UNIVERSITY OF ULSTER

FACULTY OF COMPUTING, ENGINEERING AND BUILT ENVIRONMENT

COURSEWORK SUBMISSION SHEET

This sheet must be completed in full and attached to the front of each item of assessment before submission to Module Coordinator/Instructor, Professor Girijesh Prasad, via Blackboard Learn.

Student's Name ……Tomide Victor Afolabi……

Registration No ……B00874627……..

Course Title ……MSc Data Science…….

Module Code/Title……COM736: Data Validation and Visualisation……

Instructor ……Professor Girijesh Prasad ……

Date Due …… Monday 8th May 2022…….

Submitted work is subject to the following assessment policies:

1. Coursework must be submitted by the specified date.

2. Students may seek prior consent from the Course Director to submit coursework after the official deadline; such requests must be accompanied by a satisfactory explanation, and in the case of illness by a medical certificate.

3. Coursework submitted without consent after the deadline will not normally be accepted and will therefore receive a mark of zero.

I declare that this is all my own work and that any material I have referred to has been accurately referenced. I have read the University's policy on plagiarism and understand the definition of plagiarism. If it is shown that material has been plagiarised, or I have otherwise attempted to obtain an unfair advantage for myself or others, I understand that I may face sanctions in accordance with the policies and procedures of the University. A mark of zero may be awarded and the reason for that mark will be recorded on my file.

Data Validation and Analysis An importance of machine learning using AIBL data as a case study.

Tomide Victor Afolabi
Afolabi-t@ulster.ac.uk
Faculty of Computing, Engineering and the Built Environment
Ulster University
Londonderry, United Kingdom

***Abstract*— The research aims to use data validation principles to categorize ageing data from the Australian Imaging, Biomarker & Lifestyle (AIBL) study. Specifically, the study compares the effectiveness of the data validation techniques on two subsets of AIBL data. The techniques used in this study include min-max normalization, Boruta feature selection method, Mean-feature importance plot, and correlation matrix plot. The study utilized the random forest machine learning method to model the data.**

The results emphasize the importance of each step in the data validation process, as the study analyzes the impact of normalization on AIBL data in both its presence and absence.

Keywords: Data Validation, normalization, Boruta, feature selections, random forest.

I. INTRODUCTION

The rapid growth in information technology and the resulting storage of large amounts of data, commonly known as big data, has led to the urgent need for data science to manage and store this data. Data quality is often poor due to factors such as dirtiness and incompleteness, and adhering to data science principles is necessary to ensure proper analysis and utilization of the data. Extracting meaningful information from data involves several stages, including data collection, data preprocessing, exploratory data analysis, feature engineering, model selection and training, model evaluation, and deployment. Data preprocessing is a critical component of data science that requires significant attention during the initial stages of data validation. One common challenge encountered during data preprocessing is the presence of features with larger value scales relative to other features, which can be attributed to a lack of data normalization or scaling. Normalization is essential to increase the reliability of the data (1) and is an important step in the data validation process.

The importance of data validation (DV) cannot be overstated in the field of data science, as it enables early error detection and minimizes the incorrect and overall cost involved in the analysis process (2). One of the techniques employed in data validation is feature selection, which involves filtering features based on their relevance or importance to the output.

Normalization is a crucial component of data validation and involves transforming the scale of variable values to a common scale. Common normalization techniques include min-max normalization, z-score normalization, and log transformation. Min-max normalization, for instance, involves rescaling data to a particular range such as (0,1) or (-1,1) (3).

A reliable machine learning algorithm is necessary for analyzing, classifying, and forecasting information contained in high-quality data. Machine learning studies have increasingly been applied in the healthcare sector (4), with the goal of improving productivity through early disease detection, among other applications.

Random forest (RF) is an example of a machine learning model that employs an ensemble learning approach to classify or predict an outcome. This technique involves combining independent trees, with each tree relying on an independent random sample vector with equal distribution across all trees (5).

