Data analysis for a set of university student lists using the k-Nearest Neighbors machine learning method

Data analysis for a set of university student lists using the k-Nearest Neighbors machine learning method

D Pedrozo[1], F Barajas[1], A Estupiñán[2], K L Cristiano[3], and D A Triana[3]

[1] Universidad Autónoma de Bucaramanga, Bucaramanga, Colombia
[2] Universidad de Investigación y Desarrollo, Bucaramanga, Colombia
[3] Universidad Industrial de Santander, Bucaramanga, Colombia

E-mail: aestupinan4@udi.edu.co

Abstract. A tool using machine learning which organizes, clean and analyzes data in massive quantities was developed. So that the user can make predictions about different aspects of a list of more than 2000 students, using the k-nearest neighbor machine learning method. Data were obtained from a Colombian university, which has previously it has been cleaned and organized to accommodate the input parameters into a script. The computational tool was written in python from Jupyter notebook. The script is able to perform analysis predictive of the different filters previously chosen for decision raking by the example of marketing, the tool will keep track of financial actions and of the different categories chosen for students at the university, allowing a global analysis and thus choose the best options from all data granted. Among these categories we can classify method of payment, the value paid, what is the method of most commonly used payout by students, individual averages per semester and academic programs, addresses, and academic careers among others.

1. Introduction

Let's imagine that we need to pick the best students from certain careers to carry out the most challenging projects, or even students with academic difficulties, whose averages are not the best to support them, we could also evaluate and strengthen careers whose overall averages are not very and be able to improve the university in general. All of this allows for better management of the information because seeing it from a global position allows you not only to make decisions immediate improvement, if not projecting its direct consequences into the future, between predicting how many students could enter each academic program, even students who are taking the same subjects according to where they live to form groups study or reunite students according to their place of residence whose taste for sport shared to do group games and training with the goal improve the competitive part of college and the relationships between athletes, then analyzing the new data to make a prediction of potential members to the selection of certain college teams.

Scopes of classification algorithm are so large that we can cover in the different fields of knowledge and experimental evidence that we have either in nature or the various groups of species. Then, some applications of this method will be exhibit. is used to predict cloud cover

by half of a camera in the sky that is installed in a waterproof casing with a heating and cooling system.

The assembly allows continuous operation in using the K-Nearest Neighbors method to decide if what the camera sees in the sky or cloud [1–3], it's also used in hierarchical classification of textual documents evaluation of the K-Nearest Neighbors algorithm for prediction problems with outputs a real-coded genetic algorithm for the evolution of linear transformations [4], is also used for problems as a classification of Lone lung nodules [5–7], classification of cervical cells using core [8], there is also software for treating pornographic web pages based on K-Nearest Neighbors classifier [9], machine learning algorithms for classification of pyramidal neurons affected by aging and finally the neighbor's algorithm closest applied to the filmaffinity.com website [10]. The nearest neighboring algorithm/rule (Nearest Neighbors) is the non-decision procedure simplest parametric parameter, which is assigned to the unclassified observation (test sample class/category/label of the nearest sample (using metric) in the training set, in a large sample case or in a training set number approach the probability of error to Nearest Neighbors is limited by (lower limit) R probability Bayes' probability of classification error and upper bound 2R (1 - R) the importance of the method.

In section 2, we present the theoretical framework of the numerical method, the characteristics, and advantages that the method has for this type of study, in section 3 we present the results obtained for the different kinds of variables taken in consideration. Finally, in section 4, conclusions and observations respect to benefits offered by this computational tool are shown.

2. Methods

This method is mainly used in the field of nonparametric data classification, which performs the estimation of the value of the probability density function or in some cases this method directly resorts to the subsequent probability that a random element may belong to this class. For example, Figure 1 shows the different lengths of estimates for two different values of k, where we can see that this value will depend only on the accuracy and precision shown by the data set.

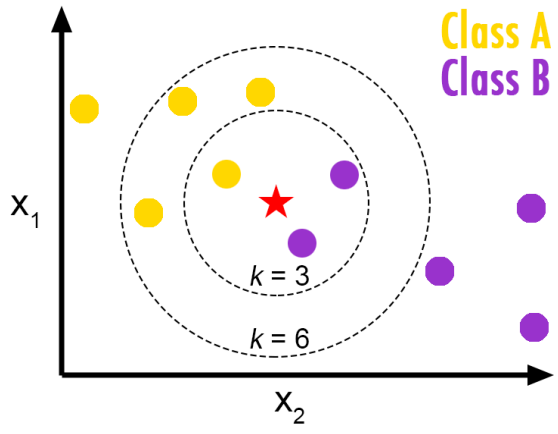

Figure 1. Classification model of the k-Nearest Neighbors method, using two classes of the population data sample, taking the values of $k = 3$ and $k = 6$.

