Zeszyty Naukowe WWSI, No 22, Vol. 14, 2020, pp. 7-21
DOI: 10.26348/znwwsi.22.7

Prediction of Missing Values in Adult Data Set of UCI Machine Learning: A Case of Study

Alejandra Luna, Mario Bello, Ana Hernández and Edmundo Bonilla*

Tecnológico Nacional de México, Instituto Tecnológico de Apizaco

Abstract

These days, not having complete data of any kind can be a big problem for different organizations when making decisions. In this article, we propose to use Shannon entropy and information gain to predict and impute missing categorical data in any data set. It is detailed with an example of how entropy is applied and knows the level of uncertainty of each attribute value. Likewise, the imputation of the missing attributes is also carried out with other imputation techniques in the Adult data set of UCI Machine Learning to denote the advantages offered by the proposed methodology.

***Keywords* —** Shannon theory, Entropy, Missing attributes, Adult dataset, UCI Machine Learning.

1 Introduction

Nowadays not counting with the full data of any kind can become a big issue for different organizations, companies or decision makers. Besides, this issue is not exclusive of any area of study in particular, this is due to the evolution of the information technologies that has caused the increase in the volume of handled information.

The missing data in a data set refer to an instance that does not carry any value. The reasons for having missing values in data sets are numerous, for example: flaws in the system that stores the data which do not fully respond to a questionnaire, the uncertainty of data that are required (when referring to a census) etc. The problem of missing data occurs in various areas. The proposed method, based on a supervised model, allows us to tackle this problem without the elimination of attributes or records with missing values even in the case of categorical data when applying statistical techniques (as generating values with appropriate distribution) is impossible.

Clearly, this problem occurs in demographic adjustments by imputation, for example: in the National Institute of Statistics, Geography and Informatics (INEGI for its initials in Spanish),

*E-mail: edbonn@walla.co.il

Manuscript received April 23, 2020

the number of households reported in 2015 was less than expected. In addition, the surveys carried out were adjusted, because they contained a large number of missing data. Therefore, there is no clear evidence of unusual household growth, forcing the INEGI to carry out the censuses, counts and surveys necessary for its measurement [1]. This institution also have other areas with the missing data issue such as poverty, disability, employment, social security, etc.

In this paper, we propose the use of the Shannon´s information theory also known as Shannon entropy to predict and impute the missing attributes of a categorical kind in any data set. For the test, the data set used was the Adult data set de UCI Machine Learning, which has 32561 instances and every instance is formed by 15 different attributes that are detailed later. This data set was made by Barry Becker from the 1994 U.S. Census Bureau database.

This document is organized as follows; Section 2 provides a brief explanation of the rise of information theory. In Section 3, the most important works are shown, where the imputation of values is addressed. In Section 4, it is detailed how the information theory is implemented for the imputation of categorical values. Section 5 presents the tests performed on the Adult data set, as well as a comparison between the technique proposed in this document with other imputation techniques. Finally, in Section 6, the conclusions derived from the results obtained and comparisons performed are explained.

2 Shannon Theory

In 1940 and 1950, Claude E. Shannon used a theorem in the simple case of process with no memory to characterize the optimum possible performance by communicating and information source through a medium or random coded channel using codes [2]. This theorem was applied directly to study the behavior of the error frequency and the average distortion in time in a communication system. Later a variation to define a mathematical entropy or information average in a process was added and pinpointing the asymptotic behavior. Therefore, the entropy is a measurement of the uncertainty and the value of the entropy allows to appreciate the uncertainty in the determination of the result of a given test or experience [3].

3 Related Works

In [4], it was exposed that the missing data are often found in the data sets used to build prediction method models. Four techniques of missing data were evaluated in the software cost modeling context, list delete (LD), media imputation (MI), similar response pattern imputation (SRPI) and full information maximum probability (FIML). By applying these techniques to ERP data sets, it was seen that FIML is the most appropriated technique when the missing data are not completely random and the LD, MI and SRPI techniques will be biased unless the data are completely random.

