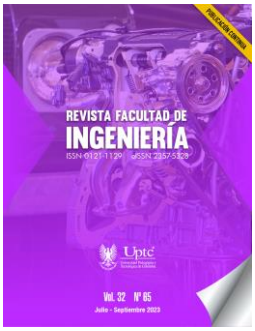
Revista Facultad de Ingeniería
Journal Homepage: https://revistas.uptc.edu.co/index.php/ingenieria



Measuring Representativeness Using Covering Array Principles

Alexander Castro-Romero[1]
Carlos-Alberto Cobos-Lozada[2]

Received: March 13, 2023 **Accepted:** September 11, 2023 **Published:** September 13, 2023

Citation: A. Castro-Romero, C.-A. Cobos-Lozada, "Measuring Representativeness Using Covering Array Principles," *Revista Facultad de Ingeniería*, vol. 32, no. 65, e15314, 2023. https://doi.org/10.19053/01211129.v32.n65.2023.15314

Abstract

Representativeness is an important data quality characteristic in data science processes; a data sample is said to be representative when it reflects a larger group as accurately as possible. Having low representativeness indices in the data can lead to the generation of biased models. Hence, this study shows the elements that make up a new model for measuring representativeness using a mathematical object testing element of coverage arrays called the "P Matrix". To test the model, an experiment was proposed where a data set is taken, divided into training and test data subsets using two sampling strategies: Random and Stratified, and the representativeness values are compared. If the data division is adequate, the two sampling strategies should present similar representativeness indexes. The model was implemented in a prototype software using Python (for data processing) and Vue (for data visualization)

[1] M. Sc. Universidad Pedagógica y Tecnológica de Colombia (Tunja-Boyacá, Colombia). alexander.castro01@uptc.edu.co. ORCID: 0000-0001-9469-5445
[2] Ph. D. Universidad del Cauca (Popayán-Cauca, Colombia). ccobos@unicauca.edu.co. ORCID: 0000-0002-6263-1911

technologies, this version of the model only allows to analyze binary data sets (for now). To test the model, the "Wines" dataset (UC Irvine Machine Learning Repository) was fitted. The conclusion is that both sampling strategies generate similar representativeness results for this dataset, although this result is predictable, it is clear that adequate representativeness of the data is important when generating the test and training datasets subsets. Therefore, as future work we plan to extend the model to categorical data and explore more complex datasets.

Keywords: classification algorithms; coverage arrays; data quality; data sets; data representativeness.

Medición de la representatividad utilizando principios de la matriz de cobertura

Resumen

La representatividad es una característica importante de la calidad de los datos en procesos de ciencia de datos; se dice que una muestra de datos es representativa cuando refleja a un grupo más grande con la mayor precisión posible. Tener bajos índices de representatividad en los datos puede conducir a la generación de modelos sesgados, por tanto, este estudio muestra los elementos que conforman un nuevo modelo para medir la representatividad utilizando un elemento de prueba de objetos matemáticos de matrices de cobertura llamado "Matriz P". Para probar el modelo se propuso un experimento donde se toma un conjunto de datos y se divide en subconjuntos de datos de entrenamiento y prueba utilizando dos estrategias de muestreo: Aleatorio y Estratificado, finalmente, se comparan los valores de representatividad. Si la división de datos es adecuada, las dos estrategias de muestreo deben presentar índices de representatividad similares. El modelo se implementó en un software prototipo usando tecnologías Python (para procesamiento de datos) y Vue (para visualización de datos); esta versión solo permite analizar conjuntos de datos binarios (por ahora). Para probar el modelo, se ajustó el conjunto de datos "Wines" (UC Irvine Machine Learning Repository). La conclusión es que ambas estrategias de muestreo generan resultados de representatividad similares para este conjunto de datos. Aunque este resultado es predecible, está claro que la representatividad adecuada de los datos es importante al generar subconjuntos de conjuntos de datos de prueba y entrenamiento, por lo tanto, como trabajo futuro, planeamos extender el modelo a datos categóricos y explorar conjuntos de datos más complejos.

