DATA SCIENCE APPLICATIONS IN
HEALTH AN SOCIAL CARE

by

MARYURI ARIELA QUINTERO QUINTERO

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Industrial Engineering at
The University of Texas at Arlington
December, 2018

Arlington, Texas

Supervising Committee:

Aera LeBoulluec, Supervising Professor
Paul Componation
Jesús González

ACKNOWLEDGEMENTS

I would like to acknowledge Dr. Aera LeBoulluec for her guidance, encouragement, and mentorship as the chair of my thesis committee. I also want to thank the faculty and staff of the Industrial, Manufacturing, & Systems Engineering Department for an enjoyable and enlightening journey. Also, thanks to the Center on Stochastic Modeling, Optimization, & Statistics for believing in my research skills and giving me the COSMOS scholarship 2018.

I must also extend my thanks to the Fulbright Organization for awarded me with a scholarship to study in the United States. This award not just allowed me to study in a high-quality university but also provided me with opportunities to experience the American culture, meet people from all over the world, and be an ambassador of my country in the United States.

The International Students, Inc. earns special thanks for giving me the opportunity to actively participate in leadership roles in this organization and becoming my family away from home. Your encouragement, support, and friendship gave me positive vibes to keep going for my goals.

My husband, Ezequiel Aguilar, deserves much more than just thanks. He was always by my side unconditionally, and I doubt this project would have been achievable without his love, care, and help during the most challenging phases of this project. Finally, I recognize my mother Telma Quintero, my father Rory Quintero, and my sisters Jessy, Amalia, and Auristela for cheering me along the way with love.

November 29, 2018

LIST OF FIGURES

Chapter 1:

Chapter 2:

LIST OF TABLES

Chapter 1:

Chapter 2:

ABSTRACT

DATA SCIENCE APPLICATIONS IN HEALTH AND SOCIAL CARE

Maryuri A. Quintero Q.

The University of Texas at Arlington, 2018

Supervising Professor: Aera LeBoulluec

Health and social care are areas of concern worldwide nowadays. Chronic diseases such as cancer and social problems such as tobacco consumption are leading risks of deaths in many countries, and preventive efforts are urgently needed to decrease the negative impact that those problems cause. Technology has made available an unprecedent amount of data in the health and social care fields, which scientists are using to achieve a better understanding of many problems that are a burden for the health and social systems globally. Although previous studies have provided approaches to analyze data, more efficient and accurate methods are needed to obtain predictive models with better performance. This thesis employs data science tools to analyze the characteristics of health and social care data and determine the best approaches to improve the accuracy and efficiency of predictive models in the studied fields. Data preprocessing is considered the key action to increase the statistical power of the data and perform valid data analyses in this study. In brief, this thesis present two researches that are focused on enhancing the data analysis by implementing data science techniques to preprocess data.

TABLE OF CONTENTS

CHAPTER 1

MISSING DATA IMPUTATION FOR ORDINAL DATA

MISSING DATA IMPUTATION FOR ORDINAL DATA

M. Quintero & A. LeBoulluec (2018). "Missing Data Imputation for Ordinal Data".

International Journal of Computer Applications 181(5):10-16. ISSN 0975 – 8887.[1]

[1] Used with permission of the publisher, 2018 where 2018 is the year permission was granted. See Appendix A.

Missing Data Imputation for Ordinal Data

Maryuri Quintero
University of Texas at Arlington
701 S. Nedderman Drive
Arlington, TX 76019
ing.maryuri.quintero@gmail.com

Aera LeBoulluec, PhD
University of Texas at Arlington
701 S. Nedderman Drive
Arlington, TX 76019
aeral@uta.edu

ABSTRACT

The treatment of missing data has become a mandatory step for performing valid data analysis in most scientific research fields. In fact, researchers have found that dealing with missing data avoids misleading data analysis and improves the quality and power of the research results [1]. According to the authors in [2,3], the missing values in a data set could be missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR), a categorization that should be taken into consideration to deal with the problem of missing data. The number of observations, the types of variables, and the percentage of missing values in a data set are also important characteristics that should be contemplated before dealing with missing values. Understanding the missing data case helps the researchers to identify the imputation techniques that best handles the missing data problem. However, the development of procedures to impute categorical data is not significantly available as the procedures focused on continuous data imputation [1]. This study compares six different imputation methods to find the one that performs the most appropriate treatment for categorical data, type ordinal, in a breast cancer dataset.

General Terms

Data imputation; missing data.

Keywords

MCAR; categorical data; ordinal data.

1. INTRODUCTION

The adequate analysis of data in all kinds of research fields is often hindered by the presence of missing information, a widespread problem that many data analysts face commonly. The occurrence of missing values arises from different reasons such as measurement errors, accidental deletion of recorded values, non-responses, and mistakes in data entry. As a result, analysts could end up drawing flawed conclusions about the data since the missing values have a detrimental effect when the data is analyzed [1]. In fact, some researchers argue that the performance of statistics on datasets with large amount of incomplete responses is significantly affected by the missing values [4]. According to the authors in [5], missingness in a dataset weakens the data analysis outcomes because the missingness brings ambiguity into the data analysis, reduces the statistical power of the data, and yields inaccurate statistical estimators such as means, variances, and percentages. The authors in [1,4] also support the idea that weak statistics, biased parameter estimates, loss of information, and inefficient standard errors result from the analysis of incomplete data. In brief, the missing values hold valuable information that is suppressed from the data analysis leading to erroneous findings.

Missingness can be appropriately handled through a variety of methods for imputing missing values. However, picking the right imputation method to treat the missing values depends on the information known by the analyst such as the causes of missingness, the type of missingness in the dataset, and the type of data.

1.1 Types of Missing Values

The presence of missing observations is common in all kind of data collection, and this missingness could show different missing data patterns. Therefore, understanding the causes and patterns of missing data is crucial to perform a valid statistical analysis and select the best data treatment method. Rubin [2] considers that randomness behavior is the primary concern when the analyst deals with missing values. In fact, the author in [2] provides a basic classification of the types of missing data based on the randomness patterns that could emerge in a data due to problems in the data collection process.

The first type of missing data occurs when the data is *missing completely at random* (MCAR). This type of missing data happens when the cause of missingness in a variable has no relation with neither the missing values in that variable or the responses in other variables. Data missing completely at random usually results when a random subset of the study sample overlooks a question unintentionally leading to missingness in the data without a systematic cause. When data is MCAR, the missingness is under the control of the researcher, and the cause of missingness is some random event [6]. A good example of data MCAR occurs when some subjects of study neglected to answer a question in a survey because they did not see the question in the back of the survey form that they were filling out. Data MCAR could weaken the statistical power in the data, but this type of missingness does not cause significant bias in the data analysis outcomes because the respondents and nonrespondents do not share systematic differences [4].

Data *missing at random* (MAR) is the second type of missingness described by Rubin [2]. When data MAR happens, there is a probability that the missing data depends on measurable characteristics of the respondents but the missingness is unrelated to the missing responses themselves. In other words, the observed data has conditions that randomly affect the missing process. The authors in [7] state that in data MAR "the subjects with missing data are a selective rather than a completely at random subset of the total study population". In similar words, the respondents that caused data MAR correspond to a group of respondents whose characteristics enhance the probability of missingness in certain variables. For instance, an elderly patient with memory deficiency has difficulties remembering a certain event, so this patient leaves unanswered questions in a clinical form. The resulting missing values are related to the age of the patient, but it is not related to the event itself [4]. The author in [4] affirms that using the proper statistical model for imputing data missing at random could consider that the missingness as *ignorable* in a particular type of inference, so the condition related to the missing values can be measured and used during the data analysis process.

Finally, data *missing not at random* (MNAR) is the third type of missingness that could emerge after the data collection process. This type of missingness occurs when the causes for missing values are unknown, and there is no way to get information about what is producing incomplete data. According to Finch [1], there is a high probability of getting data missing not at random in a variable when the responses are directly related to the value of the variable itself. For example, students who consume large amounts of cigarettes frequently are more likely to leave a question unanswered if they are asked to indicate the number of cigarettes they have consumed in the last week. This behavior results from the respondent's need of hiding their real behavior leading to serious bias in the statistical analysis. In data MNAR, the missingness cannot be ignored as in data MAR, and the treatment of missing values become more difficult [4].

Unfortunately, when the data is missing systematically because of another variable (MAR or MNAR), the analyst could have a hard time trying to figure out the type of missing values, and making assumptions is the only way to determine these types of missingness and their influence on the data analysis [2,8]. In the case of MCAR, there is no correlation between the variable with missing values and another variable, so the information about the cause of missingness is not relevant in the data analysis to control the biases [6].

Classifying the types of missing values from the data is not an intuitively task, so Myers [9] presents a schema that simplifies the differences between the types of missing values and gives an approach to classify the missingness based on the probability of missingness on a given variable *"Y"* (see Figure 1).

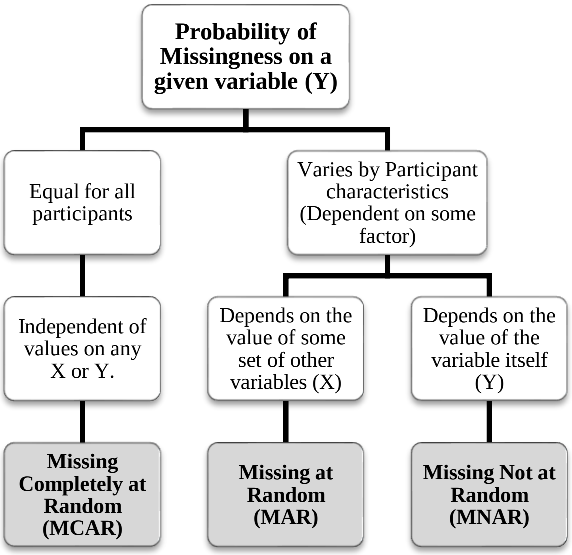

Probability of Missingness on a given variable (Y)
Equal for all participants
Varies by Participant characteristics (Dependent on some factor)
Independent of values on any X or Y.
Depends on the value of some set of other variables (X)
Depends on the value of the variable itself (Y)
Missing Completely at Random (MCAR)
Missing at Random (MAR)
Missing Not at Random (MNAR)

Fig 1: Classification of the type of missing values based on the probability of missingness on a given variable *"Y"* [9]

1.2. Types of Data

The treatment of missing data requires methods that make appropriate assumptions for the type of data used in the study. So, identifying the type of data becomes a relevant step before conducting any action or analysis on the studied data. Quantitative data and qualitative data are the two basic types of data that could be found in all kinds of research fields (see Figure 2).

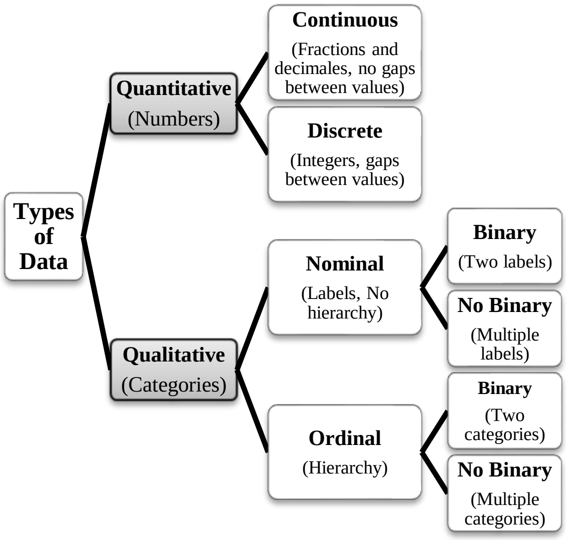

Types of Data
Quantitative (Numbers)
Continuous (Fractions and decimales, no gaps between values)
Discrete (Integers, gaps between values)
Qualitative (Categories)
Nominal (Labels, No hierarchy)
Binary (Two labels)
No Binary (Multiple labels)
Ordinal (Hierarchy)
Binary (Two categories)
No Binary (Multiple categories)

Fig 2: Classification of the types of data

Quantitative data, also known as numerical data, results from numerical measurements that have meaningful values represented as a set of numbers. There are two different types of *numerical data* based on the scale of measurement for this type of data: discrete data and continuous data [10].

The first type of numerical data is *discrete data*, a type of data whose scale is made up of a list of possible numbers with gaps between them. The discrete data are only integer values (whole numbers) that can go to infinity or be part of a fixed list of numbers. According to the authors in [11], the discrete numbers can be counted, but they cannot be subdivided meaningfully because the data cannot be broken down into meaningful smaller units. Examples of discrete data are the number of defective parts in a production batch and the number of patients waiting for examination. Neither the defective parts nor the patients can be subdivided in significant smaller units. There is no such thing as "half of a defective part" or "one third of a patient".

Continuous data is the second type of numerical data. Continuous data can take any numeric value in an interval because the measurement scale does not consider gaps between values measured. When the data is continuous, the numbers can be meaningfully subdivided into smaller parts (fractions and decimals), but the outcome values cannot be counted since there are infinite possible values that can result from the subdivision of a measured value. For instance, measurements of money, time, and temperature can be recorded and broken down into smaller parts, and the resulted numbers still have meaning. The time it takes an athlete to complete a race can be any value between a minimum and a maximum value of time, and this measure can be expressed in hours to fractions of a second.

Qualitative data, or categorical data, is the second basic type of data that could be found in research. The authors in [11] define categorical data as "data that can take on only a specific set of values representing a set of possible categories". In similar words, categorical data are recorded observations placed into categories according to certain qualitative traits. This type of data cannot be numerically measured like the numerical data type. The categorical data can be nominal, ordinal, or binary.

Nominal data is a type of qualitative data that falls into categories without any order or inherent ranking sequence. If the data is nominal, the values are represented with labels, words, letters or alphanumeric symbols that have no numerical significance. Gender and race categories are good examples of nominal data. When nominal data has two possible categories such as "Yes/No answers" or "female/male gender options", the data is nominal and binary, and it is called *dichotomous*.

Categorical values can have a significant order or ranking. If the order of the data matters, the data is classified as ordinal data. Ordinal data can be counted and ordered, but it is not possible to measure it. In other words, the ordinal values are values assigned to hierarchical categories; the occurrences of observations per category can be counted, so there is mathematical meaning, but the value of the category is not meaningful mathematically if it is measured. For example, if 100 patients are asked to provide their level of satisfaction with their health insurance company by using a numerical scale from 1 (lowest) to 5 (highest), the outcome data will be the ordinal type, and the average of the 100 answers will have meaning. Ordinal data can have multiple categories, as shown in the example above, but binary ordinal data can happen if there are just two categories in a hierarchy.