In [5], the authors focused on the image processing using Shannon's entropy or information theory taking into a count the histogram of the image of Gray level as a distribution of probability. In addition, the Tsallis entropy was applied as a generalization of the theory of information. The image threshold method used a non-extensive relative entropy for the first time. This method provided to probability functions that better represent the region corresponding to

the object and the background of the image. In the tests, it was known that said method is robust and can perform for a large variety of images with an optimum result.

In the last few decades the research in order to deal with missing data has received more attention, in [6], the use of multivariable imputation by chain equation (MICE), a comparison of MICE using several methods to deal with the missing data. A comparison of the different approaches of MICE methods based on the IRIS Data Set was made by Fisher and Anderson, which can be located in the repository of the UCI Machine Learning, where two criteria for evaluation were considered. To know the standard value and the duration of the average confidence interval.

In [7], mathematical derivations are presented, one computer simulation and an example of the real data to demonstrate the importance of properly specifying the dependency if the data with multiple imputation (MI) is used for incomplete in several levels. Displaying three different MI strategies with an approach based on a multilevel imputation model, produced the estimation of valid parameters of interclass correlation and regression coefficient in random interception models in most of the simulated conditions.

In other research areas such as epidemiology, the problem of missing data is also present. In [8], the authors mentioned that the missing data may constitute a big challenge in the analysis and interpretation of the results and potentially weaken the validity of the results and conclusions, including the analysis of full cases , the missing indicator method, the imputation of unique value and the sensibility analysis that are part of the sceneries (either for better or worse).

In document [9], an imputation method was proposed to predict absent attributes by supervised learning. This method assigns the problematic value of attribute missing in a data set by using the discretization based on the known and available attribute values. Then using the C4.5 algorithm the corresponding discrete values are predicted for the missing values of the attributes.

4 Methodology

In this document, the use of entropy to predict and facilitate the imputation of missing attributes in any data set is proposed, as long as this data are a categorical type. The final goal is to look for those missing values and replace them by using entropy minimization.

This methodology is represented in Figure 1, where the data base is represented with a $n\ x\ m$ matrix, the attributes are represented by V_n and each instance is represented by e_m, in the matrix is possible to see the missing attributes represented by '?'.

When the preprocessing starts, the extraction of the training set is made, which is formed by all the instances that are complete, in the same way the evidence set is extracted, which means, all those instances that have a missing attribute, then the Shannon entropy is applied to predict the missing attributes and finally the results are obtained. To exemplify the use of this methodology, the data set shown in Table 1 is proposed, which is formed by ten instances (I_1 to I_{10}), each instance has two attributes (X and Y) and values of missing attributes are denoted by '?', in this set the attribute Y is taken as a discriminant attribute.

As previously mentioned, this work focusses in no numeric data treatment, for this reason, it is necessary to replace said attribute with an equivalent, in this case are replaced by numbers,

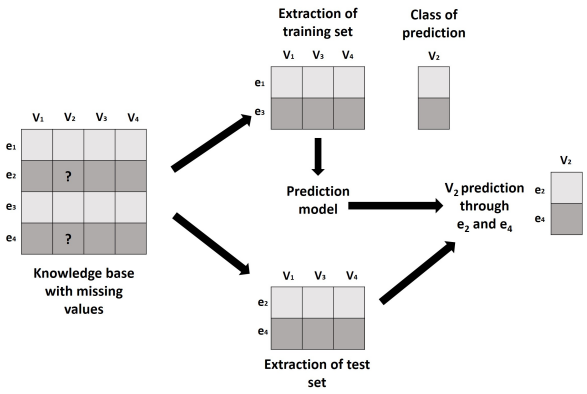

Figure 1: Graphical representation for the proposed methodology.

Table 1: Data set proposed as example.

ε	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}
X	US	KH	US	ING	US	?	KH	KH	?	US
Y	> 5	> 5	≤ 5	≤ 5	> 5	> 5	≤ 5	≤ 5	≤ 5	> 5

clarifying that this is only a label to facilitate the treatment set as: (i) 1 for the value > 5 and (ii) 2 for the value ≤ 5. As can be seen there are only two types of value in the Y_n attribute, for that reason the values 1 and 2 are assigned consequently.

