors, except for the first and κ neighbors. The purpose of weight reduction is to limit the influence of anomalies and improve classification accuracy. If the size of the training set is too large compared to the number of neighbors that are taken into account, then the presented algorithm and the majority rule achieve close results. The author claims that his proposed algorithm achieves better results on small and medium-sized training sets.

Choosing an optimal value for the parameter is difficult when using the majority rule due to the nature of data variation and the probability of classification error. The variation may be because the classification is largely due to the number of neighbors and the number of classes. Additionally, as, after a certain value depending on the size of the training set, increases, the probability of error may increase under certain circumstances.

Similar difficulties in choosing an optimal value for k do not exist when using WKNN. Therefore, it can be argued that with the use of the proposed algorithm, the selection of an optimal value for k can be made without worrying about increasing the probability of error. Several experiments proving the truth of the statement have been conducted. Doudani offers two more weighting functions – inverse weight and rank weight.

C. Pseudo Nearest Neighbor Rule (PNNR)

In [27], the proposed algorithm is based on two others – the weight calculation algorithm, WKNN, and the Local Mean Learning (LM) algorithm [28]. Conventionally, the algorithm is called Uniform. A variation of the algorithm called UWKNN is often used, which differs in that one is added to the neighbor's rank.

D. Dual K-Nearest Neighbor

In [29], a weighting function is proposed, aiming to reduce the sensitivity of the parameter by using a combination of the distance of the samples and their rank in the neighborhood.

E. Fuzzy K-Nearest Neighbor (FKNN)

The Fuzzy KNN algorithm [30] is based on the fuzzy set theory first proposed by Zadeh in 1965. [31]. The main idea of the algorithm is the calculation of the degree of belonging of each element of the training set to the classes. The degree of belonging is taken into account when classifying the unlabeled sample.

V. Implementation of the Experiments

Often, programming languages such as Python and R are used to train machine learning models. Over the years, they have proven to be some of the most suitable solutions for artificial intelligence and ML. Some of the most developed libraries are written and adapted specifically for these programming languages.

The scikit-learn library contains many predefined algorithms for ML. In addition to the basic algorithms, the library also allows the application of some modification or set of additional parameters that give some flexibility. In some cases, however, despite the availability of parameters to control a given model, there is a need for the so-called "custom changes". In such cases, the library cannot always be used.

The current study does not use a predefined library. All source code for the classification and evaluation algorithm is written in the Python programming language.

A. Datasets for the Experimental Study

The data sets used for the experimental study are selected to match the data that a small business would have – small training sets, heterogeneous data, data with anomalies, and unbalanced data. A more unusual category of data has also been added to the sets used – those that have the $n \gg p$ problem, i.e. the number of attributes is many times greater than the number of samples. Although rare, such data do occur. A typical example is medical data, where the characteristics describing a sample are much more than the samples under study.

The structure of the used datasets is presented and described in Table I. The sets are arranged in increasing order of the number of samples. Data on the balance between the number of representatives of each class are taken from the sources of the sets. The percentage of abnormalities in each set was calculated using the Z-score method (3).

$$z = \frac{x-\mu}{\sigma} \quad (3)$$

Where:

x – sample of the set

μ – average value

σ – the standard deviation

The Z-score formula finds the number of standard deviations from the mean. It is considered an anomaly if the value of the sample obtained by (3) is above three or below three. For the purposes of the experimental study, an "anomaly set" will mean a set containing more than 10% anomalies.

TABLE I. TRAINING SETS USED IN THE EXPERIMENTAL STUDY

Dataset	Attributes	Samples	Classes	No balanced	Anomaly
Lenses[1]	4	24	3	**Yes**	**100%**
Lung Cancer[2]	56	32	3	No	**78.12% (25)**
Soybean (small)[3]	35	47	4	**Yes**	0%
SRBCT[4]	2308	83	4	Ne	0%
Cryotherapy[5]	6	90	2	No	3.33% (3)
Beavers[6]	3	114	2	**Yes**	0.87% (1)
Iris[7]	4	150	3	No	0 .66% (1)
Hepatitis[8]	19	155	2	**Yes**	**92.9% (144)**
Wine[9]	13	178	3	No	5.62% (10)
Glass[10]	9	214	6	**Yes**	9.35% (20)
Thyroid[11]	5	215	3	**Yes**	8.83% (19)
Stars[12]	4	240	6	No	2.92% (7)
Algerian Forest Fires[13]	11	243	2	No	7.41% (18)
Ecoli[14]	8	335	8	**Yes**	0%
Ionosphere[15]	34	351	2	**Yes**	0%
Breast Cancer Wisconsin[16]	30	569	2	**Yes**	**13% (74)**
Absenteeism[17]	19	740	28	**Yes**	**77.84% (576)**
Pima Indians Diabetes[18]	8	768	2	No	**10.41% (80)**