The treatment of both numerical and categorical missing data has been studied for years in order to find the best imputation methods for different types of data. Since more approaches have been developed to deal with continuous data missingness [1], this study is focused on performing and comparing different imputation methods to find the one that best deals with ordinal data, a type of categorical data.

2. DATA AND METHODOLOGY

2.1 Data Characteristics

In the present study, a breast cancer dataset is used to perform and compare six different imputation methods. The dataset is a large multivariate dataset composed by 11 different variables and 699 observations whose values are integers resulted from an ordinal classification (see Table 1).

The breast cancer dataset was obtained from the University of Wisconsin Hospitals, and it was created by Dr. William H. Wolberg. This dataset can be found in the UCI Machine Learning Repository where there is available information about the data collection process and characteristics of the breast cancer database [12].

Table 1. Dataset information

Wisconsin Breast Cancer Dataset	
Data Characteristics	Multivariate
Variable Characteristics	Integer
Type of Data	Classification (Ordinal)
Number of Observations	699
Number of Variables	11 (10 predictors and 1 response variable)
Missing Values	Yes. One predictor has 16 missing values.
Variables	*Predictors* 1. Sample code number: id number 2. Clump Thickness: 1 - 10 3. Uniformity of Cell Size: 1 - 10 4. Uniformity of Cell Shape: 1 - 10 5. Marginal Adhesion: 1 - 10 6. Single Epithelial Cell Size: 1 - 10 7. Bare Nuclei: 1 - 10 *(16 missing values)* 8. Bland Chromatin: 1 - 10 9. Normal Nucleoli: 1 - 10 10. Mitoses: 1 - 10 *Response Variable* 11. Class: (2 for benign, 4 for malignant)

2.2 Methodology

The breast cancer dataset includes ten (10) predictor variables of which just one has missing values. However, the variable with missing values was not included in this study in order to compare statistical measures in a complete dataset with the dataset imputed with different approaches. Then one variable with complete data, Uniformity of Cell Size, was selected to simulate different missing values levels. The variable *Uniformity of Cell Size* has a high correlation with other variables in the dataset, which is convenient for performing better inferences in methods such as Multiple Imputation by Chained Equations that uses all the variables in a dataset to predict the missing values in the variable with missingness problems.

The missing values were introduced completely at random into the variable *Uniformity of Cell Size* leaving this variable with a percentage of missing values. One level of sample size (699 observed values) and seven levels of missing data were included in the variable of interest to analyze the performance of different imputation methods with varied percentage of missingness in the data.

The first step to simulate the missing values was to get a completely-at-random sample of observations from the total observations recorded in the variable *Uniformity of Cell Size*. After obtaining a sub sample of values from the variable of interest, these values were replaced with empty responses to produce missingness in that variable. Seven levels of incomplete data Missing Completely at Random (MCAR) were simulated: 2%, 4%, 5%, 10%, 12%, 15%, and 20%. Then the simulated missing values were treated with six different imputation methods for each level of missing data included in the study. Finally, an evaluation criterion was used to measure the performance of all the imputation methods applied to the missing values for each level of missingness. The methodology used in this study is illustrated in the Figure 3.

All the variables in the breast cancer dataset included in this study have numerical ordinal data in a range from 1 to 10. Therefore, all the imputation methods used in this study attempted to produce integer values within the given range.

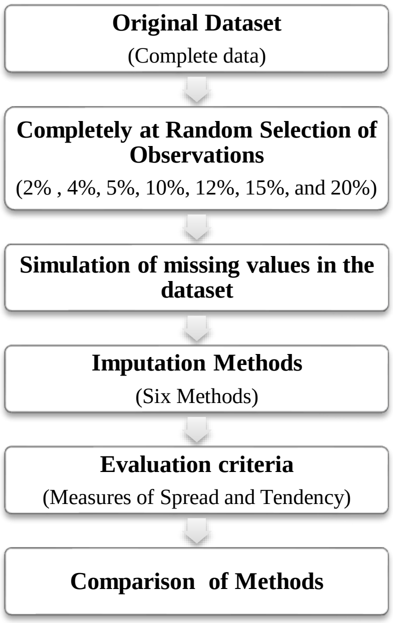

Original Dataset
(Complete data)
Completely at Random Selection of Observations
(2% , 4%, 5%, 10%, 12%, 15%, and 20%)
Simulation of missing values in the dataset
Imputation Methods
(Six Methods)
Evaluation criteria
(Measures of Spread and Tendency)
Comparison of Methods

Fig 3: Methodology used to compare different imputation methods in this study

The present study considers just those variables with complete data, so the measures performed on the imputed data can be compared with the measures made on the original complete data.

2.3 Evaluation criteria

Measures of spread and measures of central tendency were the parameters used as evaluation criteria in this study. The measures of spread are useful to analyze the similarity between the instances in the variable where missing values are imputed. Also, the measures of spread explain how scattered the observed values are in a dataset and how much these values differ from the mean value. On the other hand, the measures of central tendency produce a single value that describes all the values in the dataset and the central position within that set of data, which is useful to understand the data and its tendencies.

The variance and the standard deviation are the two measures of spread used for evaluating the different imputation methods in this study. Similarly, the mean value served as evaluation criteria to compare the central tendency between the imputed data and the original data (known values in the breast cancer database). These measures were calculated for both the original data and the imputed data for each imputation method and for each level of MCAR data involved in this study.

The percentage of error is a statistical tool that simplifies the comparison between experimental values and true values. Since the results of each imputation method aim for the original values in the data in this study, the calculation of the percentage of error was useful to determine the precision of each imputation method to predict the missing values in the variable of interest. The percentages of error closer to zero indicate that the imputation method produced values that were very close to the measures in the original dataset. The Equation 1 is used to calculate the percentage of error in this study.

$$\%Error = \left|\frac{Experimental\ value - True\ value}{True\ value}\right| \times 100 \quad (1)$$

Where:

%Error = the percentage of error
Experimental value = the value obtained from the imputation method.
True value = the known data values.

The absolute difference between the experimental values and the true values is known as the absolute error, and it can be used as a simplified way to compare measures between the results of an experiment and the true values.

3. IMPUTATION METHODS

Six different methods were used to treat the simulated missingness in the breast cancer dataset. A brief description of the assumptions for each imputation method is provided below.

3.1 The most frequent value method

This method replaces the missing instances with the most common value within a set of values in a given variable. In other words, the method imputes the missing data with the number that is most likely to occur in a set of numbers in a variable.

3.2 Mean substitution method

The Mean Substitution method consists of replacing the missing data in a variable by the mean of all known values of that variable [5,13]. The mean is usually denoted by the symbol "$\bar{x}$", and its value is equal to the sum of all the values in the variable divided by the total of observations in the variable. The mean calculation is represented in the Equation 2.

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) = \frac{x_1 + x_2 + \cdots + x_n}{n} \qquad (2)$$

Where:

$\bar{x}$ = the mean value
x_i = the observations in the variable
n = total number of observations

3.3 Random selection imputation

The Random Selection approach is a method based on randomly assigning a value to the missing data. The values randomly selected are framed in a specific range of values, which should have the same characteristics as the values in the variable with the missingness (numerical or categorical data). Each number in between the range has the same probability of being assigned to the missing data [14].

3.4 K-Nearest Neighbors classification using Euclidean distance

The k-Nearest Neighbor method (KNN) is a conventional non-parametric classifier that uses the distances between the value treated and its k-nearest neighbors to find the final output for the value treated [5,15]. The KNN method defines a set of K nearest cases from the values treated and then estimates the replacement value from these neighbor cases selected [5]. The K-NN method uses the mean value to estimate the value for continuous data and the mode value to replace the missing values when the data is categorical [16].

One of the most common functions used for calculating the distance metrics in KNN is the Euclidean distance function. This function helps to measure the distances between two data points of interest in a feature space. The authors in [15] argue that, to calculate the distance between A and B, the normalized Euclidean metric can be determined by using the following equation:

$$dist(A,B) = \sqrt{\frac{\sum_{i=1}^{m}(x_i - y_i)^2}{m}} \qquad (3)$$

Let represent *A* and *B* by feature vectors *A = (x_1, x_2, …, x_m)* and *B = (y_1, y_2, …, y_m)*, where ***m*** is the dimensionality of the feature space.

3.5 Multiple imputation by chained equations

The authors in [5] define Multiple Imputation by Chained Equations (MICE) as "an iterative algorithm based on chained equations that uses an imputation model specified separately for each variable and involving the other variables as predictors" (p.1). In other words, MICE is a method that produces multiple predictions for the missing values by considering all the variables in the dataset as predictors. This method takes into consideration the statistical uncertainty when data is imputed by addressing the missingness problem with multiple imputations and a flexible approach to handle variables of different types of data [17]. The authors in [17] mention that when the MICE method is used, each variable with missing values is conditionally modeled based on the other variables in the data and uses its own distribution when repeated interactions between variables are performed. The iterations through all the variables are repeated until the process converges and a final complete dataset results from the imputed values.

3.6 Soft-Impute: Matrix completion by iterative soft-thresholding of SVD decompositions

Soft-Impute is an algorithm that iteratively replaces the missing values with values generated from a soft-thresholded SVD (Singular Value Decomposition). The Soft-Impute method facilitates the efficient regularization of solutions by computing a low-rank SVD of a dense matrix [18]. The Soft-Impute method uses parameters that consider low dimensionality, and when this method is performed, the values of the objective function decrease with each iteration producing minimum values in the function. This method repeatedly replaces the missing values with the current estimate, and then updates the estimate by solving an algorithm.

4. RESULTS

Three main steps were performed to compare the six different imputation methods involved in this study: the calculation of evaluation metrics, the calculation of absolute errors, and the ranking of best imputation methods.

The calculation of evaluation metrics is the first step to compare the imputation methods included in this study. The mean value, the standard deviation, and the variance are the metrics defined as the evaluation criteria in this research, and they were calculated after performing each imputation method for different levels of missing values. The results from the calculation of the evaluation metrics for each imputation approach are provided in the Table 2. These same metrics were calculated for the original data before simulating missing values and applying any imputation approach. A mean equal to **3.134**, a standard deviation equal to **3.051**, and a variance equal to **9.0** are the values for the measures calculated in the original dataset. These measures are required to determine the following steps in this section.

The estimation of the absolute errors for each imputation method is the second step for the comparison of the imputation approaches included in this study. The calculation of the absolute errors was necessary to determine how well each imputation method performed in comparison to the original dataset characteristics. As it was mentioned in the section 3.3, the absolute errors result from the absolute difference between experimental values and true values. In this study, the absolute difference between the metrics of each imputation method and the metrics of the original data determines the absolute errors required in this research. For instance, if the evaluation metric is the *standard deviation*, the standard deviation of each imputation method is compared with the standard deviation of the original data. The absolute difference between those standard deviations estimates the absolute error for the studied case (see Equation 3).

An example of the absolute error calculation for the Multiple Imputation method (MICE) when the data has 2% of missing values and the evaluation metric is the standard deviation is as follow:

$$AbsError = |Experimental\ value - True\ value| \quad (3)$$

$$= |3.004 - 3.051|$$

$$= 0.047$$

The absolute error for the MICE method when there is 2% of missing data and the evaluation metric is the standard deviation resulted equal to 0.047. This value shows how similar is the standard deviation of the MICE method to the standard deviation calculated for the original data under the given conditions. The same calculation was performed for each imputation method and for each evaluation metric under different missing data levels, which is summarized in the Table 3. In brief, the absolute error technique helped to identify how close were the imputed values from the original values in the breast cancer dataset.

The ranking of the best imputation methods to treat missing data was the last step to study the performance of the imputation methods in this study. The ranking of the imputation methods was made by using the absolute errors to give positions and weights to each imputation approach.

Table 2. Evaluation criteria results after performing different imputation methods for different percentages of missingness

Method	Evaluation criteria results per percentage of missingness																				
	2%			4%			5%			10%			12%			15%			20%		
	M	StD	V	M	StD	V	M	StD	V	M	StD	V	M	StD	V	M	StD	V	M	StD	V
MFV	3.087	3.018	9.111	3.052	3.011	9.069	3.031	3.004	9.022	2.914	2.950	8.700	2.877	2.931	8.592	2.798	2.911	8.474	2.715	2.855	8.152
MS	3.127	3.004	9.022	3.132	2.982	8.894	3.132	2.967	8.805	3.114	2.880	8.293	3.117	2.848	8.112	3.099	2.811	7.903	3.116	2.724	7.418
RS	3.170	3.054	9.000	3.237	3.070	9.000	3.282	3.085	9.000	3.454	3.177	10.000	3.461	3.166	10.000	3.489	3.158	9.000	3.578	3.121	9.000
KNN	3.127	3.017	9.100	3.130	3.010	9.059	3.127	2.997	8.979	3.133	2.937	8.626	3.134	2.916	8.501	3.136	2.884	8.318	3.206	2.801	7.846
MICE	3.130	3.004	9.025	3.133	2.983	8.898	3.136	2.968	8.808	3.132	2.883	8.312	3.127	2.850	8.123	3.117	2.816	7.929	3.139	2.728	7.441
SI	3.112	3.008	9.048	3.099	2.991	8.946	3.094	2.977	8.862	3.039	2.901	8.415	3.013	2.877	8.279	2.963	2.846	8.102	2.937	2.770	7.672

Evaluation criteria values for the original data (before imputation)
M = 3.134
StD = 3.051
V = 9.000

Evaluation Criteria:
M = Mean
StD = Standard Deviation
V = Variance

Imputation Methods:
MFV = The Most Frequent Value
MS = Mean Substitution
RS = Random Selection
KNN = K-Nearest Neighbors
MICE = Multiple Imputation by Chained Equations
SI = SoftImpute

Table 3. Absolute errors between the evaluation criteria values of the original data and the imputed data for different percentages of missingness

Method	Absolute error per evaluation criteria and percentage of missingness																				
	2%			4%			5%			10%			12%			15%			20%		
	M	StD	V	M	StD	V	M	StD	V	M	StD	V	M	StD	V	M	StD	V	M	StD	V
MFV	0.047	0.033	0.111	0.083	0.040	0.069	0.103	0.048	0.022	0.220	0.102	0.300	0.258	0.120	0.408	0.336	0.141	0.526	0.419	0.196	0.848
MS	0.007	0.048	0.022	0.003	0.069	0.106	0.003	0.084	0.195	0.020	0.172	0.707	0.017	0.203	0.888	0.036	0.240	1.097	0.019	0.328	1.582
RS	0.036	0.003	0.000	0.103	0.019	0.000	0.147	0.034	0.000	0.319	0.126	1.000	0.326	0.114	1.000	0.355	0.107	0.000	0.443	0.070	0.000
KNN	0.007	0.035	0.100	0.004	0.042	0.059	0.007	0.055	0.021	0.001	0.115	0.374	0.000	0.136	0.499	0.001	0.167	0.682	0.072	0.250	1.154
MICE	0.004	0.047	0.025	0.001	0.069	0.102	0.001	0.084	0.192	0.003	0.168	0.688	0.007	0.201	0.877	0.017	0.236	1.071	0.004	0.324	1.559
SI	0.023	0.044	0.048	0.036	0.061	0.054	0.040	0.075	0.138	0.096	0.151	0.585	0.122	0.174	0.721	0.172	0.205	0.898	0.197	0.282	1.328

Table 4. Ranking of best imputation methods per overall performance and evaluation criteria for different percentages of missingness

Method	Ranking per evaluation criteria for each percentage of missingness																					Overall Performance (Ranking)
	2%			4%			5%			10%			12%			15%			20%			
	M	StD	V	M	StD	V	M	StD	V	M	StD	V	M	StD	V	M	StD	V	M	StD	V	
MFV	3	2	3	3	2	3	3	2	3	3	**1**	**1**	3	2	**1**	3	2	2	3	2	2	**2.33 (2)**
MS	2	3	2	2	3	3	2	3	3	3	3	3	3	3	3	3	3	3	2	3	3	**2.76 (5)**
RS	3	**1**	**1**	3	**1**	**1**	3	**1**	**1**	3	3	3	3	**1**	3	3	**1**	**1**	3	**1**	**1**	**1.95 (1)**
KNN	2	3	3	3	3	3	3	3	2	**1**	**2**	**2**	**1**	3	**2**	**1**	3	3	3	3	3	**2.48 (4)**
MICE	**1**	3	3	**1**	3	3	**1**	3	3	2	3	3	2	3	3	2	3	3	**1**	3	3	**2.48 (3)**
SI	3	3	3	3	3	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	**2.95 (6)**

Overall Performance: The lowest value represents the imputation method with a better approach to treat the missingness in the breast cancer dataset.

