

Use of Data Mining to Identify Preferences for Humanistic Courses and Support the Management of University Extension of the Autonomous University of Aguascalientes

Humberto Muñoz Bautista¹, Miguel Ortiz-Esparza², Jaime Muñoz Arteaga¹, Claudia Acra-Despradel³ and Klinge Villalba-Condori⁴

¹ Universidad Autónoma de Aguascalientes, Av. Universidad # 940, Ciudad Universitaria, C.P. 20100, Aguascalientes, Ags. México.

² Center for Research in Mathematics, Quantum Knowledge City, 98160, Zacatecas, Mexico

³ Universidad Nacional Pedro Henríquez Ureña, Santo Domingo, Dominican Republic

⁴ Universidad Católica de Santa María, San José S/N, Arequipa, Perú

Abstract

Humanistic courses are a requirement of the Autonomous University of Aguascalientes for all its students, who must take at least three courses from different disciplines during their career to graduate. The courses offered are a way in which students from different careers can relate to each other, and in this way, develop comprehensively. These courses are offered year after year with minimal change as to which courses and at what times they are offered. The studies regarding the real demand of the students are almost non-existent, and for the same reason the courses offered do not always serve to meet the demand of the students, who find themselves in the need to take courses that they normally do not attend. would sign up to meet the requirements. In addition to the needs of the students, the problem of course management arises, where a process is carried out for the registration of courses to be offered, publication and courses and management of written students.

Keywords

Humanist Courses, Schedule, Data mining

1. Introduction

For years the Autonomous University of Aguascalientes has aimed to create professionals with comprehensive training in all aspects of their life, which is why it has insisted that its students take courses outside their area and with students from other careers. so that in this way they have knowledge in other areas and meet people from other careers with interests in common with them. This is why the University made the decision to create humanistic training courses, which are a set of courses that seek to expand the knowledge of university students in areas that are not normally studied in their careers but that are of interest. of some students.

Every year a considerable number of courses are opened in two modalities, intensive courses, which are opened in inter-semester periods for three weeks, and extensive courses, which are courses that cover the entire semester. The investment made to be able to have this offer is too great, from hiring teachers for each course to the materials consumed during them. The courses that are opened year after year vary very little, since the study plans or the courses to be taught are rarely radically modified; This is done with little or no consideration of the general opinion of the university community, since despite the demand that this generates, the courses remain static in terms of supply. This generates a large number of problems, since many courses are offered


CITIE 2022: International Congress on Trends in Educational Innovation, December 12-04, 2023, Zacatecas, Zacatecas

✉ hmuntista@gmail.com (H. Muñoz-Bautista); miguel.ortiz@cimat.mx (M. Ortiz-Esparza); jaime.munoz@edu.uaa.mx (J. Muñoz-Arteaga); c.acra@unphu.edu.do (C. Acra-Despradel); kvillalba@ucsm.edu.pe (K. Villalba-Condori)

ORCID 0000-0003-1720-0554 (H. Muñoz-Bautista); 0000-0001-8762-5780 (M. Ortiz-Esparza); 0000-0002-3635-7592 (J. Muñoz-Arteaga); 0000-0002-6429-5675 (C. Acra-Despradel); 0000-0002-8621-7942 (K. Villalba-Condori)



© 2023 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

that are only filled with students who enter only to meet the university's requirements and not because they would really like to take the course; On the other hand, there are courses that have excessive demand and of which only one or two groups are opened per period, regardless of the fact that there are a large number of students who are left out of said courses due to lack of places to take them.

It is for this reason that the decision was made to carry out a study, through surveys of current students at the university, about which courses they would really like to take, at what times and in what modality of those offered. This is with the intention of analyzing whether the courses that are opened efficiently satisfy the demand of the student community, as well as to see which are the courses that are really demanded by the students, which are those that are opened unnecessarily, and which are They are the ones who should expand their offer. Everything is done with the aim of the university improving the current offer of humanistic training courses so that it really satisfies the demand of the students, and that this requirement really meets the purpose for which it was created.

2. Related Work

The organization and search of metadata using developed software or specialized software such as Weka is in constant development due to the growing demand for data to be processed, which is why more precise and advanced approaches are increasingly required for the efficient management of the data. information [10,12,18].

