UTM Computing Proceedings

Innovations in Computing Technology and Applications

Paper Title (use style: *paper title*)

Subtitle as needed *(paper subtitle)*

### Ji Tong Lin

### School of Computing, Faculty of Engineering

### Universiti Teknologi Malaysia (UTM)

### Selangor, klang

### jitonglin1998@gmail.com

**Abstract**—This electronic document is a “live” template. The various components of your paper [title, text, heads, etc.] are already defined on the style sheet, as illustrated by the portions given in this document. (Abstract)

Obesity, Machine Learning, Decision Tree, Random Forest, Support Vector Machine (SVM)

# Introduction

This template, modified in MS Word 2003 and saved as “Word 97-2003 & 6.0/95 – RTF” for the PC, provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. All standard paper components have been specified for three reasons: (1) ease of use when formatting individual papers, (2) automatic compliance to electronic requirements that facilitate the concurrent or later production of electronic products, and (3) conformity of style throughout a conference proceedings. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

# Literature Review

## Selecting a Template (Heading 2)

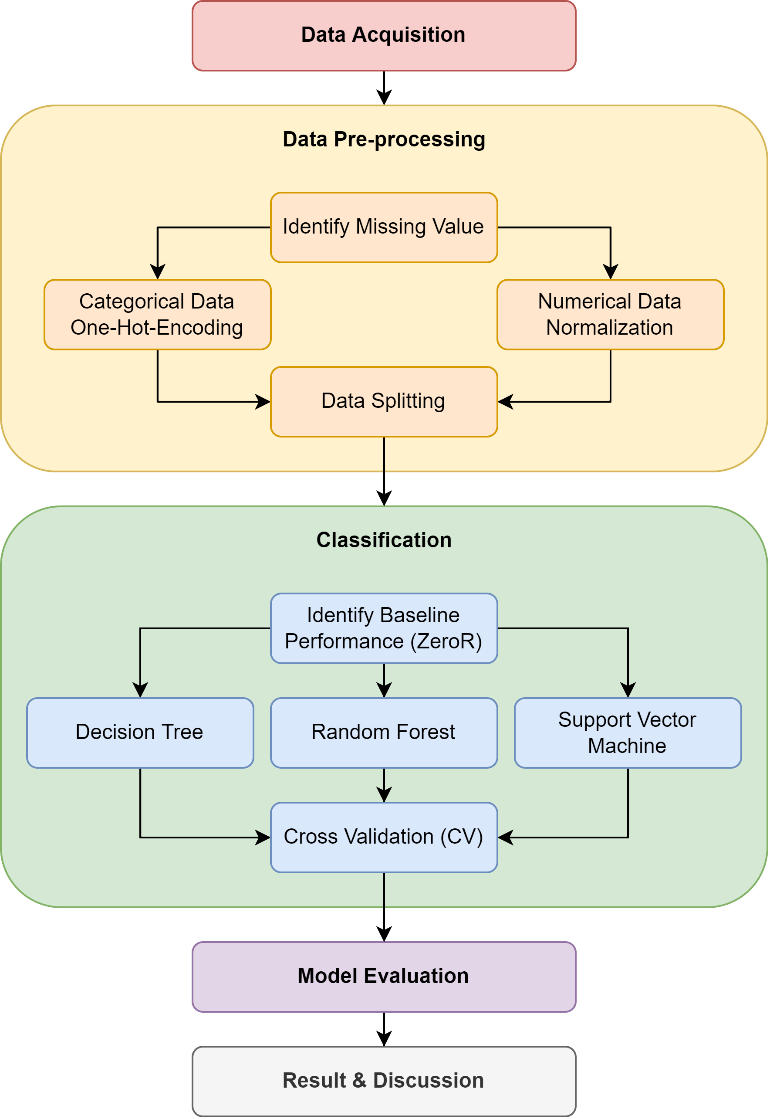
First, confirm that you have the correct template for your paper size. This template has been tailored for output on the A4 paper size. If you are using US letter-sized paper, please close this file and download the file for “MSW\_USltr\_format”.