The imputation method with a value equal to one (1) in an evaluation metric is the method with the best approach based on that criteria in a specific missingness level. The method with a criteria value equal to two (2) have the second place as a good method to approach the respective missingness case, and the methods with a criteria value equal to three (3) represent the ones that were least precise to predict the missing values. A ranking of the best imputation methods is shown in the Table 4, in which the best imputation methods are the ones with an overall performance closest to one (perfect performance level).

The Random Selection method (RS) was the imputation method with the best performance in this study. This method obtained the lowest percentage of errors for the standard deviations and variances in the majority of the missing levels cases except for 10% level of missingness, in which the method performed less precise than other methods. The Most Frequent Value method (MFV) got the second place in the ranking of the best imputation methods because it generated the second most precise values in the evaluation criteria for all the missing values levels. The Multiple Imputation by Chained Equation method (MICE) and the K-Nearest Neighbor method (KNN) performed similarly in this study and achieved the following third and fourth places in the ranking of best imputation methods. The Mean Substitution method (MS) and the Soft-Impute method (SI) were the imputation techniques with the poorest performances in this study (see Table 4).

5. CONCLUSION

In the present study, six imputation methods were performed to treat different missing values levels in a categorical dataset. These levels of missing values were simulated and introduced in a breast cancer dataset by following a completely at random assumption. Then the performances of the six imputation methods were compared based on three evaluation criteria (Mean Value, Standard Deviation, and Variance).

The results show that the Random Selection method is the method with the best performance to treat the type of categorical data in this study (see Table 4). This method provided a small percentage of error when comparing the metrics for the imputed data with the metrics calculated for the original data. Other methods such as the Most Frequent Value, Multiple Imputation by Chained Equations, and the K-Nearest Neighbor Method offer secondary approaches to treat the data in this study.

In addition, the results in the current study demonstrate that the most commonly used imputation methods such as Mean and Multiple Imputation are not necessarily the most appropriate methods to treat categorical data, type ordinal. In fact, these methods achieved low positions in the ranking of the best methods for imputing the missing data case studied.

Performing the imputation methods used in this study to treat other types of categorical data, such as nominal data and binary categorical data, could serve as a supplementary research to evaluate the performance of these imputation methods under different scenarios. Moreover, further research can be performed to find appropriate approaches to treat categorical data that has other types of missingness patterns, such as MAR and MNAR.

The presence of missing data is a common problem that affects the data analysis process in all kinds of research projects. Although some researchers have studied and provided approaches for the treatment of missing values, there are still few procedures to impute the missingness in categorical data in comparison to the methods available for imputing continuous data. There is no universal method to impute data, but the results of this study suggest that the Random Selection method provides a good approach to handle the missingness problem in ordinal data, a type of categorical data.

6. REFERENCES

[1] Finch, W. 2010. “Imputation methods for missing categorical questionnaire data: A comparison of approaches”. Journal of Data Science, vol. 8(8), pp. 361-378.

[2] Rubin, D. 1976. “Inference and missing data”. Biometrika, vol. 63(3), pp. 581-592.

[3] Little, R. and Rubin, D. 2002. “Introduction” in Statistical Analysis with Missing Data, 2nd ed., John Wiley & Sons, Inc., pp. 3-23.

[4] de Leeuw, D. and Huisman, M. 2003. “Prevention and treatment of item nonresponse”. Journal of Official Statistics, vol. 19(2), pp. 153-176.

[5] Schmitt, P., Mandel, J., and Guedj, M. 2015. “A comparison of six methods for missing data imputation”. Journal of Biometrics & Biostatistics, vol. 6(1), pp. 1-6.

[6] Graham, J. W., Hofer, S. M., Donaldson, S. I., MacKinnon, D. P., and Schafer, J. L. 1997. “Analysis with missing data in prevention research” in The science of prevention: methodological advances from alcohol and substance abuse research, vol. 1, pp. 325-366.

[7] van der Heijden, G. J., Donders, A. R., Stijnen, T., and Moons, K. G. 2006. “Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example”. Journal of Clinical Epidemiology, vol. 59(10), pp. 1102-1109.

[8] Schafer, J. L. and Graham, J. W.2002. “Missing data: our view of the state of the art”. Psychological Methods, vol. 7(2), pp. 147-177.

[9] Myers, T. A. 2011. “Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data”. Communication Methods and Measures, vol. 5(4), pp. 297-310.

[10] Bhattacharyya, G. and Johnson, R. 2014. Statistics: Principles and Methods. 7th edition. John Wiley & Sons, Inc. [E-book] Available: Safari e-book.

[11] Bruce, P. and Bruce, A. 2017. Practical Statistics for Data Scientists. 1st edition. O’Reilly Media, Inc. [E-book] Available: Safari e-book.

[12] Wolberg, W. 1992. “Breast Cancer Wisconsin (Original) Data Set”. Internet: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)

[13] Olinsky, A., Chen, S., and Harlow, L. 2003. The comparative efficacy of imputation methods for missing data in structural equation modeling. European Journal of Operational Research, vol.151(1), pp. 53-79.

[14] Shrive, F. M., Stuart, H., Quan, H., and Ghali, W. A. 2006. “Dealing with missing data in a multi-question depression scale: a comparison of imputation methods”. BMC medical research methodology, vol. 6(1), pp. 57.

[15] Hu, L., Huang, M., Ke, S., and Tsai, C. 2016. “The distance function effect on k-nearest neighbor classification for medical datasets”. SpringerPlus, vol. 5, pp.1-9.

[16] García-Laencina, P. J., Sancho-Gómez, J. L., Figueiras-Vidal, A. R., and Verleysen, M. 2009. “K nearest neighbors with mutual information for simultaneous classification and missing data imputation”. Neurocomputing, vol. 72(7-9), pp. 1483-1493.

[17] Azur, M. J., Stuart, E. A., Frangakis, C., and Leaf, P. J. 2011. “Multiple imputation by chained equations: what is it and how does it work?”. International journal of methods in psychiatric research, vol. 20(1), pp. 40-49.

[18] Mazumder, R., Hastie, T., and Tibshirani, R. 2010. “Spectral regularization algorithms for learning large incomplete matrices”. Journal of Machine Learning Research, vol. 11, pp. 2287-2322.

CHAPTER 2

EARLY DETECTION OF SMOKING BEHAVIOR AMONG YOUTH

EARLY DETECTION OF SMOKING BEHAVIOR AMONG YOUTH

M. Quintero & A. LeBoulluec (2018). “Early detection of smoking behavior among youth”. Prevention Science. ISSN 1389-4986.[2]

[2] The manuscript was submitted to the journal *Prevention Science* in 2018.

Early detection of smoking behavior among youth

Maryuri Quintero
University of Texas at Arlington
701 S. Nedderman Drive
Arlington, TX 76019
ing.maryuri.quintero@gmail.com

Aera LeBoulluec, Ph.D.
University of Texas at Arlington
701 S. Nedderman Drive
Arlington, TX 76019
aeral@uta.edu

ABSTRACT

Objective. To determine the best subset of features for the early detection of cigarette use among youth by using machine learning techniques.

Methods. This study analyzed data collected from 154,685 American students who were 21 years of age or younger in grades 6 through 12 between 1999 and 2009. Chi squared Test, Recursive Feature Elimination, and Extra-Trees Classifier were the three feature selection methods compared in this study. A logistic regression model was constructed for the outcomes of each feature selection method to predict cigarette use among youth. The model with the best performance determined the best subset of features to predict cigarette use among youth timely.

Results. The three models developed in this study achieved performances rates above 91% for all the metrics evaluated. However, the model built with the features selected by the Extra-Trees Classifier method provided slightly better outcomes, which were above 93% for all the evaluation criteria, outstanding over the other two constructed models. Therefore, the features provided by the Extra-Trees Classifier were considered the best subset of features for the early detection of cigarette use among youth.

Conclusion. Algorithms for feature selection in machine learning proved to be highly effective in the selection of the features that best contribute to the prediction of smoking behavior among youth. These techniques can be used to promote opportune anti-smoking programs.

General Terms

Logistic regression, feature selection method, smoking.

Keywords

Tobacco consumption, cigarette use, youth, categorical data.

1. INTRODUCTION

The worldwide statistics of tobacco consumption and its fatal consequences are alarming. Recent studies estimate that 15.4% of the world population (1.1 billion people) are current smokers and anticipate an increase of 45% in the number of smokers over the next twenty years [1, 2, 3]. Globally, more than six million people die every year due to health problems caused by smoking, which means that smoking kills at least one person every six seconds [1, 2, 3, 4]. Unfortunately, the current mortality rate attributable to smoking lean towards a worse scenario in which there is 30% of increase in the number of deaths (eight million deaths) caused by smoking over the next two decades [3].

Currently, cigarette smoking is the second leading preventable cause of death worldwide and the first one in the developed world [2, 5]. In the United States, a developed country, cigarette smoking remains the leading risk factor for fatalities and health problems that can be avoided [1, 6, 7]. Historical data registers 17.7 million smoking-attributable mortalities from 1964 to 2012 in the United States [8], and some studies estimate that more than 480,000 Americans die from health complications associated with smoking each year [5]. Lung cancer, cardiovascular diseases, and respiratory complications are some of the disease outcomes associated with smoking [9], and some researchers approximate that cigarette smoking is the responsible for three in ten cancer deaths nowadays [10, 11]. There is no doubt that smoking leads to risky health problems; however, many health consequences of smoking can be prevented if actions are taken in a timely manner.

Beside the serious health effects of smoking, smoking involves high medical expenditures, years of productive life lost, and productivity losses that become an economic burden for the healthcare system in the United States. For instance, 5.1 million years of life were lost between 2000 and 2004 due to premature deaths caused by smoking, and an average of $311 billion accounted for smoking-related healthcare expenditures between 2009 and 2012, including at least $133 billion in direct medical care for adults aged 18 years or more and $156 in productivity losses [1, 5, 6, 12, 13]. In recent years, cigarette smoking was associated with an estimate of 8.7% of the total healthcare costs in the United States [12], and some authors claim that over $300 billion represent the annual economic impact of illnesses caused by smoking [13]. Health insurers are good examples of entities highly affected by the healthcare costs resulted from smoking. For instance, some federal government-sponsored insurance programs such as Medicare and Medicaid have paid more than 60% of the expenses associated to smoking in the United States [12]. Indeed, facing the problems of tobacco use and decreasing the number of lives affected by smoking could influence positively on the healthcare system in the United States.

The deathtraps of smoking are not only reliant on current tobacco consumption but also on other factors that explain previous smoking behavior such as smoking initiation age, duration of smoking, and daily smoked cigarettes [9, 14]. Approximately 18.1% (42.1 million) American adults were active cigarette consumers in 2012, and most of them (33 million smokers) smoked daily. Some researchers argue that more than 80% of adult American smokers start smoking cigarettes during adolescence, around the 18 years of age, and many of them could have faced addiction to nicotine by young adulthood after the initiation of daily smoking habits

[7, 15, 16]. Each day, over 3,800 youth Americans under 18 years of age smoke a cigarette for the first time, and another 1,000 youth in similar ages convert smoking in a daily practice, although anti-tobacco programs and regulations have been promoted during the past decades [7, 16]. Furthermore, recent studies foresee that smoking will be the cause of an early death for more than five million Americans younger than 18 years of age alive today, so there is an urge for preventing youth initiation and progress to established smoking habits [5].

Prevention policies, regulations, and programs to warn about the dangers of smoking are promoted at the local, state, and national levels in the United States to face the tobacco epidemic, reduce tobacco use initiation, and promote cessation. The U.S. Public Health Service, the Centers for Disease Control and Prevention (CDC), the U.S. Food and Drug Administration (FDA), and other federal, state, and local agencies support anti-smoking movements persistently to improve prevention policies, reduce secondhand smoke impact, assist current smokers, inform about the effects of smoking, monitor tobacco advertising, and control prices on tobacco products in the United States [17]. Previous studies mentioned that tobacco controls implemented over the last half of the century have prevented around 8.0 million early deaths, saved 157 million life-years, increased life expectancy for former smokers, reduced the productivity losses caused by illnesses and deaths, and avoided a cost of about $100 billion per year in the United States [6, 8]. Even though tobacco control efforts have influenced on the public health scenario positively in the United States, continuous initiatives are needed to decelerate the increasing number of early deaths caused by smoking.