Three different values had been provided for the nationalities V_n, a value has been assigned to every nationality set as: (i) $A = UnitedStates\ (US)$, (ii) $B = Cambodia\ (KH)$ and (iii) $C = England\ (ING)$. This is a process done to not use the full name of the nationality and is much more efficient in the treatment of the attributes, is also worth mentioning that all of this labeling process is called *discretization* and any label can be used to replace the different values of the attributes in a data set for an equivalent. Once this process is done, the process is shown in Table 2, where it can be seen that all of the values had been replaced by the equivalent.

That said, for the treatment of the missing attributes two fundamental concepts are required for the imputation, entropy and information gaining processes. The entropy is an uncertainty or disorder measurement, which is represented in a mathematical way in (1).

Table 2: Data set transformed.

ε	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}
X	A	B	A	C	A	?	B	B	?	A
Y	1	1	2	2	1	1	2	2	2	1

$$H(X) = -\sum_{i=1}^{k} p_i \, log_2 \, p_i \tag{1}$$

In (1), the term X represents a recollection of objects; p_i represents the probability for possible values, while i represents the possible responses or state of the objects. The information gaining is a discrimination measurement, in this work will be used as an indicator to know the attribute that should be chosen to continue with the imputation process, this measurement is represented in (2).

$$G(\varepsilon, V_i) = H(\varepsilon) - I(\varepsilon|V_i) \tag{2}$$

In (2), $G(\varepsilon, V_i)$ belongs to the information gaining of a recollection of elements (ε), regarding the different values of the attribute in the object recollection (V_i), $H(\varepsilon)$ correspond to the entropy of all object recollection, also called global entropy, while $I(\varepsilon|V_i)$ represents the entropy of the different values of the attribute (V_i) in the object recollection (ε), also called local entropy.

The steps to follow to predict the missing attributes shows in Table 2 are: (i) knowing of the training data set $(\varepsilon = \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{10})$, the number of complete examples $(\varepsilon^c = \varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5, \varepsilon_7, \varepsilon_8, \varepsilon_{10})$ and how many examples have missing value $(\varepsilon^m = \varepsilon_6, \varepsilon_9)$, and (ii) substitution of attributes according to the entropy theory. It should be considered that there are eight complete values and 2 missing values, four with nationality A of those three correspond to class 1 and only one belongs to class 2; three with nationality B of those one belongs to class and two belong to class 2; and a nationality C, that belongs to class 2. We start with the value ε_6, assuming that the absent value is A then, the substitution of missing value is done in (1) and (3) is obtained, the result is called global entropy.

$$-\left(\frac{5}{9}\right)\left[\frac{4}{5} \times log_2 \frac{4}{5} + \frac{1}{5} \times log_2 \frac{1}{5}\right] = 0.401 \tag{3}$$

In (3), in the first fraction, number 5 is the new number of examples with the value for A and 9 is the number of full examples. In the second fraction, 4 is the number of examples with nationality A and besides belong to class 1, in the fourth fraction, 1 is the number of examples with nationality A and belongs to class 2. All values increased by 1 this due to the supposition of the missing value A, the same thing happens for the calculation of the values B and C.

The same process is done now assuming that the absent value is B the substitution of values is made in (1) and the global entropy is calculated in (4).

$$-\left(\frac{4}{9}\right)\left[\frac{2}{4} \times log_2 \frac{2}{4} + \frac{2}{4} \times log_2 \frac{2}{4}\right] = 0.444 \tag{4}$$

The same process is made again assuming that the missing value is C and the substitution of values is made in (1), the global entropy is calculated in (5).

$$-\left(\frac{2}{9}\right)\left[\frac{1}{2} \times log_2 \frac{1}{2} + \frac{1}{2} \times log_2 \frac{1}{2}\right] = 0.222 \tag{5}$$

