BOOK OF PROCEEDINGS

6th International Conference on Advances in Statistics

ICAS CONFERENCE

INTERNATIONAL CONFERENCE ON ADVANCES IN STATISTICS

October 16-18 2020

http://www.icasconference.com/

ICAS'2020

6th International Conference on Advances in Statistics

Published by the ICAS Secretariat
Editors:
Prof. Dr. Ismihan BAYRAMOGLU and Prof. Dr. Fatma NOYAN TEKELI

ICAS Secretariat
Büyükdere Cad. Ecza sok. Pol Center 4/1 Levent-İstanbul
E-mail: icasconference.academic@gmail.com
http://www.icasconference.com

ISBN: 978-605-65509-2-8

SCIENTIFIC COMMITTEE

Prof. Dr. Ayman BAKLEEZI
Qatar University – Qatar

Prof. Dr. Barry C. ARNOLD
University of California, Riverside – USA

Prof. Dr. George YANEV
The University of Texas Rio Grande Valley – USA

Prof. Dr. Hamparsum BOZDOGAN
The University of Tennessee – USA

Prof. Dr. Hamzeh TORABI
Yazd University – IRAN

Prof. Dr. Jorge NAVARRO
Universidad de Murcia – Spain

Prof. Dr. Jose Maria SARABIA
University of Cantabria – Spain

Prof. Dr. Leda MINKOVA
University of Sofia – Bulgaria

Prof. Dr. Narayanaswamy BALAKRISHNAN
McMaster University – Canada

Prof. Dr. Nikolai KOLEV
University of Sao Paulo – Brasil

Prof. Dr. Sarjinder SINGH
Texas A&M University-Kingsville – USA

Prof. Dr. Stelios PSARAKIS
Athens University of Economics & Finance – Greece

Dr. Ilham AKHUNDOV
University of Waterloo – Canada

Prof. Dr. Aydin ERAR
Mimar Sinan Fine Arts University – Turkey

Prof. Dr. I. Esen YILDIRIM
Marmara University – Turkey

Prof. Dr. Gulay BASARIR
Mimar Sinan Fine Arts University – Turkey

Prof. Dr. Mujgan TEZ
Marmara University – Turkey

Prof. Dr. Sahamet BULBUL
Istanbul Ayvansaray University – Turkey

Assoc. Prof. Dr. Baris ASIKGIL
Mimar Sinan Fine Arts University – Turkey

Assoc. Prof. Dr. Esra AKDENIZ
Marmara University – Turkey

ORGANIZATION COMMITTEE

Prof. Dr. Ismihan BAYRAMOGLU
Izmir University of Economics – Turkey
Conference Chair

Prof. Dr. Fatma NOYAN TEKELI
Yıldız Technical University – Turkey

Prof. Dr. Gulhayat GOLBASI SIMSEK
Yıldız Technical University – Turkey

Assoc. Prof. Dr. Gulder KEMALBAY
Yıldız Technical University – Turkey

Instructor PhD Ozlem BERAK KORKMAZOGLU
Yıldız Technical University – Turkey

Dear Colleagues,

On behalf of the Organizing Committee, I am pleased to invite you to participate in ***6th INTERNATIONAL E – CONFERENCE ON ADVANCES IN STATISTICS*** which will be organised **fully virtual** on dates between 16-18 October 2020 .

Conference was originally planned for April 2020 but due to the global spread of COVID-19 (Corona Virus) and The Council of Higher Education's declaration on "Measures to be Taken in Higher Education Institutions about COVID-19" (March 6, 2020) the conference is postponed to this ***current date.***

All informations are available in conference web site. For more information please do not hesitate to contact us. info@icascconference.com

We cordially invite prospective authors to submit their original papers to ICAS-2020,