Prior researches have identified numerous risk factors that incite cigarette smoking behavior in young people. For instance, Berg, Aslanikashvili, and Djibuti [2] have associated youth smoking to factors such as sociodemographic influences, substance use behaviors, attitudes toward smoking, tobacco-associated policies, exposure to smokers, community influences, family and friends influences, and pro-tobacco advertising. Furthermore, involvement in unhealthy activities, poor performance in school, low school attachment, and Internet use are another important group of risk factors that the previously mentioned researchers have related to smoking initiation and maintenance among young people [2]. Other studies have obtained similar findings about the factors linked with cigarette use among youth in the United States. For example, some studies identified that demographics, economics, culture, exposure to pro-tobacco advertising, living with a smoker, parental smoking, having friends who smoke, and low academic scores influence youth to smoke cigarettes [7, 14]. Similarly, other authors argue that frequent cigarette use among youth Americans can be explained by factors such as race, frequency of cocaine use, physically inactive/active behavior, age of smoking initiation, and feeling sad or hopeless [18]. In brief, youth are exposed to a very influential environment that makes them more vulnerable to the webs of smoking if no actions are taken to prevent youth from smoking.

The results of previous researches provide important findings to study and prevent the factors that have influence on youth behavior toward smoking and the desire for initiating smoking. Preventing youth from smoking initiation is a priority nowadays, but more parallel efforts are needed to identify current youth smokers to assist them in the process of quitting smoking, which could save many lives and reduce the detrimental health and economic consequences of smoking.

This study presents an alternative for the early detection of current cigarettes young consumers by using machine learning techniques. Furthermore, the study determines the best subset of features related to the target problem and provides a simple methodology to develop high-performance models with the help of machine learning algorithms. The models developed in this research are attempts to collaborate in the identification of active young cigarette smokers to carry out timely anti-smoking strategies and avoid nicotine dependence at early stages of life.

2. METHODOLOGY

2.1 Data Source

The present study analyses data from the 1999, 2000, 2002, 2004, 2006, and 2009 National Youth Tobacco Surveys (NYTS) provided by the Center for Disease Control and Prevention (CDC). The NYTS gathers representative data about the prevalence of tobacco use among youth in middle and high school levels and serves as a reference point to design, implement, and assess anti-tobacco strategies in the United States. The survey includes areas of interest about tobacco use among young people such as, tobacco-related beliefs, attitudes towards smoking, access to tobacco, exposure to secondhand smoke behaviors, and exposure to pro- and anti-tobacco influences. Students from both public and private schools participate in the survey, and the participation is voluntary and anonymous. Additional information regarding the methodology followed by CDC to collect the data is available in [19].

2.2 Data preprocessing

Real world data is filled with a lot of issues that can affect the outcomes of any study. In fact, data preprocessing has become a mandatory step to improve the results of any research leading to the need for using technology to develop more efficient data preprocessing techniques. These preprocessing techniques help to explore the data, clean the data, and extract the most useful information from a dataset, which are the three main steps followed when the data is preprocessed in this study. Exploring the data provides an overview of the available information while data cleaning deals with missing values and erroneous data. Feature engineering is the last preprocessing step in this research, and it obtains the best features from the data to improve machine learning models. All these three preprocessing steps are explained in detail in the following sections.

2.2.1 Data exploration

Data exploration is the first step for data preprocessing, and it involves understanding the data, identifying the type of data, discovering trends, and checking out missing values.

The data used in this research consists of 47 core set of variables and 154,685 responses from students who were 21 years of age or younger in grades 6 through 12 between 1999 and 2009 in the United States. All the variables have integer values in binary, multiclass, or ordinal classification formats, which are categorical data types.

The table 1 summarizes the general characteristics of the participants who completed the NYTS between 1999 and 2009. Similar numbers of female and male respondents are represented in the study, and most of the respondents (87%) are between 12 and 17 years of age. The race group *white* is the most prevalent among the participants with 51.1% of respondents, followed by the *Hispanic* and the *African American* groups with 24.5% and 17.4% of responses respectively. Around 7% of the participants are Asian, American Indian, Alaska Native, Native Hawaiian, or another Pacific Islanders. Overall, 47.9% of youth reported being in middle school (6th, 7th, or 8th grades), and another 51.9% of youth reported being in high school (9th, 10th, 11th, or 12th grades). The rest of the participants said that they are ungraded or registered in another grade.

The variable *Cigarette Use (CU)* was chosen as the target variable to identify smoking behavior among youth in this study. Respondents who said that they smoked cigarettes at least one day during the last 30 days were classified as *current cigarette smokers*. Of all youth, 15.1% (23,367) resulted being current cigarette smokers, including 7.1% female youth and 8.0% male youth among all the respondents (see Table 1). From these numbers, one would clearly see that around 85% of observations are labeled with the class *nonsmoker* and the remaining 15% instances are labeled with the class *smoker*, so the target variable exhibits an important inequality distribution between its classes, which is known as *imbalance data* [20].

Imbalanced data happens mostly in classification problems where anomaly discovery is critical and influences on the performance of learning algorithms. There are different methods to simplify the imbalance problem and yield more acceptable results. Some of these methods address imbalance data at the problem definition level, and other methods are focused on the data or algorithm levels [21]. In this research, a problem-definition-level method is used to learn from unbalanced data, and it consists on using appropriate evaluation metrics that provide an effective insight into the performance of the model and give more value to the minority class than the commonly used classification accuracy measure [21]. In fact, accuracy has proven to perform poorly to evaluate models built with imbalanced data because it gives more importance to the common classes [21]. Using robust metrics with more emphasis on the minority class than accuracy measures improves the evaluation of learning models that use imbalance data. The performance metrics considered in this study are explained in the section 2.5.

2.2.2 Data cleaning

After exploring the data and identifying irregularities in the data, data cleaning was essential to improve the quality of the data. Data cleaning is a process that detects inconsistencies in the data, treats incomplete values, deletes useless records, and corrects errors in the data. By using data cleaning, the data becomes more consistent, usable, and efficient, which guarantee better outcomes when the data is analyzed.

Data errors and missing values are the two main problems that data cleaning encounters. Missing values occur when the respondents leave unanswered questions, which reduces the sample size available for study. On the other hand, errors in the data happen due to data entry faults, flawed measuring instruments, or respondents who make mistakes when answering questions. Both missing values and errors in the data appear during the data collection process and reduce the reliability and power of the data. The data used in this study has been previously reviewed by CDC to detect and remove data errors, but this data includes missing values that require treatment.

The six datasets from the 1999, 2000, 2002, 2004, 2006, and 2009 National Youth Tobacco Surveys provided by CDC were pulled into one dataset in this study. The data in all the datasets was represented in categorical values in binary, multiclass, or ordinal formats depending on the feature. The datasets had different number of features and observations, so finding the core set of features was the first step for data cleaning. The following step was to compare the values of each feature between the six datasets. In this second step, adjustments in the values of the features were made if the values presented inconsistences between two or more datasets. Assuring that each dataset had the same features and values was necessary before combining all the six datasets in only one dataset. Finally, just the common features among the six datasets were set aside and pulled together to conform a single dataset with 47 features and 154,685 observations.

Levels of missingness under 6% of missing values were reported in 46 variables after a preliminary study of the data, and this missingness was considered *missing completely at random* (MCAR) for the purposes of the present research. MCAR data happens when a random group of respondents fail to answer a question unintentionally causing missing responses in the dataset that do not have a specific cause [22]. Under this condition of missingness, the researcher has more flexibility to deal with the missing data.

Table 1. Characteristics of the NYTS 1999, 2000, 2002, 2004, 2006, and 2009 participants.

Characteristic	Total	
Total participants	**154,685**	**100%**
Sex		
Female	77,454	**50.1%**
Male	77,231	**49.9%**
Race		
White	78,973	**51.1%**
Black (African American)	26,964	**17.4%**
Hispanic or Latino	37,872	**24.5%**
Asian	7,100	**4.6%**
AI/AN*	2,223	**1.4%**
NH/OPI*	1,553	**1.0%**
Age		
9 yrs old	220	**0.1%**
10 yrs old	164	**0.1%**
11 yrs old	9,021	**5.8%**
12 yrs old	21,866	**14.1%**
13 yrs old	26,525	**17.1%**
14 yrs old	23,642	**15.3%**
15 yrs old	22,219	**14.4%**
16 yrs old	20,934	**13.5%**
17 yrs old	18,940	**12.2%**
18 yrs old	9,657	**6.2%**
19 yrs old	1,031	**0.7%**
20 yrs old	120	**0.1%**
21 yrs old	346	**0.2%**
Grade		
Middle school	74,171	**47.9%**
High school	80,299	**51.9%**
Ungraded/other grade	215	**0.1%**
Total smokers	23,367	**15.1%**
Female	10,964	**7.1%**
Male	12,403	**8.0%**

Note: The values presented in this table were obtained after data imputation.
* AI/AN = American Indian/Alaska Native
* NH/OPI = Native Hawaiian/Other Pacific Islander

The most frequent value method was the technique used to impute the missing values in the dataset. This imputation technique consists on replacing the missing values in a variable with the mode value in that variable [22]. Since the percentages of missing values are low, the most frequent value is considered suitable to treat the missingness in the dataset. The random value method was also used to impute the variable called *sex* in the dataset.

The variable *sex* had two classes (female and male) with similar number of participants in each class and 0.45% of missingness. These missing values were imputed by randomly assigning one of the two classes to the incomplete cases to prevent adding more values to one sex group or creating imbalance data in the variable.

Both data exploration and data cleaning are necessary data preprocessing steps to prepare the data for feature engineering since this final step requires useable data free from errors and missing values to avoid mistakes in the creation of features and features selection. Feature engineering is explained in the following section.

2.3 Feature engineering

Feature engineering is the process of transforming raw data into understandable and representative features that assure the success of a learning model [23]. These features show the highest scores of importance among all the variables available in a dataset, and they provide information that better fit to the learning model and target problem. Having enough informative features, using an appropriate model for the type of features, and evaluating the model performance with the right metrics are necessary to produce better outcomes.

In this study, feature engineering was used to create more informative features and select a subset of most useful features from the 47 variables that conform the NYTS dataset used in this research.

2.3.1 Feature creation

Creating new features from existing information in the dataset can highlight information that is not clearly understood by using the original features. Feature creation consists on combining or decomposing variables and information from the raw data to construct a new more useful feature. The new features are placed within the dataset, and they hold new information and patterns that can be used to improve the outcomes of learning models [24]. In brief, feature creation adds more meaning to the data and fits the data to the research needs.

The response variable in this study, cigarette use, collected information about seven different frequencies of cigarette use among youth, which allowed to identify smoking behaviors among the studied population. Feature creation by binarizing the data in the target variable provided a more effective way to differentiate between smokers and nonsmokers and enhanced the analysis of the data and the learning model in this study.

The variable *cigarette use* (CU) was created based on the response to the NYTS question: “During the past 30 days, on how many days did you smoke cigarettes”. Participants of the NYTS were asked to report whether they have smoke cigarettes 1 or 2 days, 3 to 5 days, 6 to 9 days, 10 to 19 days, 20 to 29 days, all 30 days, or any day during the past 30 days. Then a threshold was used to classify the responses in a binary format [25]. All the contestants who reported that they smoked one or more days during the last 30 days were classified as *smokers* with the binary code one (1). On the other hand, the contestants who said that they did not consume any cigarette during the last 30 days were identified as nonsmokers with the code zero (0). After binarizing the data in the target variable, the new binary feature was included within the dataset for further use.

2.3.2 Feature Selection

Researchers are challenged by the large amounts of data available after the increasing development of modern technology. High-dimensional data is generated and updated at an extraordinary speed leading to the need of removing insignificant and redundant variables that affect the data analysis and the decision-making process. Recent studies support *feature selection* as an alternative for dealing with the problem of analyzing data with high dimensions. Feature selection, or variable selection, is a preprocessing technique used for selecting a subset of relevant variables from the original dataset following a specific criterion, which removes the variables that do not describe the input data efficiently and affect the prediction outcomes [26-29]. Moreover, feature selection facilitates the visualization and understanding of the data, reduces the learning time and computation requirements, enhances the efficiency and accuracy of learning algorithms, reduces the dataset dimensionality, and simplify the learning results [26-29]. Indeed, the challenges that data with large number of features carries can be approached with feature selection techniques to guarantee less complexity and more valuable outcomes from the data analysis process.

In machine learning, feature selection has an important role to boost the results of the learning algorithms. Moreover, feature selection methods can be categorized according to their relationship with learning approaches into filter, wrapper, and embedded methods [29]. In this research, these three feature selection methods are performed and compared to identify the best subset of features to predict cigarette smoking behavior among youth using logistic regression, a supervised machine learning algorithm. Improved classification performance is expected from including feature selection in the preprocessing phase of this study.

• *Filter Methods*

Filtering of features is one of the oldest approaches for feature selection, and it consists on using a ranking criterion to score variables in a dataset and select a subset of most relevant features [26-29]. All the features receive a score based on the criteria, and a threshold is used to remove the features that achieved values below the threshold [27]. The process of filtering the features is performed only once considering univariate methods that analyze each feature individually and its relationship with the response variable [26]. In addition, filter methods select the features independent of any classifier since this method is performed before implementing any learning algorithms [26,27] (see figure 1). Some authors say that filter methods provide a robust and simple approach that reduce correlation among features, avoids overfitting, works well with large data, give fast results, and performs efficiently [26-29]. Chi squared test, Fisher score, information gain, Markov blanket filter, and correlation-based feature selection are some examples of methods used for filtering features.

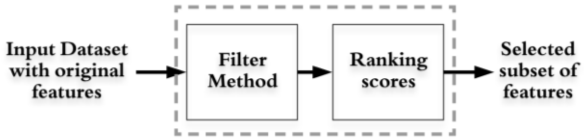

Input Dataset with original features
Filter Method
Ranking scores
Selected subset of features

Figure 1. Feature selection using a filter method algorithm.

The Chi squared test was the filter method used in this study. This filter method is a statistical test that measures the lack of independence between the target variable and every feature variable in the dataset [30, 31]. The Chi squared method in feature selection tests if a specific class (target variable) can be predicted by using the occurrence of a specific feature. In other words, the Chi squared test is applied to a set of features to assess the likelihood of correlation between them based on the frequency distribution of these features. High values of Chi square test define an incorrect hypothesis of independence, so the occurrence of the feature and the occurrence of the target class are highly correlated, which is an indicator that the feature should be selected for training the model. The Chi squared test of a feature ***t*** and the class ***c*** can be calculated from the equation 1 by using the two-way contingency table of a feature ***t*** and a class ***c*** [31].

$$x^2(t,c) = \frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)}$$
(1)

In the equation 1, A is the number of positive instances that contain the feature *t*; B is the number of negative instances that contain feature *t*; C is the number of positive instances that do not contain the feature *t*; D is the number of negative instances that do not contain feature *t*; and, N is the total number of instances. After calculating the chi square scores for all features, the top ranked features are selected to be used for training the model. All the features that are most likely to be independent of the response variable are considered irrelevant for the classification model and discarded from the model training.