As shown in Figure 1, this method essentially serves to label or group the values looking for the most similar data points (by proximity), which were previously learned in the data training stage using basically the methodology implemented in a Machine Learning, in addition to guessing new points based on that classification.

One of the most important features of this method is:

- Calculate the distance between the item to be classified and the rest of the items in the training data set.
- Select the "k" closest elements (with less distance, depending on the function used).
- Perform a majority vote among the k points: those of a class or label that dominate will decide their final classification.

One of the most important parameters, to take into account in this method, is the probability density, used to determine the Lebesgue measurement or the distance between the parameters closest to a certain value.

For this reason, it's of great importance to define a measurement density of x, with the purpose of defining an estimate of the data density using what is usually known as the Lebesgue measurement, which is defined in Equation (1), as follows:

$$f_n(x) = f_n(x; X_1, ..., X_n). \tag{1}$$

The decision to be able to take a certain density $f(x)$, is mainly based on the statistical consistency of the problem we are treating and on the data we have obtained from the sample, which determines the smoothness, optimization, confidence, and interpretability of the data to study and other criteria, in addition to being able to evaluate other criteria as shown in the references [11].

To calculate the estimate of the nearest k-neighbor method, for example, taking values $X_1, ..., X_n$ being a sequence value taken randomly from a vector and assuming that the measure of common probability μ of the sequence is continuous with respect to the measure of Lebesgue λ, with a density f.

It's possible define the function $f(x)$ using Lebesgue's differentiation (see Equation (2)) theorem as:

$$f(x) = \lim_{\rho \to 0} \frac{\mu(B(x, \rho))}{\lambda(B(x, \rho))}. \tag{2}$$

Taking into account the relationship shown in the previous equation, you can estimate the value of the function $f(x)$, for a value of k between $1 \leq k \leq n$, when $R_{(k)}(x)$ will be the distance from x to the $k - th$ nearest neighbor in data sequence, In this way, the value μ_n will be the empirical distribution taken and is given for the Equation (3),

$$\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} 1. \tag{3}$$

The density function for the $k - th$ neighbor is specified for the Equation (4).

$$f_n(x) = \frac{\mu_n(B(x), R_{(k)}(x))}{\lambda(B(x, R_{(k)}(x)))}, \tag{4}$$

The Equation (4), can be written also as the Equation (5),

$$f_n(x) = \frac{k}{n\lambda(B(x, R_{(k)}(x)))} \tag{5}$$

To have greater detail, of the mathematical and statistical study of the empirical distribution function obtained through these equations, we recommend reviewing the work developed by Biau G, *et al.* [12]. In the implementation that we have done in this code, we have used the Python Scikit-learn package, which allows us to perform the type of data with which we are

going to work, which in our case is: code, average and academic program of the student among others [13, 14].

Using the k-Nearest Neighbors Classification Method to generate predictions about the academic and financial records of members of a school (in this case a university). Through these records obtained from previous years, predictions are generated to see which is more viable, to see if there is a correlation between them, and that gives us a better analysis for each question that we can ask to get the best out of this project.

Based on this method, and having a good database, both for training and testing, new data can be inserted to be processed using this algorithm and with it give us a good prediction about what we want to know based on the data of e in addition, it intends to expand it, to take new input data to be able to fine-tune the accuracy of the model, and also, to cover an increasing number of categories that allow us to globalize under the same framework common to all students with the greatest amount possible individual parameters.

3. Results

The first thing we should do is compile the data that we have taken from the university's database. The python pandas tool was used, in order to extract the data and organized it by columns. Depending on the category in which they have been selected the corresponding information to study.

Once we have the data organized, we proceed to continue with the data analysis process, using the "Describe" instruction, we have access to statistical data information, such as: the value of the average sample, the value minimum, the maximum value, the different quarterlies and the standard deviation obtained by each drop.

In this way, we can make an analysis of the number of students who have entered the university in the last 3 years (see Figure 2), showing a slight increase in the last year, also is important to highlight that for example, the 201700 academic period refers to the first semester of 2017, and 201750 refers to the second semester of the same year 2017, as shown in Figure 2.

The behavior of each category by class can be perform as follow, in Figure 3, we can see that the most common value corresponding to the weighted average of the students, which is between 3.6 and 3.7 approximately, indicating that the average of the students is at an acceptable level.

