2019-1

An Investigation of Three Subjective Rating Scales of Mental Workload in Third Level Education

Nha Vu Thanh Nguyen
Technological University Dublin

An Investigation of Three Subjective Rating Scales of Mental Workload in Third Level Education

Nha Vu Thanh Nguyen

A dissertation submitted in partial fulfilment of the requirements of
Dublin Institute of Technology for the degree of
M.Sc. in Computing (Data Analytics)

2019

Declaration

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Data Analytics), is entirely my work and has not been taken from the work of others save and to the extent that such action has been cited and acknowledged within the text of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

The work reported in this dissertation conforms to the principles and requirements of the Institutes guidelines for ethics in research.

Signed: **Nha Vu Thanh Nguyen**

Date: February 8, 2019

Abstract

Mental Workload assessment in educational settings is still recognized as an open research problem. Although its application is useful for instructional design, it is still unclear how it can be formally shaped and which factors compose it. This paper is aimed at investigating a set of features believed to shape the construct of mental workload and aggregated together in models trained with supervised machine learning techniques. In detail, multiple linear regression and decision trees have been chosen for training models with features extracted respectively from the NASA Task Load Index and the Workload Profile, well-known self-reporting instruments for assessing mental workload. Additionally, a third feature set was formed as a combination of the two aforementioned feature sets and a number of other features believed to contribute to mental workload modeling in education. Models were trained with cross-validation due to the limited sample size. On the one hand, results show how the features of the NASA Task Load index are more expressive for a regression problem than the other two feature sets. On the other hand, results show how the newly formed feature set can lead to the development of models of the mental workload with a lower error when compared to models built with the other two feature sets and when employed for a classification task.

Keywords: Mental Workload, Cognitive Load Theory, Education, NASA-TLX, Workload Profile, Decision Trees, Multiple Linear Regression, Regression, Classification, Machine Learning, Modeling

Acknowledgments

I would like mainly to send many thanks to Dr.Luca Longo, Lecturer of Dublin Institute of Technology and a member of the Applied Intelligence Research Center. This work would not have been possible without his support. He has helped me understand my study goals and encourages me to pursue it. He has shown me, by his example, what a good scientist should be.

I want to thank Dr.Deirdre Lawless, Dr.John McAuley, Dr.Brendan Tierney. As my teachers and mentors, they have taught me more than I could ever give them credit for here. The knowledge I learned from them supported me to complete this work.

I am grateful to my Guru and friends all in the Dublin Meditation center with whom I have had the pleasure to work during this time. They taught me a great deal about love in whatever I have enthusiasm for, light in whatever I concentrate, and life in whatever I put efforts.

Nobody has been more important to me in the pursuit of this project than the members of my family. I want to thank my mother whose love and guidance are with me in whatever I pursue. She has provided me extensive personal and professional advice to complete my Dissertation.

Contents

List of Figures

List of Tables

List of Acronyms

Acc	Accuracy
ANOVA	Analysis of Variance
CLT	Cognitive Load Theory
CTML	Cognitive Theory of Multimedia Learning
DT	Decision Tree
EFS	Extended Feature Sets rating scale
GiniClas	Decision Tree GINI Classification
GiniReg	Decision Tree GINI Regression
Inf	Information
LR	Linear Regression
MAE	Mean Absolute Error
MSc	Master of Science
MWL	Mental Workload
NASA, NASA_TLX	NASA rating scale
Pre	Precision
Rec	Recall
RMSE	Root Mean Squared Error
WP	Workload Profile rating scale

Chapter 1

Introduction

1.1 Background

A person affected by Mental Workload is likely to show some psychological symptoms such as emotional stress and inability in achieving goals which is typically felt over a sustained period of time. This is often accompanied by feelings of hopelessness and inadequacy resulting in more errors in task performance and results (Miyake, 2001). Hence, Mental Workload on a student in the third level education can directly impact on the effectiveness and quality of ones learning process (Fredricks, Blumenfeld, & Paris, 2004).

The third level Education System in Ireland comprises of all training after second-level, encompassing higher education in universities and colleges, further education on Post Leaving Certificate and other courses. The degree-awarding authorities can grant awards at all academic levels [1], approved by the Government of Ireland. According to The Higher Education Authority (HEA, 2004), the 35 years from 1965 to 2000 saw the number of students in the third level education grow from 18,200 to almost 120,000 [2]. These rapidly increasing numbers reflect the number of third-level students seeking help with depression, anxiety, relationship problems and academic issues. Today these numbers have reached unprecedented levels. Members of Psychological Counsellors in

[1] https://en.wikipedia.org/wiki/Third-level_education_in_the_Republic_of_Ireland

[2] www.education.ie (A Brief Description Of the Irish Education System)

Higher Education pointed to a 40% increase in demand for counseling over the last 10 years, with waiting lists for counseling services at many colleges [3]. In 2016, the USI Student Dropout Survey found that 61.6% of students experienced burnout while attending third level and 27.6% dropped out due to stress and anxiety [4].

Nowadays, diversity appears in many fields, especially in the third level education. Many students who study together come from a different culture, different background, different mother-tongue, different ages, male and female, and often different in behavior. Considering all of this, it is necessary to value the Mental Workload through the verbal or written feedback of students. Based on self-assessment, the lecturer can predict the Mental Workload on students participated in study activities.

Previous studies have focused mainly on laborer Mental Workload in the ergonomics area of the industry and on working environments in order to improve staff productivity or performance. This study aims to investigate three subjective rating scales of Mental Workload, which is still recognized as an open research problem in higher education.

1.2 Research Project

Many studies nowadays discuss Mental Workload in the office and factory environment. How about Mental Workload in education, especially in the third level education? And the question is: can a mechanism that creates "machine learning" support a lectures strategy?

In this research, Mental Workload is measured through three self-assessment instruments: the NASA Task Load Index, the Workload Profile, well-known self-reporting instruments for assessing mental workload and a third feature set which was formed as a combination of the two above but also considering features believed to contribute to mental workload modeling in Education. The students would carry out self-assessment

[3]https://www.irishtimes.com/news/education/there-is-a-tsunami-of-third-level-students-with-mental-health-problems-1.2924516

[4]https://www.independent.ie/irish-news/education/going-to-college/coping-with-college-37239268.html

before and after their class activities. To obtain a more accurate prediction model, strong correlation variables and good control features will allow for the critical improvements in this research. Measurable attributes need to be included in designing advanced responses to problem case and reference models.

1.3 Research Objectives

Literature review is for stating the concepts and opinions on Mental Workload, which is the indicator equivalent to Cognitive Load Theory, but in education. The design of the experiment is to create a proper model learning environment for three self-assessment scales of Mental Workload. Implementation and execution of the designed solution are required in order to find the optimal model. This is done by algorithms of error-based and information-based data which, once learnt, become critical in forming the data structure of three given data sets. For the evaluation of proposed solutions for models, as a continuous target and a categorical target, are two combined indicators (RMSE, R-squared) and the combination of Accuracy, Precision and Recall, respectively.

1.4 Research Methodologies

Mental Workload is an excellent measure for designing instructional conditions and also for use a relevant indicator in predicting the learning processes. With the clarity of relevant variables and a validated ques

tionnaire, we can adopt two supervised learning algorithms which are good predictors for optimal training of categorical and continuous, namely, Decision Tree and Linear Regression, respectively.

The research methodologies are quantitative, i.e., finding the relationship between student Mental Workload and statistically significant features, learning algorithms, training models, and carrying out hypothesis tests on comparative indicators for the optimal model selection.

1.5 Scope and Limitations

The domain is a supervised process set to identify the correlated variables in the relationship between student Mental Workload and the features affecting instructional conditions.

The scope of the research is limited to one module and the number of classes in four years with only one lecturer. Thus, the sample may not be entirely representative of the whole student population base.

1.6 Document Outline

This document is organized into five chapters:

- Chapter 1 is this introduction to the dissertation which provides the overall context and answers to two important questions needed in order to proceed with the research: who cares? and what for?

- Chapter 2 introduces the theoretical frameworks, including Cognitive Load Theory and how to apply it for education through measuring Mental Workload indirectly by the subjective rating scales. This will present the practical aspects of designing instructional conditions based on individual Mental Workload score. Subsequently, this research will fill the gap in existing research by its findings.

- Chapter 3 describes the research hypothesis and makes an approach in solving the stated problem. It also attempts to clearly explain the steps to collect and prepare the data, in order to proceed with model training and testing. Lastly, selecting the optimal model is based on the best outcome of accuracy and least errors found.

- Chapter 4 focuses on the implementation of the optimal model and the results thereof. Based on this there is a critical discussion about the results and the existing research in similar contexts.

- Finally, chapter 5 concludes the contribution of the research to the existing body of knowledge.

Chapter 2

Literature review and related work

2.1 Cognitive Load Theory

Cognitive load is mainly referred to as human Mental Workload in the field of Ergonomics, used to improve working conditions (Ree et al., 2014; Weigl, Mller, Angerer, & Hoffmann, 2014; Balfe, Crowley, Smith, & Longo, 2017). Regarding that, it becomes increasingly popular as one of the first-hand indicators when taking into account learning interaction (Foo et al., 2013). J. Sweller (1988) discussed learners can absorb and retain information effectively only if it is provided in a way that does not "overload" their mental capacity. In other words, instructional design and materials of an instructor can be used to reduce cognitive load in learners (J. Sweller, 1994). An instructor would play her/his role to help learners become an expert in a given topic. An expert can categorize problems using the capacity of long-term memory; the ability to explain and perform tasks easily.

Cognitive load theory (CLT) is a contemporary educational psychology theory applying cognitive science principles to instructional design (V. M. J. Sweller J. & Paas, 1998). It has been considered among instructional designers for early years to create resources in such a way that encourages the activities of the learners and optimizes their learning performance (Jeng-Chung, 2014). Within CLT, three types of cognitive load have been conceptualized to have an adequate Mental Workload. It would be a balance between the intrinsic difficulties of a task (intrinsic load) (Galy &

Mlan, 2015), the way it is presented (extraneous load) (Blayney, Kalyuga, & Sweller, 2015; Galy & Mlan, 2015) and the amount of effort performed by the learner to integrate the new knowledge into the old one (germane load).

2.1.1 Types of cognitive load theory

Cognitive Load is understood under three types, namely, intrinsic cognitive load, extraneous cognitive load, and germane cognitive load.

- Intrinsic Cognitive Load is the term that was first used in the early 1990s by Chandler and Sweller (1991). It describes how much capacity of the working memory is used by the interactivity of the units of information being processed.

- Extraneous Cognitive Load is the term used to express some type of unnecessary (artificially induced) cognitive load which is attributed to the design of the instructional materials. Chandler and Sweller (1991) introduced this concept of extraneous cognitive load to report the results of experiments. These experiments were conducted to investigate the working memory load, such as instructors presentation, the textbook in its format and the external distractions, the internal emotional concerns, etc. Many of these experiments involved materials demonstrating the split-attention effect. They found that the format of instructional materials either promoted or limited learning.

- Germane Cognitive Load was first described by V. M. J. Sweller J. and Paas (1998). It is known as the processing, construction, and automation of schema. Therefore, a germane load is working memory of learners, which will process new information into advanced and more complex memory storage.

The three types of Cognitive Load mentioned above together have an interactive impact on learners. The more extraneous load, the less room there will be for the germane load. Hence it is necessary to have the instructional materials designed to limit the amount of extraneous load and to facilitate the increase in germane load.

2.1.2 Instructional conditions in the third level education

Third level students, who get involved in class activities will improve in various mental abilities like critical thinking, decision making, memory and analytical skills, etc. However, psychological fear and a sense of inferiority can lead to non-optimal mental workload devoted to the learning task, with higher chances of error, less productivity, and a predominant sense of uneasiness. As a driving factor in class activity, a lecturer needs to recognize and to predict the students problem and to build teaching activities and instructional materials in order to resolve the Mental Workload

In practice, interactive learning is related to instructional strategies, in-class activities, goal setting, and individual personalities (Wei, Chen, & Kinshuk, 2012) (pg 540). A good way to increase learner's interaction is by reducing the human Mental Workload when interacting on workflow. Instead of assessing learning progress, it would be have a greater impact to assess human Mental Workload (MWL), as it is proven to have an early effect on student performance. This study attempts to identify the different variations of student Mental Workload through a variet of teaching methods and lessons in class through self-assessment techniques, such as the NASA Task Load Index (NASA-TLX) and Workload Profile (WP).

Direct instructions

Direct instruction is a general term used for the explicit teaching of a skill-set to students through lectures or demonstrations of source material. It is a teacher-directed method, meaning that the teacher stands in front of a classroom and presents the information, in contrast to exploratory models such as inquiry-based learning. Direct instruction includes tutorials, participatory laboratory classes, discussion, recitation, seminars, workshops, observation, active learning, practice, or internships.

Multimedia learning

As in the context of information explosion, the studying and teaching are always relevant to the internet wholly or partly. A human can access or process only a finite

amount of information at a time, despite the huge amount of information available on the Internet. The human brain does not interpret multimedia instructions made by words, sensory information and pictures in a mutually exclusive way. As a result, the combination of direct instructions one-way to the electronic communication and extended to group activity multiple-way is the adopted way in modern times.

According to Cognitive Theory of Multimedia Learning (CTML), instructional condition is based upon three assumptions: (1) dual-channel, or the auditory and the visual channel (Wong, Castro-Alonso, C., Ayres, & Paas, 2015; Gough Young, Wodehouse, & Sheridan, 2015; Jaewon, Dongsik, & Chungsoo, 2016); (2) limited processing capacity, each channel has a finite capacity (Haji et al., 2016; Lin et al., 2017); and (3) active processing, learning is an active process including selection, filtering, organization and integration of information to prior knowledge (Macken & Ginns, 2014; Blayney et al., 2015; Agostinho et al., 2015) . The expected result is that higher learning outcomes and lower cognitive load (Yung & Paas, 2015); whatsoever the level of instructional guidance needed to match learners' levels of expertise (Yuling, Yuan, Tzu-Chien, & Sweller, 2015; Kalyuga, Chandler, & Sweller, 1998). For such instances, the designed instructional conditions should keep intrinsic load being static (Haji, Rojas, Childs, Ribaupierre, & Dubrowski, 2015), minimizing extraneous load (Jihyun, Dongsik, & Chungsoo, 2014) and promoting germane load (Leahy, Hanham, & Sweller, 2015; Young, Van Merrienboer, Durning, & Ten Cate, 2014).

Social constructivist

Social constructivist emphasizes the importance of culture and context in understanding what occurs in society and in constructing knowledge based on this understanding. Social constructivist approaches can include reciprocal teaching, peer collaboration, cognitive apprenticeships, problem-based instruction, web quests, anchored instruction and other methods that involve learning with others (Kim, 2001).

On the whole, the expected outcome is that learners can transfer learned concepts to a new context. By that way, the thinking of the group as a whole at first, with the objective of processing certain information is aimed at increasing understanding

(Orru, Gobbo, O'Sullivan, & Longo, 2018). As referred to the cognitive-effective theory of learning with media, learner's mood had an effect on germane load, extraneous load, and intrinsic motivation (Liew & Tan, 2016). These three loads when combined into computational methods, in the context of Mental Workload representation and assessment, will quantify the Mental Workload imposed on learners by social teaching activities and instructional material (Galy & Mlan, 2015; Kalyuga & Singh, 2016).

2.2 Mental Workload

In education, the main reason for assessing cognitive load or Mental Workload is to measure the mental cost of performing a learning task with the goal of predicting the learner's performance (Jimenez-Molina, Retamal, & Lira, 2018; Byrne, Tweed, & Halligan, 2014). Cognitive Load Theory (CLT) in the context of instructional design theory is one of the important indicators to measure. It not only works on the design phase but also becomes a guideline for designers in making appropriate structural changes (Foo et al., 2013). The assumption in design approaches is that the more difficult the task is, the more Mental Workload increases and the performance usually decreases (Xie et al., 2017). However, it is personal and complicated in different ways that are difficult to predict (Longo, 2015a). To construct the measurement in educational settings, the majority of research used Mental Workload in Ergonomics as the alternative one (Longo & Barrett, 2010), i.e., the experience of Mental Workload depends on each individual by way of different cognitive style, different education, and upbringing. There is no widely accepted definition of MWL in spite of the total cognitive load needed to accomplish a specific task under a finite period (Cain, 2007). As a consequence, Mental Workload (MWL) is a fundamental design concept in Human-Computer interaction (HCI) and Ergonomics (Human Factors) and sometimes is referred to as Cognitive Load, specifically in Cognitive Psychology.

There is leading research in measuring and evaluating the Mental Workload. However, how it effects instructional design or performance measurement when linked with the workload measure is still unclear (Hancock, 2017).

2.2.1 Measurement methods

To measure MWL is as necessary in predicting human performance as in designing technologies (Longo & Leva, 2017), interfaces (Longo & Dondio, 2015), information-based procedures and instructions (Longo, 2016). There are different ways proposed for measuring MWL, but categorized into three main techniques:

Self-assessment or subjective measures

This measure is based on the analysis of the subjective feedback provided by human interacting with an underlying task or system (Moustafa, Luz, & Longo, 2017). The form is often a survey or questionnaire, mostly post-survey. The common instruments are the NASA Task Load Index (NASA, NASA-TLX) (G. Hart & E. Stavenland, 1988), the Workload Profile - Multiple resource Theories (WP) (Tsang & L. Velazquez, 1996), the Subjective Workload Assessment Technique (SWAT) (Reid & Nygren, 1988) and the simplified SWAT (Luximon & Goonetilleke, 2001).

Task performance measures

This measure refers to primary and secondary task measures and is considered as an objective performance measurement. The time to complete a task, the reaction time to secondary tasks and the number of errors on the primary task are examples of measures, are concrete ways of tracking the different actions performed by a user during a primary task. However, these human performance indicators can be assessed by subjective usability (Longo, 2017, 2018).

Physiological measures

This measure performs as the analysis of physiological indicators and responses of the human body, including EEG (electroencephalogram), eye tracking and heartbeat measurements during the time of completion of the tasks.

2.2.2 Multi-dimensional and uni-dimensional measures

Concerning subjective measures, the NASA-TLX and WP are multi-dimensional because they include more than one method to assess and measure; whereas uni-dimensional measure has only one. The Rating Scale Mental Effort (RSME) is a uni-dimensional procedure that considers the exerted subject's effort and subjective ratings. These ratings are indicated across a continuous line, within the interval 0 to 150 with ticks every ten units. Labels on 'absolutely no effort', 'almost no effort', 'a little effort', 'some effort', 'rather much effort', 'considerable effort', 'great effort', 'very great effort' and 'extreme effort' are used along the line. On the one hand, the procedure is relatively simple, quick and it has shown a good degree of sensitivity. On the other hand, it has also demonstrated to have a poor diagnostic capacity (Zijlstra, 1993).