The Boruta Algorithm is another feature selection method that depends on the feature importance score of the random forest model. This algorithm involves an embedded wrapper that is built around a random forest and involves the addition of "shadow attributes" that reveal misleading or correlated features.

The study is aimed to compare the effectiveness of data validation techniques in analyzing the Australian Imaging, Biomarker & Lifestyle (AIBL) ageing data. The study involved performing data preprocessing, feature selection, and other validation techniques before modelling the data. The primary focus of the comparative analysis was to evaluate the impact of data validation on the analysis of the AIBL data.

Machine learning is of three types. Supervised, Unsupervised and Reinforcement machine learning. The study uses supervised ML for binary classification.

1. Supervised Learning (SL)

Supervised Learning (SL) is a type of Machine Learning method that involves using labelled input and output data to create a model. The model is then used to predict the output for new input data based on the input's characteristics. To train the model, the data is divided into training, validation, and testing sets. Some Machine Learning algorithms, such as Random Forest, perform both training and validation processes. SL can be divided into two types: regression and classification.

2. Binary classification (BC)

Binary classification is a method for predicting the outcome of two classes, where the target variable diagnosis will always be either 0 (healthy control) or 1 (non-healthy control). It falls under the category of supervised learning classification. BC can be applied to different types of data such as numerical, categorical, and continuous data types. However, in the case of multivariate data types, caution should be taken, and data transformation, such as hot encoding, can be employed to convert data types.

II. METHODS AND MATERIALS

The study utilized various methods including data collection, merging and cleaning, modeling, validation and evaluation of AIBL data. Outliers in Figure 1 were identified using base graphviz box plots and then replaced by the mean. Visualization tools were also used to determine if data transformation through lambda or log was necessary. Boruta Algorithm was employed for analyzing the AIBL data, and the study followed a three-phase experimental procedure: Data Preparation, Feature Selection and Data Preprocessing and Data Modeling and Validation

The AIBL data were split into normalized (Min-max) and unnormalized subsets, and the three phases were applied to both. The results of the two subsets were compared, and conclusions were drawn.

The study utilized the Random forest ML technique to perform binary classification on the two datasets. The objective was to classify Healthy Control, Alzheimer's Disease (AD), and Mild Cognitive Impairment (MCI) cases into distinct Diagnosis categories, namely Healthy Control as (0) and Non-Healthy Control as (1). The following procedures were applied:

A. Data Description

The AIBL dataset is a diverse and large dataset consisting of data from participants aged 55 to 96, which is available for research and educational purposes. The dataset has played a crucial role in the development of machine learning models for identifying and diagnosing cognitive impairment diseases. The data was collected and stored in a CSV file and initially contained 1688 observations and 36 features. After preprocessing, the dataset was reduced to 862 observations and 34 features. Further data preparation resulted in the reduction of the dataset to 862 observations and 31 features. Feature selection was then applied to arrive at 8 features and 862 observations for the normalized AIBL data, and 7 features with 862 observations for the unnormalized dataset. Data balancing was carried out using Synthetic Minority Oversampling Technique (SMOTE), which resulted in 780 observations from the initial 603 observations for the training set, and 7 or 8 features were obtained.

B. Experimental Procedures

The study conducted a comparison and analysis of the Random Forest classification accuracy between two subsets of the AIBL dataset using the following procedures:

Phase 1 - Data preparation: In this stage, the data were merged, and irrelevant and duplicate features such as RID, SITEID, VISCODE, and Exam test date were removed. The dataset was checked for missing values, and any missing data was corrected. Variables and values were adjusted, such as correcting Age values represented as "/1956" and converting them to the actual age. Negative values were replaced with the mean value of the corresponding column. An additional dataset was saved for normalization at the end of this phase.