. Applied Statistics
· Bayesian Statistics
· Big Data Analytics
· Bioinformatics
· Biostatistics
· Computational Statistics
· Data Analysis and Modeling
· Data Envelopment Analysis
· Data Management and Decision Support Systems
· Data Mining
· Energy and Statistics
· Entrepreneurship
· Mathematical Statistics
· Multivariate Statistics
· Neural Networks and Statistics
· Non-parametric Statistics
. Operations Research
· Optimization Methods in Statistics
· Order Statistics
· Panel Data Modelling and Analysis
· Performance Analysis in Administrative Process
· Philosophy of Statistics
· Public Opinion and Market Research
· Reliability Theory
· Sampling Theory
· Simulation Techniques
· Spatial Analysis
· Statistical Software
· Statistical Training
· Statistics Education
· Statistics in Social Sciences
· Stochastic Processes
· Supply Chain
· Survey Research Methodology
· Survival Analysis
· Time Series
· Water and Statistics
· Other Statistical Methods

Selected papers will be published in **Journal of the Turkish Statistical Association**. http://jtsa.ieu.edu.tr

We hope that the conference will provide opportunities for participants to exchange and discuss new ideas and establish research relations for future scientific collaborations.

Conference Website : https://icasconference.com
E Mail: icasconference.academic@gmail.com

On behalf of Organizing Committee:
Conference Chair
Prof. Dr. İsmihan BAYRAMOĞLU
Izmir University of Economics

16 OCTOBER 2020 FRIDAY

10:30 – 11:00 OPENING CEREMONY
Professor Ismihan BAYRAMOGLU / Izmir University of Economics - Turkey

11:00 – 11:30	**Keynote Speech:** **Professor Serkan ERYILMAZ /** Atılım University - Turkey *Statistical Aspects of Wind Energy*

11:30 – 11:45	**B R E AK**

SESSION A

SESSION CHAIR	**Gulder KEMALBAY**	
TIME	**PAPER TITLE**	**PRESENTER /** CO AUTHOR
11:45 – 12:00	Polya-Aeppli Process of Order k of Second Kind with an Application	**Meglena LAZAROVA** / Stefanka CHUKOVA & Leda MINKOVA
12:00 – 12:15	A Statistical Consistent Test Based on Bivariate Random Thresholds	**Aysegul EREM** / Ismihan BAYRAMOGLU

12:15 – 12:30	Performance Analysis of Ridge Deviance Control Charts for Monitoring Poisson Profiles	**Ulduz MAMMADOVA** / M. Revan OZKALE
12:30 – 12:45	Gompertz-Exponential Distribution: Record Value Theory and Applications in Reliability	**Shakila BASHIR** / Ahmad Mahmood QURESHI
12:45 – 13:00	On the Order Statistics of Dependent Random Variables Constructed from Bivariate Random Sequences	**Ismihan BAYRAMOGLU** / Omer L. GEBIZLIOGLU

13:00 – 13:30	**LUNCH BREAK**

SESSION B

SESSION CHAIR	**Fatma NOYAN TEKELI**	
TIME	**PAPER TITLE**	**PRESENTER** / CO AUTHOR
13:30 – 13:45	Robust Ranked Set Sampling Methods for One-Sample T-Test	**Yusuf Can SEVIL** / Tuğba YILDIZ
13:45 – 14:00	On Classification with Multiple Birth Support Vector Machines	**Guvenc ARSLAN**
14:00 – 14:15	New Mathematical Formulations for the Distributed Permutation Flowshop Scheduling Problem	**Alper HAMZADAYI** / Hanifi Okan ISGUDER
14:15 – 14:30	Handling Missing Values in Random Forests: An Application to Demographic Survey Data	**Duygu ICEN** / Ayse ABBASOGLU OZGOREN & Anıl BOZ SEMERCI

14:30 – 14:45	Tail Dependence Estimation Based on Estimation of Kendall Distribution Function via Rational Bernstein Polynomials	**Mahmut Sami ERDOGAN** / Selim Orhun SUSAM
14:45 – 14:50 **(Poster)**	POT Method for Ruin Probability in Infinite Time with Non-Stationary Arrivals and Heavy Tailed Distribution Claims or Loss Models	**Redhouane FRIHI** / Rassoul ABDELAZIZ