Following the research of MWL, subjects performed two laboratory tasks separately (single function) and simultaneously (dual function). The multi-dimensional procedure compared better than the uni-dimensional methods regarding sensitivity to task demands, concurrent validity with performance, and test-retest reliability. This finding strongly supports the notion that MWL is multi-dimensional in nature (Tsang & L. Velazquez, 1996).

2.2.3 Criteria for evaluating Mental Workload measures

There are different criteria for the development of MWL measurement methods and hence an array of literary terms needed to evaluate their inferential capacity (ODonnel & Eggemeier, 1986).

- Sensitivity: the measurement method should be responsive to variations in task difficulties and other factors believed to influence MWL on the task level;
- Diagnosticity: the method should be diagnostic and be capable of identifying the changes in workload variation and the causes of these changes;
- Intrusiveness: the method should not be intrusive or interfere with the primary task performance;

- Requirements: the method should demand minimum requirements to avoid influencing the performance of the person during primary task execution;
- Acceptability: the method should achieve high acceptance from the person;
- Selectivity: the method should be highly sensitive to MWL factors and not affected by other factors that are not related to MWL;
- Bandwidth and reliability: the method should be consistent or stable;
- Validity: comprising of face validity (the method covers the construct of MWL) and concurrent validity (the degree to which measures of MWL expected to be theoretically related, are related).

A measurement technique including in all the criteria above is ideal, but it is not always the case. First things first is a good construct built for Mental Workload representation & assessment (Guastello, Marra, Correro, Michels, & Schimmel, 2017). In the study of Longo (2015a) has demonstrated how the framework outperformed state-of-the-art subjective MWL assessment techniques regarding sensitivity, diagnosticity, and validity (Longo, 2018, March). So far, the criteria for evaluating MWL assessment techniques were also in the research of (Rizzo & Longo, 2017).

2.3 Subjective rating scales

In the perspective of research, human Mental Workload is measured by subjective rating scales which are easy to administer and analyze by comparison of two other measures: performance and physiological (Xiaoru, Damin, & Huan, 2014). Subjective measures provide an index of general workload, and multi-dimensional measures can determine the source of Mental Workload. The main drawback is that they can only be administered post-task, thus influencing the reliability of long tasks. In addition, meta-cognitive limitations can diminish the accuracy of reporting and cause difficulty in performing comparisons among raters on an absolute scale. Despite that, they appear to be the most appropriate types of measurement for assessing Mental Workload

because they have demonstrated high levels of sensitivity and diagnosticity (Rubio, Daz, Martn, & Puente, 2004).

To measure the Mental Workload, subjective workload techniques recently used are NASA-TLX, (Seker, 2014; Byrne et al., 2014; Longo, 2018; Foo et al., 2013; Adar & Delice, 2017; Xiaoru et al., 2014; Mitropoulos & Memarian, 2013; Weigl et al., 2014) or Workload Profile (Valdehita, Ramiro, Garca, & M. Puente, 2004; ?, ?). Both are multi-dimensional measures applied in the field of psychology or technological improvements. Furthermore, when evaluating the cognitive load, it is better to consider combined measures (Xiaoru et al., 2014; Adar & Delice, 2017). A regression equation is often applied to predict the model for Mental Workloads, such as Partial Least Square of Structural Equation Modelling (Kuo-Kuang, Chung-Ho, Shuh-Yeuan, & Wei-Jhung, 2013).

2.3.1 NASA Task Load Index

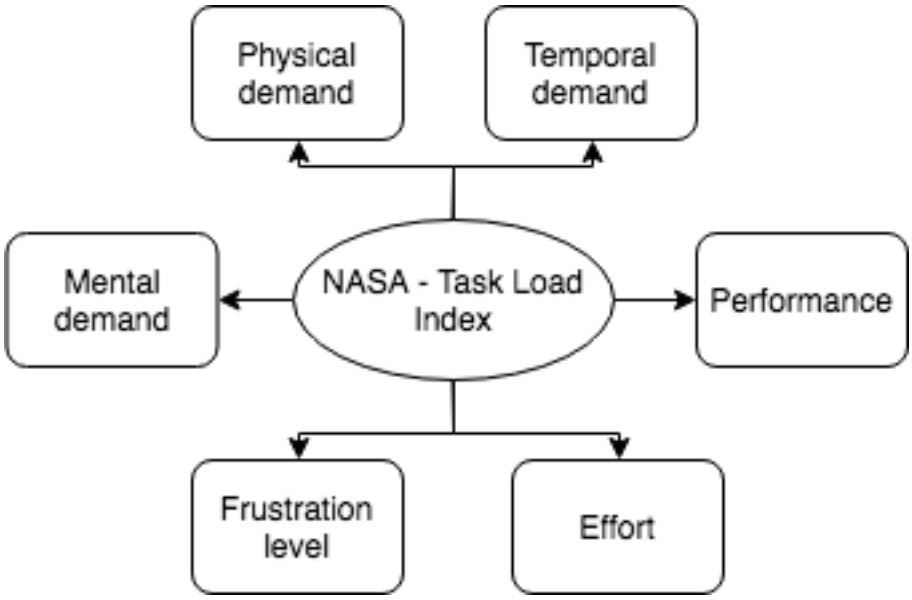

Figure 2.1: Six components of NASA Task Load Index

The NASA Task Load Index (NASA, NASA-TLX) instrument has been used far beyond its original application (aviation), in the field of requiring concentration (crew competence), health care (Colligan, Potts, Finn, & Sinkin, 2015), language (English)

and other complex socio-technical domains (G Hart, 2006). The Human Performance Group developed it at NASA's Ames Research Center over a three-year development cycle that included more than 40 laboratory simulations (G. Hart & E. Stavenland, 1988). It is a combination of six factors believed to influence MWL (figure2.1). The goal was to summarize the productivity during activities performed by the test subjects in different environments.

2.3.2 Workload Profile - Multiple resource theory

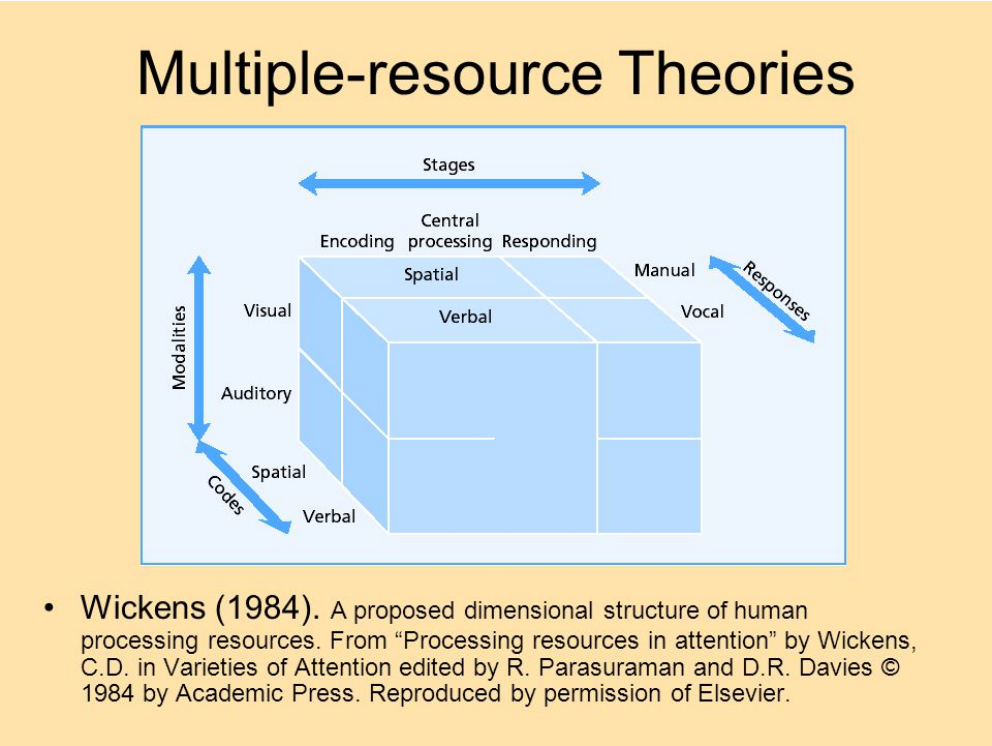

Figure 2.2: Workload Profile - Multiple Resource Theory

Tsang and L. Velazquez (1996) built the Workload Profile (WP) upon the Multiple Resource Theory proposed in the structure of Wickens (2008). The theory was shown to be partially relevant to the concept of Mental Workload, with the greatest relevance to decreased performance due to dual-task overload (Wickens, 2017). Its history comes from a computational version of the multiple resource model which was applied to multitask driving simulation data. The importance of four dimensions accounts for

task interference and the association of resources within the brain structure. In this theory, individuals have different capacities or resources related to:

- stage of information processing - central processing, response selection;
- code of information processing - spatial processing, verbal processing;
- resources - visual input, auditory input;
- response - manual response, speech response.

The most important application of the multiple resource model is to recommend design changes when conditions of multitask resource overload exist.

2.4 Summary

2.4.1 Gaps in research

The majority of models in the research predict a cognitive load score through total cognitive load by multi-criteria (Jaewon et al., 2016) or combined measurement methods:

- Adar and Delice (2017) evaluated MWL using Multi-criteria HFLTS method: a new decision-making method, Hesitant Fuzzy Linguistic Term Set, and evaluating the mental workload by employing the dimensions used in NASA-TLX. The HFLTS method, which allows qualitative and quantitative criteria is used in alternative evaluation interchangeably.
- Sewell, Boscardin, Young, Cate, and O'Sullivan (2016) measured cognitive load during procedural skills training with colonoscopy as an exemplar: the instrument (the Cognitive Load Inventory for Colonoscopy) using a multi-step process and cognitive load theory to develop a self-report instrument measured three types of cognitive load (intrinsic, extraneous and germane load).

- Naismith, Cheung, Ringsted, and Cavalcanti (2015); Ngu and Phan (2016) had evidence of correlation of intrinsic cognitive load and instructional design. Naismith et al. (2015) showed limitations of subjective cognitive load measures in simulation-based procedural training: The questionnaires appear to be interchangeable as measures of intrinsic cognitive load, but not of total cognitive load.

Some research indicated the effect of cognitive load theory on cognitive types in designing the learning conditions framework. Nevertheless, it is still unclear as to how it can formally shape a successful framework and which factors compose it:

- In the research of the germane load impact (Schwonke, 2015; Cheon & Grant, 2012), Cheon and Grant (2012) described the effects of the metaphorical interface on germane cognitive load in Web-based instruction. The results indicated that germane cognitive load positively affected learning performance despite there being no relationship between germane cognitive load and students' prior knowledge. That being said, both germane cognitive load and prior knowledge similarly contributed to learning performance. Besides, Schwonke (2015) considered a metacognitive type of load in resource-oriented theories, which was expected to have the as same effect as cognitive load theory.

- Effects of cues and real objects on learning in a mobile device supported environment (Liu, Lin, & Paas, 2013): The theoretical framework of cognitive load theory with arrow-line cues would decrease extraneous cognitive load. But there is no proof of overlap between the different sources of information used and that it affects learning, i.e., the availability of real plants would increase germane cognitive load.

- Xiaoru et al. (2014) suggested improving pilot MWL evaluation with combined measures. However, the inconsistent conclusions on the sensitivities of various MWL evaluation indices probably resulted from the different experiment tasks, which were designed to induce MWLs.

There are leading research to fill the gaps above:

- On one hand, a defeasible reasoning framework (Longo, 2015a; Blayney et al., 2015) was essential. The research provided an extensible framework built upon defeasible reasoning, and implemented with argumentation theory, in which MWL can be better defined, measured, analyzed, explained and applied with different human-computer interactive contexts. On the other hand, Kalyuga and Singh (2016) studied within the frameworks of productive failure and invention learning that has reportedly demonstrated as minimally guided tasks before explicit instruction might benefit novice learners.

- Blayney et al. (2015) found out more finely-grained methods of evaluation of learner prior experience, which required for optimal tailoring of instructional methods to levels of learner expertise. The benefit of rapid diagnostic tests is to monitor learners progress and alter the instructional techniques in real time.

- Davids, Halperin, and Chikte (2015) applied evidence-based design principles to manage cognitive load and optimize usability. It is essential to improve the educational impact of e-learning resources, especially relevant to multimedia resources.

There are experiment research in Mental Workload prediction in aviation (Xu, Xiaoru, & Damin, 2015) to optimize human factors and reduce human errors (Smith, 2017), in education to understand why negative sentence is more difficult to remember (Macbeth et al., 2014), in communication (Longo, 2015b) and networking technology (Colombi et al., 2012). This study aims to predict the total Mental Workload by subjective rating scales with statistically significant variables.

2.4.2 Research question

To what extent "*can a model of Mental Workload be built upon a set of features extracted from the literature of mental workload and applied in third-level education?*"

Chapter 3

Design and methodology

The study will be implemented in four main parts. Firstly, the phase of data understanding, which includes data collection, description, exploration and quality verification. Secondly, the phase of data preparation, which describes how to select, clean, construct, integrate and format data for the purpose of analysis. Thirdly, the phase of modeling, describing the chosen technique; how to generate test design, how the model is built and assessed. Fourthly, the evaluation phase, which describes the reliability and validity of the results, and suggests the next steps for development.

The schema of research is shown in figure 3.1. Each subjective rating scale or subjective measure in the training set and test set, contains a constant value and categorical value of output (MWL). The chosen model should reflect the type of output. If it were a case of Mental Workload score (as a continuous feature), it would be trained and tested through Linear Regression and Decision Tree GINI Regression; if it were a case of Mental Workload level classes (as a categorical feature), it would be trained and tested through Decision Tree Information Gain and Decision Tree GINI Classification.

As such, there were twelve models in total for three subjective measures.

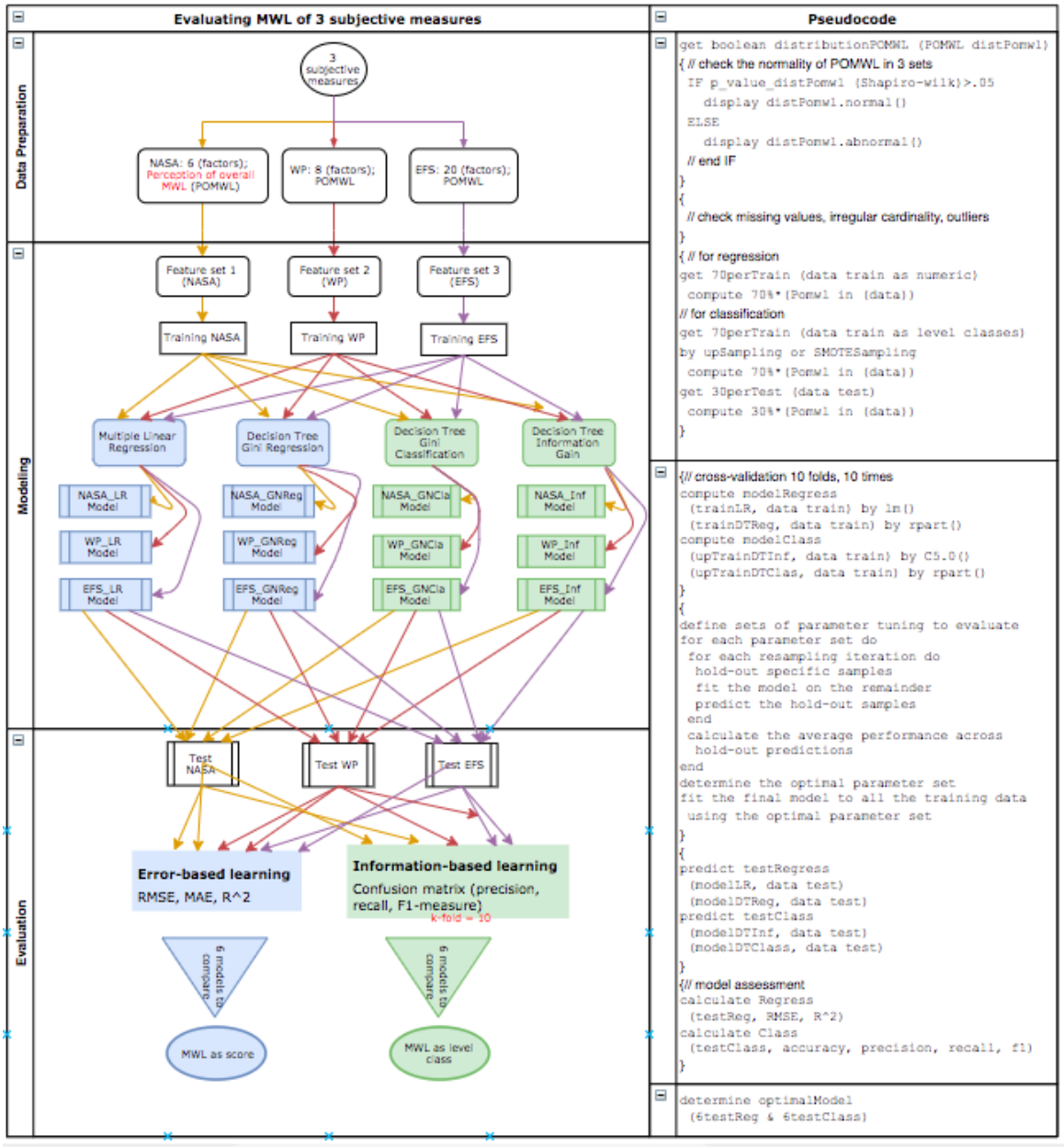

Figure 3.1: Schema of research

3.1 Business understanding

The application of Human Mental Workload in third level education will allow the prediction of student performance. The benefits of which include restricting the numbers of students at risk of failure, and allowing the design of class conditions to reduce the load on learners. In the current era of information explosion, the significant task

of continuous knowledge gathering is crucial to higher education.