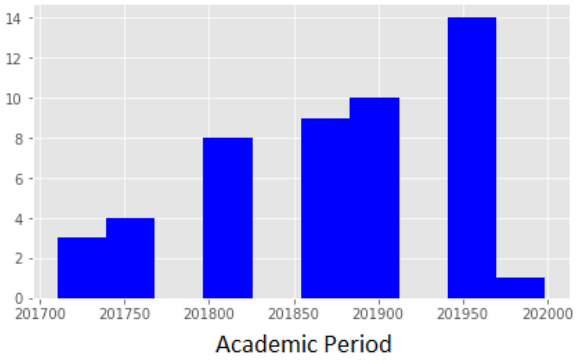

Figure 2. Histogram behaviour of the number of students in function of the academic period.

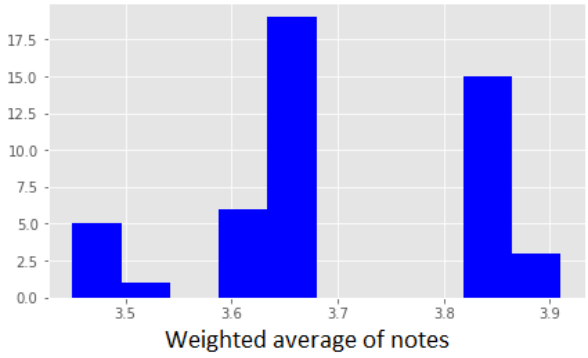

Figure 3. Histogram of the students, with respect to the weighted cumulative average of grades (maximum value is 5.0).

In the following figures, It's found more detailed information on the repetition intervals of students that have a weighted average value so far in the university (see Figure 4). We can also perform a histogram with the nine academic careers that has the university, where we detect an high number of students correspond to the career number "1", that is Industrial engineering,

and low number of students that study social communication correspond to number "5" (see Figure 5).

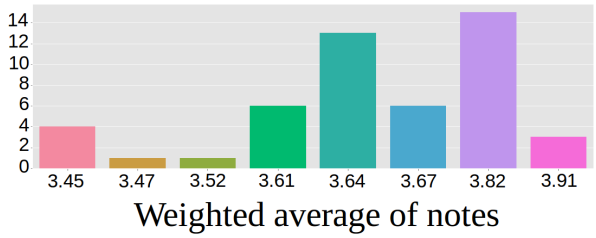

Figure 4. Detailed histogram of the weighted average of the students in the university.

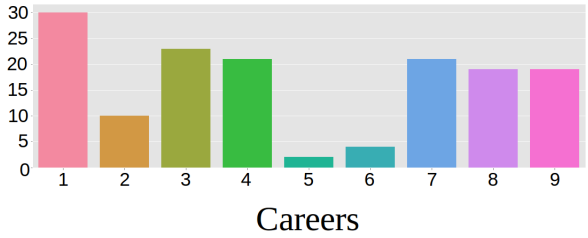

Figure 5. Detailed histogram of the number of students, for the nine different academic careers that has the university.

In this work, we have compared the variables related to the academic period, the grades of each student, the academic career and the neighborhood of residence of each student. For example, in the following graph, for a sample of 50 students, it was obtained using the method of the neighboring k, the sample of 5 different regions, with the purpose of organizing them according to the average that the students have obtained so far, it was obtained the result shown in Figure 6. Whenever we use the neighboring k method, we must verify the accuracy of the method based on the value of k chosen (see Figure 7).

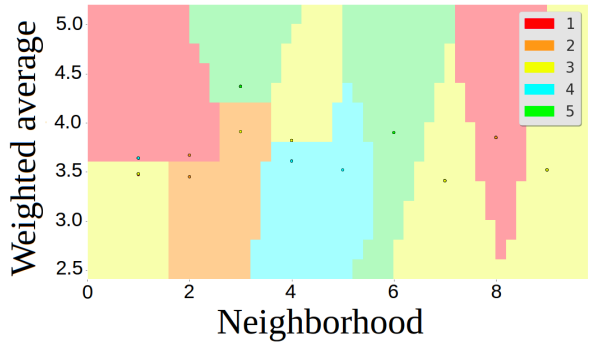

Figure 6. Classification of 5 groups of students in function of the neighborhood of residence, using the k neighbors method, based on the weighted grade point average for k = 5.