Then the sum of the obtained results in (3), (4) and (5) is made, the obtained result is calculated in (6), this result represents the entropy of the complete data set.

$$\sum = 0.401 + 0.444 + 0.222 \approx 1 \tag{6}$$

Considering the missing attribute ε_6 is A, the calculation of the entropy for value B follows using (7).

$$-\left(\frac{3}{9}\right)\left[\frac{1}{3} \times log_2 \frac{1}{3} + \frac{2}{3} \times log_2 \frac{2}{3}\right] = 0.306 \tag{7}$$

It is still assumed the missing attribute as A and the calculation of the entropy for value C follows using (8).

$$-\left(\frac{1}{9}\right)[1 \times log_2 1 + 0 \times log_2 0] = 0 \tag{8}$$

The calculation of information gaining for a follows, taking into consideration the value for global entropy for A obtained in (3) and the results obtained in (7) for B and in (8) for C. The result is calculated in (9).

$$\sum = 0.401 + 0.306 + 0 \approx 7071 \tag{9}$$

Now, it is assumed that the missing attribute ε_6 is B and the entropy is calculated for the A value, after making the substitution in (1), the obtained result is 0.3606. In the same way was the entropy for the value C is calculated, by making the substitution in (1), the result is 0. Then the information gaining is calculated for B taking into consideration the global entropy of that value, obtained in (4), and also the result obtained for $A = 0.3606$ and for $C = 0$. The result obtained of this sum is 0.8050.

Now, it is assumed that the missing attribute ε_6 is C and the entropy is calculated for the A value, after making the substitution in (1), the obtained result is 0.3606. Then the entropy for the value B is calculated and the substitution in (1) is made in the same way, the result is 0.3061. Finally, the information gaining for C is calculated, takin into consideration the global entropy value obtained in (5) and the values obtained for $A = 0.3606$ and for $B = 0.3061$. The result of the sum is 0.8889.

In accordance with the information gaining obtained for the A, B and C values, it can be seen that the value generating less uncertainty or less entropy in the information is $A = 0.7071$, this means that the missing nationality is $UnitedSates\ (US)$.

To find the missing value ε_9 the process described previously is made taking into account that the value for ε_6 is now A. Summarizing it is obtained that $A = 0.8364$, $B = 0.6859$ and $C = 0.6363$, the value that generates less entropy is C, which means that the nationality that belongs to ε_9 is England (ING). In this way the previous process is defined as algorithm reiterating that is possible to apply it to any data set with missing data.

The Adult data set is used to practice the proposed methodology and was obtained from the UCI Machine Learning repository [10]. The data set was donated by Ronny Kohavi and Barry Becker form the data base of 1994 census it has 32561 instances with a total of 15 attributes:

Table 3: Registry of the Adult Data Set.

#	A	B	G	N	O
13	23	Private	Adm-clerical	US	$\leq 50k$
14	32	Private	Sales	US	$\leq 50k$
15	40	Private	Craft-repair	?	$50k$
16	34	Private	Transport-moving	Mexico	$\leq 50k$
17	25	Self-emp	Farming-fishing	US	$\leq 50k$

age, *type of job*, *fnlwgt*, *level of education* represented in a numeric way, *marital status*, *occupation*, *relationship*, *race*, *gender*, *capital earnings*, *capital losses*, *hours per week*, *native-country* and *income*. This last one is considered a discriminant attribute.

The attributes to work with are in the columns B, G and N in Table 3, because in this are the missing attributes and the column *O* is taken as a discriminant attribute, it can be seen that the data are categorical and is needed to transform them to facilitate the treatment.

5 Test and Results

To be able to do the treatment of the missing data in the Adult data set three aspects were taken into account: (i) the source where the data set was obtained; (ii) the format in of the data set is found, in this case, it is the extension of our file, which can be type: *csv*, *.txt*, *.data*, *.arff*, etc.; and (iii) the type of data to treat, which means numerical or categorical data.