3.1.1 Extended Feature Sets

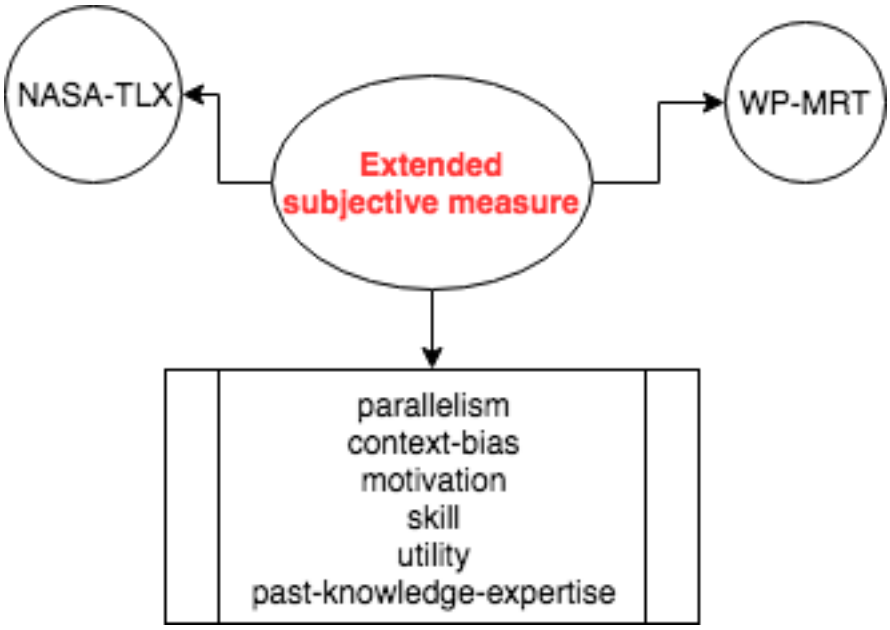

Figure 3.2: Extended Feature Sets

Human behaviour is both personal and complicated, in different ways that are difficult to predict. A working model to assess Mental Workload, comprising of task assessment (NASA) and multiple resources effects (WP) with additional education factors, is the aim of the complete subjective rating scales, which is the Extended Feature Sets (EFS). The other education factors are listed below:

- Context - distractions/ interruptions during the teaching session
- Parallelism - not engaged with teaching session or engaged in other parallel tasks
- Motivation - motivated by teaching session
- Skill - have no influence or help
- Utility - teaching session useful for learning
- Past Knowledge Expertise - experience with the session

- Arousal - sleepy tired or fully cognizant

3.1.2 Research hypothesis

This research aims to investigate the optimal model in predicting the Mental Workload score with the new subjective rating scale (EFS). The hypothesis of the research:

- H_a: A model trained with EFS lead to significantly more accurate and less error in prediction of overall Mental Workload than models trained with NASA-TLX or WP (through Decision Tree, Linear Regression).

3.2 Data understanding

3.2.1 Data collection

The dataset has been collected from delivering the Research Design & Proposal Writing module from 2015 to 2018. The training and test sets split from the data of 684 records. Two different groups of part-time and full-time students participated in the study. They attended the MSc module 'Research design and Proposal writing' in different years. Both groups spent time on four topics in Science; The scientific method; Planning research; Literature Review in two conditions (1) multimedia slides verbally presented by lecturer on a white-board, (2) multimedia video projected on a white-board. At the end of each topic, students were asked to fill in questionnaires, aimed at quantifying the Mental Workload during the class.

The NASA-TLX, the WP and the EFS are multi-dimensional and thus require participants to answer some questions. RSME (Rating Scale Mental Effort) is uni-dimensional adding one further question to compare to the other scale in the same delivered questionnaire. The formation of the two subgroups, one received the NASA-TLX, one received the WP and the other proceeded EFS randomly.

3.2.2 Data description

Table 3.1: Variable definition of 3 subjective measures

Feature	Description
MWL	- Perception of Mental Workload, multi-dimensional measure is also the target variable in 3 subjective measures (NASA-TLX, WP and EFS). - Continuous variable, range from 0 to 100; Ordinal type with 5 groups: Extreme underload [0-10], Underload [11-25], Optimal load [26-75], Overload [76-90], Extreme overload [91-100]
Features below are continuous variables and having the range from 0 to 100	
NASA_Ment	- Task demand features, Mental and perceptual activity required, the teaching session easy or demanding, simple or complex.
NASA_Phys	- Task demand features, Physical activity required, the teaching session easy or demanding, slack or strenuous.
NASA_Temp	- Task demand features, Time pressure due to the pace, the space was slow or rapid.
NASA_Perf	- Perceived performance, how successful or satisfied felt with learner's performance.
NASA_Frus	- Cognitive state, how irritated, stress, annoyed felt.
NASA_Effo	- Cognitive state, how hard (mentally, physically) to accomplish the performance.
WP_Solv	- Central Processing, attention for activities like remembering, problem-solving, decision-making and perceiving.
WP_Resp	- Response Processing, attention for selecting the proper response channel (manual or speech) and its execution.
WP_TaSpa	- Task and space, attention for spatial processing.
WP_Verb	- Verbal material, attention for reading or processing linguistic material or listening to verbal conversations.
WP_Visu	- Visual resources, attention for attending the teaching session based on the visual information.
WP_Audi	- Auditory resources, attention for attending the teaching session based on the auditory information.
WP_Manu	- Manual response, attention for manually responding to the teaching session.
WP_Spee	- Speech response, attention for producing the speech response (engaging in a conversation, answering questions).
Features comprise of the factors in NASA and WP	
EFS_nasa_Ment	- Mental demand (NASA).
EFS_Para	- Just attending teaching session or engaged in other parallel tasks (mobile browsing/ social networks, chatting, reading, conversation).
EFS_nasa_Temp	- Temporal demand (NASA).
EFS_wp_Manu	- Manual response (WP).
EFS_wp_Visu	- Visual resources (WP).
EFS_nasa_Effo	- Effort (NASA).
EFS_wp_Solv	- Central Processing (WP).
EFS_nasa_Frus	- Frustration (NASA).
EFS_Cont	- Interruptions during the teaching session distractions (mobile, noise, questions, other participants,..)
EFS_wp_TaSpa	- Task & space (WP).
EFS_Moti	- Motivated by the teaching session.
EFS_wp_Verb	- Verbal material (WP).
EFS_Skil	- Skills have no influence or help.
EFS_wp_Audi	- Auditory resources (WP).
EFS_nasa_Phys	- Physical demand (NASA).
EFS_wp_Spee	- Speech response (WP).
EFS_Util	- The teaching session was useful for learning.
EFS_Know	- How much experience knowledge with the session.
EFS_Arou	- Sleepy tired or fully activated awake.
EFS_nasa_Perf	- Performance (NASA).

3.2.3 Data exploration

Table 3.2: Data exploration of NASA

Feature set 1: NASA-TLX (N=230)										
Feature	**type**	**miss**	**n**	**min**	**1stQ**	**median**	**3rdQ**	**max**	**mean**	**sd**
RSME [0-150]	R	1	229	10	40	50	73	105	53.74	21.72
Features below range from 0 to 100										
MWL	R	1	229	10	45	55	65	85	53.65	14.84
Mental	R	0	230	5	40	50	65	100	50.20	16.99
Physical	R	0	230	5	15	25	50	100	31.10	20.75
Temporal	R	1	229	5	30	45	55	100	45.28	17.86
Performance	R	0	230	10	30	45	60	85	44.35	18.04
Frustration	R	0	230	5	25	35	55	95	38.24	19.52
Effort	R	0	230	5	35	50	60	100	49.13	19.24

R: Range, Q: quarter

There were 230 students enrolling in the NASA questionnaire. The mean and median of MWL ($\text{Mean}_{MWL} = 53.65$, $\text{Median}_{MWL} = 55$) and RSME (Mean = 53.74, Median = 50) were not so different. Two features having the opposite trend of MWL were Physical demands and Frustration. In terms of six features having the impact on MWL, the lowest score was Physical demands (mean = 31.1) and the highest one was Mental demands (mean = 50.2).

Table 3.3: Data exploration of WP

Feature set 2: Workload Profile (N=217)										
Feature	**type**	**miss**	**n**	**min**	**1stQ**	**median**	**3rdQ**	**max**	**mean**	**sd**
RSME [0-150]	R	1	216	0	35	50	70	100	50.36	20.95
Features below range from 0 to 100										
MWL	R	0	217	5	40	50	65	90	51.72	16.63
Central	R	0	217	5	45	55	70	100	55.03	19.47
Response	R	0	217	5	35	50	65	100	49.44	21.72
Spatial	R	1	216	5	25	47.5	60	100	44.13	22.85
Verbal	R	0	217	5	50	65	75	100	61.97	19.19
Visual	R	0	217	10	50	65	75	100	61.47	19.09
Auditory	R	1	216	5	55	65	75	100	64.16	18.88
Manual	R	0	217	5	30	50	65	100	46.89	24.41
Speech	R	0	217	5	25	50	65	100	46.18	24.93

R: Range, Q: quarter

The number of 230 students enrolled in the NASA questionnaire. The mean and median of MWL ($\text{Mean}_{MWL} = 53.65$, $\text{Median}_{MWL} = 55$) and RSME (Mean = 53.74, Median = 50) were not so different. Two features having the opposite trend of MWL were Physical demands and Frustration. Regarding six features having an impact on MWL, the lowest score was Physical demands (mean = 31.1), and the highest one was Mental demands (mean = 50.2).

Table 3.4: Data exploration of EFS

Feature set 3: Extended Feature Sets - EFS (N=237)										
Feature	**type**	**miss**	**n**	**min**	**1stQ**	**median**	**3rdQ**	**max**	**mean**	**sd**
Features below range from 0 to 100										
MWL	R	2	235	5	40	50	60	95	50.79	16.60
_NASA_Men	R	0	237	5	35	50	60	90	48.46	16.50
_Parallelism	R	0	237	5	10	15	35	80	25.50	21.42
_NASA_Tem	R	1	236	10	45	50	60	85	51.40	14.24
_WP_Manual	R	0	237	5	30	50	65	100	47.74	24.25
_WP_Visual	R	1	236	10	50	65	75	100	62.5	17.94
_NASA_Eff	R	4	233	10	40	55	65	100	51.83	18.74
_WP_Central	R	0	237	10	35	50	70	100	52.65	21.76
_NASA_Fru	R	1	236	5	30	50	55	90	44.13	19.07
_Context	R	0	237	5	10	20	35	95	25.74	20.31
_WP_Spatial	R	0	237	5	15	35	55	100	36.46	23.01
_Motivation	R	0	237	5	50	65	75	100	61.73	20.32
_WP_Verbal	R	1	236	5	45	60	70	100	57.92	20.20
_Skill	R	1	236	5	35	52.5	65	100	49.96	23.32
_WP_Audi	R	0	237	5	50	60	70	100	59.05	18.93
_NASA_Phy	R	0	237	5	10	25	45	80	29.43	20.72
_WP_Speech	R	0	237	5	20	40	60	100	40.61	24.06
_Utility	R	0	237	5	55	70	80	100	66.62	20.37
_PastKnow	R	1	236	5	35	55	66.25	95	52.5	20.83
_Arousal	R	0	237	10	45	55	70	100	57.81	19.38
_NASA_Per	R	0	237	10	45	65	75	100	61.71	18.82

R: Range, Q: quarter

The number of 237 students enrolled in the EFS questionnaire. The mean and median of MWL were not so different ($\text{Mean}_{MWL} = 50.79$, $\text{Median}_{MWL} = 50$). Four features having the opposite trend of MWL were two additional features (Parallelism, Context), one feature relevant to NASA (Physical demands) and one relevant to WP (Spatial Processing). In the total of twenty features having the impact on MWL, the lowest score was Parallelism (mean = 25.50), Context (mean = 25.74); and the highest one was Utility (mean = 66.62).

3.2.4 Data quality verification

The data set of assessing Mental Workload with three measures are NASA-Task Load Index (N=230), Workload Profile (N=217) and Extended Feature Sets (N=237). There are 0.5-2 % missing values, and the median of each variable values are imputed. The **descriptive features** are factors in measuring and classifying the levels of Mental Workload score. The **target feature** is the perception of Mental Workload score.

3.3 Data preparation

3.3.1 Data selection

Data is randomly divided into approximately 70:30 for training and test sets. For each data division of the three subjective scales, the target variable is Mental Workload, but the description variables can be different in quantities and characteristics. For instance, NASA-TLX has six independent factors (mental demand, physical demand, temporal demand, performance, frustration, and effort); WP has eight factors (central, response, spatial, verbal processing, visual and auditory input, manual and speech response); EFS is the combination of 6 factors in NASA-TLX as above, 7 factors in WP as listed above without response processing variable; and 7 others (parallelism, context, motivation, skill, utility, past knowledge expertise and arousal) which has 20 factors in total.

Three measures will have the list of datasets extracted from cross-validation by ten folds and ten occurences.

3.3.2 Data cleaning

The first step is testing for normality of target feature. There are two main methods of assessing normality, graphically by histogram, and numerically by Shapiro-Wilk Test or skewness and kurtosis measure for skewed data.

If seeing Mental Workload shows as an interval, the second step would be analyzing the correlation between the Mental Workload score and independent variables (descriptive features) to illustrate the factors put into the model. The Mental Workload score is an interval scale, but non-parametric as the distribution is discrete in the lowest and highest range. When the Mental Workload score is non-parametric, we carry out two tests: Kruskal-Wallis (as for nominal descriptive feature); Spearman (as for numeric descriptive feature). The correlation will be a reasonable explanation for the essential variables in models of Multiple Linear Regression and Decision Tree Gini Regression after training. If Mental Workload shows as ordinal, the second step

is resolving a class imbalance. In classification problems, a disparity in the frequencies of the observed classes can have a significant negative impact on model fitting. One technique is to subsample the **training data** in a manner that mitigates the issues. There are three types of sampling methods. Firstly, down-sampling randomly subset all the classes in the training set so that their class frequencies match the least prevalent level. Secondly, up-sampling sample randomly (with replacement) the minority class to be the same size as the majority class. Lastly, hybrid methods are techniques such as SMOTE down-sample the majority class and synthesize new data points in the minority class. This method is used for the models of Decision Tree Information Gain and Decision Tree Gini Classification.

Lastly, the best model will be the one with less error or better accuracy after comparing the training and evaluation output.

3.4 Modelling

In the prospect of matching Machine Learning approaches to projects, the pre-requisites of a project are mostly the viable creation of an accurate prediction model. Firstly, data should be split into training and test sets (70:30) to train and validate the model. With the training set, data is trained in Decision Tree and Multiple Linear Regression. The nature of the target value is similar to the Likert scale as rating scales to capture estimations of magnitude. Data from Likert scales and continuous rating scales are quantitative (Joshi, Kale, Chandel, & Pal, 2015) which is interval data. However, it can also be ordinal scale in some point of views (Joshi et al., 2015). As in the context, the Multiple Linear Regression is applied for Mental Workload score as a continuous feature; the Decision Tree is applied to train in two methods, Information Gain (applied C5.0) to deal with ordinal value and Gini index (applied CART) to deal with both viewpoint as interval or ordinal value. Therefore, there are four models (2 models for numeric target value and 2 models for categorical target value) on each instrument. Accordingly, twelve models will be produced to compare the accuracy, precision, and recall of actual and predicted value for categorical models; and RMSE, R-squared for

numeric models.

To achieve a representative sample to select for the optimal predictive model, requires cross-validation to take place. Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. To use a limited sample for estimating how the model performs predictions on data, which is the test set not applied during the training of the model. The procedure is as follows:

1. Shuffle the dataset randomly
2. Split the dataset into k groups
3. For each unit group:
    - Take the group as a hold-out or test data set
    - Take the remaining groups as a training data set
    - Fit a model on the training set and evaluate it on the test set
    - Retain the evaluation score and discard the model
4. Summarize the skill of the model using the sample of model evaluation scores

Any preparation of the data prior to fitting the model occurs on the cross-validation assigned training dataset within the loop rather than on the broader data set. The tuning of hyperparameters also applies to cross-validated training. A failure to perform these operations within the loop may result in data leakage and an optimistic estimate of the model skill ((Gareth, Daniela, Trevor, & Robert, 2013), pg 181). The results of running k-fold cross-validation are the mean of the model skill scores ((Stuart & Peter, 2016), pg 708).

The "caret" package (short for Classification and Regression Training) is a set of functions that attempt to streamline the process for creating predictive models. The 'train' function can be used to evaluate, resample, estimate the effect of model tuning parameters on performance. It is used for training models, dealing with class imbalance and testing data on the background of RStudio software. Also, "ggplot" package is also used for visualization comparison of models.

3.4.1 Error-based learning

Multiple Linear Regression

Multiple Linear Regression is fundamental to error-based learning. Linear Regression models determine the optimal values for the weights in the model, in fact, reducing the error of the weights. Proceeding the Multiple Linear Regression by multiplying the weights by the descriptive features. The weights of Linear Regression are the effect of each descriptive feature on the predictions returned by the model. In detail, regression is a method of modeling that a target value (named y) based on independent predictors (named x_1, x_2, x_3). Hence, for estimating and discovering the cause and effect relationship between variables, one would apply regression. The regression equation can differ in the number of x_i variables and the type of relationship between the x_i and y.

In general, the simple linear regression equation is y = m*X + b, where "m" is the slope and "b" is the intercept. In machine learning, the equation above will be: y(x) = w_0 + w_1*x, where w_s are the parameters of the model, x is the input, and y is the target variable. Different lines will have different values of w_0 and w_1. In real value, having the multiple input variables data set; and the Multiple Linear Regression Machine Learning model would be described as: y(x) = w_0 + w_1*x_1 + w_2*x_2 + + w_i*x_i. The question must then be asked for how well do the coefficients of "w" predict the target value? This can be determined by (1) squaring the error difference between the predicted value y(x) and the target value y-true; (2) sum over all data points; (3) divide that value by the total number of data points. This value provides the average squared error over all the data points. These three steps are called cost function also known as the Mean Squared Error (MSE) function, as:

$$J(w) = \frac{1}{n}\sum_{i=1}^{n}(y(x^i) - y^i_{true})^2 \tag{3.1}$$

The method used in "caret" package is 'lm'. The tuning parameter for the model is the intercept.

3.4.2 Information-based learning

Machine learning algorithms build predictive models using only the most informative features. The most informative features are the descriptive features whose values split the instances of the data set into similar sets for the target feature value.

Decision Tree through Information Gain

The ideal discriminatory feature will partition the data into pure subsets where all the instances in each subset have the same classification. Information Gain of a descriptive feature is a measure of the reduction in the overall entropy of a prediction task by testing on that feature. As in, entropy is a computational measure of the impurity or heterogeneity of the elements in a set.

The method used in "caret" package is 'C5.0'. C5.0 pruning technique adopts the Binomial Confidence Limit method. The tuning parameters are trials, models and winnow.

Decision Tree through Gini Index

Gini index shows the frequency of instance misclassification in a data set, if classifying it based on the distribution of classifications in the data set. The Decision Tree model uses the Gini Index as its splitting criteria.

The method used in "caret" package is 'CART.' CART can handle both nominal and numeric attributes to construct a decision tree. The tuning parameter is cp (complexity pruning) which removes excess branches from the decision tree for improved accuracy.

3.5 Evaluation

In the course of model evaluation, three other important issues to be considered are "prediction speed, capacity for retraining and interpretability" (D.Kelleher, Namee, & Darcy, 2015).

For the data cleaning step, data is imputed and transformed properly on the type of output. Training samples are randomly created by "createDataPartition" a list of indices represented for chosen instances from 70 percent of the whole data 10 times, and 10-fold cross-validation in all experiments. The correlation between the Mental Workload score and relevant feature characteristics in the subjective rating scale is tested on hypothesis tests (ANOVA, independent T-test, Pearson or Kruskal-Wallis, Mann-Whitney, Spearman) by p_{value}. If $p_{value} < 0.05$, there do exist a connection.