Regarding data search, many algorithms have been proposed that allow for agile queries of large data sets. The improvement in search speed, as well as in obtaining information, is thanks to indices that have been developed as B, B+ trees or hash tables, which are types of structures through which data is represented in an orderly manner that allows insertions. and efficient deletions thereof [9]. In the search and organization of data, the implementation of metadata is of utmost relevance, since these provide very important information such as origin, creation dates, formats, etc. Which allows the precision and efficiency of searches to increase [7].

The state of the art is constantly evolving because, as mentioned in previous points, the data handled is increasing exponentially, which is why the development of increasingly sophisticated, precise, agile and optimized algorithms is promoted. [18].

Currently, no field study has been carried out regarding the humanistic courses at the Autonomous University of Aguascalientes, the few modifications that are made in the study plans or in their offer are based solely on the registrations that are made during the previous periods. This is one of the worst ways to carry out this analysis, since many students enter courses in which they have no interest, but still enroll to fulfill the course requirements; this results in an inefficient course offering for the university community.

2.1. Assignment of Schedule Problem

The assignment of schedules problem is a recurrent situation in all educational institutions [3], where the schedule of the students is planned, assigning signatures, classrooms, and teachers, with the objective of avoiding an overlap of hours. It corresponds to optimization problems in computational complexity theory classified as NP-complete problems or NP-hard problems [5], due to its complexity, resource limitations and number of restrictions. These problems requiring the entry of applications with intelligent algorithms [17]

2.2. Weka

Weka is a software tool widely used in the field of data mining and machine learning. Weka, named after "Waikato Environment for Knowledge Analysis", was developed at the University of Waikato in New Zealand [11]. Weka provides a collection of machine learning algorithms for

classification, regression, clustering, and rule extraction tasks. Additionally, it includes tools for data preprocessing, model evaluation, visualization, and data exploration.

2.3. Data Science: Algorithmic

Within data science we have different techniques for data processing, such as classification, clustering and prediction algorithms, with the option of being able to combine algorithms with each other to obtain a better result.; however regardless of the approach taken for it, asides from the requirements of implementation to cover and the selection criteria according to the problem to solve it is important to identify beforehand the area of implementation and the intrinsic characteristics that it will provide.[6]

Within the renowned areas of computer science to cover the demands of the society the branches of Artificial Intelligence and Data Science provide the basis for the development of modern technologies and data driven based solutions [1,16]

The following table shows the algorithmic techniques taken into consideration [4, 2, 14, 15], the model they are part of, a brief description, and remarks of the thought process regarding the selection of the supervised learning model and the unsupervised model:

Table 1
Algorithmic techniques

Learning	Model	Description	Remarks
Supervised	Decision Tree	Weighted decision trees.	Use the principle of Shannon's information theory.
Supervised	Random Forests	Group of trees with characteristics.	High accuracy and training complexity. They usually use the result of the decision trees for the input of the next tree.
Unsupervised	Expectation Maximization algorithm (EM)	Iterative method for maximum likelihood estimation of parameters.	Ensures convergence of the likelihood of function. Helps when variables are missing and in poorly conditioned problems.
Unsupervised	K-means	Grouping by distance between data.	Mainly the Euclidean distance is used to obtain the clusters by characteristics.
Unsupervised	Hierarchical Clustering	Create ascending or descending group hierarchy.	Generally, uses a greedy algorithm.
Unsupervised	Gaussian Mixture Models	Probabilistic model that representing normally distributed data	Suffers in data scalability, making it unsuitable for large data sets

3. Problem Outline

A valuable tool in the field of data science is Weka software. Weka provides a wide range of machine learning algorithms and tools for data preprocessing and analysis. Its user-friendly environment and extensibility make it a popular choice for researchers and professionals in the field of data mining. [8] To begin the study, a survey was first planned that could, with the smallest number of questions, gather the necessary data to know the trends of university students regarding humanistic courses, from which are the most requested to which are the most popular. trends by major and sex for each course.