14:50 – 15:00	**B R E AK**

SESSION C

SESSION CHAIR	**Jale ORAN**	
TIME	**PAPER TITLE**	**PRESENTER** / CO AUTHOR
15:00 – 15:15	SHARP: A State-Space HAR Model Using Particle Gibbs Sampling	**Aya GHALAYINI** / Marwan IZZELDIN & Mike TSIONAS
15:15 – 15:30	Inflation Targeting, Credibilty and Taylor Rule: The Estimation of Monetary Policy Reaction Function for the Central Bank of Turkey	**Gozde YILDIRIM** / Ahmet TIRYAKI
15:30 – 15:45	Machine Learning Extension of the Simulated Method of Moments for Estimation of Agent-Based Models	**Jiri KUKACKA**
15:45 – 16:00	The Impact of Periodicity on Volatility-Volume Relations	**Sherry LUO** / Zhen WEI & Marwan IZZELDIN

16:00 – 16:15	**B R E AK**

16:15 – 16:45	**Keynote Speech:** **Professor Jale ORAN /** Marmara University - Turkey *Behavioral Finance: A Retrospective*

17 OCTOBER 2020 SATURDAY

10:30 – 11:00	**Keynote Speech:** **Professor Rza BASHIROV /** Eastern Mediterranean University - North Cyprus *Statistical Comparison of Modelling Approaches Demonstrated for Biomedical Networks*

11:00 – 11:15	**B R E AK**

SESSION D

SESSION CHAIR	**Guvenc ARSLAN**	
TIME	**PAPER TITLE**	**PRESENTER /** CO AUTHOR
11:15 – 11:30	Bell Marginal Models for Longitudinal Count Outcomes	**Hatice Tul Kubra AKDUR**
11:30 – 11:45	The Interplay Between Determinism, Stochasticity And Fuzzyness Illustrated For P16-Mediated Pathway	**Nimet Ilke AKCAY** / Rza BASHIROV
11:45 – 12:00	Modelling Pre-service Mathematics Teachers Reasoning Under Uncertainty in the Egyptian Context	**Samah Gamal Ahmed ELBEHARY**
12:00 – 12:15	Hypogeometric Distribution and Related Discrete Time Point Process	**Silvana PARALLOJ /** Stefanka CHUKOVA & Leda MINKOVA
12:15 – 12:30	Fusion of Geometric and Texture Features for Side View Face Recognition	**Salman Mohammed JIDDAH, Main ABUSHAKRA, Kamil YURTKAN**

12:30 –13:30	**LUNCH BREAK**

SESSION E

SESSION CHAIR	**Gulhayat GOLBASI SIMSEK**	
TIME	**PAPER TITLE**	**PRESENTER /** CO AUTHOR
13:30 – 13:45	Different Similarity Measures for the Clustering of Time Series	**Yamina KHEMAL-BENCHEIKH** / Assia BOUIZANE
13:45 – 14:00	A Method for Constructing and Interpreting Some Weighted Premium Principles	**Gema PIGUEIRAS** / Antonia CASTAÑO-MARTÍNEZ & Fernando LÓPEZ-BLAZQUEZ & Miguel A. SORDO
14:00 – 14:15	A Bayesian Model of COVID19 New Cases	**Marta SANCHEZ-SANCHEZ** / Alfonso SUÁREZ-LLORENS & Ángel BERIHUETE
14:15 – 14:20 **(POSTER)**	A Survey Of Groundwater Quality in Suburb of Ulaanbaatar City, Mongoliahydrochemical Investigation of Groundwater in 14 Th Khoroo of Khan-Uul District Using Multivariate Statistical Techniques	**Enkhbayar JAMSRANJAV /** Dagvasuren GANBOLD & Gerelt-Od DASHDONDOG & Munkhtsetseg ZORIGT

14:20 – 14:30	**BREAK**

14:30 – 15:00	**Keynote Speech:** **Professor Agamirza BASHIROV /** Eastern Mediterranean University - North Cyprus *Wide Band Noises: Theory and Applications*

SESSION F

SESSION	**Aysegul EREM**	

CHAIR		
TIME	**PAPER TITLE**	**PRESENTER /** CO AUTHOR
15:00 – 15:15	Evaluating the Effects of Outliers in Bootstrap	**Ugur BINZAT** / Engin YILDIZTEPE
15:15 – 15:30	An Alternative P Chart For Monitoring High-Quality Processes Based on Improved Estimator	**Senem SAHAN VAHAPLAR** / Ozlem EGE ORUC
15:30 – 15:35 **(Poster)**	Estimating the Gini Index for Income Loss Distributions Under Random Censoring	**Bari AMİNA /** Abdelaziz RASSOUL & Ould Rouis HAMID