Equally important testing the results of 12 models is, test sets are from 30 percent of the whole data, which has different instances from 10 training sets. Then, the comparisons of (RMSE, R-squared), and (Accuracy, Precision, Recall) were counted on ten results of these 10 test sets through hypothesis tests and visualized illustration. For evaluating the optimal model, there are two types by testing the difference of predicted and actual values within each instrument and between instruments through RMSE, R-squared, Accuracy, Precision, and Recall. Within instruments, Friedman ANOVA or chi-squared are performed for the hypothesis of statistically significant difference between observed (actual values) and predicted values in one random test set. Between instruments, to decide a stronger predictive relationship between the Mental Workload models, density plots and box plots are for visualization comparison and ANOVA or Kruskal-Wallis are for testing the statistically significant difference of Accuracy, Precision, Recall as the categorical target in the average of 10 results on each model. Similarly, RMSE and R-squared of the continuous target is also compared in the average of 10 results on each model. The threshold of significant difference between models is $p_{value} < 0.05$.

The purpose of an evaluation is threefold: determine which is the most suitable model for a task, estimate how the model will perform, and convince users the model will meet their needs. For a categorical target, those three indicators are Accuracy, Precision, and Recall. Accuracy or classification accuracy is the number of correct classifications in the total number that reflects the optimal predictive model. Precision captures when a model makes a positive prediction (correct prediction). Recall defines how confident that all the instances with a positive target level found. Besides, F1 is

the central tendency measuring of precision and recall that takes the average of a set of values but is less sensitive to outliers. For a continuous target, two evaluated indicators are RMSE (Root Mean Squared Error) and R-squared. R-squared is a statistical measure of how close the data are to the fitted regression line in the range of 0 and 1 that is the larger the number, the better the model fits data. RMSE emphasizes large individual errors while MAE is the absolute of RMSE that the smaller the number, the more exact the model.

3.6 Strength and limitation

The key step in any predictive analytics project is deciding which type of model to use. There are two approaches to learning in this research: error-based (Multiple Linear Regression) and information-based (Decision Tree). The first distinction between models is the distinction between parametric and non-parametric models. It generally describes whether the size of the domain representation to define a model by the number of features in the domain or by the number of instances in the data set. In a parametric model, the size of the domain representation is independent of the number of instances in the data set, whereas, in a non-parametric model, the number of parameters (the domain representation) used by the model increases as the number of instances increases. Multiple Linear Regression is the parametric model, and Decision Tree is non-parametric. Generally, parametric models make stronger assumptions about the underlying distributions of the data in a domain. Non-parametric models are more flexible but can struggle with large data sets; however, it runs into time and space complexity issues as the number of instances grows.

The other important distinction made between classification models is whether they are generative or discriminative. In terms of this, Linear Regression and Decision Tree are discriminative models, which learn the boundary between classes rather than the characteristics of the distributions of the different classes. The Regression models learn a soft boundary, which considers the distance from the boundary. Decision Trees are induced by recursively partitioning the feature space into regions belonging to the

different classes.

The primary requirement of a project is to create an accurate prediction model. Besides that, interpretability is the important driver for decision making in a business scenario. Decision Trees and Linear Regression models are easy to interpret in comparison of support vector machines or ensembles.

Chapter 4

Results and discussion

This chapter is structured to describe, give evaluations and discuss relevant studies in depth:

- the data structure of three subjective rating scales (subjective measures)
- the results of four training models for each subjective rating scales
- model comparison in each subjective rating scales
- lastly model selection within each subjective rating scales (4 models for each) and between subjective rating scales (6 models using Mental Workload as the continuous feature and 6 models using Mental Workload as the categorical feature).

4.1 Data description

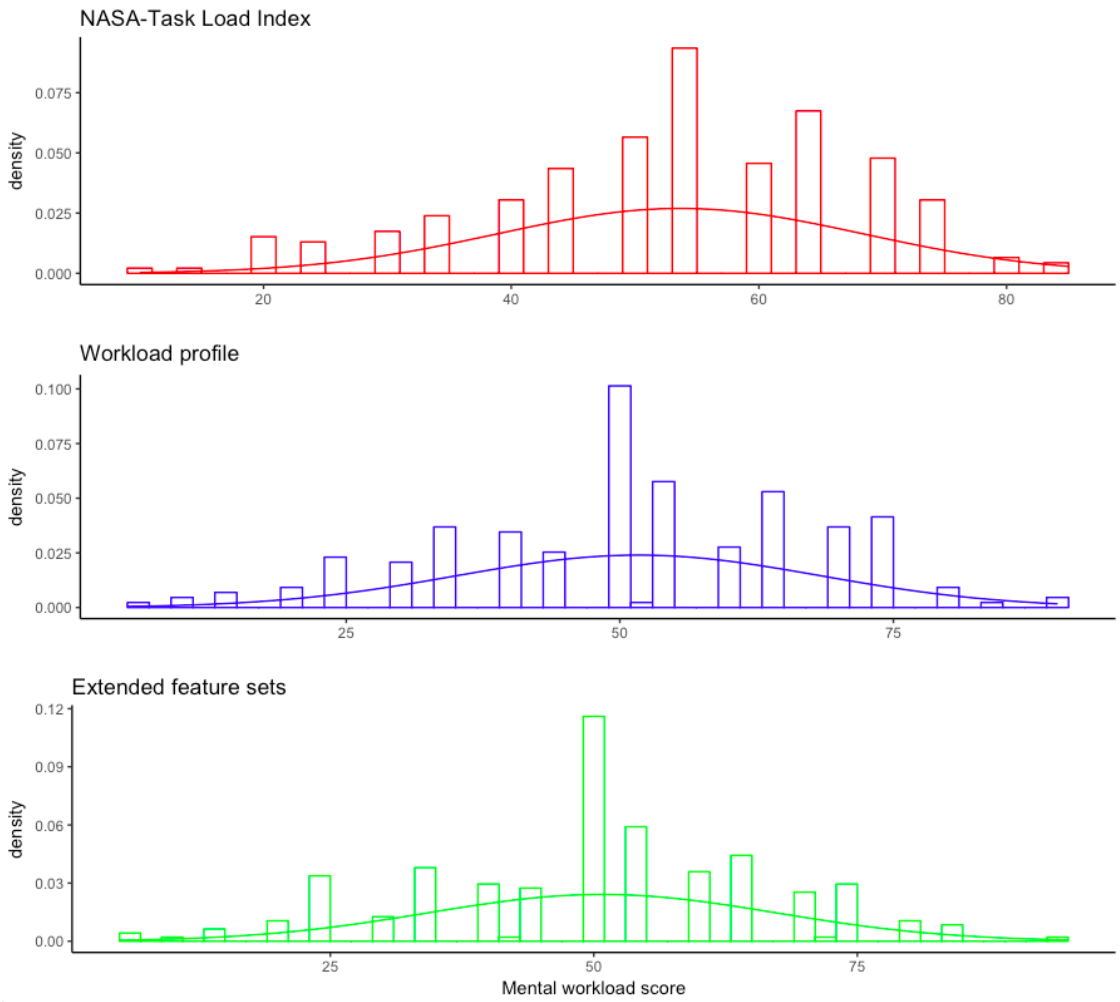

Figure 4.1: Histogram of Mental Workload score in three subjective rating scales

Three histograms of NASA-TLX (N=230), WP (N=217) and EFS (N=237) showed that they seem discrete, despite, the histograms not giving a strong indication of non-normality. There were slightly shorter right tail in three of them: more students in underload than overload level (figure 4.1).

4.1.1 NASA Task Load Index

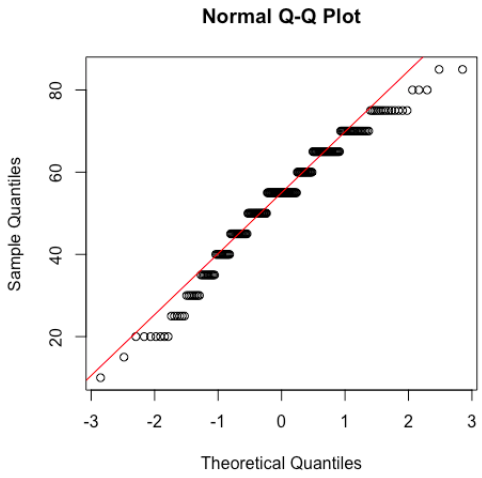

Figure 4.2: Q-Q plot of NASA dataset (N=230)

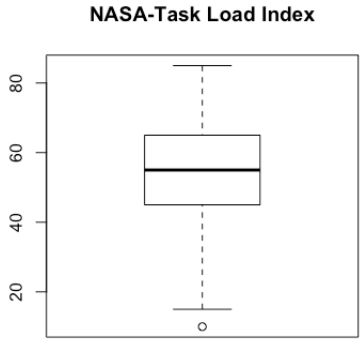

Figure 4.3: Boxplot of NASA dataset (N=230)

If the data is normally distributed, the points in the QQ-normal plot will lie on a straight diagonal line. The normal distribution having a skewness of 0, or between -0.5 and 0.5 is relatively symmetrical. For the kurtosis looks at the combined size of the tails will have the value of 0 of a normal distribution. So, if a dataset has a positive kurtosis, it has more in the tails than the normal distribution. If a dataset has a negative kurtosis, it has less in the tails than the normal distribution.

The figure 4.2 showed the upper right and bottom left points not fitting with the line. The points in the range 40 to 70 were on the line. A Shapiro-Wilk test

was performed to confirm NASA set distribution explicitly, (figure A.1) p_{value} <0.001 rejecting the hypothesis that this data was independently drawn from a standard normal distribution. This distribution was apparent both in the box plot (figure 4.3 which exhibited a short up-trend; and in the histogram, which advocated a moderately left skewed tail (-0.507) and light-tailed on kurtosis (-0.108).

The target feature treated as a categorical value would show the order below:

Table 4.1: MWL as categorical feature in NASA

Levels (N=230)	**Frequency (6 levels)**	**%**	**Frequency (4 levels)**	**%**
extreme underload	1	0.435		
underload	14	6.087	15	6.522
optimal load 1	79	34.348	79	34.348
optimal load 2	131	56.957	131	56.957
overload	5	2.174	5	2.174
extreme overload	0	0.000		

Table 4.1 showed there was a substantial imbalance in underload and overload levels compared to optimal load levels. The percent of highest level was optimal load 2, and the least was overload. Students often had self-assessment in the bottom and up levels less than in the middle levels.

4.1.2 Workload Profile

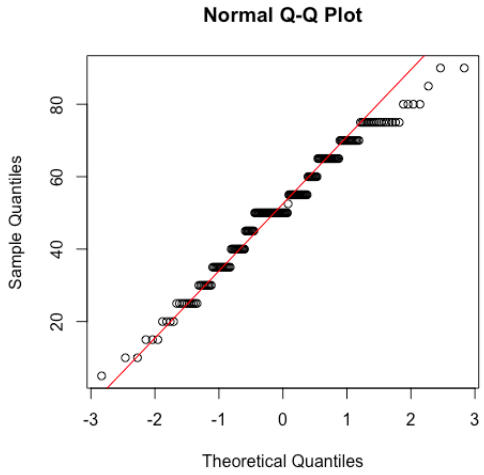

Figure 4.4: Q-Q plot of Workload Profile dataset (N=217)

Figure 4.5: Box plot of WP dataset (N=217)

The figure 4.4 showed fitted points from 10 to 70 on the diagonal line, and above points (above 70) were under the line. The median of WP in the box plot (figure 4.5 was lower than the median of NASA Mental Workload score. The Shapiro-Wilk test for data distribution performed p_{value} <0.001, as means of the data were not normally distributed. The histogram described WP data was fairly symmetrical through skewness(-0.285) and light-tailed kurtosis (-0.272).

The target feature treated as a categorical value would show the order below:

Table 4.2: MWL as categorical feature in WP

Levels (N=217)	**Frequency (6 levels)**	%	**Frequency (4 levels)**	%
extreme underload	3	1.382		
underload	17	7.834	20	9.217
optimal load 1	95	43.779	95	43.779
optimal load 2	95	43.779	95	43.779
overload	7	3.226	7	3.226
extreme overload	0	0.000		

Table 4.2 showed there was a substantial imbalance in underload and overload levels compared to optimal load levels. The percent of highest level was as optimal load 1 as optimal load 2, and the least was overload. Students often had self-assessment in the bottom and up levels less than in the middle levels.

4.1.3 Extended Feature Sets

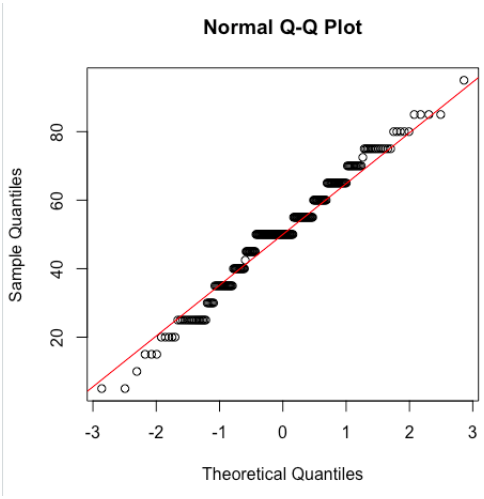

Figure 4.6: Q-Q plot of EFS dataset (N=237)

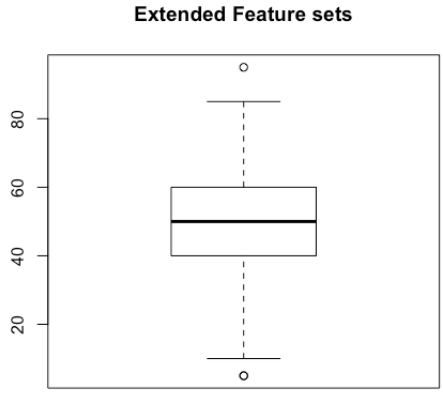

Figure 4.7: Box plot of EFS dataset (N=237)

The figure 4.6 showed QQ plot of EFS having fitted points from 40 to 80 on the diagonal line and bottom points were under the line. The median of EFS in boxplot was as same as NASA (figure 4.7. The Shapiro-Wilk test confirmed EFS distribution with p_{value} <0.001 meaning the data not normally distributed. This data was symmetrical skewness (-0.207) and very light kurtosis tailed (-0.091) which was close to 0.

The target feature treated as a categorical value would show the order below:

Table 4.3: MWL as categorical feature in EFS

Levels (N=237)	**Frequency (6 levels)**	**%**	**Frequency (4 levels)**	**%**
extreme underload	3	1.266		
underload	24	10.127	27	11.392
optimal load 1	107	45.148	107	45.148
optimal load 2	93	39.241	93	39.241
overload	9	3.797	10	4.219
extreme overload	1	0.422		

Table 4.3 showed there was a strong imbalance in underload and overload levels compared to optimal load levels. The percent of highest level was optimal load 1 (differing from NASA set was optimal load 2), and the least was overload. Students often had self-assessment in the bottom and up levels less than in the middle levels.

On the whole, Mental Workload score has been measured by the Likert scale. To be seen as a continuous or categorical variable, it depends on the target and the context of research. Likert items are often done in attitude surveys, and regarded as a true ordinal scale; although it is wise to report both mean/SD and % of response in the two highest categories. If the score is as a continuous value, showing that scores differ when considering variety group of participants. If the score is treated as a categorical value, highlighting how response patterns vary across subgroups, then item scores as the discrete choice among a set of answer options and look for item-response models or statistical model that allows coping with polytomous items.

The Mental Workload score was skewed in three of rating scales; as if its levels were an imbalance when being divided into four. Sub-sampling the training data was applied by upSample or SMOTE methods to increase effectively learning models. Up-Sampling randomly sampled (with replacement) the minority class to be the same size as the majority class; SMOTE drew artificial samples by choosing points that lie on the line connecting the rare observation to one of its nearest neighbors in the feature space. Both techniques outperformed over and under-sampling. However, in the nature of subjective rating scales there was lack of samples in two-tailed of data, Up-sampling showed more effectively than SMOTE sampling (figure A.5, A.6).

4.2 Data exploration

4.2.1 Correlation between Mental Workload score and its factors of three rating scales in whole datasets

NASA Task Load Index

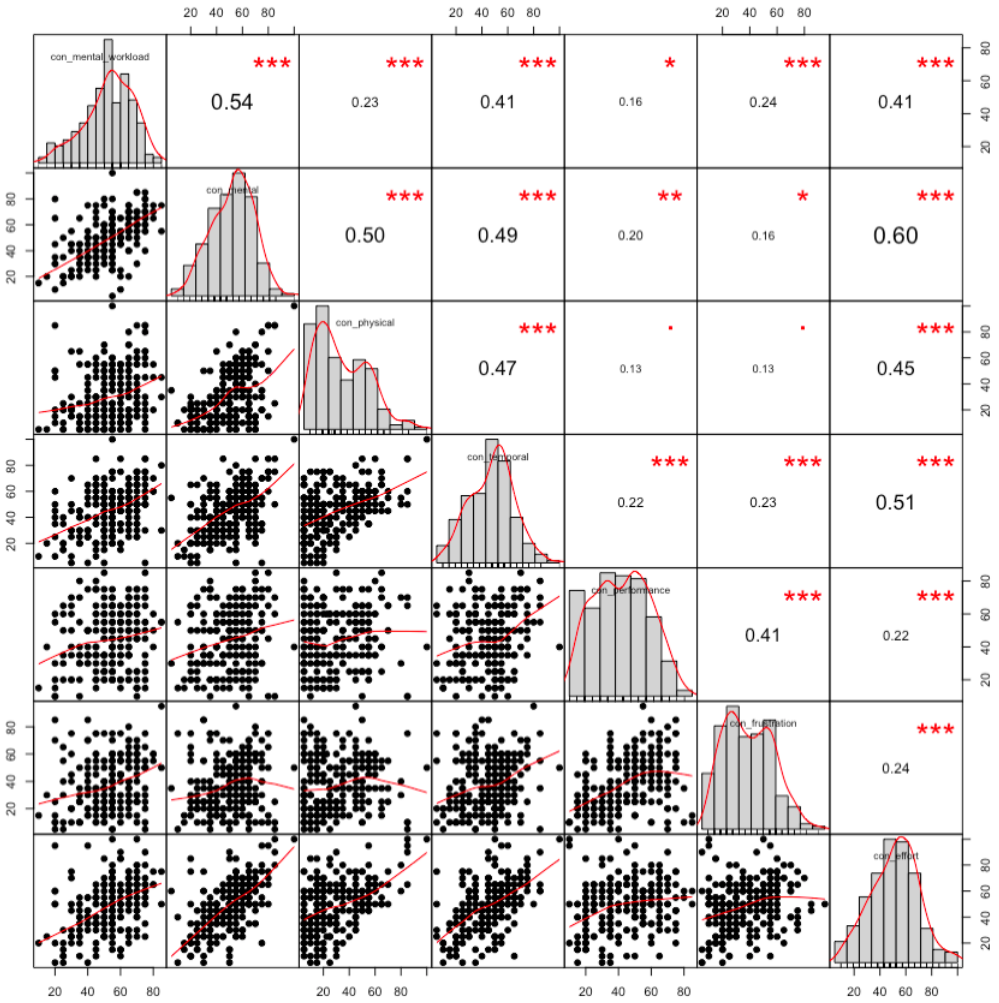

Figure 4.8: NASA Scatter plot matrix (N=230)

The distribution of each variable was shown on the diagonal line in scatter plot matrix (figure 4.8, 4.9, 4.10, 4.11, 4.12). On the bottom of the diagonal: the bivariate scatter plots with a fitted line were displayed. On the top of the diagonal: the value of the correlation plused the significance level as stars. Each significance level was associated

to a symbol: p_{value} (0.001, 0.01, 0.05, 0.1, 1) as symbols ("***","**","*","."," ").