Palabras clave: algoritmos de clasificación; calidad de los datos; conjuntos de datos; matrices de cobertura; representatividad de los datos.

Medindo a representatividade usando os princípios da matriz de cobertura

Resumo

A representatividade é uma característica importante da qualidade dos dados nos processos de ciência de dados; Uma amostra de dados é considerada representativa quando reflete um grupo maior com a maior precisão possível. Ter baixos índices de representatividade nos dados pode levar à geração de modelos viesados, portanto, este estudo mostra os elementos que compõem um novo modelo para medir a representatividade utilizando um elemento de teste de objetos matemáticos de matrizes de cobertura denominado “Matriz P”. Para testar o modelo foi proposto um experimento onde um conjunto de dados é retirado e dividido em subconjuntos de dados de treinamento e de teste utilizando duas estratégias de amostragem: Aleatória e Estratificada, por fim, os valores de representatividade são comparados. Se a divisão dos dados for adequada, as duas estratégias de amostragem deverão apresentar índices de representatividade semelhantes. O modelo foi implementado em software protótipo utilizando tecnologias Python (para processamento de dados) e Vue (para visualização de dados); Esta versão permite apenas analisar conjuntos de dados binários (por enquanto). Para testar o modelo, foi ajustado o conjunto de dados “Wines” (UC Irvine Machine Learning Repository). A conclusão é que ambas as estratégias de amostragem geram resultados de representatividade semelhantes para este conjunto de dados. Embora este resultado seja previsível, fica claro que a representatividade adequada dos dados é importante ao gerar subconjuntos de conjuntos de dados de treinamento e teste, portanto, como trabalho futuro, planejamos estender o modelo para dados categóricos e explorar conjuntos de dados maiores e complexos.

Palavras-chave: algoritmos de classificação; qualidade dos dados; conjuntos de dados; matrizes de cobertura; representatividade dos dados.

I. INTRODUCTION

Data science leverages data to support decision making and occupies a more important place within organizations every day. However, the data on which it is based are not always of adequate quality, thus promoting inadequate decision making. In this sense, Srivastava et al. [1] state that "high quality data is critical for effective data science".

Now, as Clarke [2] mentions, "the Big data literature, both academic and professional, focuses heavily on opportunities. Less attention has been paid to the threats that arise from reusing data, consolidating data from multiple sources, applying analytical tools to the resulting collections, drawing inferences and acting on them." In other words, much progress has been made in algorithms and data processing techniques, but aspects such as data quality have been neglected.

It is important to note that in any data-driven project, data quality verification becomes relevant. These specific tasks are mainly related to the validation of data ranges, the handling of missing data, the detection and handling of outliers in each attribute, among others [3]. However, at this stage there is no explicit emphasis on the analysis of the representativeness of the data initially collected in relation to the reality or the universe of the business that is expected to be modelling, a fact that can have negative consequences in the generated models.

Representativeness is defined as the level of participation of individuals as a combinatorial element of characteristics in the sample (or in the data available for the data science process). In more precise terms, it seeks to measure whether in a population, where two or more characteristics are studied, there is at least a minimum number of representatives of each combination of characteristics. For example, if the characteristics are type (private or public) and location (urban or rural), there is at least a certain number of elements of each combination: urban private, rural private, urban public, and rural public.

Thus, errors in the collection, construction, treatment, or sampling of the data set can have implications on the representativeness of the data, and these can lead to problems in the generated models. For instance, those outlined by Yapo and Weiss [4]: racism in the results of image searches, sexism in how advertising is displayed to people, and even discrimination against minorities in criminal risk software, which in general results in consequences that should be highlighted as a social problem. Supporting the above and from a more technical point of view, Polyzotis et al. [5]

indicate that "for the development of reliable, robust and understandable machine learning models... it is necessary to build the model using high quality training data... the data supplied to the model at the time of service must be similar in distribution (and in features) to the training data, otherwise the accuracy of the model will decrease". It can then be concluded that quality data is required to obtain consistent and fair results in a social context.