Second Phase - Feature Selection and Data Preprocessing

The second phase of the analysis involved selecting relevant features and preparing the data for modelling. The first step in this phase was to visualize the data to obtain graphical information such as identifying outliers, non-normalized distributions and imbalanced data. The target variable was logically formed to contain two classes of values, where Healthy Control was represented by zero while Non-Healthy Control was represented by one. To check for normality, Shapiro-Wilk's normality test was employed to test for the p-value. A histogram was used to check for normal distribution, while a box-cox plot was used to find lambda for transformation. The normalization dataset was subjected to min-max normalization.

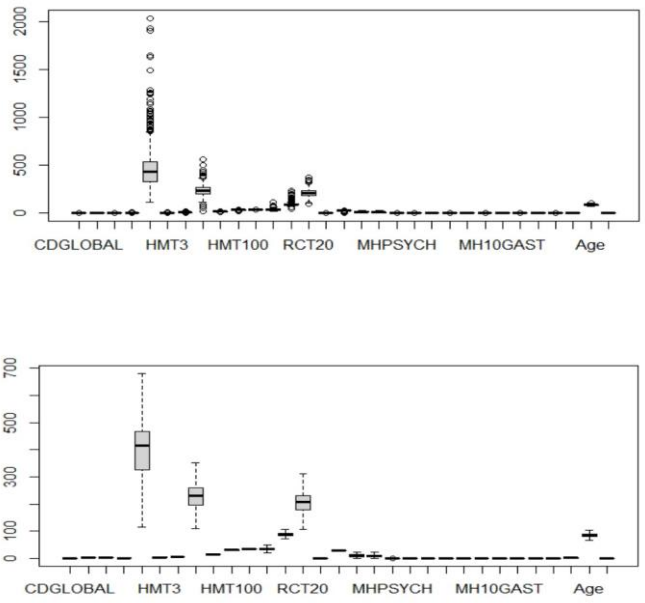
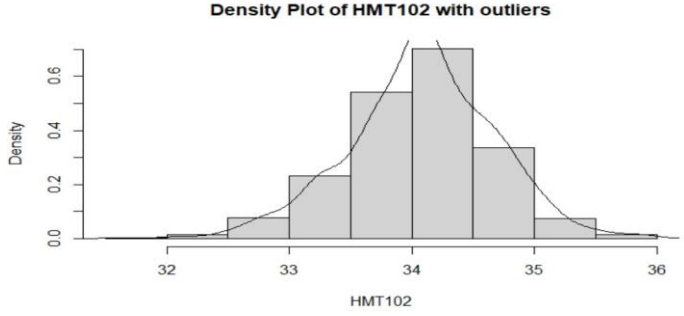


Figure 1. Boxplot for Outliers



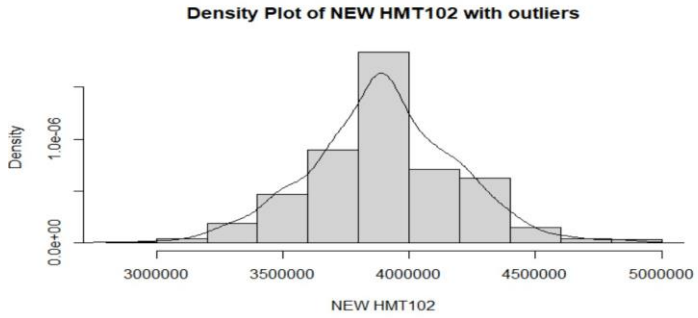


Figure 2. Histogram for HMT 102 transformation

To select the most important features, the study utilized the Mean variable importance and the Boruta algorithm, resulting in the selection of 6 explanatory variables for the unnormalized dataset and 7 explanatory variables for the normalized dataset. These findings were illustrated in Figures 3 and 4. The data modelling process was applied to the training set, which contained 7 or 8 features from the two datasets, while the testing set was used for evaluation.