To carry out this study, 680 surveys were carried out randomly among students from all possible centers and careers to form a statistical sample about which humanistic courses have a real demand. The survey was designed to collect the following types of data:

- Edad [18-25]
- Sexo [M - F]
- Carrera [Tipificados]
- Turno [M - V]
- Horario del curso [M V S]
- Tipo del curso [I E A]
- Curso que tomaría [Tipificados]

General information about the students is requested, such as age; obtaining a range of between 18 and 25 years among the respondents; sex, career; Students could select from the university's list of majors to avoid a problem in capturing the major's name and a subsequent problem in data cleaning; shift; morning (M) or afternoon (F) according to your class schedule; course schedule; The university offers morning (M), afternoon (V) and Saturday (S) schedules for humanist courses; type of course; Within the courses there is an intensive (I) and extensive (E) modality, so students could select the modality of their preference or both (A); and the course they would take; the list of courses offered by the university is taken.

The surveys were carried out within the university city to cover the current university population. It was sought that most of the students interviewed were within the first semesters of their degrees, thus ensuring that the answers were as close to reality as possible, since during this period, most students are studying or finishing to complete this requirement.

Once the 680 surveys were completed, the information obtained was cleaned, since there were several entries with missing data, which were filled with the average data, and the information was emptied into a database.

Once the complete database was available, we proceeded to choose which data were necessary and which would only get in the way when carrying out a data mining process; It was determined that the data chosen initially were those necessary to correctly carry out the study that was wanted to be carried out and therefore no fields were eliminated.

The database was then moved to a file with an. arff extension, which briefly described the type of data used and all the instances in the registry.

Using the weka program, the data was analyzed. The knowledge base was clustered using the EM algorithm, because when this procedure was carried out through KMEANS the error obtained was too large, which is why it was determined that the best way to classify this base was through the EM algorithm.

Attribute	Cluster						
	0 (0.06)	1 (0.15)	2 (0.23)	3 (0.17)	4 (0.16)	5 (0.19)	6 (0.03)
edad							
mean	20.5308	19.3487	20.2022	19.7162	18.7102	18.3782	20.5129
std. dev.	1.2434	0.491	0.4226	2.001	0.5787	0.4852	0.5043
sexo							
F	1.0025	28.196	37.2628	18.7861	11.288	35.4648	3.9998
M	12.0106	1.048	6.9271	13.2412	19.5721	1.0799	4.1211
[total]	13.0131	29.244	44.1899	32.0273	30.8601	36.5448	8.1209
semestre							
mean	5.7493	3.3118	3.2149	2	2	1.5662	7.9963
std. dev.	2.5013	0.8321	0.4109	1.6187	1.6187	0.4957	0.1226
carrera							
Administracion_de_Empresas	1.0109	1	1	1	1	1	6.9891
Administracion_Financiera	2.0361	1.0087	1	1.1338	2.8168	7.0046	1
Computacion_Inteligente	1.0005	1.9418	1.0009	1.2998	15.508	3.2489	1
Contador_Publico	1	1.0046	1	1.0191	3.9756	10.0007	1
Diseno_de_Modas	1.0001	1.018	18.9817	1	1	1.0001	1.0001
Electronica	6.8891	1.0016	3.9793	1	1	1	1.13
Ensenanza_del_Ingles	1.0002	3.9789	7.0205	1	1	1.0002	1.0001
Estomatologia	1.0002	15.6323	1.016	4.2602	1.0282	1.0631	1
Gestion_Turistica	1.0004	1.0076	1.0008	26.4707	2.8048	3.7157	1
Ingenieria_Bioquimica	1	1.0141	1	1.2258	1.0043	12.7557	1
Lenguas_Hispanicas	1.0002	3.857	1.0004	1.4044	6.9829	3.7551	1
Mercadotecnia	2.0469	1.0001	1	1.2135	2.7395	1	1
Psicologia	4.0285	6.7792	16.1903	1	1	1.0004	1.0015
[total]	24.0131	40.244	55.1899	43.0273	41.8601	47.5448	19.1209
turno							
0	1.1227	28.2208	43.1372	30.68	26.3038	35.5313	1.0042
1	11.8904	1.0232	1.0527	1.3473	4.5563	1.0135	7.1166
[total]	13.0131	29.244	44.1899	32.0273	30.8601	36.5448	8.1209
horario							
0	1.0263	13.3045	24.7369	2.5024	17.2135	4.2234	6.9931
1	1.0004	4.047	13.0043	13.6116	1.7473	29.5894	1.0001
2	2.0361	6.7894	1.012	1.6366	2.8873	3.6386	1
3	10.9504	7.1031	7.4368	16.2767	11.0119	1.0934	1.1276
[total]	15.0131	31.244	46.1899	34.0273	32.8601	38.5448	10.1209
tipo							
0	6.2007	1.0821	33.8476	17.4742	18.1233	4.265	7.0071
1	3.9231	19.0929	9.831	14.0766	1.1073	26.9677	1.0015