Likewise, in big data contexts where it is thought that having a lot of data means that the whole population is represented, it is generally not clear how sampling is performed or if it is done at all. Rojas et al. [6] state that "sampling appears to be an important approach for exploring large datasets. However, as far as has been identified, there is no evidence in the literature on how data scientists use sampling techniques with Big Data, nor how those sampling techniques affect the quality or focus of their insights. Based on the lack of Big Data sampling tools available today, we maintain the hypothesis that data scientists are using random sampling..."

Following this line of thought, Schönberger and Cukier [7] mention that "random sampling has been a tremendous success and is the backbone of modern-scale measurement. But it is still a shortcut, a second-order alternative to collecting and analyzing the entire data set. It brings with it a few inherent weaknesses... If there are systematic biases in the way the data are collected that can make the extrapolated results grossly inaccurate". So, it is possible to conclude that, although sampling techniques are important elements in scientific studies, it is key to measure the level of representativeness of the constructed data sets to avoid possible biases in the generated models.

Currently, the mentioned problems, in terms of data quality and especially representativeness, have been addressed from statistics, through mathematical models, which generally compare data sets with external and much larger ones (such as censuses, seen as reference data) or rely on third party information to detect missing elements. However, in the context of Big Data, this task has become even more challenging. Considering all of the above, this paper presents the progress in the development of a new model to measure representativeness in datasets.

II. METHODOLOGY

In the methodology, two aspects are considered: materials that are divided into model, software, and data; and method that narrates the design of the experiment.

A. Tools

To measure representativeness in data sets, a model that was implemented through prototype software and will be tested with a recognized data set is proposed. Each of these elements will be explained in the following sections.

1) Model. The main algorithm uses a Coverage Array (CA) element called "P Matrix", Torres and Izquierdo [8] mention that CAs "are combinatorial objects that have been successfully used to automate the generation of software test cases. CAs have the characteristics of being of minimum cardinality (i.e., they minimize the number of test cases), and of maximum coverage (i.e., they guarantee the coverage of all combinations of a certain size among the input parameters)".

The above concept could be represented by the coverage array CA (5; 4, 2, 2, 2) presented in Table 1. This covering array has 5 rows (N), 4 parameters (k), it is binary (alphabet 2 -v- for the 4 columns), and strength 2 (t). Since it is a strength 2 array, combinations of values {0-0}, {0-1}, {1-0}, {1-1}, {1-1}, appear at least once for any combination of columns or parameters, as highlighted by Timaná et al. [9].

Table 1. Coverage array with configuration (5; 4, 2, 2, 2).

C1	C2	C3	C4
1	**1**	1	0
0	**0**	0	0
0	**1**	**0**	1
1	0	0	1
0	0	**1**	**1**

To verify that the coverage array complies with the minimum combinations established by the parameter, an element called Matrix P is used, as shown in Table 2. Each row counts the possible combinations with respect to the combinatorics between columns, the idea is that at least one record complies with the combinatorics.

Table 2. P Matrix.

Column Merge vs. Data Merge	00	01	10	11
C1, C2	2	1	1	1
C2, C3	2	1	1	1
C1, C3	2	1	1	1
C1, C4	1	2	1	1
C2, C4	1	2	1	1
C3, C4	1	2	1	1

For now, the model only evaluates binary type data and allows a general assessment of representativeness based on the data in the P matrix (Figure 1).