5.1 Treatment of the missing attribute values with the entropy

Once the data set was explored, 583 missing data were obtained, just for the *native-country* attribute, besides, this attribute is formed by 41 different nationalities. For the Adult treatment, it was necessary to develop a program that allows to process an enormous amount of information, and that will be making the process, explaining Section 4. Said program was developed in Java. This language was chosen, since it is a multi-platform language, and it exists in an endless variety of libraries and/or utilities created to extend functionality, and the previous experience with this language is the most important reason for it.

As a first step, the program carries out the discretization process on the values of the *income* and *native-country* attributes. For the *native-country* attribute, the program assigns a numerical data for each different value, this is managed in regards of the order in that they appear, which means, if the value of *United-States* shows up first, then value 1 is assigned, and if value *Cuba* shows up next, value 2 will be assigned, when continuing with the next value, if *United-States* appears again, 1 will be assigned again. For this process is carried out until the program is done examining all of the values for the *native-country* attribute. Afterwards, the program processes the *income* attribute. The values that were found are two: ≤ 50 and > 50, thus, number 1 or 2 are assigned according on the order in which they show up.

Table 4: Imputation result for the native-country attribute.

Instance	Imputed value
15	Iran
39	Iran
52	Outlying-US(Guam-USVI-etc)
62	Outlying-US(Guam-USVI-etc)

Table 5: Result of the entropy calculation for the type of job attribute.

Instance	Imputed value
28	Self-emp-inc
62	Without-pay
70	Without-pay
78	Without-pay

Once the program is done discretizing, the program starts to run all the data set until it finds the value of a missing attribute, it proceeds to calculate the global and local entropy, so it, later on, determines the information gaining, and, at last, the value that generates less uncertainty is the one that is imputed. In Table 4, some results obtained from this process are displayed. It can be seen that the program identifies the first missing value on instance 15, which is confirmed in Table 3, denoted by '?', after making all the calculations, it imputes the *Iran* value.

It is important to clarify that the entire previously described process was made just for predicting the missing value on instance 15. Once the program is done processing said instance, then it proceeds looking for the next missing value, and it carries out the same calculations until completing the 583 missing attributes. As a summery, for the 583 missing values in the *native-country* attribute, the program imputes 143 attributes with the value of *Iran* and 440 values are imputed with the value of *Outlying-US(Guam-USVI-etc.)*. The execution time for the imputation of the 583 missing values of said attribute was achieved in five seconds.

Later one, the same procedure is applied for the *type of job attribute*, where the income is still seeing as the discriminating attribute. The *type of job attribute* has 1836 missing values. This is a smaller amount in comparison to the 41 different values for the *native-country* attribute. However, the number of instances with missing attributes is larger in comparison to the *native-country* attribute.

The next step followed by the program is to read each instance in the data set until it finds the first missing attribute to, then, calculate the entropy, whether global or local, just as the information gaining, and to impute the value that generates less uncertainty. In Table 5, it can be seen that the program detects the first missing value for the type of job attribute on instance 28, *Self-emp-inc* value is imputed.

Likewise, all of this process is carried out for the 1836 missing values of the *type of job* attribute. As a summary, for this attribute, the program imputes 191 values with the name of

Table 6: Calculation result of the entropy for the attribute occupation.

Instance	Imputed value
28	Exec-managerial
62	Priv-house-serv
70	Priv-house-serv
78	Priv-house-serv

Table 7: Results of the imputation in the primary biliary data set.

Instance	Imputed value
314	Yes
315	Yes
316	Yes
317	Yes

Self- emp-inc, and it assigns 1645 values with the name of *Without-pay*. The time of execution for the imputations of the 1836 missing attributes was two seconds.

At last, the program carries out the entire procedure for the *occupation* attribute. Said attribute has 1843 missing values and 14 different values. Just as it was done with the *type of job* and *native-country* attributes, the *income* attribute was used as a discriminating attribute. The steps carried out by the program are the same as the ones described for the *type of job* and *native-country* attributes. In Table 6, some results of the program execution are displayed. Instance 28 has a missing attribute. After calculating both entropies and information gaining, the program determines that the value to be imputes is *Exec-managerial*.