[1] https://archive.ics.uci.edu/ml/datasets/lenses
[2] https://archive.ics.uci.edu/ml/datasets/lung+cancer
[3] https://archive.ics.uci.edu/ml/datasets/soybean+(small)
[4] https://rdrr.io/cran/plsgenomics/man/SRBCT.html
[5] https://archive.ics.uci.edu/ml/datasets/Cryotherapy+Dataset+
[6] https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/beavers.html
[7] https://archive.ics.uci.edu/ml/datasets/iris
[8] https://archive.ics.uci.edu/ml/datasets/hepatitis
[9] https://archive.ics.uci.edu/ml/datasets/wine
[10] https://archive.ics.uci.edu/ml/datasets/glass+identification
[11] https://archive.ics.uci.edu/ml/datasets/thyroid+disease
[12] https://www.kaggle.com/brsdincer/star-type-classification
[13] https://archive.ics.uci.edu/ml/datasets/Algerian+Forest+Fires+Dataset++
[14] https://archive.ics.uci.edu/ml/datasets/ecoli
[15] https://archive.ics.uci.edu/ml/datasets/ionosphere
[16] https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28original%29
[17] https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work
[18] https://www.kaggle.com/uciml/pima-indians-diabetes-database

The Z-score formula finds the number of standard deviations from the mean. It is considered an anomaly if the value of the sample obtained by (3) is above three or below three. For the purposes of the experimental study, an "anomaly set" will mean a set containing more than 10% anomalies.

B. Data Description

The data used in this experimental study varied in volume, the number of classes, the percentage of anomalies, and the imbalance. Some training sets are of greater interest for research because they contain a larger number of anomalies or are unbalanced. They are Lenses, Lung Cancer, Soybean, SRBCT, Beavers, Hepatitis, Glass, Thyroid, Ecoli, Ionosphere, Breast Cancer Wisconsin, and Absenteeism.

The *Lenses* training set contains 24 samples with 4 attributes each, divided into 3 classes. It is unbalanced and all samples are considered an anomaly. The data is used to classify patients who need contact lens fitting. The potential business application is in optics, where patients are examined and the need for contact lens fitting is determined.

The Lung Cancer set contains information about people who have characteristics of lung cancer patients. The classification task is to correctly diagnose new patients, based on the data of 32 patients (samples), 56 indicators (attributes), and 3 possible diagnoses (classes). Anomalies account for 78% of the data. A potential business application of such a dataset is in laboratories examining patient samples.

Soybean contains information on soybean diseases. It contains 47 samples with 35 attributes classified into 4 classes. It is defined as unbalanced, without anomalies. A set containing information on cereals, legumes, and similar crops can find application in any farm growing them.

SRBCT contains data from 83 patients and 2308 attributes (genes) classified into 4 classes. The data is intended to aid in the correct classification of various childhood cancers. The set is balanced, has no anomalies, and is considered an $n \gg p$ problem. It can be used in laboratories, research, and scientific centers.

Cryotherapy is a set containing data for 90 samples with 6 attributes. Unbalanced and free of anomalies.

Iris is a well-known and researched set for the classification of flowers of the Iris species. It is used as the basis for the study of many machine learning algorithms. It contains 150 samples having 4 attributes and samples are divided into 3 categories. It is balanced and has no anomalies. The application of such a set is in the field of the flower business - a flower shop can recognize the types of the received goods, and greenhouses can more accurately classify the flowers grown.

Hepatitis contains information on 155 samples, each described by 19 attributes and classified into 2 categories. The set is not balanced and has a high percentage of anomalies – 92.9%. Contains information for patients diagnosed with hepatitis. The classification problem it solves is to answer the question of whether the patient will live. It can be used in doctor's offices.