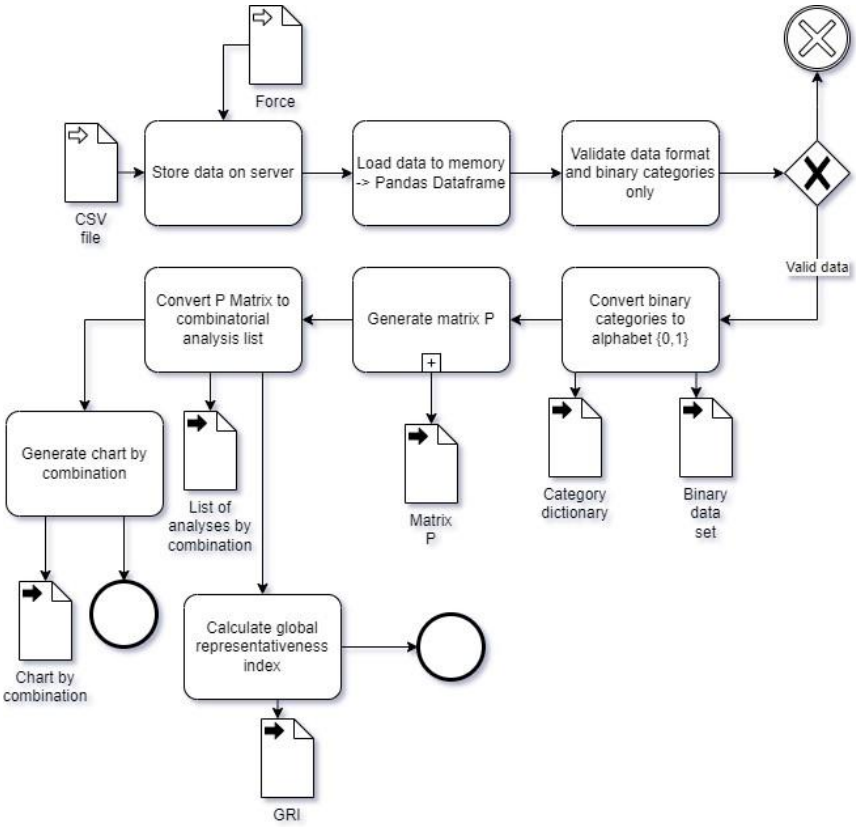


Fig. 1. Representativeness analysis model.

The representativeness index, i.e., the average of the percentage of the normalized incident count using MinMax, is used for the measurement.

2) Software. For the development of this project, an application called "Representativeness Meter" was developed (software available at https://github.com/alexander-castro/representativeness-https://github.com/alexander-castro/representativeness-front) (Table 3).

Table 3. Application technologies.

Software	Version
Python	3.10
Flask	2.2.2
Scikit-learn	0.20

Software	Version
Pandas	1.5.0
Numpy	1.23.4
Vue	3.2.45
ChartJs	3.9.1

3) Data. The "Wine Data Set" [10] was used with the following adjustments: only two classes were taken, with a total of 130 records that were converted to binary considering whether they were above or below the mean. The columns used were "Alcohol, Malic acid, Ash, Magnesium, Color intensity, Class". Available at https://github.com/alexander-castro/representativeness/blob/main/uploads/paper_data.csv.

B. Method

The experiment to be performed consists of comparing the representativeness of the data after using two classifiers—one random “train_test_split” and one stratified “StratifiedShuffleSplit”—and a tree classification algorithm “DecisionTreeClassifier” and the average accuracy metric, as shown in Figure 2.