• *Wrapper Methods*

Wrapper methods select a subset of features by following a search process that includes a given learning algorithm for the feature selection [29]. During the search process, different combinations of features are assessed and compared to other combinations, and a specific predictive model is used as a black box to evaluate each combination of features and calculate a score for each feature based on their predictive power [28]. After several iterations, the features with better scores and usefulness levels for the learning algorithm are selected in a cross-validation assessment [26, 27, 29] (see figure 2). High classification accuracy, detection of dependencies among features, better classifier interaction, smaller subset size, and optimization of the classifier performance are some of the benefits of using wrapper methods for feature selection [26-29]. On the other hand, wrapper methods have poor capability for generalization, require more computational time and resources, show more complexity, tend to overfit on small training datasets, and is less scalable for large datasets [26].

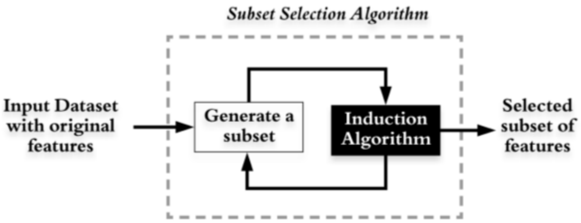

Subset Selection Algorithm
Input Dataset with original features
Generate a subset
Induction Algorithm
Selected subset of features

Figure 2. Feature selection using a wrapper method algorithm.

Some examples of wrapper methods are sequential forward selection, best-first search, genetic algorithms, beam search method, and recursive feature elimination method [26]. This research uses the recursive feature elimination method to select a subset of features from the studied dataset. The recursive feature elimination method trains the classifier, computes the ranking criterion for all features, and removes the feature with the lowest scores from the current set of features after implementing the ranking criterion [32, 33]. This process of training the model, calculating scores, and removing a group of useless features is recursively repeated until obtaining the desired number of features [33].

- *Embedded Methods*

Embedded methods are feature selection methods that combine the learning process with the feature selection step to identify which features provides higher accuracies to the machine learning model [26, 34, 35]. In these feature selection methods, the learning algorithm uses its own process to select features as part of learning and perform classification at the same time. Additionally, embedded methods do not require splitting the training data into training set and testing set, and they measure the ‘‘usefulness” of feature subsets [26]. Similar to wrapper methods, embedded methods are performed by using a specific learning algorithm and cross-validation assessments [26,36], but embedded approaches have the advantage that they avoid retraining the predictive model over again for every feature subset studied such as in wrapper methods [28] (see figure 3). Furthermore, embedded methods are less expensive and complex than wrapper methods, offer better classier interaction, identify dependencies between features effectively, yield faster solutions, avoid over-fitting, and give better use of the available data [26-28, 36]. Although embedded approaches have advantages over other feature selection methods, they are specific to the learning model used, have poor generality, and select features based on hypothesis made by the classifier [26]. Decision tree-based algorithms are among the common embedded methods used for feature selection, and some examples are the ID3 algorithm, CART, C4.5, and random forest [35]. Other examples of embedded methods are the multinomial logistic regression algorithms, artificial neural networks, weighted naïve Bayes, and some regularization models such as those methods based on Lasso or Elastic Net that include linear classifiers like Support Vector Machines [35].

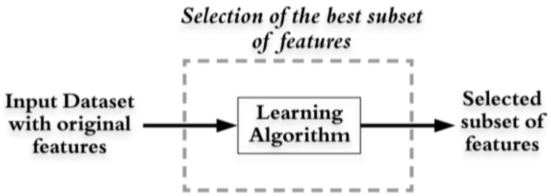

Selection of the best subset of features
Input Dataset with original features
Learning Algorithm
Selected subset of features

Figure 3. Feature selection using an embedded method algorithm.

Extremely randomized Trees or Extra-Tress Classifier was the third feature selection method performed in this study. This method is similar to the Random Forest algorithm because it determines the best split by selecting a random subset of K features at each node [37]. However, the Extra-Trees Classifier splits nodes by choosing both features and cut-points randomly instead of using some criterions while developing a tree. Multiple trees are trained to train the algorithm, and each of these trees are built to generate an ensemble model by using all the training data without bootstrap copying to grow the trees. Besides using the complete learning data to build each tree, a single threshold selected at random is assign to each feature in each node to define the split.

In the Extra-Trees algorithm, it is important to know the M number of trees in the ensemble model, the parameter K that denotes the number of features randomly selected at each node, and the parameter n_{min}, which is the minimum sample size for splitting a node. The parameter K provides information about the strength of the feature selection process, n_{min} exposes the strength of averaging output noise, and M determines the strength of the variance reduction of the ensemble model aggregation [37]. The features that yield the highest importance scores are selected from the random splits.

2.4 Logistic Regression

Regression methods are model building techniques that attempt to describe the relationship between a response variable and a set of predictor variables. When the outcome variable is categorical, logistic regression models facilitate the regression analysis by transforming a non-linear function into a linear form where the log-odds for the positive values of the response variable is a linear combination of a set of *n* independent variables [38]. Logistic regression (LR) is a supervised machine learning algorithm used for classification tasks, and it can be expressed in the equation 2 with the general form of the log-odds (*ln*). The equation 3 shows the model for the natural logarithm of the odds for the outcomes in the response variable followed by the inverse of the logit transformation of the equation 3 [38, 39].

$$\begin{aligned}\ln(Y) &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \\ &= \beta_0 + \textstyle\sum_{j=1}^{n} \beta_j X_j \qquad (2)\end{aligned}$$

$$ln\,\frac{P(Y|X_1,X_2,\ldots,X_n)}{1-\,P(Y|X_1,X_2,\ldots,X_n)} = \beta_0 + \textstyle\sum_{j=1}^{n} \beta_j X_j \qquad (3)$$

$$\begin{aligned}P(Y|X_1,X_2,\ldots,X_n) &= \frac{e^{\beta_0+\sum_{j=1}^{n}\beta_j X_j}}{1+e^{\beta_0+\sum_{j=1}^{n}\beta_j X_j}} \\ &= \frac{1}{1+e^{-(\beta_0+\sum_{j=1}^{n}\beta_j X_j)}}\end{aligned}$$

Where Y is the response variable whose positives and negatives outcomes are represented with the number one (1) and zero (0) respectively, and X_1, X_2, …, X_n denote the set of *n* explanatory variables in the model. β_0 is the y-intercept or expected value of Y when the explanatories variables are zero (0), and β_j is the regression coefficient for each variable X_j.

In this study, the target variable *Cigarette Use (CU)* is a categorical variable with two classes *smokers* and *nonsmokers*, and the goal was to find the best fitting model for this classification problem. From the dataset, 70% randomly selected observations were used to build up the models, and the other 30% of the dataset was used to test the constructed models. One logistic regression model was constructed for each of the three feature selection methods using the features selected by each method. Then the performances of the three resulted models were compared to determine the best model based on the features used for its construction.

2.5 Evaluation criteria

The logistic regression models developed in this study were evaluated with a combination of singular-based metrics and curve-based assessment metrics to provide a more complete evaluation of the imbalance learning. In addition, the confusion matrix was defined to obtain a representation of the classification performance of the model, and it provides the correct predictions and the types of incorrect predictions on the given dataset, which were used to calculate the performance metrics in this study (see figure 4).

		Predicted Values	
		1	***0***
Actual Values	***1***	True Positive	False Negative
	0	False Positive	True Negative

Figure 4. Confusion matrix.

As shown in the figure 4, the positive cases in the data are labeled with the number one (1), and the negatives cases receive the label number zero (0). In this study, one means *smoker* and zero states for *nonsmoker*. In the confusion matrix, the true positive (TP) values are the positive values in the data that were correctly predicted by the model. Similarly, the true negatives (TN) outcomes are the negative values in the original data that the model predicted correctly as negative values. The positive values in the data that were not properly predicted as positives are called *false negatives* (FN), and they are type II errors. On the other hand, the negatives values in the original data that the model did not predict correctly as negatives are *false positives* (FP), which means they are type I errors.

Precision, recall, F1 score, specificity, Matthews correlation coefficient, and the diagnostic odds ratio of a test were the singular-based metrics used to evaluate the models in this study. *Precision* is a measure of exactness that calculates the proportion of positive cases that were predicted correctly among all the positives results, while *Recall* or *Sensitivity* measures the capability of the test to classify actual positive cases as positive in the test (see equations 5 and 6). The *F1 score* determines the weighted average of the precision and recall as a measure of effective classification and test's accuracy, whereas *specificity* measures the proportion of actual negative cases that the test classified correctly as negatives (see equation 7 and 8) [1, 21, 26].

Another singular-based metric used in this study was the *Matthews Correlation Coefficient (MCC)*, an evaluation metric used in machine learning to assess the quality of binary classification problems, and it gives values between -1 and +1 where +1 means perfect prediction and -1 means total disagreement between the actual and predicted values. This metric is considered robust to describe the confusion matrix and evaluate classification models built with imbalance data (see equation 9) [40]. The *diagnostic odds ratio (DOR)* of a test was another single-based metric included in this research. This metric was used to measure the effectiveness of the classification test by calculating the ratio of the odds of the test being positive when the event was positive with respect to the odds of the test being positive when the event was negative (see equation 10) [41]. The DOR achieves values from zero to infinity, and values greater than one indicate useful classification tests.

The Area Under the Receiver Operating Characteristics (AUROC) curve was the curve-based metric considered in this study. The ROC analysis is one of the most common metrics used when learning from unbalanced data, and it provides a representation of what proportion of events are correctly classified for a given false positive rate for every possible classification threshold [21]. The AUROC metric explains the ROC curve in a single number that indicates the effectiveness of a classifier. Values near to one (1) describe good discrimination capacity to distinguish between positive and negatives cases, while values equal or under 0.5 show poor performance to classify cases [21]. AUROC can be calculated with the equation 11 where $\boldsymbol{n_0}$ are given points of class zero, $\boldsymbol{n_1}$ points of class 1, and $\boldsymbol{S_0}$ the sum of ranks of class zero events [21]. The performance metrics included in this study can be calculated with the following equations:

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

$$Specificity = \frac{TN}{FP+TN} \tag{8}$$

$$MCC(\theta) = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(FP+FN)(TN+FP)(TN+FN)}} \tag{9}$$

$$DOR = \frac{TP}{FP}\Big/\frac{FN}{TN} = \frac{LR(+)}{LR(-)} = \frac{sensitivity \times specificity}{(1-sensitivity)\times(1-specificity)} \tag{10}$$

$$AUROC = \frac{2S_0 - n_0(n_0+1)}{2n_0 n_1} \quad (11)$$

The figure 5 shows the main steps involved in this study to determine the best subset of features to predict smoking behaviors among youth. First, the data was preprocessed to prepare the data for feature selection. Then three feature selection methods were performed to generate one subset of best features for each method. The next step in the study was performing logistic regression to construct one predictive model for each subset of features. Finally, the model with the best performance allowed to identify the subset of features that best fit the studied problem.

3. RESULTS

The results of this research are presented according to the outcomes of each feature selection method performed in this study, which includes the subset of best features, the fitted model, and the performance metrics for each feature selection method.

3.1 Filter Method: Chi squared test

The filter method used in this study was the Chi squared test method (CHI), which is an univariate feature selection method that selects the best features by testing the dependence between features in the dataset. The Chi squared method ranks the features based on their relationship with the output variable where highest scores are preferred because they show strongest relationships.

As shown in the figure 6, the five features with highest scores correspond to the NYTS questions related to the number of cigarettes that the subjects have smoked in their entire life (CL), the need for smoking cigarettes (FNC), the access to cigarettes in the past 30 days (GC), the preference for cigarette brands in the past 30 days (BOC), and the number of cigarettes smoked per day in the past 30 days (CPD). These features were used to fit a logistic regression model to predict cigarette use (CU) as shown in the equation 12.

$$\begin{aligned}\ln(CU) = {} & -9.1147 + 0.2270\,\mathrm{CL} + 0.1865\,\mathrm{FNC} \\ & + 0.3683\,\mathrm{GC} + 0.3360\,\mathrm{BOC} \\ & + 2.2608\,\mathrm{CPD}\end{aligned} \quad (12)$$

or,

$$P(CU) = \frac{1}{[1 + e^{-(-9.11 + 0.23\,CL + 0.19\,FNC + 0.37\,GC + 0.34\,BOC + 2.26\,CPD)}]}$$

The fitted model to predict cigarette use (Y) obtained 95.74% precision, 93.87% sensitivity, 94.79% F-measure, and 99.27% specificity. In addition, the Matthews correlation coefficient resulted in 93.90%, and the AUROC was 96.57%. The effectiveness of the classification test measured with the diagnostic odds ratio gave a value of 2076.3.

3.2 Wrapper Method: Recursive Feature Elimination

The Recursive Feature Elimination (RFE) method was applied in this study to determine the best performing feature subset from the perspective of a wrapper method. In this study, the RFE method used logistic regression to constructs a model with the available features, determines the importance of each feature, and removes the least important feature from the current set of features. This procedure was recursively repeated until the top five features were obtained, which were. marked with choice 1 in the ranking of features (see figure 7).