Wine is a dataset much like Iris - well-known, used, and studied. Contains data for 178 samples, 13 attributes, and 3 categories. The data is balanced and the anomaly rate is not high. A similar set can be used by winemakers.

Glass contains data for 214 samples, 9 attributes, and 6 classes. The set is unbalanced, but there are no anomalies. The classification problem is the recognition of types of glass, specifically glass shards found at crime scenes. It can be used in laboratories and glass manufacturers.

Thyroid described 215 patients with 5 attributes and divided them into 3 classes. The data is unbalanced and there are no anomalies. They are used to classify the action of the thyroid gland and therefore such a set can be used in doctors' offices and laboratories.

Stars dataset aims to categorize different celestial bodies into 6 categories, according to 4 attributes. There are 240 samples examined, and the set is considered balanced and contains no anomalies. The application could be in business organizations involved in space exploration.

Algerian Forest Fires contains information on 243 fires in Algeria. Each fire is classified into 2 categories and described with 11 characteristics. The set is balanced and has no anomalies. It can be used in applications for early fire warning, by conservation organizations, fire departments, for research purposes, etc.

Ecoli is a set that contains information about 335 studied samples, described in 8 categories and characterized by 8 attributes. There are no anomalies detected, but the data is unbalanced. It is used to detect specific proteins. The business application of such a set is in laboratories, doctor's offices, and scientific and research centers.

Ionosphere contains information on 351 signals passing through the Earth's ionosphere. The attributes describing them are 34, and the classes classifying them are 2. There are no detected anomalies, but the data is unbalanced. The data can be used in any organization involved in space and earth research.

Breast Cancer Wisconsin contains information on 569 patients described by 30 attributes and classified into 2 categories. The set is unbalanced and the anomaly rate is 13%. The data is used to detect malignant tumors in patients. The application can be in doctors' offices and laboratories.

Absenteeism contains information about 740 employees in different companies. Each employee is described with 19 indicators and 28 different degrees of possibility for the employee to be absent from work are defined. The high number of possible classes into which a sample can be classified causes the set to have a high rate of anomalies and to be unbalanced. Training a model with such data makes it possible to use it anywhere regardless of the subject area of the business.

Pima Indians Diabetes is a set that has 768 samples classified into two categories and described by 8 attributes. Contains data for patients who have indicators similar to those of patients with diabetes. The data is balanced, but anomalies are present. The application of such a set could be in doctor's offices.

All experimental studies were done with m-fold cross-validation, where m=10. According to [32, 33, 34, 35], to avoid the chance of random results, each model should be trained more than once and the average of the training to be accepted, for this purpose many authors suggest the number of training iterations to be ten, this is the accepted number in the current paper. The results of each iteration were recorded and the average value of the model's accuracy was obtained. For each model training, all available attributes were used without further attribute processing.

For the k-parameter studies, all training sets were trained with the traditional neighborhood method in combination with the traditional nearest neighborhood method and the weighted neighborhood method. The automatic calculation of the k parameter is done in two ways: the training set size method, described with an equation (2), and the number of classes method - equation (1) (Table II).

In the training set method the number of the classes is taken into consideration. For every m_i training iteration around 70% of the data is used and around 30% is used for testing of the model, meaning that the number of the samples is not a constant.

Table II shows the results from the experimental study. The second column of the table shows the number of samples with which the k value is determined, and the following columns shows the highest achieved result for every trained model.

Every neighborhood method is experimented, and for every method, three types of k determination are done – a manual and two automatic methods. In every cell, the accuracy percentage is noted and in brackets is the number of the neighbors (k value) that has achieved this accuracy score.

TABLE II. HIGHEST PERCENT ACCURACY OBTAINED WITH MANUAL AND AUTOMATED K VALUE SELECTION

Dataset	*N*	KNN		WKNN		DWKNN		UWKNN		Dual		Uniform		Inverse	
		Manual method	$k=\sqrt{N}$	*Manual method*	$k=\sqrt{N}$	*Manual method*	$k=\sqrt{N}$	*Manual method*	$k=\sqrt{N}$	*Manual method*	$k=\sqrt{N}$	*Manual method*	$k=\sqrt{N}$	*Manual method*	$k=\sqrt{N}$
			$k=c+1$		$k=c+1$		$k=c+1$		$k=c+1$		$k=c+1$		$k=c+1$		$k=c+1$