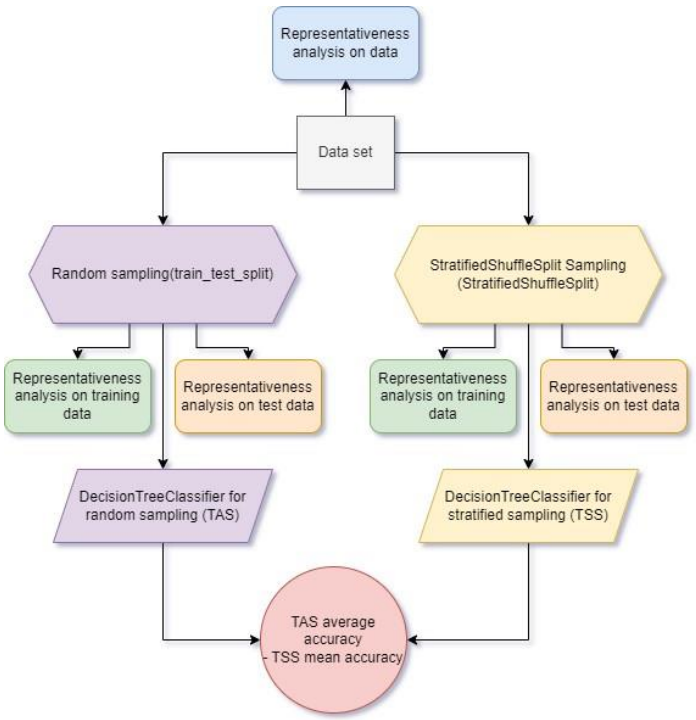


Fig. 2. Experiment design.

III. RESULTS

The data set was transformed to binary data with conversion dictionary, as shown in Figure 3.

Category\Value	0	1
Alcohol	0	1
Ash	0	1
Class	0	1
Color intensity	0	1
Magnesium	0	1
Malic acid	0	1

Fig. 3. Binary dictionary of categories.

Data analysis was performed with the complete data set, thus obtaining the P Matrix shown in Figure 4.

Row/Column	000	001	010	011	100	101	110	111
Alcohol,Malic acid,Ash	0,0,0 31	0,0,1 10	0,1,0 11	0,1,1 12	1,0,0 19	1,0,1 32	1,1,0 6	1,1,1 9
Alcohol,Malic acid,Magnesium	0,0,0 32	0,0,1 9	0,1,0 15	0,1,1 8	1,0,0 19	1,0,1 32	1,1,0 3	1,1,1 12
Alcohol,Malic acid,Color intensity	0,0,0 36	0,0,1 5	0,1,0 21	0,1,1 2	1,0,0 10	1,0,1 41	1,1,0 2	1,1,1 13
Alcohol,Ash,Magnesium	0,0,0 32	0,0,1 10	0,1,0 15	0,1,1 7	1,0,0 14	1,0,1 11	1,1,0 8	1,1,1 33
Alcohol,Ash,Color intensity	0,0,0 38	0,0,1 4	0,1,0 19	0,1,1 3	1,0,0 5	1,0,1 20	1,1,0 7	1,1,1 34
Alcohol,Magnesium,Color intensity	0,0,0 43	0,0,1 4	0,1,0 14	0,1,1 3	1,0,0 7	1,0,1 15	1,1,0 5	1,1,1 39
Malic acid,Ash,Magnesium	0,0,0 34	0,0,1 16	0,1,0 17	0,1,1 25	1,0,0 12	1,0,1 5	1,1,0 6	1,1,1 15
Malic acid,Ash,Color intensity	0,0,0 31	0,0,1 19	0,1,0 15	0,1,1 27	1,0,0 12	1,0,1 5	1,1,0 11	1,1,1 10
Malic acid,Magnesium,Color intensity	0,0,0 34	0,0,1 17	0,1,0 12	0,1,1 29	1,0,0 16	1,0,1 2	1,1,0 7	1,1,1 13
Ash,Magnesium,Color intensity	0,0,0 34	0,0,1 12	0,1,0 9	0,1,1 12	1,0,0 16	1,0,1 7	1,1,0 10	1,1,1 30

Fig. 4. P Matrix.

This model can be a little difficult to analyze, so the top three combinations of the three most representative and least representative forces are shown. Figure 5 presents the results; the last column shows the percentage of normalized representativeness.

Top	Combination	Values	Count	Percentage	Normalized percentage
1	Alcohol-Magnesium-Color intensity	0-0-0	35	3.37%	100%
2	Alcohol-Malic acid-Color intensity	1-0-1	32	3.08%	91.43%
3	Alcohol-Malic acid-Color intensity	0-0-0	31	2.98%	88.57%
78	Alcohol-Malic acid-Color intensity	0-1-1	2	0.19%	5.71%
79	Alcohol-Malic acid-Color intensity	1-1-0	2	0.19%	5.71%
80	Malic acid-Magnesium-Color intensity	1-0-1	2	0.19%	5.71%

Fig. 5. Result by combinatorial force 3.