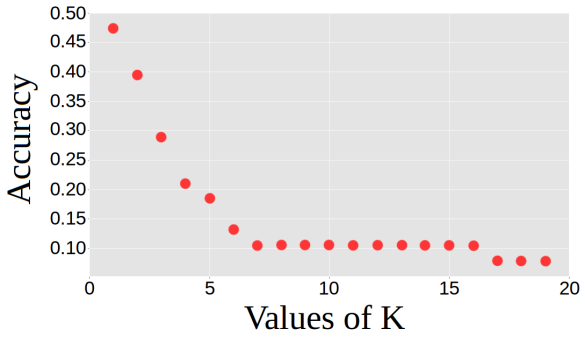

Figure 7. Accuracy of the neighboring k method, where a sample of 50 students has been taken for the first 20 k values.

One of the most relevant results of this study is to find the ratio of the average of the students according to the career they are studying in the university, we have listed the different academic programs of 1-9, (see Figure 8).

From Figure 8, we can see how the music academic program which is labeled with number 3 shows us that it is very likely that new students of this program will get an average of 4.3. On the other hand, an academic program chosen by the students in this university is the financial engineering, it is very interesting to see how the students of this career have a weighted average between 3.4 - 4.0, we can also deduce that the career in the which is very likely that a new student with an average of 3.67 will belong to the market career, identified with the number 2 (see Figure 8).

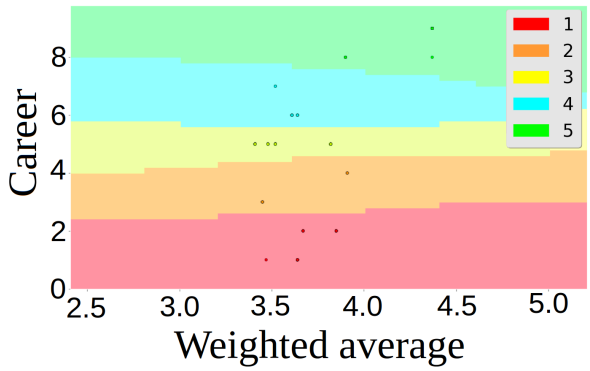

Figure 8. Career distribution, based on the weighted average of each student, where you can see 5 regions classified to catalog students 1-5 with a value k = 1.

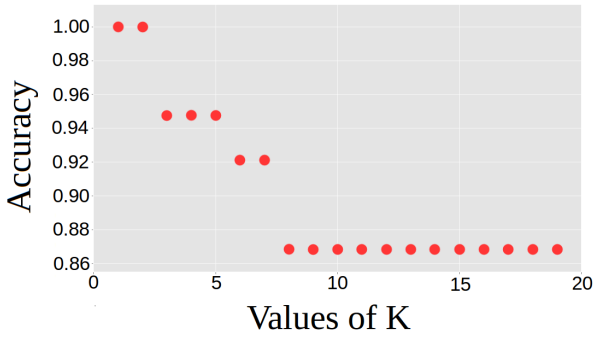

Figure 9. Accuracy of the method for the first 20 K values for a sample of 150 students.

From the neighboring k method, for the study of the student's career relationship based on their weighted average, we have obtained a precision relationship for different values of k, which is shown in Figure 9. A relationship that we were also interested in obtaining, was the weighted average of each student according to the neighborhood where he lives and the university career he studies, Figure 10 shows the results obtained for a sample of 150 students.

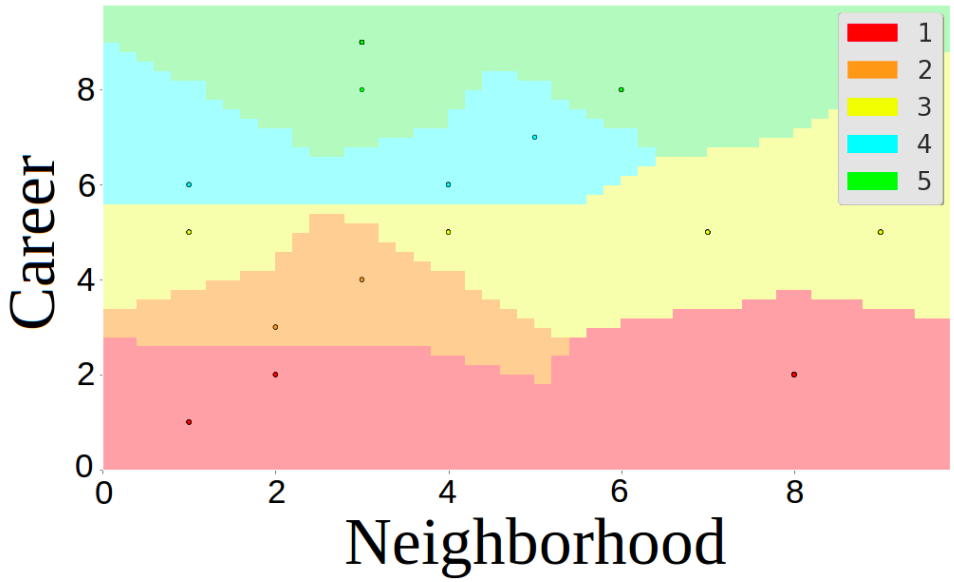

Figure 10. Distribution of the 5 regions by weighted average, depending on the neighborhood and the student's career.