Table 4.4: Correlation of Mental Workload score & factors in NASA set

	MWL	Ment	Phys	Temp	Perf	Frus	Effo
MWL	1	**0.54**	**0.23**	**0.41**	0.16	**0.24**	**0.41**
Ment		1	**0.50**	**0.49**	0.20	0.16	**0.60**
Phys			1	**0.47**	0.13	0.13	**0.45**
Temp				1	**0.22**	**0.23**	**0.51**
Perf					1	**0.41**	**0.22**
Frus						1	**0.24**
Effo							1

MWL: Mental Workload, Ment: Mental, Phys: Physical, Temp: Temporal, Perf: Performance, Frus: Frustration, Effo: Effort

Linear regression is the extent to which two variables have a straight line relationship. The closer to $+/-1$ the Cohen's effect size is, the stronger the relationship is. If the **absolute** value of the Cohen's effect size is higher than the range [-0.5,0.5] it is a strong relationship, higher than [-0.3,0.3] is moderate and higher than [-0.1,0.1] is weak.

Table 4.4 and figure 4.8 showed a strong relationship between Mental Workload and Mental demands, a moderate one with Temporal demands and Effort and a weak one with Physical demands, Frustration.

Workload Profile

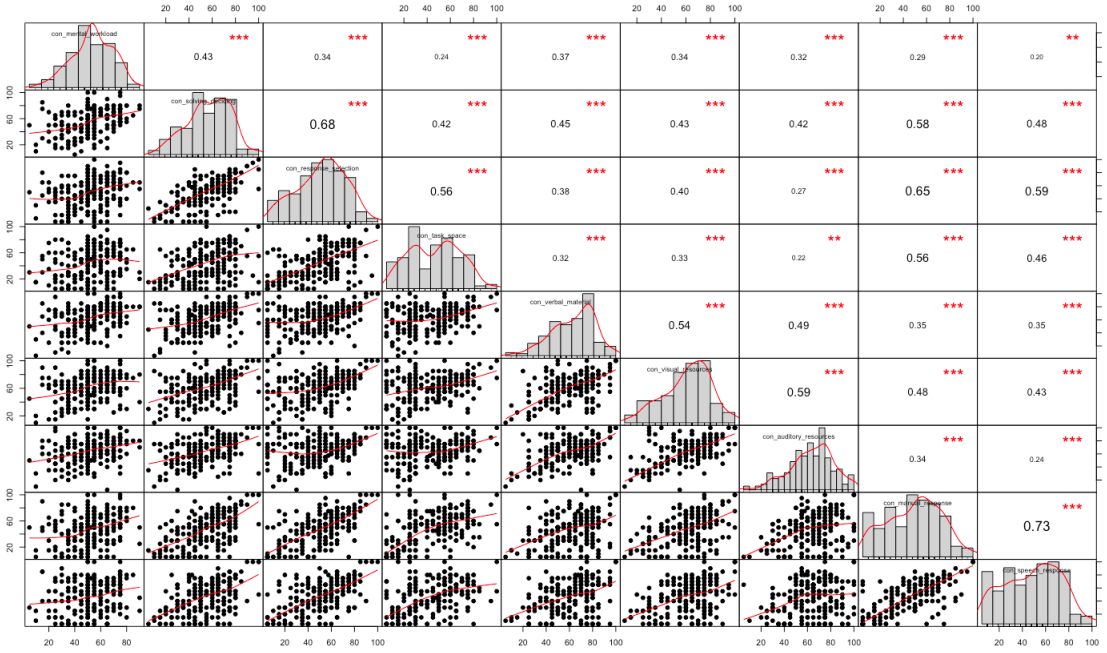

Figure 4.9: Workload Profile Scatter plot matrix (N=217)

Table 4.5: Correlation of Mental Workload score & factors in WP set

	MWL	Solv	Resp	TaSpa	Verb	Visu	Audi	Manu	Spee
MWL	1	**0.43**	**0.34**	**0.24**	**0.37**	**0.34**	**0.32**	**0.293**	0.20
Solv		1	**0.68**	**0.42**	**0.45**	**0.43**	**0.42**	**0.58**	**0.48**
Resp			1	**0.56**	**0.38**	**0.40**	**0.27**	**0.65**	**0.59**
TaSpa				1	**0.32**	**0.33**	**0.22**	**0.56**	**0.47**
Verb					1	**0.54**	**0.49**	**0.35**	**0.35**
Visu						1	**0.59**	**0.48**	**0.44**
Audi							1	**0.34**	**0.24**
Manu								1	**0.73**
Spee									1

MWL: Mental Workload, Solv: Solving, Resp: Response, TaSpa: Task&Space, Verb: Verbal, Visu: Visual, Audi: Auditory, Manu: Manual, Spee: Speech

Table 4.5 and figure 4.9 showed a moderate relationship between Mental Workload and Central Processing (Solv), Response Processing (Resp), Verbal Processing (Verb),

Visual Input (Visual), Auditory Input (Audi); and a weak one with Spatial Processing (TaSpa), Manual Response (Manu) and Speech response (Spee).

Extended Feature Sets

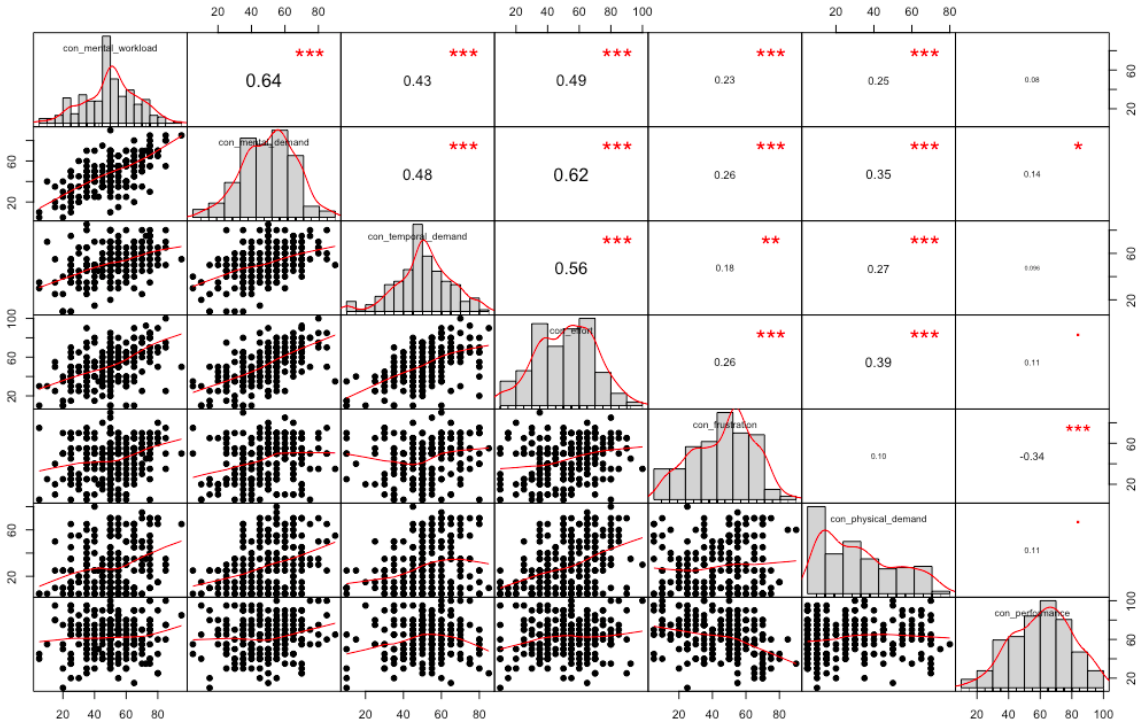

Figure 4.10: EFS Scatter plot matrix with NASA factors (N=237)

Table 4.6: Correlation of Mental Workload score & NASA factors in EFS set

	MWL	Ment	Phys	Temp	Perf	Frus	Effo
MWL	1	**0.65**	**0.25**	**0.43**	0.08	**0.23**	**0.49**
Ment		1	**0.35**	**0.48**	0.14	**0.26**	**0.62**
Phys			1	**0.27**	0.11	0.10	**0.39**
Temp				1	0.10	0.18	**0.56**
Perf					1	**-0.34**	0.11
Frus						1	**0.26**
Effo							1

MWL: Mental Workload, Ment: Mental, Phys: Physical, Temp: Temporal, Perf: Performance, Frus: Frustration, Effo: Effort

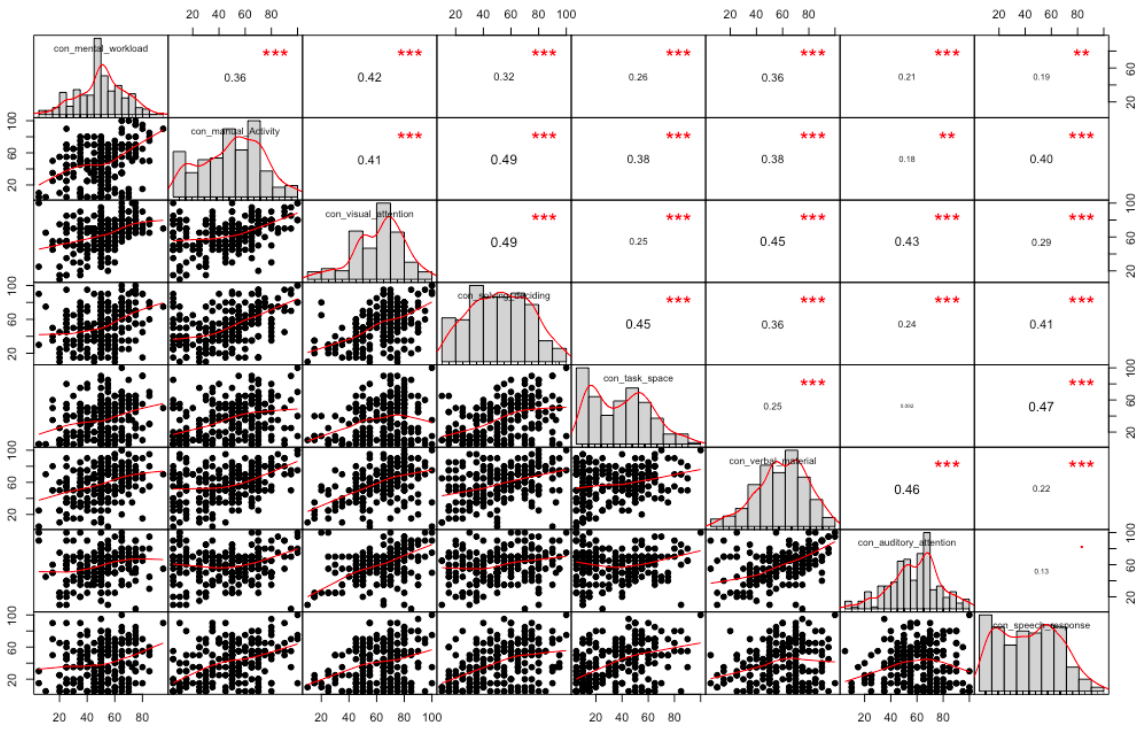

Figure 4.11: EFS Scatter plot matrix with WP factors (N=237)

Table 4.7: Correlation of Mental Workload score & WP factors in EFS set

	MWL	Solv	TaSpa	Verb	Visu	Auditory	Manu	Spee
MWL	1	**0.32**	**0.26**	**0.36**	**0.42**	**0.21**	**0.36**	0.19
Solv		1	**0.45**	**0.36**	**0.49**	**0.24**	**0.49**	**0.41**
TaSpa			1	**0.25**	**0.25**	0.09	**0.38**	**0.47**
Verb				1	**0.45**	**0.47**	**0.38**	**0.22**
Visu					1	**0.43**	**0.41**	**0.29**
Audi						1	0.18	0.13
Manu							1	**0.40**
Spee								1

MWL: Mental Workload, Solv: Solving, TaSpa: Task&Space, Verb: Verbal, Visu: Visual, Audi: Auditory, Manu: Manual, Spee: Speech

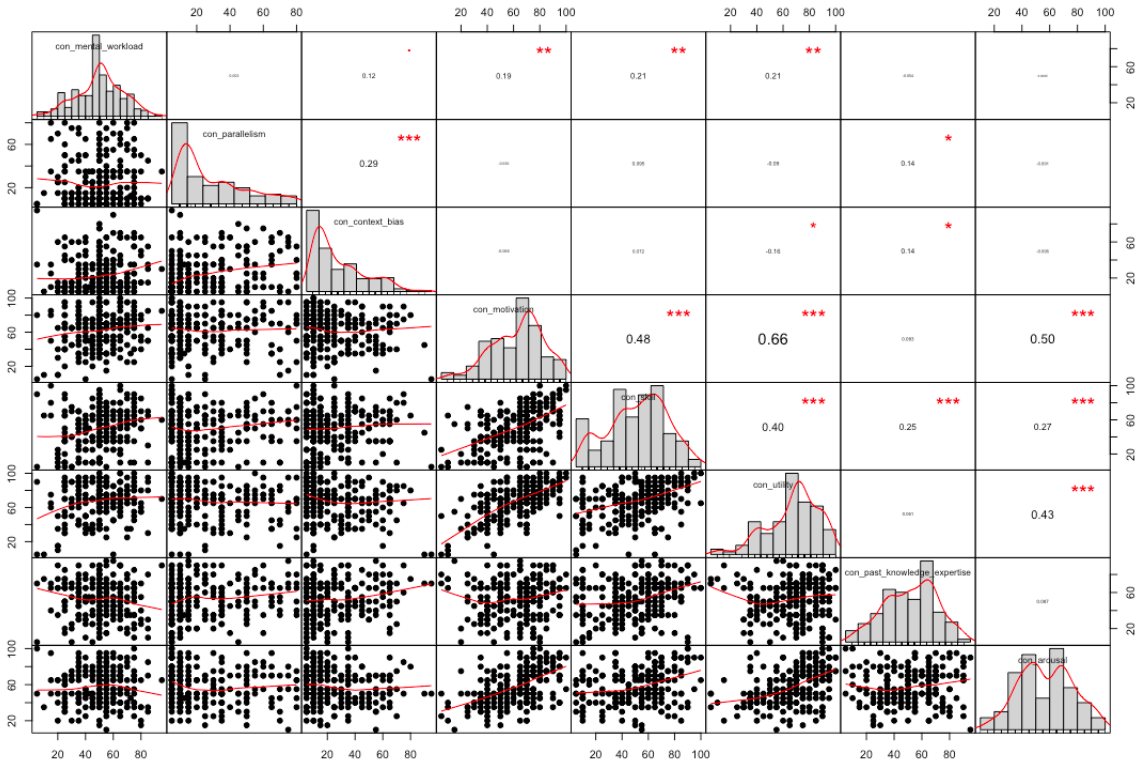

Figure 4.12: EFS Scatter plot matrix with additional factors (N=237)

Table 4.8: Correlation of Mental Workload score & additional factors in EFS set

	MWL	Para	Cont	Moti	Skil	Util	Know	Arou
MWL	1	0.003	0.12	0.19	0.21	0.21	-0.05	0.01
Para		1	**0.29**	-0.04	0.10	-0.09	0.14	-0.09
Cont			1	-0.07	0.07	-0.17	0.14	-0.10
Moti				1	**0.48**	**0.66**	0.08	**0.50**
Skil					1	**0.40**	**0.25**	**0.27**
Util						1	0.06	**0.43**
Know							1	0.09
Arou								1

MWL: Mental Workload, Para: Parallelism, Cont: Context, Moti: Motivation, Skil: Skill, Util: Utility, Know: Knowledge, Arou: Arousal

For the relevant factors in NASA, table 4.6 and figure 4.10 showed a strong relationship between Mental Workload and Mental demands, a moderate one with Temporal demands and Effort and a weak one with Physical demands, Frustration.

For the relevant factors in WP, table 4.7 and figure 4.11 showed a moderate relationship between Mental Workload and Central Processing (Solv), Verbal Processing (Verb), Visual Input (Visual), Manual Response (Manu); and a weak one with Spatial Processing (TaSpa), Auditory Input (Audi) and Speech response (Spee).

For additional factors, table 4.8 and figure 4.12 showed a weak relationship between Mental workload and "interruptions during the teaching session" (Cont), "motivated by teaching session" (Moti), "skill have no influence or help" (Skil), "teaching session useful for learning" (Util).

Above all, in consideration of numeric models, there was a strong relationship between Mental Workload and Mental demand, moderate relationship with Temporal demand, Effort, Frustration in NASA set (table 4.4) and in EFS which inherited a part of NASA set (table 4.6). This result was in line with the discovery in the paper (Reid & Nygren, 1988), MWL can be largely explained by three component factors: Time Load, Mental Effort Load, and Psychological Stress Load.

To WP set (table 4.5), a moderate relationship was found between Mental Workload and Central processing, Response processing, Verbal processing, Visual input and Auditory input; the same as in EFS set with relevant factors to WP (table 4.7), and Manual response. The additional factors of EFS set had no relationship with Mental Workload score (table 4.8).

4.2.2 Difference between Mental Workload score of three rating scales in training datasets

Does the difference in training datasets of each rating scales help to overrule the variation of datasets in building learning models? If the training sets were not different, repeated sampling would give more samples for training and the test sets would be the same for ten samples.