Analyzing all the records, the average percentage of representativeness was 34.76%. It is interesting to see how each combination behaves, for example, that of "Alcohol, Malic acid, Color intensity" can be seen in Figure 6.

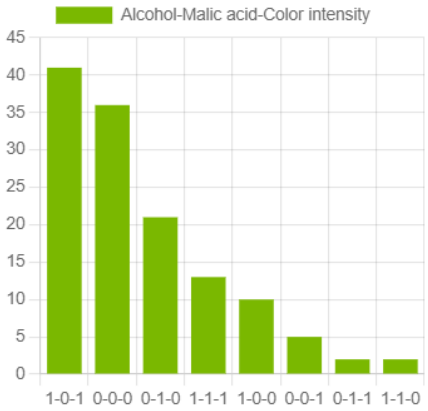


Fig. 6. Combinatorial count result "Alcohol, Malic acid, Color intensity".

Finally, the experiment is carried out as shown in Table 4 and Table 5.

Table 4. Representativeness result - stratified sampling.

<table>
<tr><th>Train</th><th>Test</th></tr>
<tr><td>
<table>
<tr><th>Fila\Columna</th><th>000</th><th>001</th><th>010</th><th>011</th><th>100</th><th>101</th><th>110</th><th>111</th></tr>
<tr><td>Alcohol,Malic acid,Ash</td><td>24</td><td>8</td><td>9</td><td>9</td><td>16</td><td>27</td><td>5</td><td>6</td></tr>
<tr><td>Alcohol,Malic acid,Magnesium</td><td>26</td><td>6</td><td>12</td><td>6</td><td>15</td><td>28</td><td>3</td><td>8</td></tr>
<tr><td>Alcohol,Malic acid,Color intensity</td><td>28</td><td>4</td><td>17</td><td>1</td><td>9</td><td>34</td><td>2</td><td>9</td></tr>
<tr><td>Alcohol,Ash,Magnesium</td><td>25</td><td>8</td><td>13</td><td>4</td><td>11</td><td>10</td><td>7</td><td>26</td></tr>
<tr><td>Alcohol,Ash,Color intensity</td><td>30</td><td>3</td><td>15</td><td>2</td><td>5</td><td>16</td><td>6</td><td>27</td></tr>
<tr><td>Alcohol,Magnesium,Color intensity</td><td>35</td><td>3</td><td>10</td><td>2</td><td>7</td><td>11</td><td>4</td><td>32</td></tr>
<tr><td>Malic acid,Ash,Magnesium</td><td>26</td><td>14</td><td>15</td><td>20</td><td>10</td><td>4</td><td>5</td><td>10</td></tr>
<tr><td>Malic acid,Ash,Color intensity</td><td>25</td><td>15</td><td>12</td><td>23</td><td>10</td><td>4</td><td>9</td><td>6</td></tr>
<tr><td>Malic acid,Magnesium,Color intensity</td><td>29</td><td>12</td><td>8</td><td>26</td><td>13</td><td>2</td><td>6</td><td>8</td></tr>
<tr><td>Ash,Magnesium,Color intensity</td><td>28</td><td>8</td><td>7</td><td>11</td><td>14</td><td>6</td><td>7</td><td>23</td></tr>
</table>
</td><td>
<table>
<tr><th>Fila\Columna</th><th>000</th><th>001</th><th>010</th><th>011</th><th>100</th><th>101</th><th>110</th><th>111</th></tr>
<tr><td>Alcohol,Malic acid,Ash</td><td>7</td><td>2</td><td>2</td><td>3</td><td>3</td><td>5</td><td>1</td><td>3</td></tr>
<tr><td>Alcohol,Malic acid,Magnesium</td><td>6</td><td>3</td><td>3</td><td>2</td><td>4</td><td>4</td><td>0</td><td>4</td></tr>
<tr><td>Alcohol,Malic acid,Color intensity</td><td>8</td><td>1</td><td>4</td><td>1</td><td>1</td><td>7</td><td>0</td><td>4</td></tr>
<tr><td>Alcohol,Ash,Magnesium</td><td>7</td><td>2</td><td>2</td><td>3</td><td>3</td><td>1</td><td>1</td><td>7</td></tr>
<tr><td>Alcohol,Ash,Color intensity</td><td>8</td><td>1</td><td>4</td><td>1</td><td>0</td><td>4</td><td>1</td><td>7</td></tr>
<tr><td>Alcohol,Magnesium,Color intensity</td><td>8</td><td>1</td><td>4</td><td>1</td><td>0</td><td>4</td><td>1</td><td>7</td></tr>
<tr><td>Malic acid,Ash,Magnesium</td><td>8</td><td>2</td><td>2</td><td>5</td><td>2</td><td>1</td><td>1</td><td>5</td></tr>
<tr><td>Malic acid,Ash,Color intensity</td><td>6</td><td>4</td><td>3</td><td>4</td><td>2</td><td>1</td><td>2</td><td>4</td></tr>
<tr><td>Malic acid,Magnesium,Color intensity</td><td>5</td><td>5</td><td>4</td><td>3</td><td>3</td><td>0</td><td>1</td><td>5</td></tr>
<tr><td>Ash,Magnesium,Color intensity</td><td>6</td><td>4</td><td>2</td><td>1</td><td>2</td><td>1</td><td>3</td><td>7</td></tr>
</table>
</td></tr>
<tr><td>Representativeness percentage: 35.29%.</td><td>Representativeness percentage: 40.63%.</td></tr>
<tr><td colspan="2">Accuracy: 0.88</td></tr>
</table>