## Maintaining the Integrity of the Specifications

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

# Methodology

The experimental workflow is depicted in Figure X. The obesity data is first acquired from [], followed by data pre-processing steps of identifying missing value, categorical data one-hot-encoding, and numerical data normalization. The cleaned Obesity dataset undergoes train-test split with ratio 70:30. In classification, the ZeroR algorithm is first used to identify the baseline performance of the models by using the test dataset. Then, 3 different classification models are built, namely decision tree (DT), random forest (RF), and support vector machine (SVM), to obtain 3 different classification scenarios. Cross validation (CV) is done on all 3 models to obtain a more generalizable result. Both the classification result on the test set and the CV result are used for model evaluation. Finally, the result obtained are discussed with justifications.



1. The experimental workflow

## Data Acquisition

The study starts with the acquisition of the obesity data from []. The acquired obesity dataset consists of 2111 samples with 17 features including the target feature "NObeyesdad". According to [], the obesity dataset is cleaned without any missing values. Besides that, SMOTE was also performed on the dataset by the authors to handle the class imbalance issue.

## Data Pre-processing

The acquired obesity data undergoes data-preprocessing. Even though the dataset is claimed to be cleaned without missing values, the study still proceeds the pre-processing step with identifying missing values. Indeed, it was found that there are no missing values in the obesity dataset.

The obesity data contains both the numerical data and categorical data, which comprises of 8 features for both type of data, excluding the target feature. Both data types undergo different pre-processing steps. For the categorical data, one-hot-encoding (OHE) is used to encode each unique value in the categorical features into new features. The original categorical feature is discarded and n number of new features will be introduced if the column contains n number of unique values. For instance, the "gender" feature has 2 unique values "male" and "female", therefore 2 new features are introduced and the original "gender" feature is discarded. Table X summarizes the number of unique values in each categorical feature (excluding target feature) and the total number of the categorical features after OHE. The obesity dataset now has a total of 31 features after the OHE on the categorical features (including the target feature).

1. One-Hot-Encoding on the Categorical Features

| **No.** | Feature Name | No. of Unique Value |
| --- | --- | --- |
| 1 | Gender | 2 |
| 2 | family\_history\_with\_overweight | 2 |
| 3 | FAVC | 2 |
| 4 | CAEC | 4 |
| 5 | SMOKE | 2 |
| 6 | SCC | 2 |
| 7 | CALC | 4 |
| 8 | MTRANS | 5 |
| **Total categorical features after OHE** | | **23** |

The target feature "NObeyesdad" is a multi-class categorical feature with 7 class labels. These class labels are also encoded into an ordinal scale. The result of encoding for the target feature is summarized in Table X.

1. Encoding of the Target Feature

| **No.** | Label | Encoded Label |
| --- | --- | --- |
| 1 | Insufficient\_Weight | 0 |
| 2 | Normal\_Weight | 1 |
| 3 | Obesity\_Type\_I | 2 |
| 4 | Obesity\_Type\_II | 3 |
| 5 | Obesity\_Type\_III | 4 |
| 6 | Overweight\_Level\_I | 5 |
| 7 | Overweight\_Level\_II | 6 |
| **Total class label** | | **7** |

The numerical data on the other hand undergo data normalization. The values are scaled within 0 to 1.

After OHE on the categorical data and normalization on the numerical data, the obesity data is split into train and test set with a ratio of 70:30 with stratification. 70% of the data will be used for the training of the selected classification models while the remaining 30% will be used to obtain the final performance of the models. The class label distribution for both the train and test set is tabulated in Table X.