Dataset	*N*	KNN		WKNN		DWKNN		UWKNN		Dual		Uniform		Inverse	
Lenses	≅ 21	0.85% (2)	0.750% (4) 0.683% (4)	0.8% (10)	0.750% (4) 0.767% (4)	0.833% (6)	0.783% (4) 0.883% (4)	0.817% (2)	0.700% (4) 0.733% (4)	0.817% (2)	0.733% (4) 0.750% (4)	0.8% (1)	0.717% (4) 0.733% (4)	0.783% (1)	0.717% (4) 0.750% (4)
Lung Cancer	≅ 28	0.6% (8)	0.533% (5) 0.442% (4)	0.567% (9)	0.533% (5) 0.450% (4)	0.608% (10)	0.533% (5) 0.433% (4)	0.633% (5)	0.600% (5) 0.517% (4)	0.558% (6)	0.483% (5) 0.517% (4)	0.575% (3)	0.492% (5) 0.467% (4)	0.6% (7)	0.467% (5) 0.483% (4)
Soybean Small	≅ 42	1.0% (3)	0.980% (6) 0.980% (5)	1.0% (4)	1.000% (6) 1.000% (5)	1.0% (1)	1.000% (6) 1.000% (5)	1.0% (3)	0.980% (6) 0.980% (5)	1.0% (6)	0.980% (6) 0.98% (5)	0.98% (1)	0.98% (6) 0.98% (5)	1.0% (4)	0.980% (6) 0.980% (5)
SRBCT	≅ 74	0.943% (3)	0.893% (8) 0.868% (5)	0.965% (16)	0.942% (8) 0.918% (5)	0.975% (18)	0.975% (8) 0.928% (5)	0.976% (16)	0.939% (8) 0.954% (5)	0.953% (28)	0.914% (8) 0.926% (5)	0.976% (23)	0.918% (8) 0.942% (5)	0.951% (3)	0.940% (8) 0.917% (5)
Cryotherapy	≅ 81	0.933% (2)	0.722% (9) 0.844% (3)	0.933% (4)	0.833% (9) 0.933% (3)	0.944% (9)	0.922% (9) 0.933% (3)	0.933% (2)	0.822% (9) 0.856% (3)	0.933% (4)	0.922% (9) 0.933% (3)	0.911% (3)	0.844% (9) 0.933% (3)	0.944% (7)	0.933% (9) 0.911% (3)
Beavers	≅ 102	0.949% (9)	0.947% (10) 0.948% (3)	0.949% (13)	0.948% (10) 0.939% (3)	0.949% (11)	0.947% (10) 0.922% (3)	0.949% (28)	0.948% (10) 0.948% (3)	0.949% (10)	0.939% (10) 0.931% (3)	0.948% (16)	0.948% (10) 0.939% (3)	0.949% (7)	0.930% (10) 0.947% (3)
Iris	≅ 135	0.973% (12)	0.973% (11) 0.960% (4)	0.973% (18)	0.960% (11) 0.960% (4)	0.973% (16)	0.960% (11) 0.960% (4)	0.967% (7)	0.973% (11) 0.960% (4)	0.967% (22)	0.960% (11) 0.953% (4)	0.967% (8)	0.960% (11) 0.960% (4)	0.98% (13)	0.960% (11) 0.960% (4)
Hepatitis	≅ 139	0.795% (26)	0.776% (11) 0.715% (3)	0.799% (20)	0.740% (11) 0.669% (3)	0.795% (30)	0.734% (11) 0.671% (3)	0.8% (24)	0.767% (11) 0.703% (3)	0.737% (27)	0.703% (11) 0.665% (3)	0.779% (29)	0.742% (11) 0.698% (3)	0.795% (23)	0.761% (11) 0.715% (3)
Wine	≅ 160	0.795% (26)	0.686% (12) 0.714% (4)	0.786% (3)	0.707% (12) 0.754% (4)	0.78% (4)	0.720% (12) 0.729% (4)	0.775% (2)	0.719% (12) 0.752% (4)	0.781% (22)	0.752% (12) 0.753% (4)	0.781% (22)	0.775% (12) 0.74% (4)	0.782% (18)	0.771% (12) 0.73% (4)
Glass	≅ 192	0.743% (1)	0.608% (13) 0.649% (7)	0.734% (1)	0.668% (13) 0.687% (7)	0.739% (4)	0.682% (13) 0.719% (7)	0.727% (2)	0.697% (13) 0.691% (7)	0.762% (8)	0.748% (13) 0.725% (7)	0.739% (3)	0.687% (13) 0.692% (7)	0.742% (1)	0.667% (13) 0.660% (7)