NASA Task Load Index

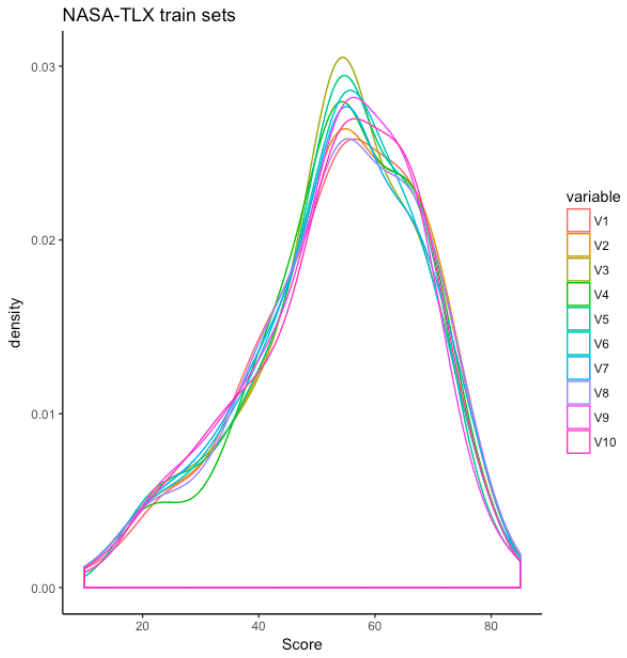

Figure 4.13: Histogram of 10 NASA training sets (N=154)

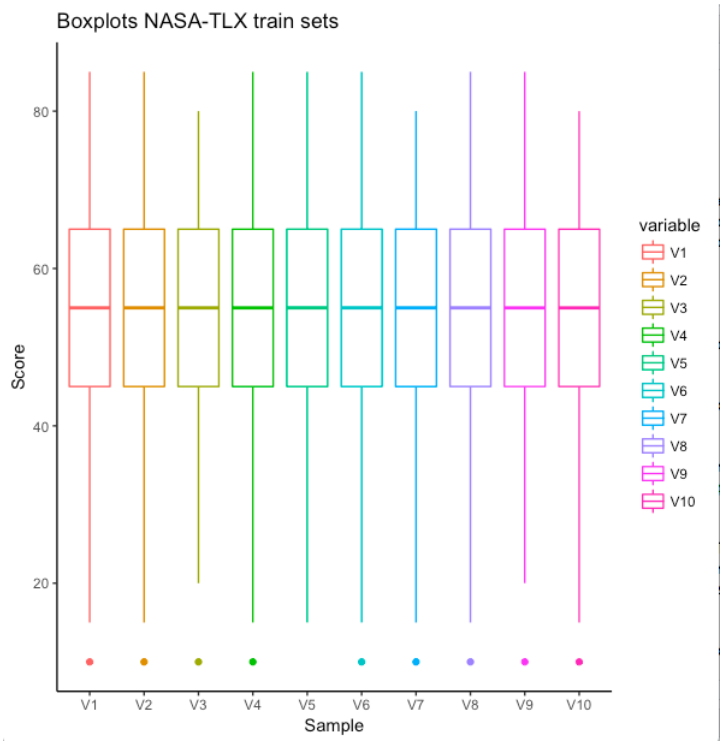

Figure 4.14: Box plot of 10 NASA training sets (N=154)

The distribution of NASA set was non-normality (section 4.1.1). So far, the whole data which divided into the training set and test set with the ratio 70:30 was as left-skewed as non-normality (figure 4.13). The differences of 10 training sets were not significant difference shown in boxplot (figure 4.14)

Workload Profile

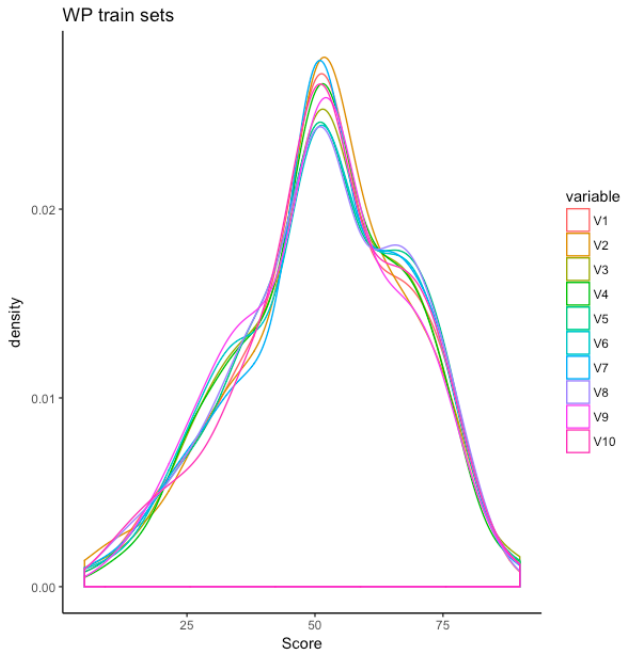

Figure 4.15: Histogram of 10 WP training sets (N=153)

As described above (section 4.1.2), the distribution of WP set was non-normality. So, the whole data which divided into the training set and test set as 70:30 would be slightly right-skewed non-normality (figure 4.15). The difference of 10 training sets were not significant, as per the difference shown in boxplot (figure 4.16)

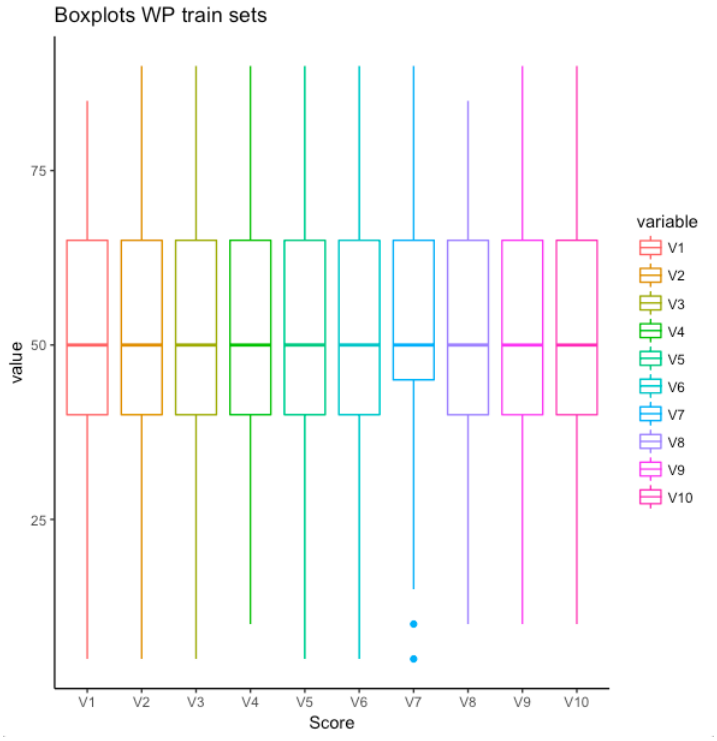

Figure 4.16: Box plot of 10 WP training sets (N=153)

Extended Feature Sets

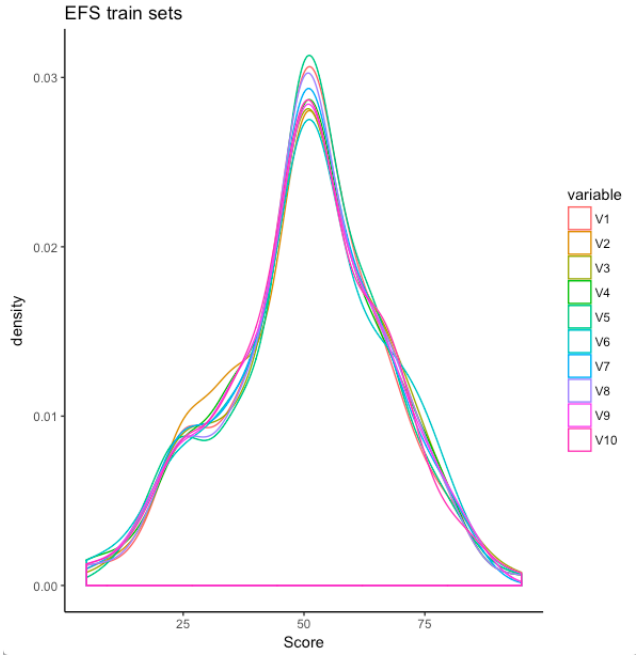

Figure 4.17: Histogram of 10 EFS training sets (N=155)

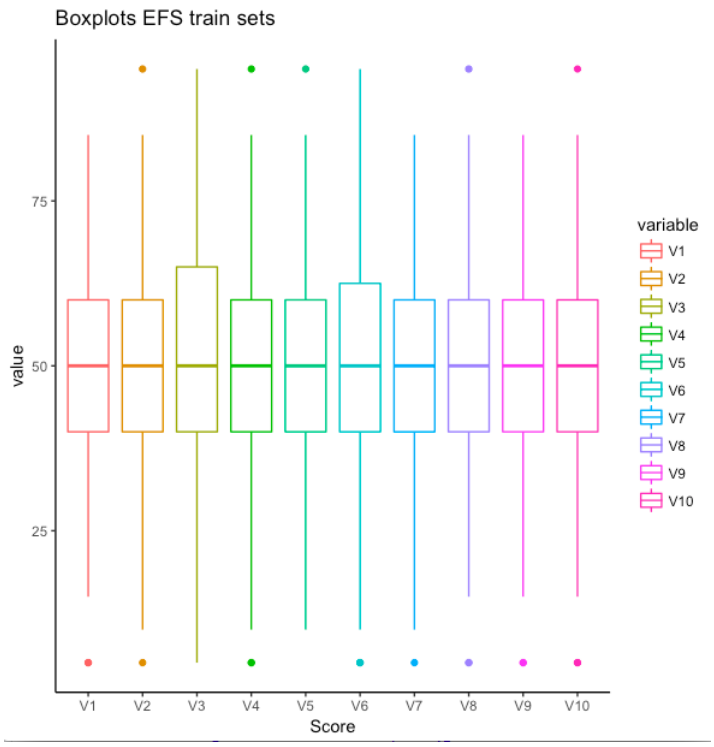

Figure 4.18: Box plot of 10 EFS training sets (N=155)

As stated above (section 4.1.3), the distribution of EFS set was non-normality. So far, the whole data, which divided into training set and test set as 70:30, were left-skewed non-normality (figure 4.17). The difference of 10 training sets was not a significant difference, as shown in boxplot (figure 4.18).

4.3 Model training

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. A single parameter called "k" splits a given data sample into the number of groups. As such, the procedure is often called k-fold cross-validation, such as k=10 becoming 10-fold cross-validation. A value of k=10 is prevalent in the field of applied machine learning that has been found through experimentation to generally result in a model skill estimate with low bias a modest variance (Max & Kjell, 2013). However, there are some variations on the k-fold cross-validation. They are commonly known as [1] train/test split, [2] LOOCV (leave-one-out cross-validation),

[3] stratified - the splitting of data into folds by criteria such as each fold has the same proportion of observations with a given categorical value/ class outcome value, and [4] repeated - the k-fold cross-validation repeated n times where the sample is shuffled prior to each repetition resulting in a different split of the sample.

These models trained below had the 10-fold and 10-repeated times cross-validation which meant ten groups split and repeated ten times in each below model. However, the "stratified" method of k-fold cross-validation should be taken into account for improvisation as there were still missing values during the training time in some cases (table 4.14, 4.16, 4.17, 4.34). So that, the repeated cross-validation may lead to inaccurate learning models on categorical outcome value.

4.3.1 NASA Task Load Index

Multiple linear regression

```
> summary(nasa_reg_cv)

Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-29.6408  -6.4733   0.3215   7.4853  22.6795

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     26.70127    3.35762   7.952 4.47e-13 ***
con_mental       0.40272    0.07137   5.643 8.34e-08 ***
con_physical    -0.07846    0.05778  -1.358   0.1766
con_temporal     0.11241    0.06660   1.688   0.0936 .
con_performance -0.06090    0.06213  -0.980   0.3285
con_frustration  0.05625    0.05666   0.993   0.3225
con_effort       0.11390    0.06508   1.750   0.0822 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.71 on 147 degrees of freedom
Multiple R-squared:  0.4119,    Adjusted R-squared:  0.3879
F-statistic: 17.16 on 6 and 147 DF,  p-value: 5.628e-15
```

Figure 4.19: NASA training result of sample 1 cross-validation (10 times, 10 folds)

Table 4.9: Summary of NASA training result of 10 samples cross-validation (10 times, 10 folds)

Sample	R-squared	$_{Adj}$R-squared	RMSE	Residual SE	Significant vars
#1	0.41	0.39	11.86	11.71	Mental demand
#2	0.39	0.36	12.44	12.14	Mental, Temporal demand
#3	0.36	0.34	12.03	11.85	Mental, Temporal demand
#4	0.39	0.37	11.85	11.67	Mental demand
#5	0.38	0.35	11.75	11.64	Mental demand, Effort
#6	0.34	0.31	12.55	12.35	Mental demand, Frustration
#7	0.36	0.33	12.52	12.34	Mental, Temporal demand
#8	0.35	0.33	12.67	12.45	Mental demand
#9	0.31	0.28	12.65	12.44	Mental demand
#10	0.45	0.43	11.44	11.3	Mental demand, Frustration

The overall significance (p_{value}) of training sets are all <0.001

In Linear Regression, sample train 10 had the highest R-squared and the lowest RMSE (R-squared = 0.45, RMSE = 11.44, table 4.9).

The model was trained by multiple linear regression with cross-validation resampling (10 folds, repeated 10 times). Based on the result, first of all, the selected model as linear regression was correct. The constant in a regression model guaranteed that the residuals had a mean of zero and prevented the bias of regression coefficients and predictions. That was because the constant value (intercept) had a statistically significant difference and Mental demand was the statistically significant factor on the model ($p_{value} < 0.001$) which followed the equation of linear form. The p_{value} for the F-test of overall significance tests was less than the significance level (0.05) affirmed rejecting the null-hypothesis and concluding that the model provided a better fit than the intercept-only model.

Table 4.9 showed the average result of R-squared ranged from 0.31 to 0.45 or $_{Adj}$R-squared from 0.28 to 0.43, RMSE ranged from 11.44 to 12.67. R-squared explained the impact of the linear model on average of 31 to 45% of the Mental Workload score variation. In other words, the Mental demand had 41% impact on Mental Workload score (training sample 1, tab4.9).

Table 4.10: Variable importance of NASA in model

Variables	Overall (%)	Variables	Overall (%)
Mental demand	100.00	Frustration	0.27
Temporal demand	15.17	Performance	0.00
Physical demand	8.10		

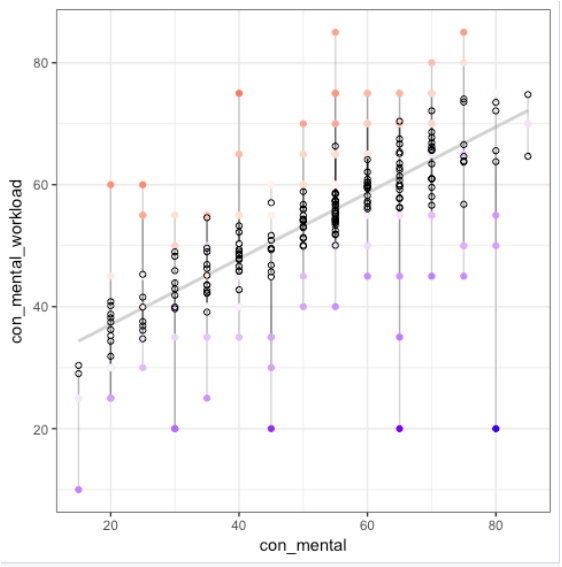

Figure 4.20: Correlation of MWL and Mental demand in NASA set

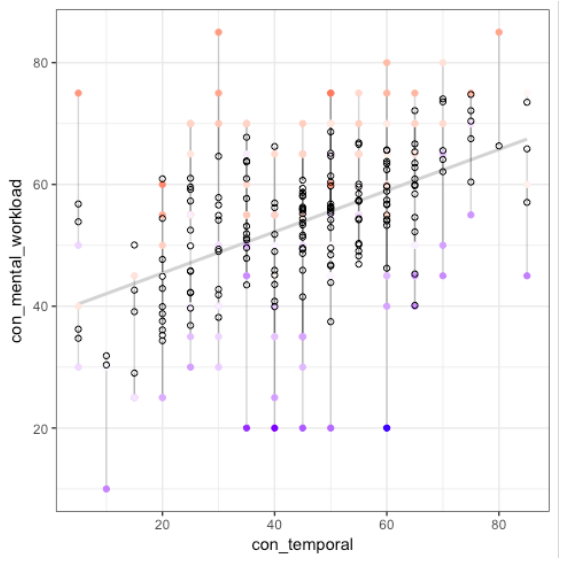

Figure 4.21: Correlation of MWL and Temporal demand in NASA set

In 10 training sets, the relationship between Mental Workload score and significant variables was investigated using a Pearson correlation. The model yielded a high correlation between model predictions and important variables, in other words, the models with more important variables were higher R-squared 4.10. A strong positive correlation was found (R-squared=0.41, n=153, $p<0.001$, training sample 1, table 4.9).

R-squared alone cannot determine whether the coefficient estimates and predictions are biased, which need to assess the residual plots using residual plots to determine if the difference between the expected value and the observed. The difference must be unpredictable. The residuals (observed error) should not be either systematically high or low. So, the residuals should be in the center of zero throughout the range of fitted values. In other words, the model is correct on average for all fitted values. Further, in the context of ordinary least squares, random errors are assumed to produce residuals that normally distributed. Therefore, the residuals should fall in a symmetrical pattern and have a constant spread throughout the range.

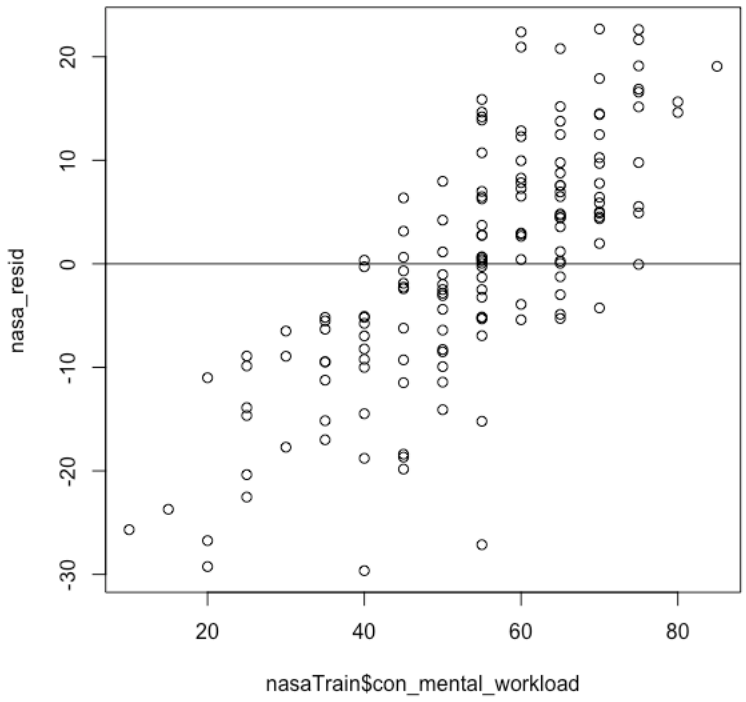

Figure 4.22: Training of NASA in Residual plot

Looking at figure 4.22, first of all, there was the difference between the expected

value and the observed; secondly, there were some points the residuals was in the range of 40 to 75 (median=55); lastly, the residuals were not normally distributed and not spread continuously throughout the range (seen histogram in the view of vertical axis).