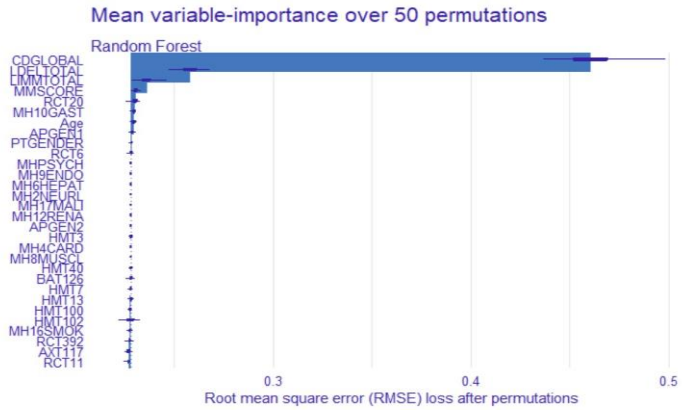


Figure 3. Mean-variable importance plot

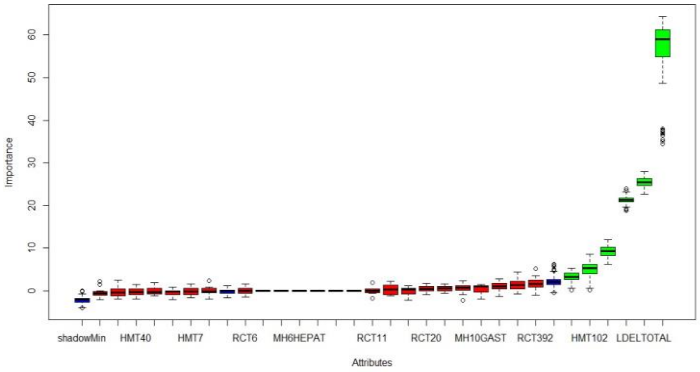


Figure 4. Boruta algorithm feature selection plot

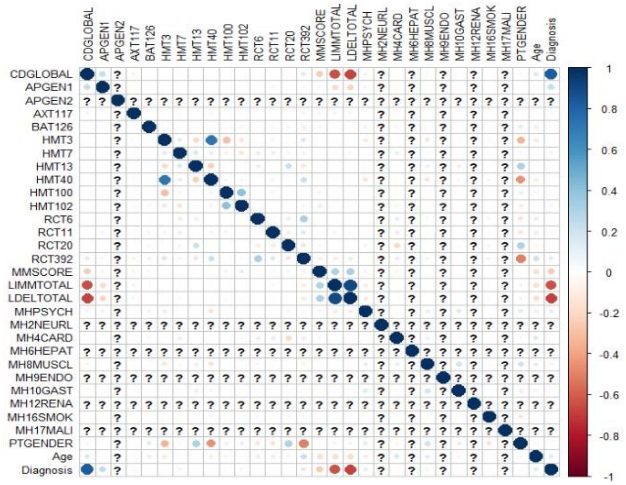


Figure 5. Correlation matrix of 31 variables

C. Evaluation Metrics

• Confusing Matrix: also called a matching matrix in unsupervised learning, is a tabular representation of a classification problem of the predicted and actual outcomes. It renders a summary visualization of the performance of the classification ML model by showing the different prediction outcomes (7).

• Classification Accuracy: is the proportional rate of correct prediction. It is calculated by equation 1

$$ACC = TR + TN/TP+TN+FP+FN \quad (1)$$

• True positive (TP): displays the outcome value of features that truly represent a condition. In this case, TP is a percentage of people who truly are HC (0). And it is calculated through a true positive rate or recall. The opposite of TP is False positive (FP), which is an error called type 1 error where the predicted outcome value ought to represent a true condition, but represent a false condition,

• True negative (TN): shows the result values of features that truly demonstrate the absence of a condition. It is a percentage of people who are truly Non-HC (1). To calculate it, TNR or specificity is used: An error attached to TN is False negative (FN) also called type 2 error.

• AUR0C (Area Under the Receiver Operating Characteristic): is a metric that reveals the performance of a model that correctly ranks the observations in the data using the ROC curve. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) (8). A higher AUROC value indicates that the classifier is better at accurately ranking the observations, while a lower AUROC value indicates a poorer classifier.