Figure 1: Clustering of analyzed data

After carrying out said clustering, the results were analyzed using the graphs that were shown to us as a result. Of these graphs, there are 3 that attract attention. The first is the one that shows the sex (Y) with respect to the course they choose (X).

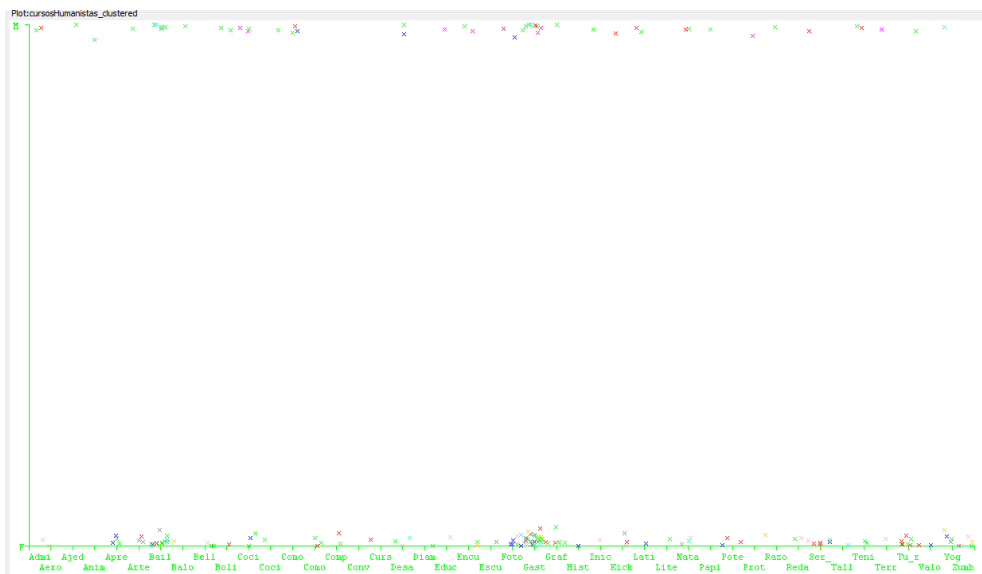


Figure 2: Data visualization of course selection by gender

From this graph it can be determined that the courses: Ballroom Dancing, Photography, Gastronomy, Graphology, Yoga and Zumba, are the focuses of interest for the female population of the University, since it is where about 75% of those interviewed said that I would take one of those courses if possible. On the other hand, men only have a very weak focus of interest in the Gastronomy course, since their population is distributed more homogeneously in all courses, without having much inclination for anyone.

The next graph of interest was the one that shows the career (Y) with respect to the course they choose (X).