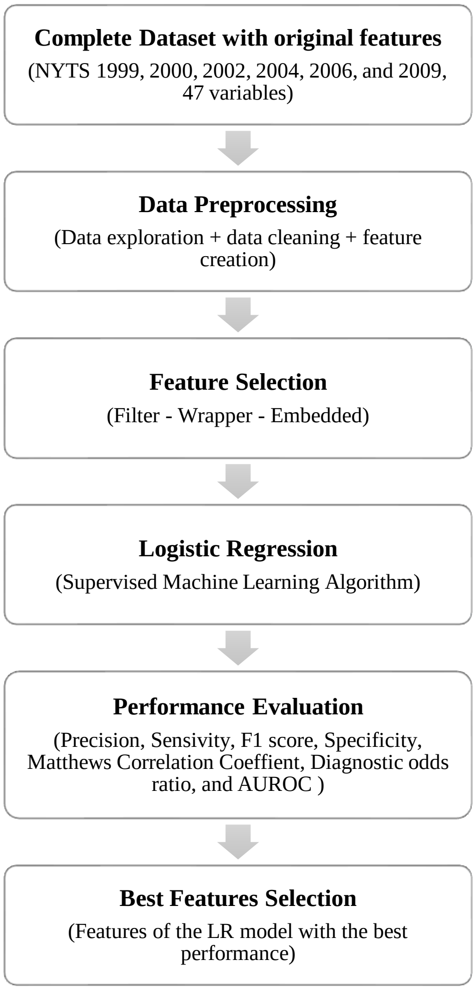

Complete Dataset with original features
(NYTS 1999, 2000, 2002, 2004, 2006, and 2009, 47 variables)
Data Preprocessing
(Data exploration + data cleaning + feature creation)
Feature Selection
(Filter - Wrapper - Embedded)
Logistic Regression
(Supervised Machine Learning Algorithm)
Performance Evaluation
(Precision, Sensivity, F1 score, Specificity, Matthews Correlation Coeffient, Diagnostic odds ratio, and AUROC)
Best Features Selection
(Features of the LR model with the best performance)

Figure 5. Determine the best subset of features to predict current cigarette use among youth.

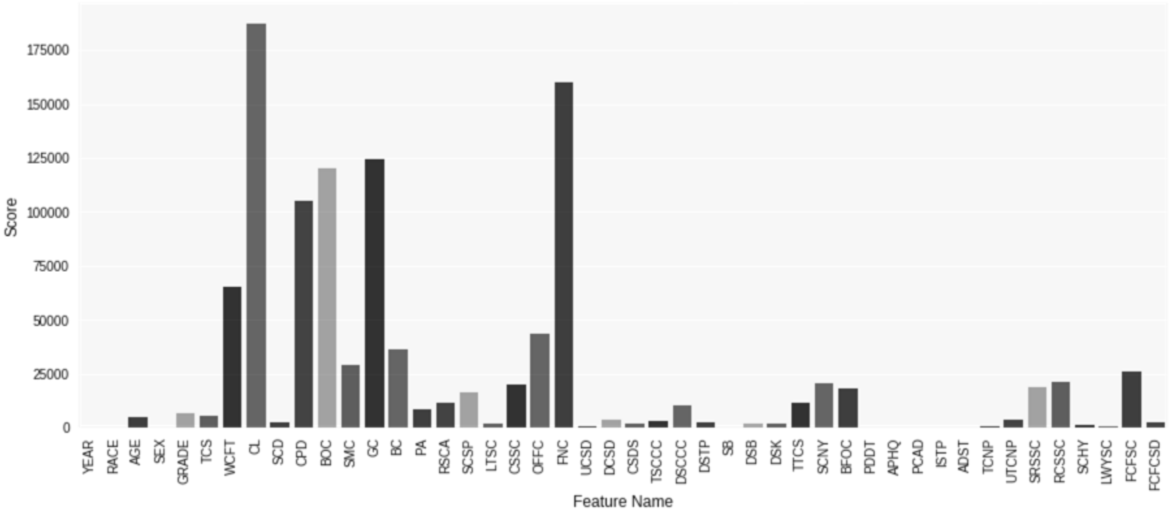


Figure 6. Features ranking using the Chi squared test method.

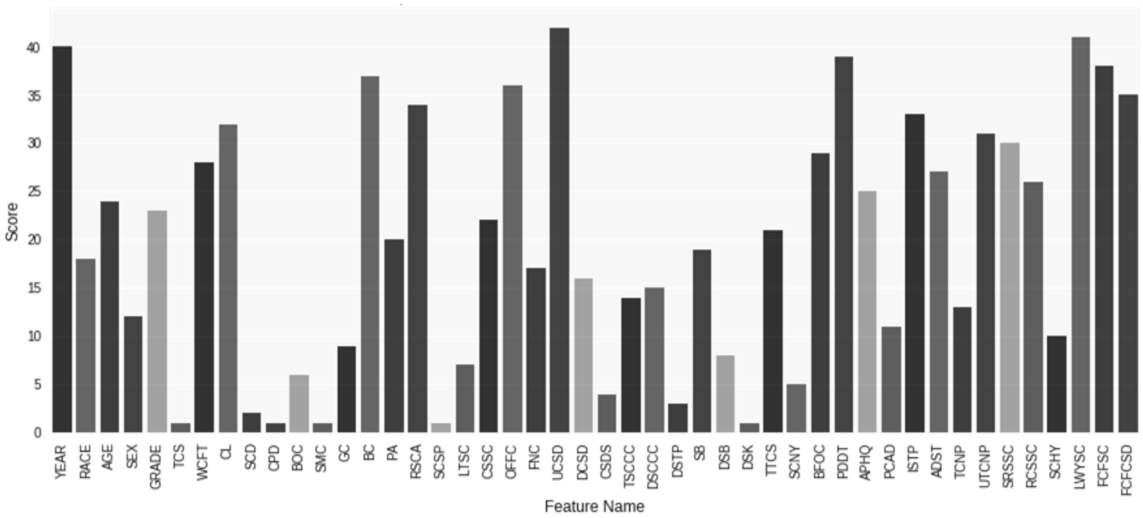


Figure 7. Features ranking using the Recursive Feature Elimination method.

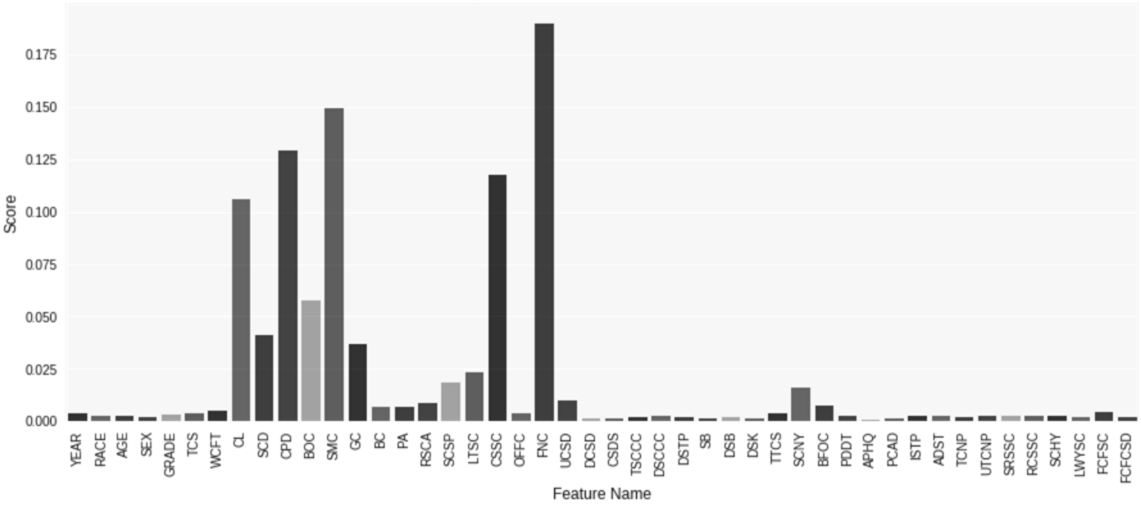


Figure 8. Features ranking using the Extra-Trees Classifier method.

The five features selected from the NYTS questionnaire by using the RFE method were associated with the previous intention of trying smoking cigarettes (TCS), the number of cigarettes smoked per day in the past 30 days (CPD), the preference for menthol cigarettes (SMC), smoking cigarettes on school property in the past 30 days (SCSP), and smoking kreteks in the past 30 days (DSK). The fitted logistic regression model constructed with the features selected by the RFE method is shown in the equation 13.

$$\begin{aligned}\ln(CU) = \; & -4.5401 - 4.0895\ \mathrm{TCS} + 2.9137\ \mathrm{CPD} \\ & + 1.1497\ \mathrm{SMC} + 1.2645\ \mathrm{SCSP} \\ & - 0.8451\ \mathrm{DSK}\end{aligned} \tag{13}$$

or,

$$P(CU) = \frac{1}{[1 + e^{-(-4.54 - 4.09\ TCS + 2.91\ CPD + 1.15\ SMC + 1.26\ SCSP - 0.84\ DSK)}]}$$

The model in the equation 13 achieved 96.79% precision, 91.70% sensitivity, 94.18% F-measure, and 99.47% specificity. On the other hand, the Matthews correlation coefficient resulted equal to 93.90%, the AUROC measured 96.57%, and the diagnostic odds ratio measured 2066.1 for the effectiveness of the tested model.

3.3 Embedded Method: Extra-Tress Classifier

Extremely randomized Trees or Extra-Trees Classifier (ETC) was the third feature selection method performed in this study, which is an embedded method. This method uses an estimator that fits a number of randomized decision trees to obtain a subset of best features ranked according to an importance score. The features with larger scores are the most important features in the dataset from the perspective of the Extra-Tress Classifier (see figure 8).

The Extra-Tress Classifier selected the features associated with the need for smoking cigarettes (FNC), the preference for menthol cigarettes (SMC), the number of cigarettes smoked per day in the past 30 days (CPD), the desire for completely stopping smoking cigarettes (CSSC), and the number of cigarettes that the subjects have smoked in their entire life (CL).

$$\begin{aligned}\ln(Y) = \; & -9.8500 + 0.2407\ \mathrm{FNC} + 1.2358\ \mathrm{SMC} \\ & + 2.6018\ \mathrm{CPD} + 0.0247\ \mathrm{CSSC} \\ & + 0.1654\ \mathrm{CL}\end{aligned} \tag{14}$$

or,

$$P(CU) = \frac{1}{[1 + e^{-(-9.85 + 0.24\ FNC + 1.24\ SMC + 2.60\ CPD + 0.02\ CSSC + 0.16\ CL)}]}$$

The fitted model represented in the equation 14 reached 96.88% precision, 93.98%sensitivity, 95.41% F-measure, and 99.47% specificity. The performance in terms of the Matthews correlation coefficient resulted equal to 94.64%%, the AUROC achieved 96.73%, and the diagnostic odds ratio measured 2935.9.

4. DISCUSSION

Searching for the best subset of features by using different feature selection methods allowed to identify a list of potential aspects to consider when detecting smoking behavior among youth. The selected features appear to be associated with intrapersonal factors, substance use preferences, and access to tobacco.

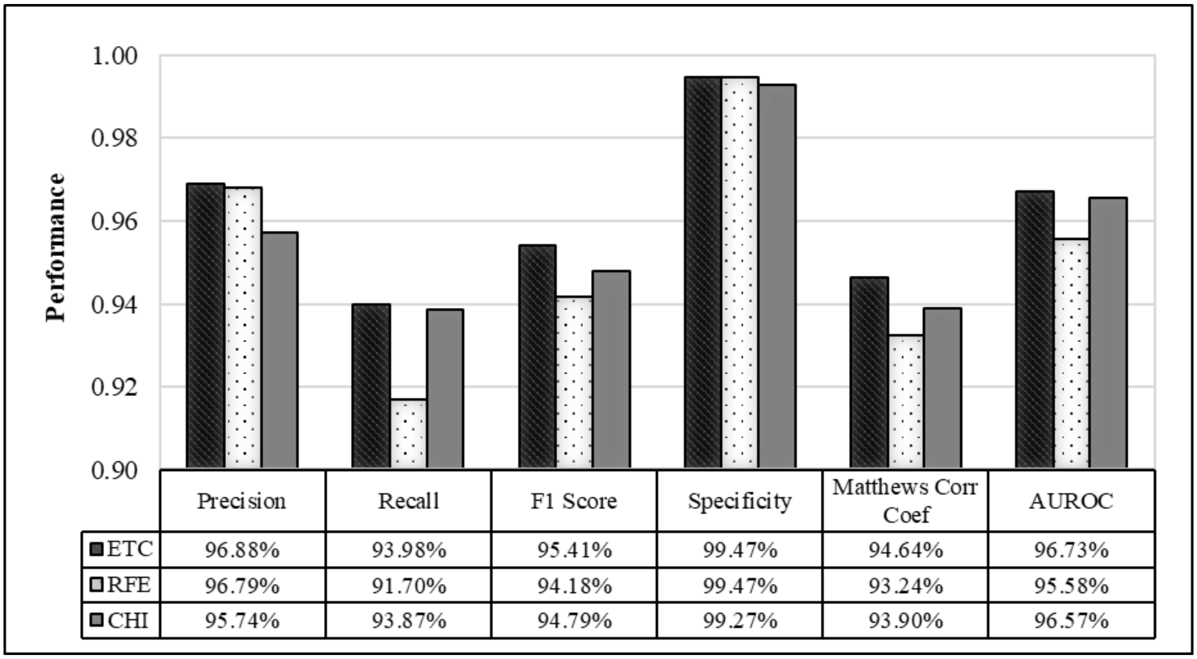
The need for smoking cigarettes and the desire for completely stopping smoking cigarettes were two selected features that highlight current smoking behaviors. Previous smoking intention was another chosen feature from the dataset that showed attitudes toward smoking, and it has been identified in former researches as a risk factor for smoking initiation and maintenance among young people [7]. Three other features explained the cigarette use behavior showing the subject's preferences for cigarette brands, menthol flavor in cigarettes, and consumption of other kind of smoking products, such as kreteks. Current cigarette use was also explained by features that provide information about the access to cigarettes and the frequency of cigarette use in the past and currently. Another interesting feature selected from the dataset was smoking cigarettes on school property. This last feature may be associated with community factors that promote cigarette use, such as the prevalence of smoking in the community in which the youth interact [2]. It is interesting that features related to race, second hand smoke, and pro-tobacco advertising in the dataset were not taken into consideration by the feature selection algorithms, as studies have shown that these factors might influence smoking on youth [2, 7, 14, 18].

The table 2 summarizes the list of features selected by each feature selection method from the NYTS 1999, 2000, 2002, 2004, 2006, and 2009 dataset. The number of cigarettes smoked per day during the past 30 days (CPD) was selected by the three feature selection methods implemented in this study. Similarly, the need for smoking cigarettes (FNC), the preference for menthol cigarettes (SMC), and the number of cigarettes that the subjects have smoked in their entire life (CL) were among the most selected features.

Table 2. Features selected per feature selection method.

Selected Features	**Feature Selection Methods**		
	ETC	RFE	CHI
FNC	**x**		**x**
SMC	**x**	**x**	
CPD	**x**	**x**	**x**
CSSC	**x**		
CL	**x**		**x**
BOC			**x**
GC			**x**
SCSP		**x**	
TCS		**x**	
DSK		**x**	

After selecting subsets of features with the filter, wrapper, and embedded methods, a logistic regression model was fitted for each method to predict youth smoking behavior. The evaluation criteria showed that the three fitted models in this research achieved performances above 91% for all the singular-based and curve-based metrics used to assess the capacity of the models to classify current cigarette use into *smokers* and *nonsmokers*, which means that the models can accurately classified smoking behaviors (see figure 9).