15:35 – 16:15	**B R E AK**

SESSION G

SESSION CHAIR	**Umut UYAR**	
TIME	**PAPER TITLE**	**PRESENTER /** CO AUTHOR
16:15 – 16:30	Regression Discontinuity Design in the Analysis of South African Social Development Praxis	**Doug ENGELBRECHT** / Joshua ENGELBRECH
16:30 – 16:45	Assessing The Impact of Microfinance: Findings From a Survey of Microfinance Participants in Akole Taluka of Maharashtra, India	**Amita YADWADKAR**
16:45 – 17:00	Currency Devaluation Versus Tariff – A Trade War Simulation	**Jen-CHI CHENG** / Bryce ENGELLAND
17:00 – 17:15	Effects of The Epidemic on The Bist Network Structure	**Deniz SUKRUOGLU**
17:15 – 17:30	Predictive Power of Exchange Rates and Interest Rates for Capacity Utilization	**Sıtkı SONMEZER** / Ismail

	and Real Sector Confidence in Turkey	Erkan CELIK

CONTENTS

HANDLING MISSING VALUES IN RANDOM FORESTS: AN APPLICATION TO DEMOGRAPHIC SURVEY DATA

Duygu İÇEN[1] – Ayşe ABBASOĞLU ÖZGÖREN[2] – Anıl BOZ SEMERCİ[3]

[1] Hacettepe University, Department of Statistics, duyguicn@hacettepe.edu.tr
[2] Hacettepe University Institute of Population Studies, Department of Demography, ayseabs@gmail.com
[3] Hacettepe University, Department of Business Administration, annilboz@gmail.com

Abstract
The purpose of this study is to examine how missing values should be handled when a classification is made with a random forest algorithm to the most recent Turkey Demographic and Health Survey (2018 TDHS) data. The main idea of ensemble learning methods is to create a better model, each solving the same problem, with more accurate and reliable predictions or decisions than using a single model [1]. As being one of the ensemble methods, Random Forests (RFs) is developed by Leo Breiman in 2001 and has been increasingly used in the field of data science since then [2]. Some important advantages of the random forest method are that it handles a large number of input variables and that it is speedy [3]. The inevitable problem of the data scientist is that s/he faces missing values in almost all areas of science. We first focus on the 2018 Turkey Demographic and Health Survey (2018 TDHS) data that has some missing values. We use different imputation methods for the missing values of this data [4]. Finally, the best imputation method for 2018 TDHS data is determined in the classification problem using Random forests.
Key Words: *Random Forest, Missing Values, 2018 TDHS Data*

1. Introduction

Missing data is a common problem in nearly every area of scientific analysis that researchers have to deal with. This yields scientists to focus on missing value imputation techniques for every area of investigation. There are several reasons how the data become missing in scientific research. Deterioration of the technical tool used for data recording, an individual refusing to answer the questionnaire or data entry errors can cause these kinds of problems.

In the literature, there have been plenty of approaches on how to deal with the missing values in research data [5]. The easiest way to handle missing data is working with the cases which are complete, in other words deleting all the cases that have missing values. However, a large amount of valuable information embedded in the data is lost as a result of discarding the missing cases. Therefore, the imputation of missing data plays a very important role in every area of science.

Random Forests (RFs) algorithm is one of the popular supervised learning methods. It creates decision trees on randomly selected data samples, gets a prediction from each tree, and selects the best solution by means of voting for the classification [1]. Moreover, RFs provide variable importance rankings and can handle a variety of data structures.

The purpose of this study is to examine how missing values should be handled when a classification is made with RFs algorithm to selected variables of the most recent Turkey Demographic and Health Survey (2018 TDHS) data. This data has indispensable importance in international comparison because it has an international standard to monitor the developments overtime in Turkey's demography and health.

2. Imputation Methods for Missing Data

In general, imputation is the process of replacing missing values in the handled data set with new estimated values. Based on this definition, various imputation techniques based on different bases have been proposed in the literature and continue to be recommended. Since a single method that provides superiority to each other among all these methods has not yet been revealed, studies in this field continue at a great pace.