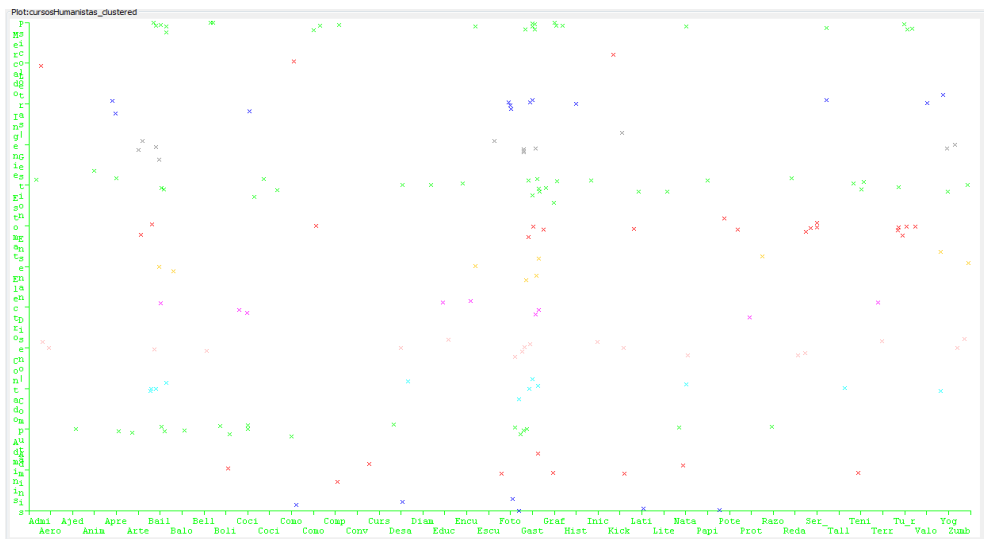


Figure 3: Data visualization of course selection by carrier

In this graph it can be seen that the majors from the Economics and Administration center have a tendency to enter the Gastronomy and Photography courses, while those from the other centers have a greater tendency to enter the Ballroom Dance, Yoga courses. and Zumba.

Finally, the last graph of interest is the graph that shows the type of course (Y) with respect to the course they choose (X).

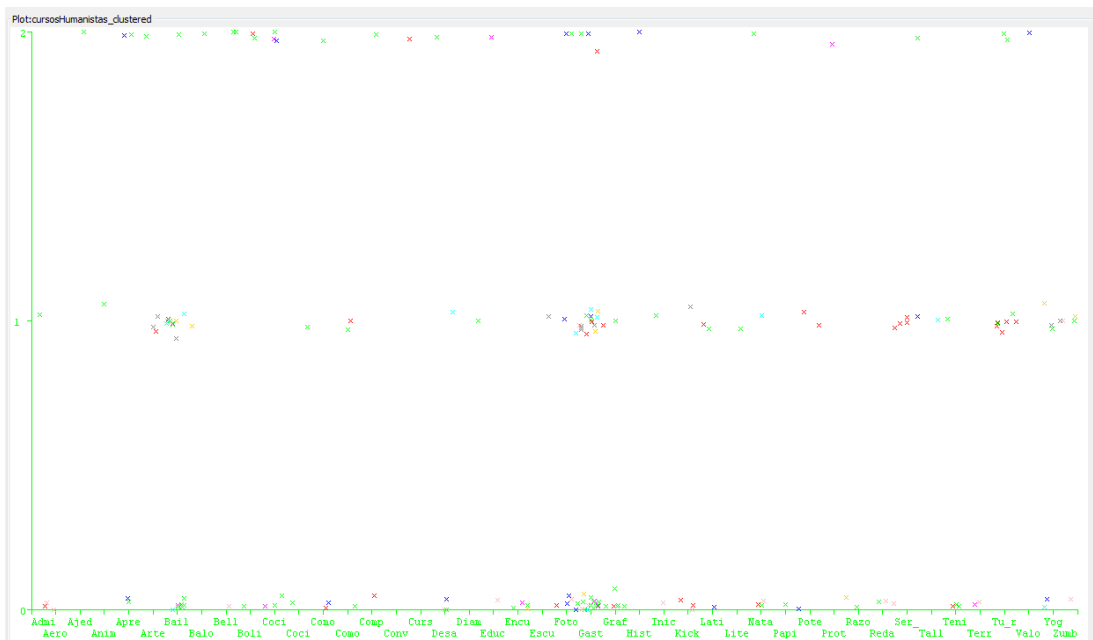


Figure 4: Data visualization of course selection by modality

In this graph you can see how the most requested type of course, for Gastronomy, Ballroom Dancing, Zumba and Yoga, is extensive and not intensive as would have been expected. However, for the other courses, the intensive period is the most requested to carry out.