	Precision	Recall	F1 Score	Specificity	Matthews Corr Coef	AUROC
■ETC	96.88%	93.98%	95.41%	99.47%	94.64%	96.73%
□RFE	96.79%	91.70%	94.18%	99.47%	93.24%	95.58%
▨CHI	95.74%	93.87%	94.79%	99.27%	93.90%	96.57%

Figure 9. Comparison of the performance metrics for the three fitted models in the study

.

The model built with the features selected by the Extra-Trees Classifier displayed the highest performance among the models for all the metrics evaluated in this study. Therefore, the best subset of features to address the early detection of smoking behavior among young people are presented by the Extra-Trees Classifier method (see table 2). The Chi squared test method produced the model that achieved the second better performance in sensitivity, f-measure, Matthews Correlation coefficient, AUROC, and diagnostic odds ratio among the three models evaluated. On the other hand, the Recursive Feature Elimination method presented a model with better scores for precision and specificity than the Chi squared test model.

5. CONCLUSION

Cigarette smoking is the second leading preventable cause of death worldwide and the first one in developed countries [1, 2, 5-7]. The increasing fatalities rate and the health and economic burdens caused by smoking can be avoided if the problem of smoking is addressed at an early stage. This research attempts to provide an alternative to detect smoking behaviors among young people since it is at young age that many people develop smoking habits.

Logistic regression, a supervise machine learning algorithm, was proposed to predict cigarette use among youth, and three different feature selection methods were used to select the most relevant features from a large dataset. Specifically, the feature selection approaches used were the Extra-Trees Classifier, the Recursive Feature Elimination, and the Chi squared test method, and one model was developed for each of these methods. Indeed, the learning models benefited from the feature selection process since the process enabled the selection of relevant features that improved the performance of the models. Furthermore, the outcomes of this study show that the three fitted models resulted with powerful capacities to predict the cigarette use among youth, which is the aim of this study. These findings suggest that machine learning approaches may be useful to support the identification of current young cigarette smokers and carry out timely anti-smoking strategies among youth.

Interesting features were selected from the dataset studied, but the best subset of features includes the need for smoking cigarettes (FNC), the preference for menthol cigarettes (SMC), the number of cigarettes smoked per day in the past 30 days (CPD), the desire for completely stopping smoking cigarettes (CSSC), and the number of cigarettes that the subjects have smoked in their entire life (CL). These features were obtained through the Extra-Trees Classifier, which is an embedded method. The model built with the features selected by the Extra-Trees Classifier achieved the highest performance among the assessed models.

The limitation in the current study was the absence of features associated to environmental, contextual, and social factors that affect the individual behavior of the studied subjects such as academic achievement, involvement in physical activities, participation in extracurricular activities, community characteristics, public policies, and emotion al state. Additionally, this study approaches imbalance data issues at the problem definition level by evaluating the models with robust metrics that do not focused on the majority class as the accuracy metric does. However, supplementary research can be done by using other methods to enhance the imbalance problem in the studied dataset, such as data-level strategies (i.e. information acquisition and sampling methods) and algorithm-level methods (i.e. methods that favor rare classes and avoid greed and recursive partitioning).

Further research is needed to deeply understand the problem of smoking among young people. Supplementary studies to measure the effectiveness of anti-tobacco programs among youth, study the prevalence of alternative smoking products among youth, and analyze the association between smoking and intrapersonal factors can support efforts to protect this vulnerable population. Certainly, this study offers insights about the smoking problem among youth that serve as a guidance for potential research in this field in the future.

6. COMPLIANCE WITH ETHICAL STANDARDS

6.1 Funding

Not applicable.

6.2 Conflicts of interest

The authors declare that they have no conflict of interest.

6.3 Ethical approval

This study is based on secondary data taken from the National Youth Tobacco Survey (NYTS) 1999, 2000, 2002, 2004, 2006, and 2009 provided by the Center for Disease Control and Prevention (CDC). Therefore, no ethics approval is needed.

This article does not contain any studies with animals performed by any of the authors.

6.4 Informed consent

Informed consent was obtained from all participants in the study.

7. REFERENCES

[1] Dumortier, A., Beckjord, E., Shiffman, S., & Sejdić, E. (2016). Classifying smoking urges via machine learning. Computer methods and programs in biomedicine, 137, 203-213.

[2] Berg, C. J., Aslanikashvili, A., & Djibuti, M. (2014). A cross-sectional study examining youth smoking rates and correlates in Tbilisi, Georgia. BioMed research international, 2014.

[3] World Health Organization. (2013). *WHO report on the global tobacco epidemic, 2013: enforcing bans on tobacco advertising, promotion and sponsorship. World Health Organization*.

[4] Iqbal, N., Irfan, M., Ashraf, N., Awan, S., & Khan, J. A. (2015). Prevalence of tobacco use among women: a cross sectional survey from a squatter settlement of Karachi, Pakistan. *BMC research notes*, 8(1), 469.

[5] US Department of Health and Human Services. (2014). The health consequences of smoking—50 years of progress: a report of the Surgeon General. *Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health*, 17.

[6] Maciosek, M. V., Xu, X., Butani, A. L., & Pechacek, T. F. (2015). Smoking-attributable medical expenditures by age, sex, and smoking status estimated using a relative risk approach. *Preventive medicine*, 77, 162-167.

[7] Bunnell, R. E., Agaku, I. T., Arrazola, R. A., Apelberg, B. J., Caraballo, R. S., Corey, C. G., ... & King, B. A. (2015). Intentions to smoke cigarettes among never-smoking US middle and high school electronic cigarette users: National Youth Tobacco Survey, 2011–2013. Nicotine & Tobacco Research, 17(2), 228-235.

[8] Holford, T. R., Meza, R., Warner, K. E., Meernik, C., Jeon, J., Moolgavkar, S. H., & Levy, D. T. (2014). Tobacco control and the reduction in smoking-related premature deaths in the United States, 1964-2012. *Jama*, 311(2), 164-171.

[9] Oza, S., Thun, M. J., Henley, S. J., Lopez, A. D., & Ezzati, M. (2011). How many deaths are attributable to smoking in the United States? Comparison of methods for estimating smoking-attributable mortality when smoking prevalence changes. *Preventive medicine*, 52(6), 428-433.

[10] Jacobs, E. J., Newton, C. C., Carter, B. D., Feskanich, D., Freedman, N. D., Prentice, R. L., & Flanders, W. D. (2015). What proportion of cancer deaths in the contemporary United States is attributable to cigarette smoking?. *Annals of epidemiology*, 25(3), 179-182.

[11] US Department of Health and Human Services. (2004). The health consequences of smoking: a report of the Surgeon General.

[12] Xu, X., Bishop, E. E., Kennedy, S. M., Simpson, S. A., & Pechacek, T. F. (2015). Annual healthcare spending attributable to cigarette smoking: an update. *American journal of preventive medicine*, 48(3), 326-333.

[13] Schaffer, P. & Oppenlander, J. (2017). Data Management and Analysis Using JMP: Health care case studies. *SAS Institute.*

[14] Ezzati, M., & Lopez, A. D. (2003). Estimates of global mortality attributable to smoking in 2000. *The lancet*, 362(9387), 847-852.

[15] Hersch, J. (1998). Teen smoking behavior and the regulatory environment. *Duke Law Journal*, 47, 1143-1170.

[16] Department of Health and Human Services. (2012). Preventing Tobacco Use Among Youth and Young Adults: A Report of the Surgeon General.

[17] Centers for Disease Control and Prevention. (2010). How tobacco smoke causes disease: the biology and behavioral basis for smoking-attributable disease: a report of the surgeon general.

[18] Adwere-Boamah, J. (2011). Multiple Logistic Regression Analysis of Cigarette Use among High School Students. Journal of Case Studies in Education, 1.

[19] Center for Disease Control and Prevention. National Youth Tobacco Survey, 1999, 2000, 2002, 2004, 2006, and 2009.

[20] He, H., & Garcia, E. A. (2008). Learning from imbalanced data. *IEEE Transactions on Knowledge & Data Engineering*, (9), 1263-1284.

[21] He, H., & Ma, Y. (Eds.). (2013). *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons.

[22] Quintero, M. & LeBoulluec, A. (2018). Missing Data Imputation for Ordinal Data. *International Journal of Computer Applications* 181(5):10-16.

[23] Zheng, A., & Casari, A. (2018). *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists.* O'Reilly Media, Inc.

[24] Ozdemir, S., & Susarla, D. (2018). *Feature Engineering Made Easy*. Packt Publishing.

[25] Baka, B. (2017). *Python Data Structures and Algorithms*. Packt Publishing.

[26] Jain, D., & Singh, V. (2018). feature selection and classification systems for chronic disease prediction: a review. *Egyptian Informatics Journal*. https://doi.org/10.1016/j.eij.2018.03.002

[27] Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1), 16-28.

[28] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar), 1157-1182.

[29] Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, 300, 70-79.

[30] Jin, X., Xu, A., Bie, R., & Guo, P. (2006, April). Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles. In *International Workshop on Data Mining for Biomedical Applications* (pp. 106-115). Springer, Berlin, Heidelberg.

[31] Yang, Y., & Pedersen, J. O. (1997, July). A comparative study on feature selection in text categorization. In *Icml* (Vol. 97, pp. 412-420).

[32] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3), 389-422.

[33] Johannes, M., Brase, J. C., Fröhlich, H., Gade, S., Gehrmann, M., Fälth, M., ... & Beißbarth, T. (2010). Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients. *Bioinformatics*, 26(17), 2136-2144.

[34] Lal, T. N., Chapelle, O., Weston, J., & Elisseeff, A. (2006). Embedded methods. In Feature extraction (pp. 137-165). Springer, Berlin, Heidelberg.

[35] Jović, A., Brkić, K., & Bogunović, N. (2015, May). A review of feature selection methods with applications. In *Information and Communication Technology, Electronics and Microelectronics* (MIPRO), 2015 38th International Convention on (pp. 1200-1205). IEEE.

[36] Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19), 2507-2517.

[37] Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63(1), 3-42.

[38] Rush, S. (2001). *Logistic regression: The standard method of analysis in medical research* (Vol. 3). Technical Report Mathematics.

[39] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

[40] Boughorbel, S., Jarray, F., & El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PloS one*, 12(6), e0177678.

[41] Glas, A. S., Lijmer, J. G., Prins, M. H., Bonsel, G. J., & Bossuyt, P. M. (2003). The diagnostic odds ratio: a single indicator of test performance. *Journal of clinical epidemiology*, 56(11), 1129-1135.

CHAPTER 3

CONCLUSIONS

CONCLUSIONS

The quality of data plays a critical role for a valid data analysis. In fact, this thesis proved that data analysis produces enhanced outcomes when the data is previously prepared for analysis and is free from missing values, erroneous information, and irrelevant data. Extracting the most useful information also improves the results of the data analysis and supports decision-making. In this study, two researches allowed to evaluate the performance of techniques to improve the quality of the data for data analysis.

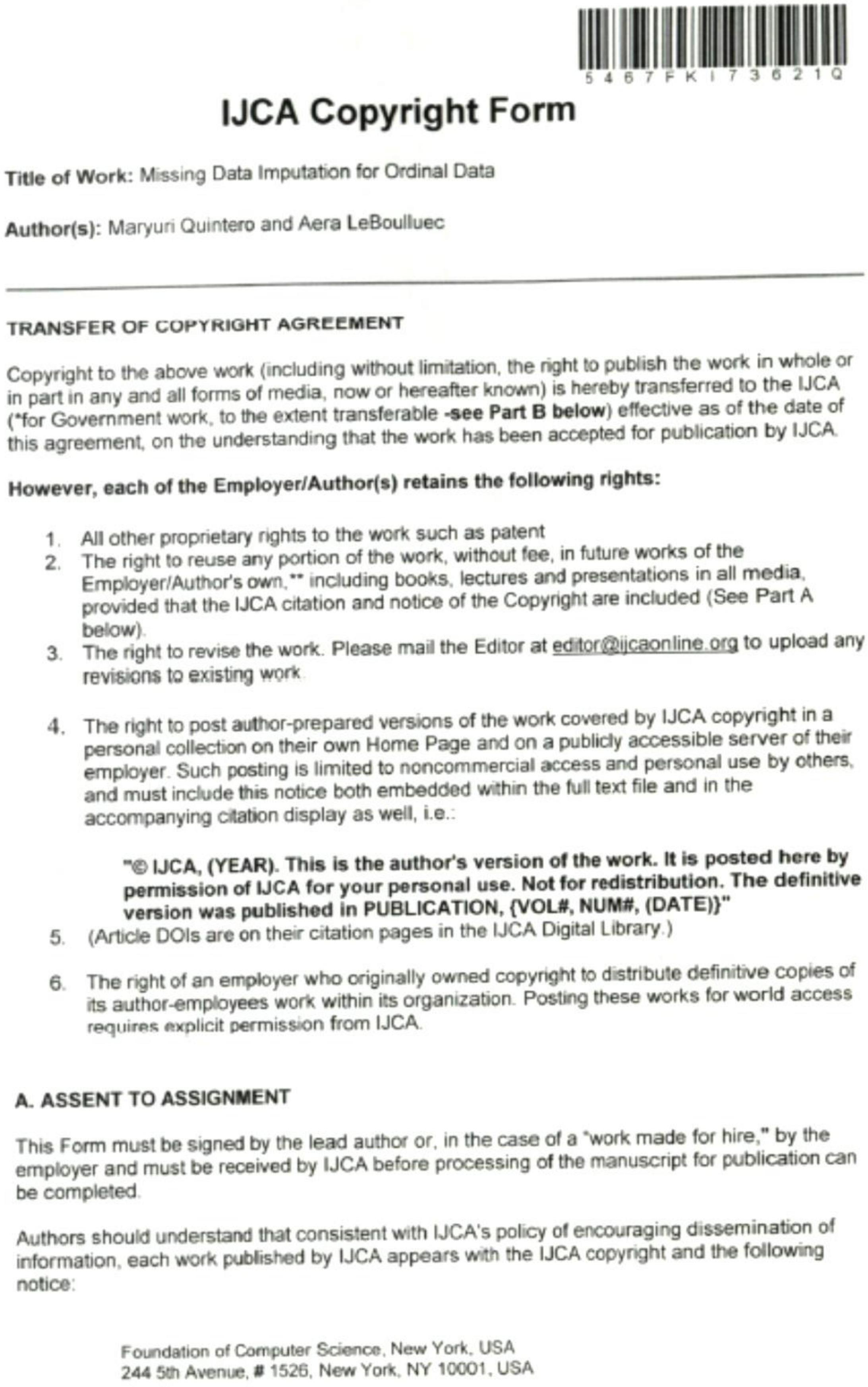
The first research included in this thesis presented a comparison of methods for the imputation of missing values in ordinal data. Six different imputation methods were performed to treat missingness in ordinal data, and the Random Selection method was the method with the best performance to treat the missing data in the studied dataset. This method produced the smallest percentage of errors for different levels of missingness and kept the tendencies and spread of the data. The Most Frequent Value method, Multiple Imputation by Chained Equations, and the K-Nearest Neighbor method offered secondary approaches to treat ordinal data.