Dataset	*N*	KNN		WKNN		DWKNN		UWKNN		Dual		Uniform		Inverse	
Thyroid	≅ 193	0.949% (1)	0.888% (13)	0.949% (2)	0.930% (13)	0.953% (6)	0.940% (13)	0.954% (1)	0.921% (13)	0.958% (8)	0.939% (13)	0.958% (2)	0.930% (13)	0.954% (1)	0.926% (13)
			0.926% (4)		0.958% (4)		0.931% (4)		0.935% (4)		0.944% (4)		0.935% (4)		0.934% (4)
Stars	≅ 216	0.725% (1)	0.621% (14)	0.742% (2)	0.638% (14)	0.725% (2)	0.654% (14)	0.725% (1)	0.654% (14)	0.737% (6)	0.721% (14)	0.717% (9)	0.687% (14)	0.712% (3)	0.654% (14)
			0.604% (7)		0.683% (7)		0.667% (7)		0.654% (7)		0.721% (7)		0.721% (7)		0.688% (7)
Algerian Forest Fires	≅ 218	0.942% (4)	0.901% (14)	0.931% (6)	0.893% (14)	0.93% (1)	0.926% (14)	0.935% (5)	0.889% (14)	0.947% (1)	0.927% (14)	0.938% (2)	0.917% (14)	0.938% (1)	0.914% (14)
			0.930% (3)		0.931% (3)		0.934% (3)		0.909% (3)		0.934% (3)		0.943% (3)		0.926% (3)
Ecoli	≅ 301	0.425% (18)	0.424% (17)	0.425% (15)	0.424% (17)	0.424% (1)	0.424% (17)	0.425% (3)	0.424% (17)	0.425% (2)	0.424% (17)	0.425% (6)	0.425% (17)	0.425% (3)	0.424% (17)
			0.424% (9)		0.424% (9)		0.424% (9)		0.424% (9)		0.424% (9)		0.424% (9)		0.423% (9)
Ionosphere	≅ 315	0.863% (1)	0.838% (17)	0.872% (9)	0.852% (17)	0.872% (1)	0.846% (17)	0.872% (2)	0.826% (17)	0.875% (15)	0.869% (17)	0.866% (3)	0.849% (17)	0.863% (2)	0.832% (17)
			0.853% (3)		0.857% (3)		0.866% (3)		0.826% (3)		0.869% (3)		0.846% (3)		0.843% (3)
Breast Cancer Wisconsin	≅ 512	0.937% (10)	0.924% (22)	0.94% (22)	0.94% (22)	0.94% (22)	0.937% (22)	0.938% (7)	0.932% (22)	0.931% (27)	0.924% (22)	0.94% (25)	0.933% (22)	0.935% (17)	0.933% (22)
			0.930% (3)		0.917% (3)		0.921% (3)		0.930% (3)		0.909% (3)		0.91% (3)		0.93% (3)
Absenteeism	≅ 666	0.345% (5)	0.284% (25)	0.376% (7)	0.324% (25)	0.369% (11)	0.331% (25)	0.369% (26)	0.353% (25)	0.362% (28)	0.345% (25)	0.376% (23)	0.365% (25)	0.372% (29)	0.357% (25)
			0.276% (28)		0.315% (29)		0.334% (29)		0.350% (29)		0.359% (28)		0.361% (29)		0.351% (28)
Pima Indians Diabetes	≅ 691	0.755% (13)	0.751% (26)	0.75% (28)	0.742% (26)	0.749% (17)	0.741% (26)	0.75% (24)	0.733% (26)	0.726% (21)	0.712% (26)	0.738% (29)	0.720% (26)	0.749% (19)	0.725% (26)
			0.699% (3)		0.685% (3)		0.669% (3)		0.702% (3)		0.672% (3)		0.673% (3)		0.701% (3)

VI. CONCLUSIONS

From the obtained results, it can be concluded that the automatic determination of the k value can give results close to the optimal, but in some cases, the difference in the percentage of accuracy is about 10-15% lower, compared to manual methods.