4. Conclusions

After an exhaustive analysis of the different results obtained through the weka program to the knowledge base, it was determined that currently the most requested courses by the university community are the Gastronomy, Ballroom Dance and Zumba courses, but these courses They do not open a sufficient number of places to satisfy the large number of demand they have, so students have to look for other courses of less interest but that are also filled, such as Yoga, Graphology and Art with clay; However, even with these courses, student demand continues to be greater than the number of places needed, so finally students resort to courses that are not of interest to them but still have places to take them.

If in the future it is planned to open more courses, it should first be considered to open more groups of the most requested courses. If this is not possible due to lack of facilities or teachers, courses should be opened that have activities related to the most requested courses for this purpose. way to meet student demand.

References

- [1] Ahmed, Z., Mohamed, K., Zeeshan, S., Dong, X.Q.: Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. Database: The Journal of Biological Databases and Curation 2020 (2020). <https://doi.org/10.1093/DATABASE/BAAA010>
- [2] Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., Aljaaf, A.J.: A systematic review on supervised and unsupervised machine learning algorithms for data science. Supervised and unsupervised learning for data science pp. 3–21 (2020)
- [3] Baldonado, M., Chang, C.-C.K., Gravano, L., Paepcke, A.: The Stanford Digital Library Metadata Architecture. *Int. J. Digit. Libr.* 1 (1997) 108–121
- [4] Berry, M.W., Mohamed, A., Yap, B.W.: Supervised and unsupervised learning for data science. Springer (2019)
- [5] Bruce, K.B., Cardelli, L., Pierce, B.C.: Comparing Object Encodings. In: Abadi, M., Ito, T. (eds.): Theoretical Aspects of Computer Software. Lecture Notes in Computer Science, Vol. 1281. Springer-Verlag, Berlin Heidelberg New York (1997) 415–438
- [6] Calderón-Reyes, J. E., Muñoz-Bautista, H., Alvarez-Rodriguez, F. J., Barba-Gonzalez, M. L., & Cardona-Reyes, H. (2022, October). Data Science Based Methodology: Design Process of a Correlation Model Between EEG Signals and Brain Regions Mapping in Anxiety. In International Conference on Software Process Improvement (pp. 141-151). Cham: Springer International Publishing.
- [7] García, E. R., & García, F. J. E. (2017). Minería de Datos. Pearson Educación.
- [8] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1), 10-18. <https://doi.org/10.1145/1656274.1656278>
- [9] Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.
- [10] Li, W., & Özsu, M. T. (2018). *Encyclopedia of Database Systems*. Springer.
- [11] Martínez Abad, F. (2018). Aplicación de técnicas de minería de datos con software Weka.
- [12] Oliva Córdova, L. M., Amado-Salvatierra, H. R., & Villalba Condori, K. O. (2019). An experience making use of learning analytics techniques in discussion forums to improve the interaction in learning ecosystems. In *Learning and Collaboration Technologies. Designing Learning Experiences: 6th International Conference, LCT 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part I* 21 (pp. 64-76). Springer International Publishing.
- [13] Paucar-Curasma, R., Villalba-Condori, K., Arias-Chavez, D., Le, N.-T., Garcia-Tejada, G., & Frango-Silveira, I. (2022). Evaluación del Pensamiento Computacional utilizando cuatro

- robots educativos con estudiantes de primaria en Perú. *Education in the Knowledge Society (EKS)*, 23. <https://doi.org/10.14201/eks.26161>
- [14] Pinto, R. C., & Engel, P. M. (2015). A fast incremental gaussian mixture model. *PloS one*, 10(10), e0139931.
- [15] Sammaknejad, N., Zhao, Y., & Huang, B. (2019). A review of the expectation maximization algorithm in data-driven process identification. *Journal of process control*, 73, 123-136.
- [16] Sharma, S., Toshniwal, D.: Mr-ovntsa: a heuristics based sensitive pattern hiding approach for big data. *Applied Intelligence* 50, 4241– 4260 (12 2020). <https://doi.org/10.1007/S10489-020-01749-6>, <https://link.springer.com/article/10.1007/s10489-020-01749-6>
- [17] Van Leeuwen, J. (ed.): *Computer Science Today. Recent Trends and Developments. Lecture Notes in Computer Science*, Vol. 1000. Springer-Verlag, Berlin Heidelberg New York (1995)
- [18] Witten, I. H., Frank, E., & Hall, M. A. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.