Scatter plots of Actual vs. Predicted are one of the most productive forms of data visualization. All points should be more or less close to a regressed diagonal line. So, if the Actual is 20, the predicted should be reasonably close to 20. If the Actual is 80, the anticipated should also be reasonably close to 80. So, drawing such a diagonal line within the graph to check out where the points lie. The lower the R-squared, the weaker the Goodness of fit of the model, the foggier or dispersed the points are (away from this diagonal line).

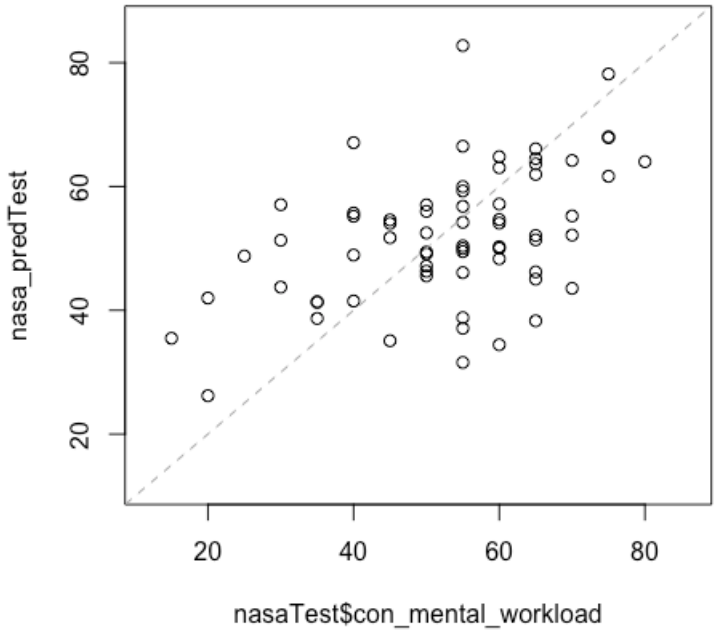

Figure 4.23: Training of NASA in comparison of Actual & Predicted values

The model above had an average R-squared (0.4) which meant 40 percent of the points would be close to this diagonal line. The model had three subsections of performance. The first one was where Actuals <40. Within this zone, the model did not look as same as the Actuals. It significantly overestimated the Actual values. The second one was when Actuals were between 40 and 75, the model mostly concentrated within

this zone whereas some points were still random. The third zone was for Actuals >75 having no data for prediction. There was virtually no relationship between model's predicted values and Actuals.

Decision Tree Information Gain

Table 4.11: Summary of NASA training result of 10 samples cross-validation (10 times, 10 folds)

Sample	Accuracy	Tuning parameters		
		trials	model	winnow
#1	0.645	1	tree	F
#2	0.580	10	tree	T
#3	0.624	1	rules	F
#4	0.643	10	rules	F
#5	0.582	20	tree	F
#6	0.579	1	rules	F
#7	0.616	1	rules	F
#8	0.565	10	rules	F
#9	0.644	1	rules	F
#10	0.627	10	tree	F

The best tuning parameters among the ten samples was sample 1 (trials=1, model=tree, winnow=F) having the highest accuracy = 0.645.

Table 4.12: Summary of NASA Information Gain training result of 10 Up-sampling cross-validation (10 times, 10 folds)

Sample	AUC	Accuracy	F1	Precision	Recall	True Pos Rate	True Neg Rate
#1	0.89	0.735	0.71	0.73	0.74	0.73	0.92
#2	0.91	0.755	0.74	0.75	0.76	0.75	0.92
#3	0.83	0.73	0.71	0.73	0.73	0.73	0.92
#4	0.92	0.784	0.77	0.79	0.78	0.79	0.93
#5	0.92	0.754	0.74	0.76	0.75	0.76	0.92
#6	0.85	0.76	0.74	0.76	0.76	0.76	0.93
#7	0.84	0.729	0.71	0.73	0.73	0.73	0.92
#8	0.91	0.784	0.77	0.79	0.78	0.79	0.93
#9	0.85	0.758	0.74	0.76	0.76	0.76	0.93
#10	0.90	0.782	0.77	0.79	0.78	0.79	0.93

AUC: Area Under the Curve,

True Pos Rate: True Positive Rate, True Neg Rate: True Negative Rate

In Decision Tree Information Gain, sample train 8 was the most optimal (Accuracy = 0.784, Precision = 0.793, Recall = 0.784, table 4.12).

Table 4.13: Summary of NASA Information Gain training result of 10 Up-sampling cross-validation on each class

Sample	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Underload										
Accuracy	0.76	0.83	0.76	0.81	0.75	0.78	0.75	0.78	0.78	0.73
Precision	0.69	0.79	0.68	0.77	0.71	0.74	0.72	0.76	0.72	0.73
Recall	1	0.99	1	0.99	0.92	0.96	0.92	0.94	0.99	0.99
Optimal load 1										
Accuracy	0.60	0.63	0.6	0.64	0.63	0.63	0.59	0.64	0.61	0.64
Precision	0.62	0.63	0.59	0.68	0.66	0.62	0.54	0.7	0.67	0.7
Recall	0.40	0.53	0.42	0.57	0.51	0.54	0.4	0.53	0.45	0.53
Optimal load 2										
Accuracy	0.68	0.62	0.63	0.65	0.65	0.64	0.65	0.68	0.64	0.67
Precision	0.66	0.63	0.68	0.71	0.7	0.71	0.66	0.71	0.66	0.74
Recall	0.71	0.5	0.5	0.58	0.58	0.54	0.6	0.67	0.59	0.6
Overload										
Accuracy	0.92	0.9	0.93	0.95	0.92	0.94	0.94	0.94	0.9	0.935
Precision	0.91	0.9	0.92	0.94	0.92	0.93	0.94	0.94	0.94	0.93
Recall	1	1	1	1	1	1	1	1	1	1

The Accuracy, Precision, and Recall in underload and overload levels were quite high, but the samples in these levels were up-sampled. So, the results may be affected. The optimal load 2 with the three indicators evaluated was generally higher than optimal load 1.

Decision Tree Gini Regression

Gini impurity is a measure of misclassification, which applies in a multiclass classifier context. "The complexity parameter (cp) which is a tuning parameter, is also used to control the size of the decision tree and to select the optimal tree size. If the cost of adding another variable to the decision tree from the current node is above the value of cp, then tree building does not continue".

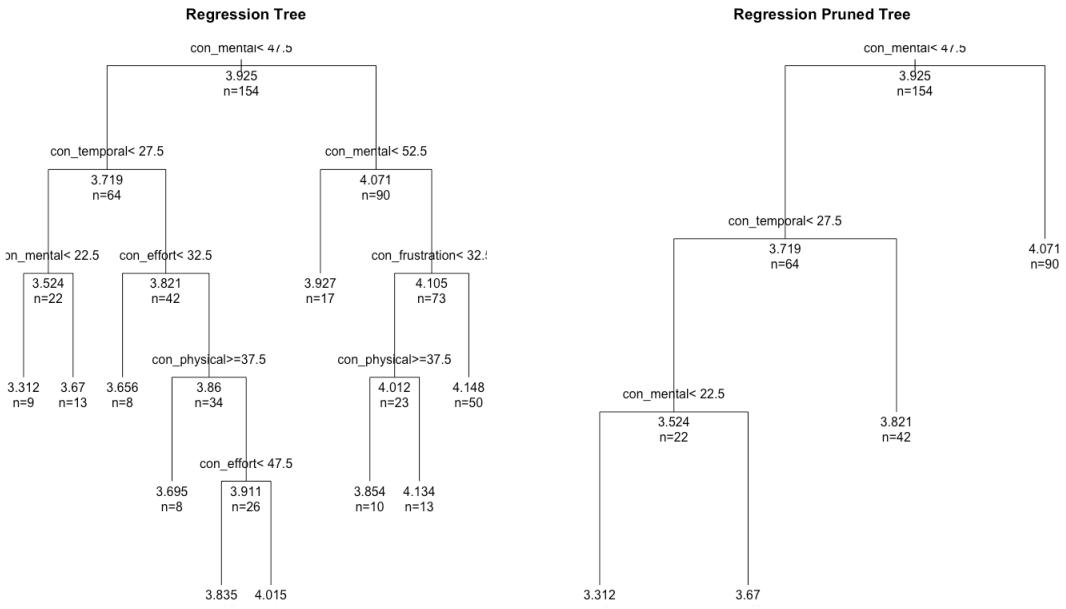

Figure 4.24: NASA decision tree Gini Regression pruned at cp=0.04157 (N=154)

The regression tree of NASA was pruned from 9 branches into 3 branches which was split by the main variables Mental demands and Temporal (illustrated by training sample 1).

Table 4.14: Summary of NASA Gini Regression training result of 10 samples cross-validation (10 times, 10 folds)

Sample	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
cp	0.04	0.08	0.05	0.05	0.04	0.07	0.05	0.03	0.07	0.05
R-squared	0.38	0.31	0.26	0.35	0.25	0.26	0.28	0.31	0.22	
RMSE	12.05	12.96	12.94	12.11	12.91	13.32	13.30	13.05	13.37	12.77
MAE	9.66	10.39	10.51	9.94	10.42	10.70	10.63	10.36	10.67	10.25

cp: complexity parameter (taken exactly 7 decimal points to prune the tree)

In Decision Tree Gini Regression, sample 1 (R-squared = 0.38, RMSE = 12.05, table 4.14) was the optimal model.

Decision Tree Gini Classification

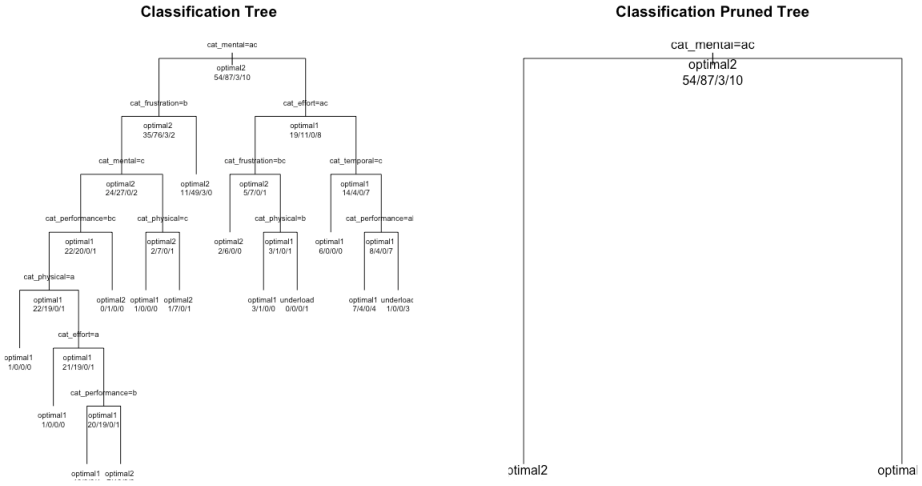

Figure 4.25: NASA decision tree Gini Classification pruned at cp=0.02985 (N=154)

The classification tree of NASA was pruned from 13 branches into 1 branch which was split by the main variables Mental demands category (illustrated by training sample 1).

Table 4.15: Summary of NASA Gini Classification training result of 10 samples cross-validation (10 times, 10 folds)

Sample	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
cp	0.04	0	0.02	0.02	0.12	0	0.03	0	0.02	0.02
Accuracy	0.65	0.58	0.61	0.60	0.55	0.57	0.59	0.55	0.59	0.63
Kappa	0.27	0.17	0.20	0.18	-0.003	0.14	0.15	0.13	0.13	0.26

cp: complexity parameter (taken exactly 7 decimal points to prune the tree)

The best tuning parameters among the ten samples was sample 1 (cp=0.038) having the highest accuracy = 0.649.

Table 4.16: Summary of NASA Gini Classification training result of 10 Up-sampling cross-validation (10 times, 10 folds)

Sample	AUC	Accuracy	F1	Precision	Recall	True Pos Rate	True Neg Rate
#1	0.82	0.619	0.58	0.61	0.62	0.61	0.89
#2	0.88	0.712	0.69	0.70	0.71	0.70	0.91
#3	0.85	0.692	0.67	0.70	0.69	0.70	0.91
#4	0.88	0.709	0.69	0.72	0.71	0.72	0.91
#5	0.69	0.450	//	//	0.45	//	0.85
#6	0.86	0.718	0.70	0.73	0.72	0.73	0.92
#7	0.85	0.653	0.64	0.68	0.65	0.68	0.89
#8	0.88	0.734	0.72	0.74	0.73	0.74	0.92
#9	0.86	0.690	0.66	0.70	0.69	0.70	0.91
#10	0.88	0.714	0.69	0.71	0.71	0.71	0.91

AUC: Area Under the Curve,

True Pos Rate: True Positive Rate, True Neg Rate: True Negative Rate

In Decision Tree Gini Classification was sample 8 having the most accurate (Accuracy = 0.734, Precision = 0.735, Recall = 0.734, table 4.16).

Table 4.17: Summary of NASA Gini Classification training result of 10 Up-sampling cross-validation on each class

Sample	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Underload										
Accuracy	0.69	0.77	0.71	0.72	0.60	0.74	0.66	0.73	0.71	0.73
Precision	0.58	0.74	0.62	0.65	0.39	0.66	0.60	0.67	0.60	0.66
Recall	0.93	0.93	0.94	0.92	0.74	0.95	0.73	0.90	0.97	0.92
Optimal load 1										
Accuracy	0.57	0.60	0.58	0.59	0.51	0.62	0.58	0.60	0.58	0.59
Precision	0.46	0.59	0.52	0.52	0.30	0.64	0.41	0.59	0.60	0.57
Recall	0.34	0.44	0.39	0.43	0.06	0.48	0.45	0.45	0.36	0.40
Optimal load 2										
Accuracy	0.55	0.61	0.61	0.63	0.50	0.61	0.61	0.66	0.60	0.64
Precision	0.61	0.59	0.71	0.72	//	0.69	0.72	0.73	0.64	0.69
Recall	0.22	0.47	0.44	0.48	0	0.45	0.43	0.58	0.43	0.53
Overload										
Accuracy	0.80	0.86	0.91	0.93	0.68	0.88	0.93	0.91	0.91	0.90
Precision	0.74	0.83	0.90	0.93	0.53	0.86	0.93	0.90	0.90	0.88
Recall	1	1	1	1	1	1	1	1	1	1

The Accuracy, Precision, and Recall in underload and overload levels were quite high, but the samples in these levels were up-sampled. So, the results may be affected. Besides, the Accuracy and Recall of the optimal load 2 were as same as the optimal load 1, but the higher Precision in optimal load 2.

4.3.2 Workload Profile

Multiple Linear Regression

```
> summary(wp_reg_cv)

Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-43.744  -7.577  -0.683  11.081  27.180

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)            23.322604   5.547451   4.204 4.58e-05 ***
con_solving_deciding    0.165724   0.090926   1.823   0.0704 .
con_response_selection  0.042937   0.092420   0.465   0.6429
con_task_space          0.003794   0.070084   0.054   0.9569
con_verbal_material     0.133417   0.086265   1.547   0.1242
con_visual_resources    0.087036   0.095624   0.910   0.3642
con_auditory_resources  0.067041   0.091492   0.733   0.4649
con_manual_response     0.034845   0.088929   0.392   0.6958
con_speech_response    -0.058296   0.075758  -0.770   0.4429
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.1 on 144 degrees of freedom
Multiple R-squared:  0.1825,    Adjusted R-squared:  0.1371
F-statistic: 4.018 on 8 and 144 DF,  p-value: 0.0002483
```

Figure 4.26: WP training result of sample 1 cross-validation (10 times, 10 folds)

Table 4.18: Summary of WP training result of 10 samples cross-validation (10 times, 10 folds)

Sample	R-squared	$_{A}dj$R-squared	RMSE	Residual SE	Significant vars
#1	0.18	0.14	15.45	15.10	
#2	0.25	0.21	14.91	14.68	Verbal Material
#3	0.26	0.21	15.23	14.81	Auditory resources
#4	0.23	0.19	14.90	14.48	Solving Deciding
#5	0.19	0.15	15.68	15.37	Solving Deciding
#6	0.24	0.20	15.32	15.01	Solving Deciding
#7	0.28	0.24	14.82	14.52	Solving Deciding
#8	0.23	0.19	15.33	14.96	
#9	0.24	0.20	14.85	14.59	Solving Deciding
#10	0.23	0.19	15.16	14.85	

The overall significance (p_{value}) of training sets are all <0.001

In Linear Regression, sample train 7 had the highest R-squared and the lowest RMSE (R-squared = 0.28, RMSE = 14.82, table 4.18).

Table 4.18 showed the average result of R-squared ranged from 0.18 to 0.28 or $_{Adj}$R-squared from 0.14 to 0.24, RMSE ranged from 14.82 to 15.68. R-squared explained the impact of linear model on average of 18 to 28% of the Mental Workload score variation. In other words, the Central Processing (Solving Deciding) had 23% impact on Mental Workload score (training sample 4, table 4.18).

In 10 training sets, the relationship between Mental Workload score and significant variables was investigated using a Pearson correlation. The model yielded a high correlation between model predictions and important variables. In other words, the models with important variables were higher R-squared (table 4.19). A strong positive correlation was found (R-squared=0.23, n=153, p<0.001, training sample 4, table 4.18)

Table 4.19: Variable importance of WP in model

Variables	Overall (%)	Variables	Overall (%)
Solving & Deciding	100.00	Auditory resources	38.37
Verbal material	84.39	Response selection	23.21
Visual resources	48.41	Manual response	19.10
Speech response	40.45	Task space	0.00

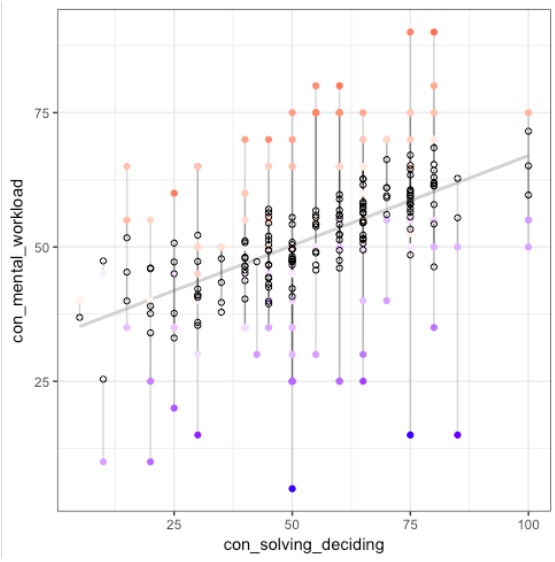

Figure 4.27: Correlation of MWL and Central Processing (Solving&Deciding) in WP

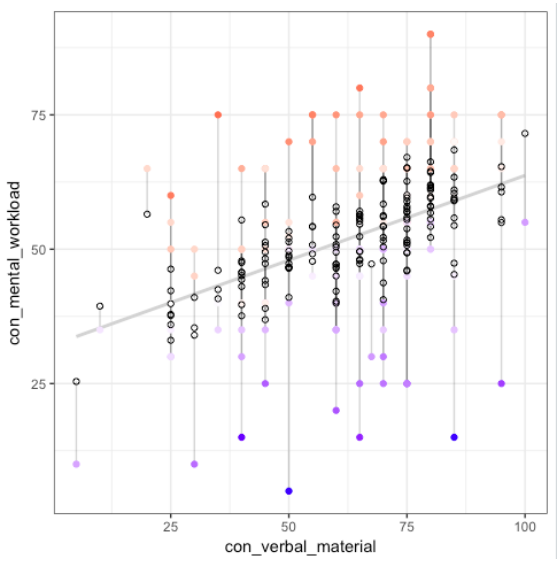

Figure 4.28: Correlation of MWL and Verbal material in WP

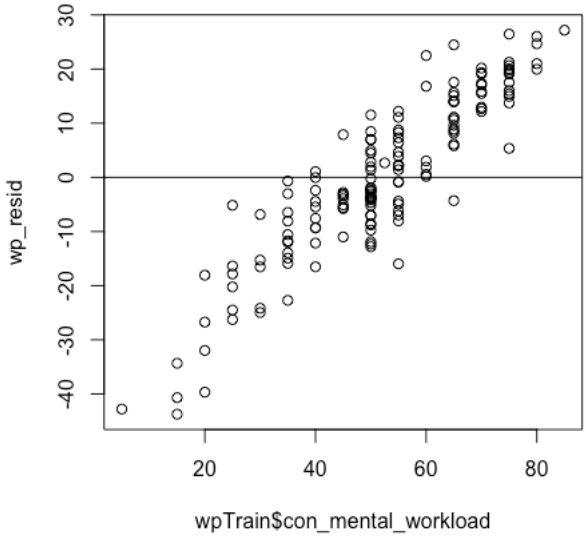

Figure 4.29: Training of WP in Residual plot

Looking at figure 4.29, there were three main points. Firstly, there was a difference between the expected value and the observed. Secondly, there were only some points the residuals centered on zero (median=55) in the range of 35 to 60 (narrower than in NASA residual plot, figure 4.22. Lastly, the residuals did not normally distribute but

slightly constant spread throughout the range (seeing histogram in the vertical axis view).