Determining the best features to predict the use of cigarettes among youth in the United States was the second research included in this thesis. In this research, three different feature selection procedures were used, but the Extra-Trees Classifier was the one that selected the best features to predict the target problem. The need for smoking cigarettes (FNC), the preference for menthol cigarettes (SMC), the number of cigarettes smoked per day in the past 30 days (CPD), the desire for completely stopping smoking cigarettes (CSSC), and the number of cigarettes that the subjects have smoked in their entire life (CL) were the top five features chosen by the Extra-Trees Classifier, which yielded a predictive model with high levels of precision and efficiency. Additionally, the outcomes of this study showed that the three feature selection methods resulted in fitted models with powerful capacities to predict the cigarette use among youth, so using feature selection methods influenced positively on the model building process. In brief, the second research in this document proved that learning models benefited from the feature selection methods since these methods enabled the selection of relevant features that improved the performance of the models.

Dealing with data challenge the performance of all kind of research. However, methodologies have been developed to facilitate the analysis of data, but still more research is needed to identify the techniques with better applicability for certain data problems. In the studies involved in this thesis, the treatment of missing values and feature selection methods were two steps that resulted beneficial to boost the data analysis and build high-performance predictive models. Moreover, the evaluation metrics used in this thesis were also crucial to analyze the performance of different techniques and compare methods appropriately. Missing data imputation, feature selection, and model evaluation can ensure better results from the data analysis if they are performed correctly during the data analysis process.

APPENDIX A

IJCA COPYRIGHT FORM



5467FKI73621Q

IJCA Copyright Form

Title of Work: Missing Data Imputation for Ordinal Data

Author(s): Maryuri Quintero and Aera LeBoulluec

TRANSFER OF COPYRIGHT AGREEMENT

Copyright to the above work (including without limitation, the right to publish the work in whole or in part in any and all forms of media, now or hereafter known) is hereby transferred to the IJCA (*for Government work, to the extent transferable **-see Part B below**) effective as of the date of this agreement, on the understanding that the work has been accepted for publication by IJCA.

However, each of the Employer/Author(s) retains the following rights:

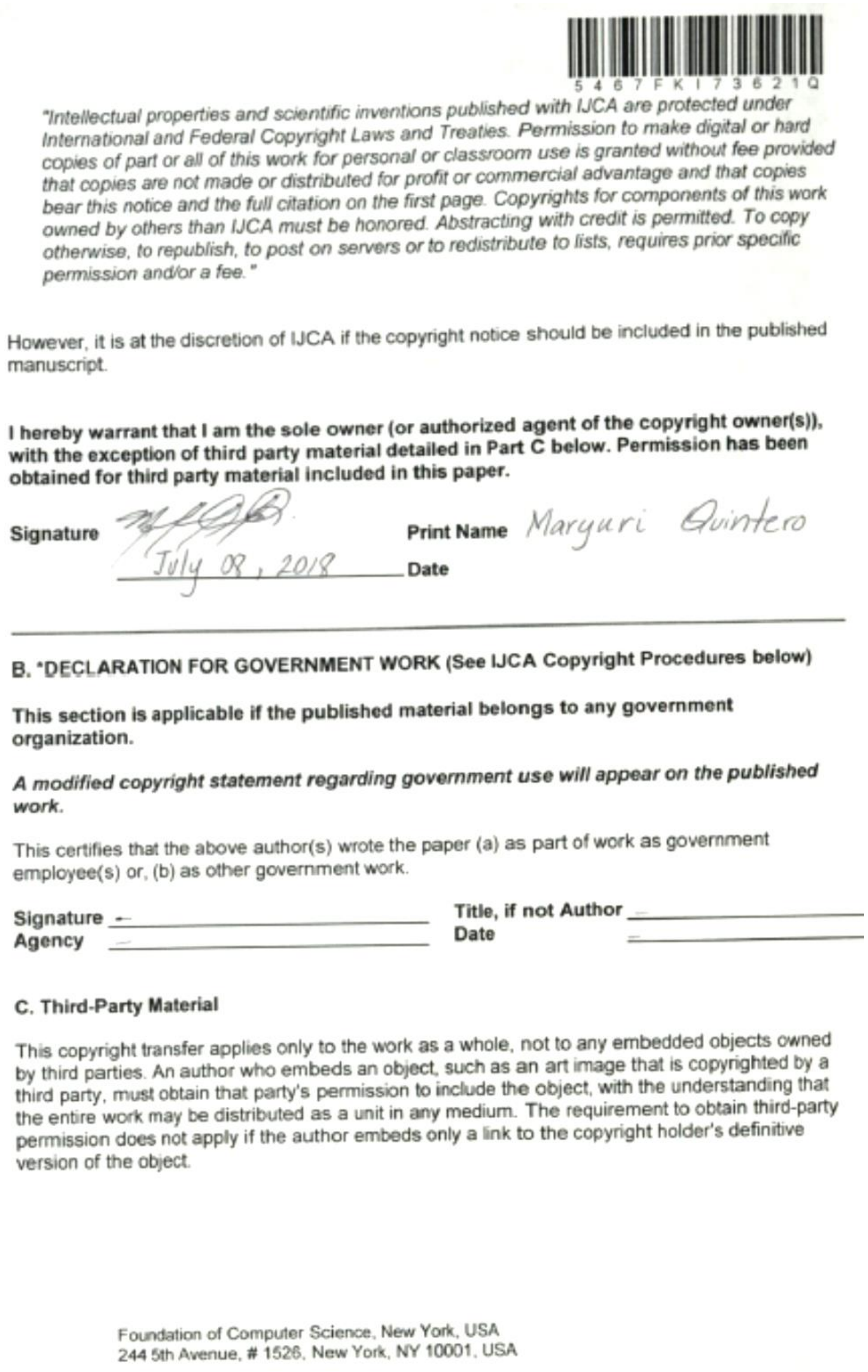
1. All other proprietary rights to the work such as patent
2. The right to reuse any portion of the work, without fee, in future works of the Employer/Author's own,** including books, lectures and presentations in all media, provided that the IJCA citation and notice of the Copyright are included (See Part A below).
3. The right to revise the work. Please mail the Editor at editor@ijcaonline.org to upload any revisions to existing work.
4. The right to post author-prepared versions of the work covered by IJCA copyright in a personal collection on their own Home Page and on a publicly accessible server of their employer. Such posting is limited to noncommercial access and personal use by others, and must include this notice both embedded within the full text file and in the accompanying citation display as well, i.e.:

   "© IJCA, (YEAR). This is the author's version of the work. It is posted here by permission of IJCA for your personal use. Not for redistribution. The definitive version was published in PUBLICATION, {VOL#, NUM#, (DATE)}"
5. (Article DOIs are on their citation pages in the IJCA Digital Library.)
6. The right of an employer who originally owned copyright to distribute definitive copies of its author-employees work within its organization. Posting these works for world access requires explicit permission from IJCA.

A. ASSENT TO ASSIGNMENT

This Form must be signed by the lead author or, in the case of a "work made for hire," by the employer and must be received by IJCA before processing of the manuscript for publication can be completed.

Authors should understand that consistent with IJCA's policy of encouraging dissemination of information, each work published by IJCA appears with the IJCA copyright and the following notice:



5467FKI73621Q

"Intellectual properties and scientific inventions published with IJCA are protected under International and Federal Copyright Laws and Treaties. Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than IJCA must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee."

However, it is at the discretion of IJCA if the copyright notice should be included in the published manuscript.

I hereby warrant that I am the sole owner (or authorized agent of the copyright owner(s)), with the exception of third party material detailed in Part C below. Permission has been obtained for third party material included in this paper.

Signature **Print Name** Maryuri Quintero

July 08, 2018 **Date**

B. *DECLARATION FOR GOVERNMENT WORK (See IJCA Copyright Procedures below)

This section is applicable if the published material belongs to any government organization.

A modified copyright statement regarding government use will appear on the published work.

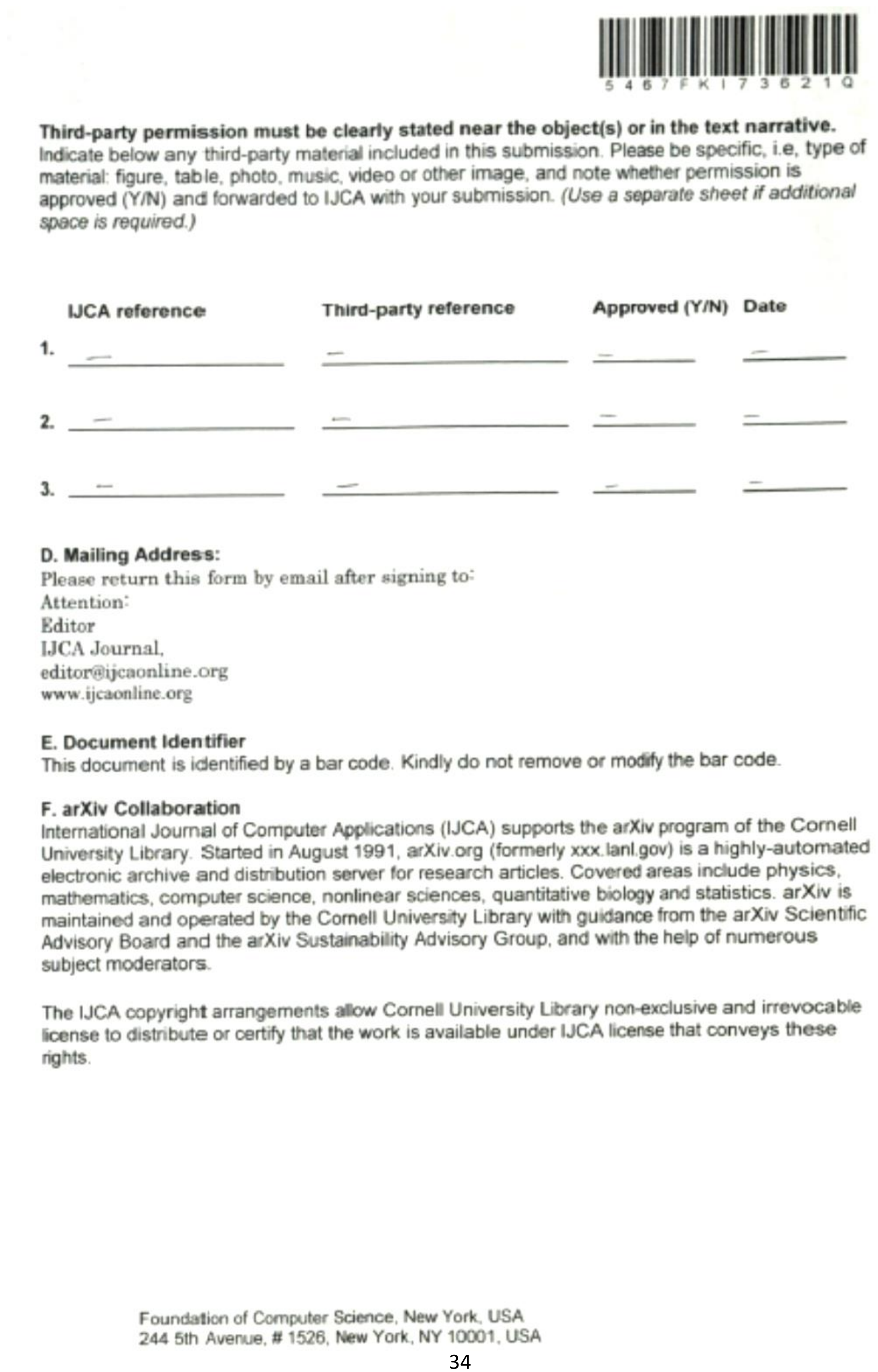
This certifies that the above author(s) wrote the paper (a) as part of work as government employee(s) or, (b) as other government work.

Signature – ____________________ **Title, if not Author** – ____________________

Agency – ____________________ **Date** – ____________________

C. Third-Party Material

This copyright transfer applies only to the work as a whole, not to any embedded objects owned by third parties. An author who embeds an object, such as an art image that is copyrighted by a third party, must obtain that party's permission to include the object, with the understanding that the entire work may be distributed as a unit in any medium. The requirement to obtain third-party permission does not apply if the author embeds only a link to the copyright holder's definitive version of the object.



5467FK173621Q

Third-party permission must be clearly stated near the object(s) or in the text narrative. Indicate below any third-party material included in this submission. Please be specific, i.e, type of material: figure, table, photo, music, video or other image, and note whether permission is approved (Y/N) and forwarded to IJCA with your submission. *(Use a separate sheet if additional space is required.)*

	IJCA reference	Third-party reference	Approved (Y/N)	Date
1.	–	–	–	–
2.	–	–	–	–
3.	–	–	–	–

D. Mailing Address:

Please return this form by email after signing to:
Attention:
Editor
IJCA Journal,
editor@ijcaonline.org
www.ijcaonline.org

E. Document Identifier

This document is identified by a bar code. Kindly do not remove or modify the bar code.

F. arXiv Collaboration

International Journal of Computer Applications (IJCA) supports the arXiv program of the Cornell University Library. Started in August 1991, arXiv.org (formerly xxx.lanl.gov) is a highly-automated electronic archive and distribution server for research articles. Covered areas include physics, mathematics, computer science, nonlinear sciences, quantitative biology and statistics. arXiv is maintained and operated by the Cornell University Library with guidance from the arXiv Scientific Advisory Board and the arXiv Sustainability Advisory Group, and with the help of numerous subject moderators.

The IJCA copyright arrangements allow Cornell University Library non-exclusive and irrevocable license to distribute or certify that the work is available under IJCA license that conveys these rights.