As summary, for the occupation attribute, the program imputes 191 values with the *Exec-managerial* and imputes 1652 values with the *Priv-house-serv* name. The execution time for the 1843 missing attributes was four seconds.

In the same manner, the imputation of missing values in the Primary Biliary Cirrhosis Data Set is made, in order to demonstrate the proposed methodology it can be applied in any data set with missing values of the categorical type. The data of this set comes from a clinical trial to evaluate the treatment of the primary biliary cirrhosis (PBC) in the liver, in the book Multiple Imputation in Practice with Examples Using IVEware [11] is explained and is used a testing set for different imputation methods shown in this book.

For the test the sex attribute is taken as a discriminant attribute. The attribute called *spider* is the one to imputes this last attribute, indicates if the patient presents vascular spiders. The attribute to impute has 106 missing values and each value can take two vales either *yes* or *not*. In Table 7, the results for the imputation can be seen, the first column corresponds to the number of the instance in which a missing attribute was found, and the second column belongs to the imputed value.

For the spiders' attribute, the program imputes 106 attributes with the value *yes* and the

Table 8: Data analysis using Weka software after loading the Adult dataset.

Native-country		Type: Nominal	
Missing attributes	583 (2%)	Different=41	Unique=1 (0%)
No.	Label	Count	Weight
1	United-states	29170	29170.0
2	Cuba	95	95.0
3	Jamaica	81	81.0

Table 9: Data analysis using Weka software after assigning missing attributes in the Adult dataset.

Native-country		Type: Nominal	
Missing attributes	583 (2%)	Different=41	Unique=1 (0%)
No.	Label	Count	Weight
1	United-states	29753	29753.0
2	Cuba	95	95.0
3	Jamaica	81	81.0

execution time was less than one second, this is due to the short amount of missing attributes when compared to the Adult data set.

5.2 Missing data treatment with WEKA

Now, the imputation of missing values is made in the Adult data set using the replace missing value method, which is included in the WEKA free software [12].

In Table 8, the analysis that the software makes when loading a data set is shown, it can be observed that the *native-country* attribute is nominal, and also that there are 41 different values for this attribute. It is also specified that there are 583 missing attributes, the column Label shows in detail the different values than the *native-country* attribute can take, while the column Count shows the times in which each *native-country* value appears in the data set.

In Table 9, the result of the imputation of the missing attributes using this software is displayed, it can be observed that the treated attribute was *native-country* and is of the nominal type, there still are 41 different values that can take that attribute, the column Label contains the different values for the data set and the column Count shows the times that each value is in the data set. The most remarkable thing is that there are already missing attributes, however, Weka imputes the value *United-States* to the 583 missing attributes. The execution time for the 583 missing attributes was less than one minute and ten.

Table 10: Analysis of the data set with RStudio.

Native-country values	Quantity
United-States	29170
Mexico	643
Philippines	198
Germany	137
Canada	121
?	583

Table 11: Results of the imputation process with the k-nearest neighbor imputation method.

Native-country values	Quantity
United-States	29723
Mexico	662
Philippines	199
Germany	137
Canada	121
Puerto Rico	116
Others	1603

5.3 Missing data treatment with K-Nearest Neighbors method

In the same Adult Data Set, the imputation of the missing values is made with the k-Nearest Neighbor Imputation (kNN) method, included in the software RStudio. This method uses the kNN (K is the nearest neighbor) to do this imputation.

In Table 10, an analysis made by this software to the nationality attribute is depicted. The nationality with the amount of times that is present in the data set is shown, in the same way the 583 missing attributes denoted by '?' can be observed.

In Table 11, the results of imputation done with k-Nearest Neighbor Imputation are shown, and each value with the new number of times present in the data set, is also worth mentioning that his method imputes 553 attributes with *United-States* value, 19 with *Mexico* value and 11 with other values.