Generally, imputation methods are divided into two main categories: statistical and machine learning approaches [6]. Statistical methods include mean or mode imputation. Although these methods are very easy to apply and interpret, they actually cause a big bias problem. However, these methods are still being used to a

greater extent. Machine learning-based methods, on the other hand, take into account the information buried inside the data. K-Nearest Neighbor (KNN) imputation method is one of the most popular methods that is used in the literature. On the fly imputation techniques impute the missing values while simultaneously growing the random forest trees.

Due to the importance of imputation methods and being an open research area for every field of science, we intend to investigate the imputation methods listed below for the classification by RFs algorithm in 2018 TDHS data. The methods used in this study for the missing value imputation are:

- Pre-Imputation methods
  - Mode imputation
  - Hot-Deck imputation
- Machine learning imputation
  - K-Nearest Neighbor
- On-the-fly imputation
  - RFs algorithm

Mode imputation means replacing the missing values of the categorical variable with its most common value in the data set. Hot-Deck imputation is a method that aims to find and assign the most similar value for the missing case. In this method, all the observations in the data set are divided into groups that have similar characteristics. Then a random observation is chosen from the group containing that data to be assigned. The value in this selected observation is assigned to the missing data. These two methods are also placed under the title of single imputation techniques. K-nearest neighbor imputation is based on the K-nearest algorithm. Distances of each observation in the data are calculated, then the nearest observation value is assigned to the missing value case. In this study, the value of k expressing the number of neighbors is taken as 3-5 and 7. For the imputation on the fly data is imputed simultaneously growing the random forest. Therefore, each case with missing values is distributed over the tree nodes with specific weights [2,3,8].

In this study, a classification prediction is made by RFs algorithm for 2018 TDHS data. Thereby, the imputation methods explained above are implemented and compared to each other. For each imputation method, the performance metrics of binary classifiers given below are calculated.

Table 1 Evaluation of binary classifiers

		Predicted Class		
		Positive	**Negative**	
Actual Class	**Positive**	True Positive (TP)	False Negative (FN)	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP)	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

The classification performance values are given in Table 1. Sensitivity, specificity, and accuracy values are calculated. Furthermore, F1 and Kappa values given in the following equations are also regarded for the performance of classification.

$$F1\ Score = 2 * \frac{Precision * Sensitivity}{Precision + Sensitivity} \quad (1)$$

$$Kappa = \frac{total\ accuracy - random\ accuracy}{1 - random\ accuracy} \quad (2)$$

3. Data

For our study, 2018 TDHS data which includes missing values are used for the classification prediction by RFs. The significance of this data is not only monitoring the developments overtime in Turkey's health and demography, but also comparing Turkey's situation with the other countries of the world. Moreover, these nation-wide demography surveys have been conducted by Hacettepe University Population Studies every five years since 1968. The variables taken into account in 2018 TDHS data and their categories are detailed in Table 2.

Table 2 2018 TDHS Data: Selected Variables (Women of Age 15-49)

Variable	Explanation	Categories
V013	Age-5-years group	15-19/20-24/25-29/30-34/35-39/40-44/45-49
V024	Region	Central/East/North/South/West
V025	Type of place of residence	Rural/Urban
V190	Wealth index	Middle/Poorer/Poorest/Richer/Richest
V501	Current Marital Status	Married/Widowed/Divorced/Not living together
S118	Mother literate	No / Yes
S119	Mother ever attended to school	No Educ-primary incomplete / primary complete/ secondary complete or high education
S120	Father literate	Yes / No
S121	Father ever attended to school	No Educ-primary incomplete / primary complete/ secondary complete or high education
S122	Parents related	Yes / No
SEDUC	Educational categories	No Educ.-Primary incomplete / First level primary / Second level primary / High school and higher
CEB	Total children ever born	0 / 1 / 2 / 3 / 4 or over
mig	Number of migrations since childhood	Never migrated since childhood / 1 / 2 / 3 or over
Sect2	Employment status by sector	Nonemployed last 12 months / Agriculture / Non-agriculture
Rejectall	**Wife beating not justified: all cases** *Cases:* *If she neglects the children* *If she answers him back* *If she refuses to have sex with him* *If she burns the food*	**No / Yes**

In this data set, we intend to define the profile of attitudes towards violence among the selected variables for all answers of the survey. Therefore, we aim to estimate the "Rejectall" variable by using RFs algorithm. In other words, we want to estimate women's attitude towards violence against women in terms of intolerance, based on her demographic characteristics. R programming language (Version 4.0.2) is used. In this data, we have 7,343 observations over 15 variables. Unfortunately, as encountered in experiments and measurements carried out in almost all fields of science, there are missing observations in this data set, too. Figure 1 expresses the variables in the data and the number of missing value observations these variables have.