For example, with Lenses, the obtained classification accuracy with the manual method is 85%, while with the automated selection method – 68%. Training with a manually found k value of the Lung Cancer set with the DWKNN method can reach 60% classification accuracy. The same set and method (DWKNN), but with an automatically selected k value, reaches only 43% accuracy. Considering the nature of the training set, this percentage is unsatisfactory.

There are quite a few cases where the automatic selection of the neighborhood size gives close to optimal results. In some cases, the difference between the obtained values is less than a percentage, and in addition, the number of neighbors used is less than with the manual method. The use of a relatively small, but not too small, size of the neighborhood is one of the set criteria for accepting the algorithm as optimal (item 3.2).

In the Thyroid set trained with the Dual method, manual selection can classify with 95.8% accuracy at neighborhood size k=8. Using the automatic neighborhood of size k=4, i.e. half, achieves 94.4% accuracy. The difference between the two methods is negligibly small and it can be argued that in such cases the automatic selection of the k parameter is appropriate, efficient, and optimal in terms of the required resources for classification.

Known modifications of weighted nearest neighbors (WKNN, DWKNN, UWKNN, Dual, Uniform, and Inverse) do not behave stably and deviations in accuracy rate are observed with the increasing neighborhood for some training sets.

The performance of a model depends on many factors regarding the training set – size, number of unique classes, data balance, missing data anomalies, etc. Therefore, it cannot be expected that some model parameters, e.g. k=5, will be optimally applicable to all sets.

REFERENCES

[1] Stoyanov, I. S., Iliev, T. B., Mihaylov, G. Y., Evstatiev, B. I., & Sokolov, S. A. (2018, October). Analysis of the cybersecurity threats in smart grid University of telecommunications and post, Sofia, Bulgaria. In 2018 IEEE 24th International Symposium for Design and Technology in Electronic Packaging(SIITME) (pp. 90-93). IEEE.

[2] Mladenova, T., & Valova, I. (2022, September). Fake news detection from Bulgarian Facebook pages. In AIP Conference Proceedings (Vol. 2449, No. 1, p. 040013). AIP Publishing LLC.

[3] Evstatiev, B. (2013). Evaluation of thermal diffusivity of soil near the surface: methods and results. Bulgarian journal of agricultural science, Agricultural academy, 19(3), 467-471.

[4] Fix, E. and J. Hodges, "An important contribution to nonparametric discriminant analysis and density estimation," Int. Stat. Rev., vol. 3, no. 57, pp. 233–238, 1951.

[5] Fix, E. and J. L. Hodges, "Nonparametric discrimination: Consistency properties," Randolph Field, Texas, Proj., pp. 21–49, 1951.

[6] Fix, E. and J. L. Hodges Jr, "Discriminatory analysis-nonparametric discrimination: Small sample performance," 1952.

[7] Cover, T. and P. Hart, "Nearest neighbor pattern classification," IEEE Trans. Inf. theory, vol. 13, no. 1, pp. 21–27, 1967.

[8] Qin, Z., A. T. Wang, C. Zhang, and S. Zhang, "Cost-sensitive classification with k-nearest neighbors," in International Conference on Knowledge Science, Engineering and Management, 2013, pp. 112–131.

[9] Zhang, S., X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient kNN classification with different numbers of nearest neighbors," IEEE Trans. neural networks Learn. Syst., vol. 29, no. 5, pp. 1774–1785, 2017.

[10] Mladenova, T., & Valova, I. (2021, June). Analysis of the KNN classifier distance metrics for Bulgarian fake news detection. In 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA) (pp. 1-4). IEEE.

[11] Zhang, S., D. Cheng, Z. Deng, M. Zong, and X. Deng, "A novel kNN algorithm with data-driven k parameter computation," Pattern Recognit. Lett., vol. 109, pp. 44–54, 2018.

[12] Meng, L., "Efficient M-fold Cross-validation Algorithm for KNearest Neighbors," 2010.

[13] Qin, Y., S. Zhang, X. Zhu, J. Zhang, and C. Zhang, "Semi-parametric optimization for missing data imputation," Appl. Intell., vol. 27, no. 1, pp. 79–88, 2007.