Based on the comparison made in [13] and the test done in this work it can be seed in Table 12, in which the techniques mean/mode are detailed, used for the subsection 5.2 by the Weka free software, and K-Nearest Neighbors, used in the subsection 5.3 with the RStudio software. Is also worth mentioning that the execution times in the proposed methodology are shorter than the mean/mode and kNN.

As can be seen in Table 12, the disadvantage of the methodology proposed in this document is that it only Works with categorical attributes. To fix said disadvantage, fuzzy logic can be

Table 12: Comparison between kNN, mean/mode and the entropy.

Imputation technique	Type of data	Advantage	Disadvantage
Mean Mode	Numeric, categorical	• Easy and simple, • Fast	• Does not generate good classification, • Correlation is biased negatively.
kNN	Numeric, categorical, combinated	• Multiple missing values are easy to handle. • Improves the prediction of the classification.	Is difficult to choose the distance function and the number of neighbors.
Entropy	Categorical	• Fast, • Takes into account all the values for the imputation	Only works with categorical attributes

used discretizing the universe of the speech interval [0,1], using (10) [14], for later divide that interval using triangular and trapezoidal belonging functions to indicate the number of partitions (3,5,7), then applying the proposed methodology and making the prediction of missing numerical values.

$$\frac{X - min(X)}{max(X) - min(X)} \tag{10}$$

Likewise, tests were carried out with methods such as Random Imputation, kNN and Drop, taken from [15], with the aim of making a comparison of the results of the different methods used in this work. For the *occupation* attribute, entropy imputes 191 *Exec-managerial* values and 1652 *Priv-house-serv* values. kNN imputes 142 *Exec-managerial* values, 17 *Priv-house-serv* values and the remaining 1684 values are mixed values. Random-Imputation imputes 241 *Exec-managerial* values, 9 *Priv-house-serv* values, 251 *Prof-specialty* values and the remaining 1342 values impute other values. Mode imputes 1843 *Prof-specialty* values; Drop removes the 1843 instances where the missing values are found. Figure 2, shows the results obtained from all the methodologies.

In this attribute, some similarities were presented in the imputation of values, for example: for the *Exec-managerial* value, between the 191 values imputed by entropy and the 142 values imputed by kNN, 36 similarities were found. Regarding the *Priv-house-serv* value, between the

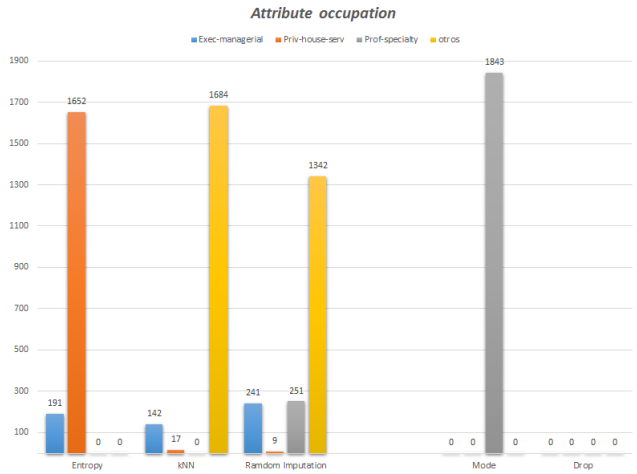

Figure 2: Results of all methodologies for the occupation attribute.

1652 values imputed by entropy and the 17 values imputed by kNN, 16 similarities were found. Similarities were also found with the Ramdom-Imputation methodology, with 33 matches for the *Exec-managerial* value and 8 for the *Priv-house-serv* value.

6 Conclusions

The proposed methodology uses the entropy missing value substitution model for categorical data with an approach of supervised classification which means that a data set with categorical missing values that belong to a binary class or multiclass needs to be provided. With the study done in this work it is possible to see that it is not difficult to process missing categorical data as exposed in Section 4. However, because of working with a data set with a great number of instances with the different missing values, it becomes a very complex process. That is the reason why the proposed methodology becomes very useful and fast to do the prediction work, which allows this methodology to be applied in different knowledge areas as mentioned in Section 1.