Table 5. Representativeness result - random sampling.

Train

Fila\Columna	000	001	010	011	100	101	110	111
Alcohol,Malic acid,Ash	25	10	8	11	15	24	5	6
Alcohol,Malic acid,Magnesium	26	9	12	7	14	25	3	8
Alcohol,Malic acid,Color intensity	31	4	17	2	7	32	2	9
Alcohol,Ash,Magnesium	23	10	15	6	12	8	5	25
Alcohol,Ash,Color intensity	30	3	18	3	5	15	4	26
Alcohol,Magnesium,Color intensity	35	3	13	3	6	11	3	30
Malic acid,Ash,Magnesium	26	14	14	20	9	4	6	11
Malic acid,Ash,Color intensity	26	14	12	22	9	4	10	7
Malic acid,Magnesium,Color intensity	28	12	10	24	13	2	6	9
Ash,Magnesium,Color intensity	26	9	9	9	15	5	7	24

Representativeness percentage: 33.33%.

Test

Fila\Columna	000	001	010	011	100	101	110	111
Alcohol,Malic acid,Ash	6	0	3	1	4	8	1	3
Alcohol,Malic acid,Magnesium	6	0	3	1	5	7	0	4
Alcohol,Malic acid,Color intensity	5	1	4	0	3	9	0	4
Alcohol,Ash,Magnesium	9	0	0	1	2	3	3	8
Alcohol,Ash,Color intensity	8	1	1	0	0	5	3	8
Alcohol,Magnesium,Color intensity	8	1	1	0	1	4	2	9
Malic acid,Ash,Magnesium	8	2	3	5	3	1	0	4
Malic acid,Ash,Color intensity	5	5	3	5	3	1	1	3
Malic acid,Magnesium,Color intensity	6	5	2	5	3	0	1	4
Ash,Magnesium,Color intensity	8	3	0	3	1	2	3	6

Representativeness percentage: 36.11%.

Accuracy: 0.92

IV. Discussion

The representativeness analysis based on the method of checking the coverage arrays shows poorly attended combinations within a data set, it is an exhaustive task that can be time consuming according to the strength to be analyzed. But it is interesting to see how common combinations can be either over represented or under represented. For example, Figure 4 shows how the combination of "Alcohol, Magnesium, Color intensity" for the values "0,0,0" i.e., the three below the overall average, has forty-three elements but the combination "0,1,1" i.e., Alcohol below the average and "Magnesium, Color intensity" above has only three. The question for researchers would be Is this combination really uncommon? Or maybe the way the sample was taken had to do with the low representativeness? (for example, if it was taken in a specific region where these indicators have a certain tendency).