[14] Guo, G., H. Wang, D. Bell, Y. Bi, and K. Greer, "An kNN model-based approach and its application in text categorization," in International Conference on Intelligent Text Processing and Computational Linguistics, 2004, pp. 559–570.

[15] Guo, G., H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN model-based approach in classification," in OTM Confederated International Conferences, 2003, pp. 986–996.

[16] Li, B., Y. W. Chen, and Y. Q. Chen, "The nearest neighbor algorithm of local probability centers," IEEE Trans. Syst. Man, Cybern. Part B, vol. 38, no. 1, pp. 141–154, 2008.

[17] Gates, G., "The reduced nearest neighbor rule (corresp.)," in IEEE transactions on information theory, 1972, vol. 18, no. 3, pp. 431–433.

[18] Li, B. L., Yu, S. W., & Lu, Q. (2003). An improved k-Nearest neighbor algorithm for text categorization. In International Conference on Computer Processing of Oriental Languages [ICCPOL], 2003.

[19] Zhang, S., "Challenges in KNN Classification," IEEE Trans. Knowl. Data Eng., 2021.

[20] Hastie, T. and R. Tibshirani, "Discriminant adaptive nearest neighbor classification and regression," in Advances in neural information processing systems, 1996, pp. 409–415.

[21] Domeniconi, C., J. Peng, and D. Gunopulos, "Locally adaptive metric nearest-neighbor classification," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 9, pp. 1281–1285, 2002.

[22] Mladenova, T., & Valova, I. (2022, October). Comparative analysis between the traditional K-Nearest Neighbor and Modifications with Weight-Calculation. In 2022 International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) (pp. 961-965). IEEE.

[23] Chaudhuri, B. B., "A new definition of neighborhood of a point in multi-dimensional space," in Pattern Recognition Letters, 1996, vol. 17, no. 1, pp. 11–17.

[24] Mladenova, T., & Valova, I. (2022, June). Comparative Analysis Between the Traditional and Nearest Centroid Methods in the K-Nearest Neighbor Algorithm. In 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA) (pp. 1-4). IEEE.

[25] Dudani, S. A., "The distance-weighted k-nearest-neighbor rule," in IEEE Transactions on Systems, Man, and Cybernetics, 1976, no. 4, pp. 325–327.

[26] Gou, J., L. Du, Y. Zhang, T. Xiong, and others, "A new distance-weighted k-nearest neighbor classifier," J. Inf. Comput. Sci, vol. 9, no. 6, pp. 1429–1436, 2012.

[27] Zeng, Y., Y. Yang, and L. Zhao, "Pseudo nearest neighbor rule for pattern classification," in Expert Systems with Applications, 2009, vol. 36, no. 2, pp. 3587–3595.

[28] Mitani, Y. and Y. Hamamoto, "A local mean-based nonparametric classifier," in Pattern Recognition Letters, 2006, vol. 27, no. 10, pp. 1151–1159.

[29] Gou, J., T. Xiong, and Y. Kuang, "A Novel Weighted Voting for K-Nearest Neighbor Rule.," J. Comput., vol. 6, no. 5, pp. 833–840, 2011.

[30] Keller, J. M., M. R. Gray, and J. A. Givens, "A fuzzy k-nearest neighbor algorithm," in IEEE transactions on systems, man, and cybernetics, 1985, no. 4, pp. 580–585.

[31] Zadeh, L. A., "Fuzzy sets," in Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers by Lotfi A Zadeh, World Scientific, 1996, pp. 394–432.

[32] Gu, Q., L. Zhu, and Z. Cai, "Evaluation measures of the classification performance of imbalanced data sets," in International symposium on intelligence computation and applications, 2009, pp. 461–471.

[33] Hockenmaier, J., "Introduction to Machine Learning, Lecture 9: Evaluation." University of Illinois, 2015. [Online]. Available: http://courses.engr.illinois.edu/cs446

[34] Hossin, M. and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," Int. J. data Min. & Knowl. Manag. Process, vol. 5, no. 2, p. 1, 2015.

[35] Watt, J., R. Borhani, and A. Katsaggelos, Machine learning refined: foundations, algorithms, and applications. Cambridge University Press, 2020.