1. Class Distribution of the Train and Test Set

| **Class Label** | Dataset | | |
| --- | --- | --- | --- |
| Train Set | Test Set | Total Samples |
| 0 | 190 | 82 | 272 |
| 1 | 201 | 86 | 287 |
| 2 | 245 | 106 | 351 |
| 3 | 208 | 89 | 297 |
| 4 | 227 | 97 | 324 |
| 5 | 203 | 87 | 290 |
| 6 | 203 | 87 | 290 |
| **Total Samples** | 1477 | 634 | 2111 |

## Classification

The baseline performance of the models is first obtained by using the ZeroR algorithm. The ZeroR algorithm determines the baseline performance of the models by predicting the class label while ignoring the features. By predicting the class label with highest frequency, the baseline accuracy can be obtained via ZeroR.

Next, the study proceeds with the classification of the obesity data using 3 different classification model, namely decision tree (DT), random forest (RF) and support vector machine (SVM). The implementation of the classification models is by using the SciKit Learn (sklearn) package in python. The hyperparameters used for the development of the 3 classification models are kept as default. The classification models are built using the train set.

## Model Evaluation

To evaluate the model, the test set is used to obtain he classification result from the 3 models. The prediction of the class labels is compared to the actual class labels and the confusion matrix is plotted for each model. From the confusion matrix, several metrics such as the accuracy, precision, recall and F1 score are obtained.

To obtain a more generalizable result, cross validation (CV) is done on the 3 models. The setting used for the CV is repeated stratified K fold with 10 repeats and 10 folds.

## Result & Discussion

The results obtained from the classification of the obesity dataset using the 3 models are recorded. Discussion on the result is done by comparing the performance of the 3 models using the calculated metrics and the CV result.

# Findings

The results and findings of the study is compiled in this chapter. Firstly, the baseline performance for the 3 selected models is determined using the ZeroR algorithm. Then, the 3 classification models are developed separately using the same train set. The test set is used to obtain the performance of the 3 models. Finally, the CV result for each model is obtained.

## Baseline Performance (ZeroR)

To obtain the highest baseline accuracy, the ZeroR algorithm predicts the majority class label, which is "Obesity\_Type\_I" or "2" at 106 samples in the test set. The baseline accuracy obtained is 0.1672 or 16.72%.

## Decision Tree (DT)

## Random Forest (RF)

## Support Vector Machine (SVM)

# Discussion

# Conclusion

## Authors and Affiliations

The template is designed so that author affiliations are not repeated each time for multiple authors of the same affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization). This template was designed for two affiliations.

### For author/s of only one affiliation (Heading 3): To change the default, adjust the template as follows.

#### Selection (Heading 4): Highlight all author and affiliation lines.

#### Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select “1 Column” from the selection palette.

#### Deletion: Delete the author and affiliation lines for the second affiliation.

#### For author/s of more than two affiliations: To change the default, adjust the template as follows.

#### Selection: Highlight all author and affiliation lines.

#### Change number of columns: Select the “Columns” icon from the MS Word Standard toolbar and then select “1 Column” from the selection palette.

#### Highlight author and affiliation lines of affiliation 1 and copy this selection.

#### Formatting: Insert one hard return immediately after the last character of the last affiliation line. Then paste down the copy of affiliation 1. Repeat as necessary for each additional affiliation.

#### Reassign number of columns: Place your cursor to the right of the last character of the last affiliation line of an even numbered affiliation (e.g., if there are five affiliations, place your cursor at end of fourth affiliation). Drag the cursor up to highlight all of the above author and affiliation lines. Go to Column icon and select “2 Columns”. If you have an odd number of affiliations, the final affiliation will be centered on the page; all previous will be in two columns.

## Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1”, “Heading 2”, “Heading 3”, and “Heading 4” are prescribed.

## Figures and Tables

### Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

1. Table Type Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

a. Sample of a Table footnote. (Table footnote)

1. Example of a figure caption. (figure caption)

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord “Format” pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

##### Acknowledgment

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression, “One of us (R. B. G.) thanks . . .” Instead, try “R. B. G. thanks”. Put sponsor acknowledgments in the unnum-bered footnote on the first page.

##### References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first . . .”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. (references)
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
3. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
4. K. Elissa, “Title of paper if known,” unpublished.
5. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
6. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
7. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.