It can be complex to answer sampling questions, but this can be the key to having more robust models, especially when looking for algorithms to be more inclusive. Now, the overall representativeness indicator is still too naive and does not give a useful approximation to the researcher, but it is an interesting starting point to understand how complete a dataset is without taking into account reference data.

Finally, it is clear that there are several ways to select data to train and evaluate an algorithm, e.g., for classification. However, the way this separation is done may depend on the results of the model, it does not mean that a higher representativeness

makes a more accurate model, but it can be a fairer starting point for algorithms that take into account most possible scenarios.

V. Conclusions

The representativeness of the data is often not taken into account in the data science process, and the consequences can be serious, especially when the algorithms are intended to have an equitable approach. Therefore, it is necessary to perform early assessments of data quality to avoid having biased models due to incomplete data. In this sense, problems such as class imbalance have been studied in depth, but the analysis of the remaining data requires more work.

A naive model that requires more iterations but gives an initial taste of how to measure representativeness in data sets, which is not so common, was proposed. The challenge lies in evolving the model to support categorical, continuous data, and, above all, establish metrics that can be interpreted by scientists and allow them to make better decisions regarding the sampling of their data.

Finally, it is important to emphasize that the creation of indicators must go hand in hand with visual aids that allow understanding different aspects of the data analyzed; the use of colors and different types of graphics to understand the analyses proposed here has been positive.

Authors' Contributions

Alexander Castro Romero: research, data analysis, software, implementation, writing - original draft.

Carlos Alberto Cobos Lozada: research, supervision, methodology, writing – review and editing.

References

[1] D. Srivastava, M. Scannapieco, T. C. Redman, “Ensuring high-quality private data for responsible data science: Vision and challenges,” *Journal of Data and Information Quality*, vol. 11, no. 1, pp. 1–9, 2019. https://doi.org/10.1145/3287168

[2] R. Clarke, “Big data, big risks,” *Information Systems Journal*, vol. 26, no. 1, pp. 77–90, 2016. https://doi.org/10.1111/isj.12088

[3] A. Alsudais, *Incorrect Data in the Widely Used Inside Airbnb Dataset*, 2020. http://arxiv.org/abs/2007.03019.

[4] A. Yapo, J. Weiss, “Ethical Implications of Bias in Machine Learning,” in *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018. https://doi.org/10.24251/hicss.2018.668

[5] N. Polyzotis, S. Roy, S. E. Whang, M. Zinkevich, “Data lifecycle challenges in production machine learning: A survey,” *SIGMOD Record*, vol. 47, no. 2, pp. 17–28, 2018. https://doi.org/10.1145/3299887.3299891

[6] J. A. Rojas, M. Beth Kery, S. Rosenthal, A. Dey, “Sampling techniques to improve big data exploration”, in *7th Symposium on Large Data Analysis and Visualization*, 2017, pp. 26–35. https://doi.org/10.1109/LDAV.2017.8231848

[7] V. Mayer-Schönberger, K. Cukier, *Big data: La revolución de los datos masivos*, Turner, 2013.

[8] J. Torres-Jimenez, I. Izquierdo-Marquez, “Survey of covering arrays,” in *15th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, 2013, pp. 20–27. https://doi.org/10.1109/SYNASC.2013.10

[9] J. Adriana Timaná-Peña, C. Alberto Cobos-Lozada, J. Torres-Jimenez, “Metaheuristic algorithms for building Covering Arrays A review,” Revista Facultad de Ingeniería, vol. 25, no. 43, pp. 31–45, 2016. https://doi.org/10.19053/01211129.v25.n43.2016.5295

[10] C. L. Blake, C. J. Merz, *UCI Repository of Machine Learning Databases*, University of California, Oakland, 1998. https://archive.ics.uci.edu/ml/datasets/wine