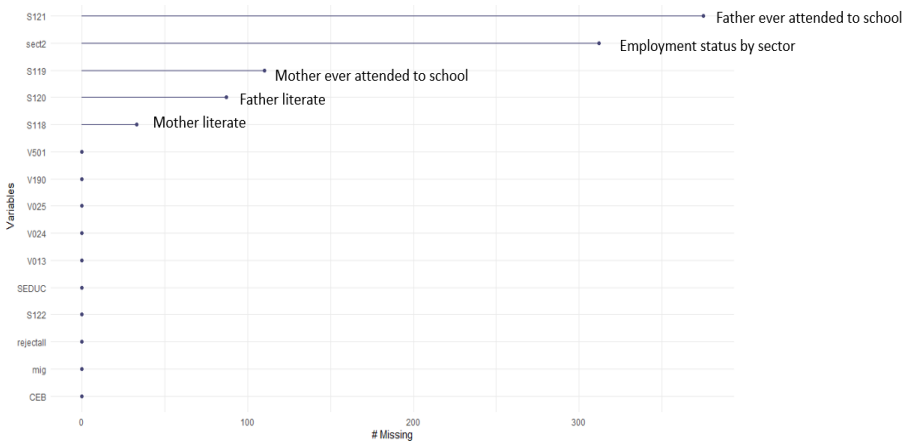


Fig.1. Missing data cases and variables

According to Figure 1, "Father ever attended to school" variable has the most missing values. This variable is followed by "Employment status by sector", "Mother ever attended to school", "Father literate", and "Mother literate" variables, respectively. In order to see the patterns of missing values or the combination of the missing cases amongst variables, Figure 2 is given below.

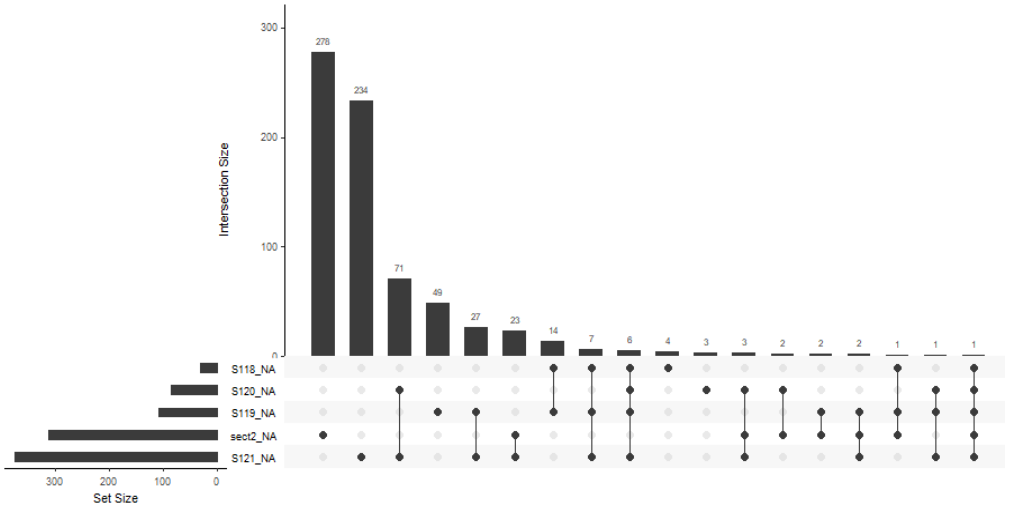


Fig.2. Missing data graph of observations and variables

It is seen from Figure 2 that, there are 71 cases where both "father literate" and "father ever attend to school" variables have missing values in the data set. Also, there is only one case that all the variables have missing values over this data. In order to see the all cells of the data matrix by rectangles, Figure 3 is drawn.

