COMPARISON BETWEEN IMPUTATION METHOD FOR HANDLING MISSING DATA

AYUNIE BINTI EZADIN
NUR IZZATY BINTI CHUMIN
SITI NUR IZZATULNISA BINTI SALIT

BACHELOR OF SCIENCE (HONS.) STATISTICS
FACULTY OF COMPUTER AND MATHEMATICAL SCIENCES

JANUARY 2021

ABSTRACT

This paper presents imputation method for the National Institute of Diabetes and Digestive and Kidney Diseases data from Arizona, United States. Missing data occurs in this data for five variables which are plasma glucose concentration, diastolic blood pressure, triceps skin fold thickness, serum insulin intake and body mass index (BMI). Missing data leads to problem that can cause bias and invalid conclusions to be made. This research objectives are to improve the data by filling the missing value and to compare which imputation method is better to handle missing value in a data set. In this research, imputation method and evaluation of the performance are applied for this data using Rstudio software. Five imputation methods used in this paper are Mean imputation method, K-Nearest Neighbour (KNN) imputation method, Multiple imputation method, Hot-Deck imputation method and Regression imputation method. The performance of these methods are evaluated using statistical analysis, coefficient of determination (R^2), mean-squared eror (MSE), root of mean square error (RMSE), mean absolute error (MAE), index of agreement (d) and bias (B). Based on the result obtained from this research, it can be concluded that K-Nearest Neighbour imputation method is the best method among the five methods that are applied to handle the missing value. Conclusions are made as K-Nearest Neighbour (KNN) imputation method shows the best performance and has the lowest error value compared to other methods.

ACKNOWLEDGEMENT

IN THE NAME OF ALLAH, THE MOST GRACIOUS, THE MOST MERCIFUL

First and foremost, praises and thanks to the God, the Almighty, for his showers and blessings throughout our journey in completing the research work smoothly and successfully.

We would like to express our very great appreciation and sincere gratitude to our research supervisor, Madam Zaitul Anna Melisa Md Yasin, lecturer of University Teknologi MARA Negeri Sembilan, campus Seremban 3 for providing us this research topic and for giving us opportunity to do this research. We feel blessed as Madam Zaitul Anna Melisa chose our group to do this research and giving us chance to explore and learn new thing. She provides the guidance about the overall picture of this research and it really helps us in understanding and completing the research. We also would like to thank her for pointing out our mistakes, giving feedback on what we have done in this research and monitoring our progress weekly to prevent any difficulties towards the end of the due for the submission. Our supervisor has been such a great help to us in our journey and we cannot complete this research smoothly without her guide. We are also happy to work under our supervisor as she is someone that is approachable, and we are comfortable to work on this research with her.

Next, we would like to thank our lecturer who taught us code subject STA650 for last semester and teach us code subject MSP660 for this semester, Dr. Nurul Nisa' Khairol Azmi. She has been teaching us for 2 semester on how to complete our proposal and our paper work for final year project. Dr Nurul Nisa' has been guiding us on our final year project from chapter 1 to chapter 5, which includes introduction, literature review, methodology, results and discussion, and conclusion and future recommendation. We are grateful for her guidance as we can complete our paper work successfully. We also want to thank her for the template of latex that has been provided and teaching us on how to do the coding of the latex.

We are also extremely grateful to our parents for their prayers and love towards us. We are thankful for their sacrifice and effort for providing us the needs to further our study. Our family understanding to our busy schedule also have been a help for us in completing this research. They have always been supporting us physically and mentally in completing this research and always encourage us to do our best for this research. Lastly, our thank goes to all people who have help and supported us in this research directly or indirectly.

TABLE OF CONTENTS