This methodology cannot be directly compared with an imputation algorithm such as KNN, because of the different process in its nature. First, kNN takes a numerical base for the neighbors to be able to be annualized and replaces the missing value by the evaluation of the values in the neighborhood taking as a base the number of neighbors (k). Second, kNN does not discriminate between attributes and instances; this method pays more attention to the attributes and instances. In Adult Data Set, the nationality attribute cannot be substituted by the missing value of an instance form the neighbors such as age, education occupation or marital status. This is due to the type of data in each attribute (numerical, nominal, categorical). This model can only be used to impute values for attributes that present missing data; this can be later processed efficiently

with an algorithm for data mining or machine learning.

This methodology can be used in a pre-processing stage without eliminating attributes or records with missing values in order to ensure the efficient performance of a supervised classification model

Acknowledgments

Gratitude to the National Council of Science and Technology (CONACYT for its initials in Spanish) for the scholarship granted to continue with our studies for master's degree, which allows to do this research and present this paper.

References

[1] B. D. Romo, "Ajuste demográfico por imputación," *Reality, Data and Space International Journal of Statistics and Geography*, vol. 9, pp. 27–60, 2018.

[2] R. M. Gray, *Entropy and Information Theory*. Springer New York, Dordrecht Heidelberg London, 2011.

[3] C. A. Gonzalo, *Teoría de la información, codificación y lenguajes*. Ministerio de Educación y Ciencia, 1975.

[4] I. Myrtveit, E. Stensrud, and U. H. Olsson, "Analyzing data sets with missing data: An empirical evaluation of imputation methods and likelihood-based methods," *IEEE Transactions on Software Engineering*, vol. 27, no. 11, pp. 999–1013, 2001.

[5] M. P. de Albuquerque, I. A. Esquef, and M. P. de Albuquerque, "Image segmentation using nonextensive relative entropy," *IEEE Latin America Transactions*, vol. 6, no. 5, pp. 477–483, 2008.

[6] G. Chhabra, V. Vashisht, and J. Ranjan, "A Comparison of Multiple Imputation Methods for Data with Missing Values," *Indian Journal of Science and Technology*, vol. 10, no. 19, pp. 1–7, 2017.

[7] O. Lüdtke, A. Robitzsch, and S. Grund, "Multiple imputation of missing data in multilevel designs: A comparison of different strategies," *Psychological methods*, vol. 22, no. 1, pp. 141–165, 2017.

[8] A. B. Pedersen, E. M. Mikkelsen, D. Cronin-Fenton, N. R. Kristensen, T. M. Pham, L. Pedersen, and I. Petersen, "Missing Data and Multiple Imputation in Clinical Epidemiological Research," *Clinical Epidemiology*, vol. 9, pp. 157–166, 2017.

[9] S. Rawal, S. C. Gupta, and S. Singh, "A Proposal for Predicting Missing Values in a Dataset Using Supervised Learning," *International Journal of Advanced Research in Computer Science*, vol. 8, no. 8, pp. 562–567, 2017.

[10] R. Kohavi and B. Becker, "UCI Machine Learning Repository: Adult Data Set," Avaliable: https://archive.ics.uci.edu/ml/machine-learning-databases/adult/, 1996, [Online. Accessed on 15-12-2019].

[11] T. Raghunathan, P. A. Berglund, and P. W. Solenberger, *Multiple Imputation in Practice: With Examples Using IVEware*. Chapman and Hall/CRC, 2018.

[12] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[13] T. Aljuaid and S. Sasi, "Proper Imputation Techniques for Missing Values in Data Sets," in *2016 International Conference on Data Science and Engineering (ICDSE)*. IEEE, 2016, pp. 1–5.

[14] D. T. Larose and C. D. Larose, *Data Mining and Predictive Analytics*. Wiley, 2015.

[15] J. Poulos and R. Valle, "Missing Data Imputation for Supervised Learning," *Applied Artificial Intelligence*, vol. 32, no. 2, pp. 186–196, 2018.