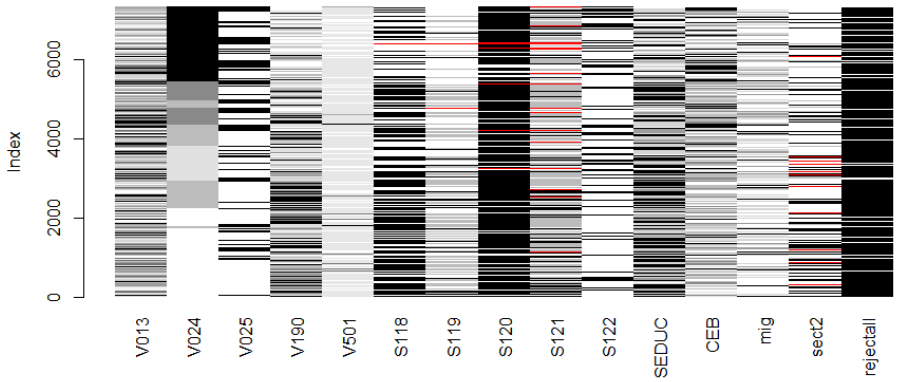


Fig.3. Missing Data Cases of 2018 TDHS Data

Figure 3 indicates how the missing data is structurally distributed among the data set. Observed data is shown in a gray-black color scheme (the darkest the color the higher the value) while missing values are highlighted in red.

After the investigation of missing values in the data, we apply the imputation methods and try to estimate the violent opinion. Firstly, we split the data into two parts. The first part consisting the 20% of the data is the test data. Then the second part consisting of 80% of the data is the train data. Here only the train data involves all the missing cases. Secondly, we apply the imputation methods we described in Section 2 on the train data then make a prediction using RFs algorithm. After that, we calculate the RFs algorithm classification performances using the test data. We make it by considering the tenfold cross-validation case. Hence, we repeated this procedure ten times. We take into account the classification performance metrics given in Section 2. The results are given in Table 3 below.

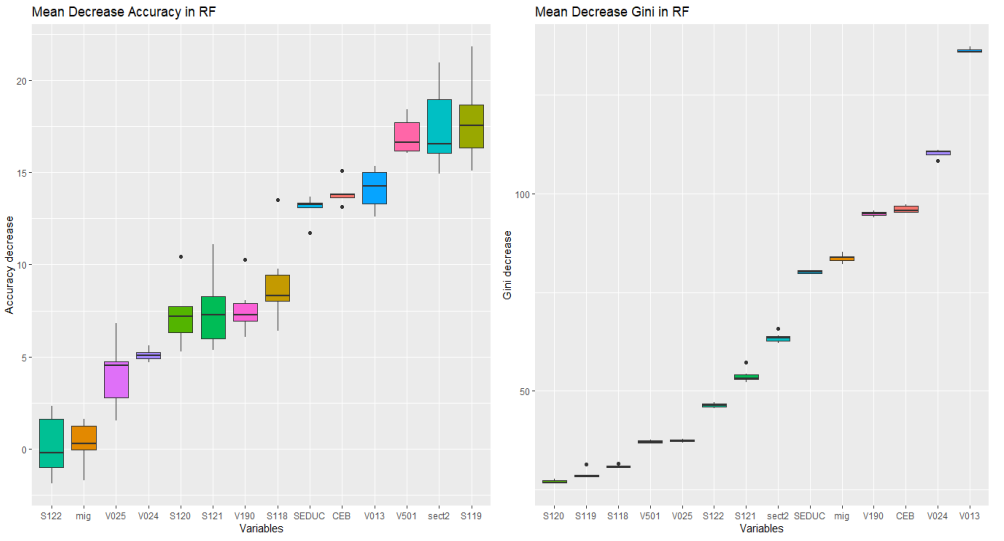
Table 3 Results of imputation methods

Imputation Method	**Accuracy**	**Sensitivity**	**Specificity**	**Kappa**	**F_1 Score**
Mode	0.88930518	0.04728653	0.98692626	0.05419803	0.08117830
Hot Deck	0.88923706	0.04487947	0.98715546	0.05092346	0.07732831
KNN-3	0.88903270	0.04439663	0.98693410	0.04972931	0.07643431
KNN-5	0.88910082	0.04259198	0.98723610	0.04748147	0.07357689
KNN-7	0.88950954	**0.04786315**	0.98708315	**0.05543079**	**0.08229096**
Random Forest	**0.88971390**	0.04708089	**0.98737899**	0.05471897	0.08108936