• Boruta Feature importance plot: a graph that ranks the features according to their importance towards target variables. It is based on the idea that by randomly adding new variables to the existing ones and getting their results, one can identify variables that are really important to the target variable. It has a wrapper based on the random forest technique (9).

III Evaluation

The dataset used in the study was divided into normalized and unnormalized subsets, each containing 862 observations. The unnormalized dataset had 7 features while the normalized dataset had 8 features. A split ratio of 7:3 was used to create a training set of 603 observations and a testing set of 259 observations. The training set was balanced using Synthetic Minority Oversampling Technique (SMOTE) to obtain 780 observations for modelling and evaluation. The performance of the models was then compared between the normalized and unnormalized datasets.

AIBL Data	ACC	AUROC	Sensitivity	Specificity
Normalized	0.9421	0.9398	0.9454	0.9342
Unnormalized	0.9344	0.9266	0.9454	0.9079

Table 1. Performance results of normalized data and unnormalized dataset

Table 1 displays the effect of a data validation concept on the performance of the datasets. Upon applying various data validation concepts, except normalization, the normalized dataset outperformed the unnormalized dataset, with an ACC of 0.9421, AUROC of 0.9398, and Specificity of 0.9342. Although both datasets had a sensitivity of 0.9454, the unnormalized dataset had a lower specificity of 0.9079 compared to the normalized data.

IV CONCLUSION

The correlation plot revealed that certain variables, such as CDGLOBAL, were highly correlated with other variables and thus needed to be dropped. However, this could result in the loss of information when working with the 7 or 8 selected features out of the initial 31 features suggested by the Boruta feature selection method.

The study concludes that the normalization step is a crucial aspect of data validation and justifies the high accuracy achieved in classification using ML techniques. It also highlights the importance of each step in the data validation process.

V LIMITATIONS

After working with an unnormalized dataset, the study found that data transformation was effective in improving data values before handling outliers, but it did not yield the desired results for data without outliers. This suggests that replacing outliers with the mean may not be the best approach for data transformation.

VI RECOMMENDATIONS

Examining other models in every stage can provide more valuable insights into the impact of normalization and other data validation concepts. Therefore, it would be interesting to compare the performance of various models throughout the entire process.

It is important to emphasize the importance of collecting and validating data. Following data validation rules can significantly reduce the amount of dirty data. While it may be challenging to completely clean data, efforts should be made to make data reliable.

VII REFERENCES

[1] H. Huang, L. Qin 2018, Empirical evaluation of data normalization methods for molecular classification. [online]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5899419/

[2] N. Ahmad, K. Dias 2021, Validating "Data Validation". [online]. Validating "Data Validation" | IEEE Journals & Magazine | IEEE Xplore

[3] G. Prasad et al, 2022, Multiple Cost Optimisation for Alzheimer's Disease Diagnosis [online]. https://www.medrxiv.org/content/10.1101/2022.04.10.22273666v1.full.pdf+html

[4] Peshawa J. Muhammad A, 2022, Investigating the Impact of Min-Max Data Normalization on the Regression Performance of K-Nearest Neighbor with Different Similarity Measurements. [online] View of Investigating the Impact of Min-Max Data Normalization on the Regression Performance of K-Nearest Neighbor with Different Similarity Measurements (koyauniversity.org)

[5] L. Breiman, 2001, RANDOM FORESTS. [online]. randomforest2001.pdf (berkeley.edu)

[6] R. Trifonov, 2017. BINARY CLASSIFICATION ALGORITHMS. [online]. 10934.pdf (journalijdr.com)

[7] C. Sweeney et al, 2022. How Machine Learning Classification Accuracy Changes in a Happiness Dataset with Different Demographic Groups. [online].

[8] K. Tilaki, 2013, Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation, [online]. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation - PMC (nih.gov)

[9] Miron B. Kursa, Witold R. Rudnick, 2010. Feature Selection with the Boruta Package. [online]. (PDF) Feature Selection with Boruta Package (researchgate.net)