In Figure 10, we can see that a student who studies medicine and lives in "Cabecera del Llano, Bucaramanga, Colombia" has an average of 3.5 on his career average. We make a free use provision and implementation of our code written in python. The code location is in the GitHub repository: `https://github.com/alexestupinan123/Machine-Learning.git` (retrieved 7/11/2019).

4. Conclusion

A data classification tool with a database of approximately 2000 students was developed. Here, it has been possible to show the level of precision of the method, depending on the number of data taken, in addition to the type of data variables that have been taken. In this work, we show the implementation of the method of k-Nearest Neighbor, for the organization and prediction of a list of students in a university, the code showed a precision above 94% for the value of k under 5, giving quite reliance to apply the algorithm to this study.

Acknowledgments

The author Alex Estupiñán, wants to give special thanks to the Universidad de Investigación y Desarrollo (UDI), Bucaramanga, Colombia, for providing the computational resources and the space to develop this research. The authors David A. Pedrozo and Fabian A. Barajas, want thanks to For all your support and patience during the writing and research period of this article to Oscar Leonardo Barajas Castro, María Edilia Castro Quintero, Isabel Cristina Lopez Lopez, Natalia Sanchez Gomez.

References

[1] A Cazorla, F J Olmo y L Alados-Arboledas 2005 Estimación de la cubierta nubosa en imágenes de cielo mediante el algoritmo de clasificación KNN *XI Congreso Nacional de Teledetección* (Tenerife: Asociación Española de Teledetección) p 335

[2] Leonardo Cavalheiro Langie, Vera Lúcia Strube de Lima 2003 Classificação Hierárquica de Documentos Textuais Digitais usando o Algoritmo kNN *1º Workshop em Tecnologia da Informação e da Linguagem Humana (TIL)* (São Carlos: Interinstitutional Center for Computational Linguistics)

[3] González H, Santos G, Campos F, & Morell Pérez, C 2016 Evaluación del algoritmo KNN-SP para problemas de predicción con salidas compuestas *Revista Cubana de Ciencias Informáticas* **10(3)** 119

[4] López Díaz José Carlos 2010 Un algoritmo genético con codificación real para la evolución de transformaciones lineales (Madrid: Universidad Carlos III de Madrid)

[5] Rivero Castro A, Cruz Correa L M and Artiles Lezcano J 2016 Selección de un algoritmo para la clasificación de Nódulos Pulmonares Solitarios *Revista Cubana de Informática Médica* **8(2)** 166

[6] Cristina García Cambronero, Irene Gómez Moreno 2006 Algoritmos de aprendizaje: Knn & kmeans *Intelgencia en Redes de Comunicación* (Madrid: Universidad Carlos III de Madrid)

[7] T M Cover and P E Hart 2018 Nearest neighbor pattern classification *IEEE Transactions on Information Theory* **13(1)** 21

[8] Rodríguez Vázquez, S., Borges, M., Vidal, A., Ginori, L., & Valentín, J 2016 Clasificación de células cervicales mediante el algoritmo KNN usando rasgos del núcleo *Revista Cubana de Ciencias Informáticas* **10(1)** 95

[9] Rodríguez R, Jorge E, Barrera F, Harry A, Bautista M, Sandra P 2011 Software para el filtrado de páginas web pornográficas basado en el clasificador KNN-UDWEBPORN *Revista Avances en Sistemas e Informática* **8(1)** 43

[10] Delgado Castillo D, Martín Pérez R, Hernández Pérez L, Orozco Morález R, & Lorenzo Ginori J 2016 Algoritmos de aprendizaje automático para la clasificación de neuronas piramidales afectadas por el envejecimiento *Revista Cubana de Informática Médica* **16(3)** 559

[11] Fix E, & Hodges Jr J L 1951 Discriminatory analysis-nonparametric discrimination: Consistency properties (Berkeley: University of California)

[12] Biau G, & Devroye L 2015 Lectures on the nearest neighbor method (Switzerland: Springer International Publishing)

[13] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, & Vanderplas J 2011 Scikit-learn: Machine learning in Python *Journal of Machine Learning Research* **12** 2825

[14] Garreta R, & Moncecchi G 2013 Learning scikit-learn: Machine learning in python (Birmingham: Packt Publishing Ltd.)