It is seen that all the values are very close to each other. The highest accuracy and sensitivity values are calculated when using the on-the-fly imputation method with RFs algorithm. On the other hand, the highest sensitivity, Kappa, and F_1 measures are obtained when the KNN - 7 imputation method is used for the imputation method on estimating violence attitudes by the RFs algorithm. As a result, it is proposed to use KNN with 7 neighbor imputation methods for handling these values in 2018 TDHS data.

It is also mentioned in Section 1 that RFs algorithm helps us to reveal the important independent variables while making the classification. Therefore, the mean decrease accuracy and the mean decrease Gini values are also calculated in our application in order to reveal the important variables. We calculate these values while

making predictions for the “Wife beating not justified: all cases” dependent variable with RFs algorithm considering ten-fold cross-validation. Figure 4 gives the box-plots of these values for each variable.



(a) Mean decrease accuracy (b) Mean decrease Gini

Fig.4. Variable importance in RFs

We take the intersection of these variables that are important for these two values. The variables “Age-5-years group”, “Educational categories” and “total children ever born” are found to be the most important and effective variables in order to predict the variable "wife-beating not justified: all cases" with RFs algorithm.

4. Conclusion and Future Works

In this study, the missing data structure is examined in selected variables of 2018 TDHS data for the first time. Popular imputation methods are applied in order to handle the missing values in 2018 TDHS data. Random Forests algorithm is applied for estimating the attitudes of women towards domestic violence against women. It has been revealed that instead of deleting the missing data, the KNN imputation method with 7 neighbours should be used when predicting the attitude towards domestic violence among women with the RFs algorithm. It has been revealed that a woman's age, educational level, and the number of children are important variables in determining the intolerant attitude towards violence against women.

It is deduced that missing data methods give better results when handled by the missingness mechanism [5,8,9]. Hence, considering 2018 TDHS data by addressing the internal structure of the missing mechanism is a potential future work. Because the missing data imputation methods differ from each other under several missing data mechanism circumstances. Last but not least, we intend to create a synthetic data that has the same characteristics as 2018 TDHS data. Missing values are planned to be created randomly on this synthetic data. Also a deeply investigation for the classification performances of RFs algorithm while considering this missingness structure of the data is planned.

References

[1] Hastie T, Tibshirani R, Friedman J . *The Elements of Statistical Learning Data Mining Inference and Prediction.* Springer; 2009.

[2] Hapfelmeier, A.. *Analysis of missing data with random forests*.2012 (Doctoral dissertation, lmu).

[3] Tang, F, Ishwaran, H. Random forest missing data algorithms. *Stat. Anal. Data Min*.2017; 10, 363–377.

[4] R Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL http://www.R-project.org/.
[5] Gangadharan N., Turner R., Field R., Oliver S. G., Slater N., Dikicioglu D. Metaheuristic approaches in biopharmaceutical process development data analysis. *Bioprocess Biosyst. Eng*. 2019; 42, 1399–1408. 10.1007/s00449-019-02147-0
[6] Twala, B., & Cartwright, M. Ensemble missing data techniques for software effort prediction. *Intelligent Data Analysis*, 2010: *14*(3), 299-331.
[7] Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputations be used for handling missing data in randomised clinical trials - a practical guide with flowcharts. *BMC Med Res Methodol*. 2017 Dec 6;17(1):162. doi: 10.1186/s12874-017-0442-1. PMID: 29207961; PMCID: PMC5717805.
[8] Yesilova, A., Kaya, Y., Almali, M. A comparison of hot deck imputation and substitution methods in the estimation of missing data. *Gazi University Journal of Science*, 2011: 24(1), 69–75
[9] Augustsson, M, Güner, L, Zidic, T. Comparing approaches for handling missing values in random forests. 2019. 10.13140/RG.2.2.25683.43041.