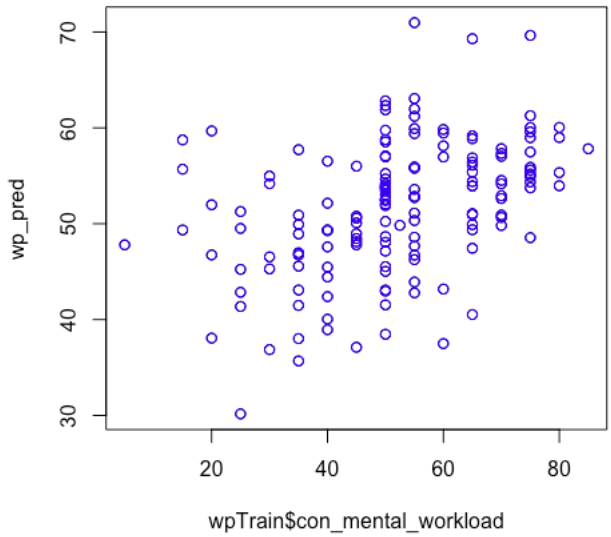

Figure 4.30: Training of WP in comparison of Actual & Predicted values

The model above had an average R-squared (0.23) which meant 23 percent of the points would be close to the diagonal line. The model had three subsections of performance. The first one was where Actuals <35. Within this zone, the model did not look as same as the Actuals. It significantly overestimated the Actual values. The second one was when Actuals between 40 and 75, the model mostly concentrated within this zone whereas some points were still random. The third zone was for underestimated Actuals >75. There was virtually no relationship between model's predicted values and Actuals.

Decision Tree Information Gain

Table 4.20: Summary of WP training result of 10 samples cross-validation (10 times, 10 folds)

Sample	Accuracy	Tuning parameters		
		trials	model	winnow
#1	0.453	10	rules	F
#2	0.516	10	rules	T
#3	0.456	1	tree	T
#4	0.498	1	tree	F
#5	0.495	10	rules	T
#6	0.508	10	tree	T
#7	0.491	1	rules	F
#8	0.522	10	rules	F
#9	0.518	1	rules	T
#10	0.495	10	tree	T

The best tuning parameters among the ten samples was sample 8 (trials=10, model=rules, winnow=F) having the highest accuracy = 0.522.

Table 4.21: Summary of WP training result of 10 Up-sampling cross-validation (10 times, 10 folds)

Sample	AUC	Accuracy	F1	Precision	Recall	True Pos Rate	True Neg Rate
#1	0.86	0.689	0.67	0.67	0.69	0.67	0.91
#2	0.85	0.655	0.63	0.62	0.66	0.62	0.90
#3	0.83	0.649	0.63	0.63	0.65	0.63	0.89
#4	0.82	0.642	0.62	0.62	0.64	0.62	0.89
#5	0.85	0.648	0.63	0.63	0.65	0.63	0.89
#6	0.85	0.657	0.64	0.64	0.66	0.64	0.89
#7	0.79	0.648	0.63	0.62	0.65	0.62	0.89
#8	0.86	0.669	0.66	0.66	0.67	0.66	0.90
#9	0.80	0.680	0.67	0.67	0.68	0.67	0.90
#10	0.84	0.650	0.64	0.62	0.65	0.62	0.89

AUC: Area Under the Curve,

True Pos Rate: True Positive Rate, True Neg Rate: True Negative Rate

In Decision Tree Information Gain, sample train 1 was the most optimal (Accuracy = 0.689, Precision = 0.665, Recall = 0.688, table 4.21).

Table 4.22: Summary of WP training result of 10 Up-sampling cross-validation on each class

Sample	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Underload										
Accuracy	0.79	0.78	0.75	0.73	0.72	0.71	0.73	0.75	0.75	0.76
Precision	0.75	0.75	0.71	0.68	0.68	0.68	0.67	0.72	0.73	0.71
Recall	0.98	0.95	0.90	0.89	0.86	0.84	0.90	0.88	0.8	0.958
Optimal load 1										
Accuracy	0.58	0.57	0.58	0.57	0.58	0.58	0.58	0.57	0.58	0.57
Precision	0.49	0.44	0.45	0.43	0.48	0.47	0.48	0.45	0.48	0.45
Recall	0.4	0.33	0.42	0.38	0.42	0.37	0.40	0.36	0.42	0.33
Optimal load 2										
Accuracy	0.58	0.57	0.56	0.56	0.56	0.59	0.56	0.59	0.59	0.57
Precision	0.53	0.46	0.45	0.48	0.45	0.55	0.43	0.52	0.54	0.45
Recall	0.372	0.35	0.28	0.30	0.31	0.42	0.28	0.43	0.41	0.33
Overload										
Accuracy	0.88	0.84	0.86	0.86	0.87	0.85	0.88	0.89	0.88	0.86
Precision	0.88	0.80	0.84	0.84	0.85	0.82	0.87	0.88	0.87	0.84
Recall	1	1	1	1	1	1	1	1	1	1

The Accuracy, Precision, and Recall in underload and overload levels were quite high, but the samples in these levels were up-sampled. So, the results may be affected. The optimal load 2 had the Accuracy and Precision were equal or higher than optimal load 1 but lower in Recall.

Decision Tree Gini Regression

Table 4.23: Summary of WP Gini Regression training result of 10 samples cross-validation (10 times, 10 folds)

Sample	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
cp	0.05	0.04	0.05	0.05	0.05	0.05	0.04	0.05	0.05	0.06
R-squared	0.17	0.22	0.26	0.18	0.24	0.27	0.23	0.18	0.21	0.17
RMSE	15.12	15.04	14.84	15.01	14.91	14.83	15.09	15.48	14.85	15.49
MAE	12.22	12.18	11.90	11.96	11.91	11.79	12.21	12.61	12.02	12.45

cp: complexity parameter (taken exactly 7 decimal points to prune the tree)

In Decision Tree Gini Regression, sample 6 was the optimal model (R-squared = 0.27, RMSE = 14.83, table 4.23).

Decision Tree Gini Classification

Table 4.24: Summary of WP Gini Classification training result of 10 samples cross-validation (10 times, 10 folds)

Sample	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
cp	0.02	0.02	0.02	0.04	0.01	0.03	0.03	0.02	0.03	0.01
Accuracy	0.44	0.51	0.46	0.48	0.46	0.48	0.53	0.51	0.50	0.47
Kappa	0.01	0.13	0.03	0.08	0.04	0.07	0.16	0.13	0.11	0.06

cp: complexity parameter (taken exactly 7 decimal points to prune the tree)

The best tuning parameters among the ten samples was sample 7 (cp=0.029) having the highest accuracy = 0.527.

Table 4.25: Summary of WP Gini Classification training result of 10 Up-sampling cross-validation (10 times, 10 folds)

Sample	AUC	Accuracy	F1	Precision	Recall	True Pos Rate	True Neg Rate
#1	0.76	0.537	0.52	0.54	0.54	0.54	0.86
#2	0.60	0.605	0.58	0.59	0.61	0.59	0.88
#3	0.78	0.605	0.58	0.61	0.61	0.61	0.88
#4	0.69	0.421	0.36	0.33	0.42	0.33	0.82
#5	0.80	0.533	0.51	0.52	0.53	0.52	0.86
#6	0.79	0.576	0.57	0.58	0.58	0.58	0.87
#7	0.74	0.499	0.49	0.48	0.50	0.48	0.85
#8	0.80	0.596	0.59	0.61	0.60	0.61	0.87
#9	0.78	0.551	0.54	0.57	0.55	0.57	0.86
#10	0.81	0.600	0.59	0.59	0.60	0.59	0.88

AUC: Area Under the Curve,

True Pos Rate: True Positive Rate, True Neg Rate: True Negative Rate

In Decision Tree Gini Classification was the optimal sample 2 (Accuracy = 0.605, Precision = 0.594, Recall = 0.606, table 4.25).

Table 4.26: Summary of WP Gini Classification training result of 10 Up-sampling cross-validation on each class

Sample	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Underload										
Accuracy	0.60	0.71	0.68	0.58	0.61	0.62	0.61	0.64	0.62	0.66
Precision	0.52	0.66	0.60	0.42	0.50	0.65	0.48	0.57	0.61	0.57
Recall	0.51	0.84	0.79	0.45	0.57	0.49	0.58	0.65	0.51	0.75
Optimal load 1										
Accuracy	0.58	0.57	0.58	0.56	0.56	0.57	0.56	0.56	0.59	0.59
Precision	0.41	0.44	0.45	0.33	0.39	0.45	0.37	0.49	0.50	0.48
Recall	0.45	0.34	0.42	0.38	0.32	0.35	0.32	0.29	0.45	0.45
Optimal load 2										
Accuracy	0.54	0.55	0.55	0.50	0.55	0.58	0.52	0.58	0.57	0.54
Precision	0.44	0.41	0.59	0.14	0.36	0.42	0.27	0.43	0.40	0.43
Recall	0.20	0.24	0.20	0.02	0.25	0.46	0.09	0.45	0.38	0.20
Overload										
Accuracy	0.75	0.79	0.78	0.65	0.79	0.79	0.73	0.84	0.71	0.82
Precision	0.67	0.73	0.71	0.51	0.73	0.74	0.63	0.81	0.66	0.78
Recall	0.99	1	1	0.83	1	1	1	1	0.87	1

The Accuracy, Precision, and Recall in underload and overload levels were higher than the other, but the samples in these levels were up-sampled. So, the results may be affected. The Accuracy and Precision of the optimal load 2 were as same as the optimal load 1, but the higher Recall of the optimal load 1.

4.3.3 Extended Feature Sets

Multiple Linear Regression

```
> summary(efs_reg_cv)

Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-32.680  -7.687   0.456   7.527  29.040

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 9.307658   7.785119   1.196   0.2340
con_mental_demand           0.542525   0.092598   5.859 3.43e-08 ***
con_parallelism             0.049521   0.057162   0.866   0.3879
con_temporal_demand         0.132901   0.100413   1.324   0.1879
con_manual_Activity         0.101294   0.053595   1.890   0.0609 .
con_visual_attention        0.150069   0.079282   1.893   0.0605 .
con_effort                 -0.009539   0.084780  -0.113   0.9106
con_solving_deciding       -0.061488   0.069849  -0.880   0.3803
con_frustration             0.049103   0.067288   0.730   0.4668
con_context_bias           -0.017429   0.057585  -0.303   0.7626
con_task_space              0.012849   0.055347   0.232   0.8168
con_motivation              0.071994   0.083602   0.861   0.3907
con_verbal_material         0.010880   0.070232   0.155   0.8771
con_skill                  -0.027651   0.058822  -0.470   0.6391
con_auditory_attention     -0.016934   0.065720  -0.258   0.7971
con_physical_demand        -0.046777   0.064071  -0.730   0.4666
con_speech_response        -0.036613   0.050690  -0.722   0.4714
con_utility                -0.029761   0.071479  -0.416   0.6778
con_past_knowledge_expertise  0.057799   0.056112   1.030   0.3048
con_arousal                -0.066757   0.071924  -0.928   0.3550
con_performance            -0.041644   0.070794  -0.588   0.5574
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.58 on 134 degrees of freedom
Multiple R-squared:  0.463,     Adjusted R-squared:  0.3828
F-statistic: 5.776 on 20 and 134 DF,  p-value: 1.334e-10
```

Figure 4.31: EFS training result of sample 1 cross-validation (10 times, 10 folds)

Table 4.27: Summary of EFS training result of 10 samples cross-validation (10 times, 10 folds)

Sample	R-squared	$_{A}djR$-squared	RMSE	Residual SE	Significant vars
#1	0.46	0.38	13.96	12.58	Mental demand
#2	0.47	0.40	14.21	12.86	Mental demand, Solving Deciding
#3	0.52	0.45	13.45	12.29	Mental demand
#4	0.51	0.44	13.88	12.60	Mental demand
#5	0.50	0.42	13.43	12.18	Mental demand, Visual attention, Motivation
#6	0.57	0.50	13.34	12.16	Mental, Temporal demand, Parallelism, Solving Deciding
#7	0.46	0.38	14.28	12.80	Mental demand
#8	0.50	0.42	13.58	12.40	Mental demand
#9	0.52	0.45	13.52	12.34	Mental, Temporal demand
#10	0.52	0.44	13.55	12.42	Mental demand, Visual attention

The overall significance (p_{value}) of training sets are all <0.001

In Linear Regression, sample train 6 had the highest R-squared and the lowest RMSE (R-squared = 0.57, RMSE = 13.34, table 4.27).

Table 4.27 showed the average result of R-squared range from 0.46 to 0.57 or $_{Adj}$R-squared from 0.38 to 0.50, RMSE range from 13.34 to 14.28. R-squared explained the impact of linear model on average of 46 to 57% of the Mental Workload score variation. In other words, the Mental demand had 57% impact on Mental Workload score (training sample 6, tab4.27).

Table 4.28: Variable importance of EFS in model

Variables	Overall (%)	Variables	Overall (%)
Mental demand	100.00	Frustration	10.74
Visual attention	30.98	Speech response	10.61
Manual activity	30.93	Performance	8.28
Temporal demand	21.07	Skill	6.22
Past knowledge	15.97	Utility	5.29
Arousal	14.19	Context bias	3.31
Solving Deciding	13.36	Auditory attention	2.53
Parallelism	13.12	Task space	2.08
Motivation	13.03	Verbal material	0.74
Physical demand	10.75	Effort	0.00

In 10 training sets, the relationship between Mental Workload score and significant variables was investigated using a Pearson correlation. The model yielded a high correlation between model predictions and important variables. In other words, the models with more important variables were higher R-squared (table 4.28). A strong positive correlation was found (R-squared=0.57, n=155, p<0.001, training sample 6, table 4.27).

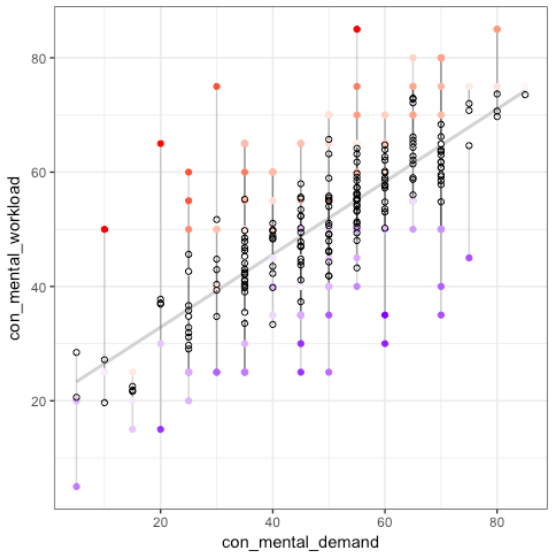

Figure 4.32: Correlation of MWL and Mental demand in EFS set

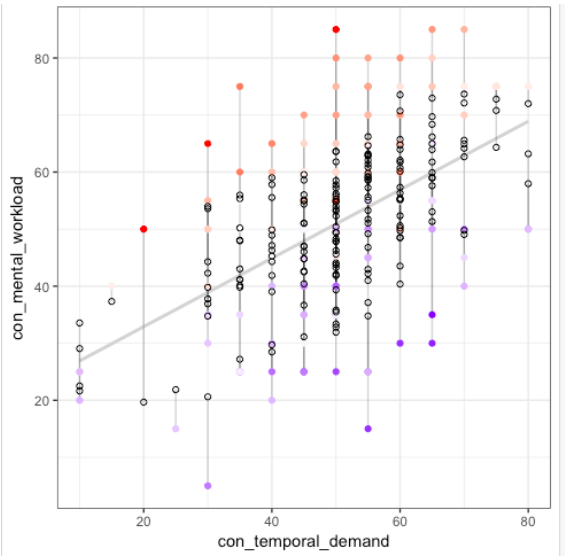

Figure 4.33: Correlation of MWL and Temporal demand in EFS set

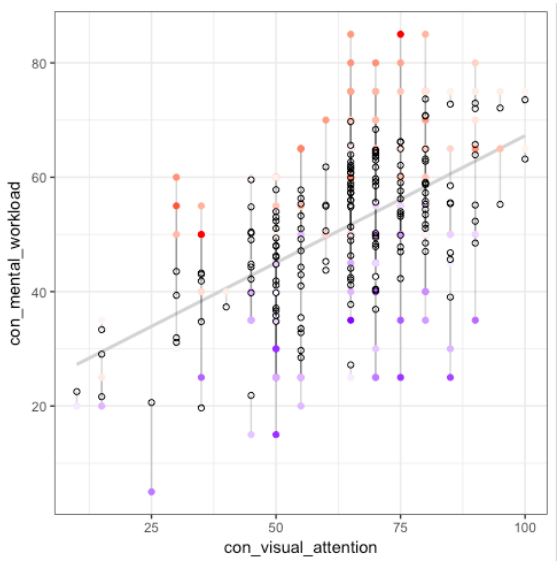

Figure 4.34: Correlation of MWL and Visual attention in EFS set

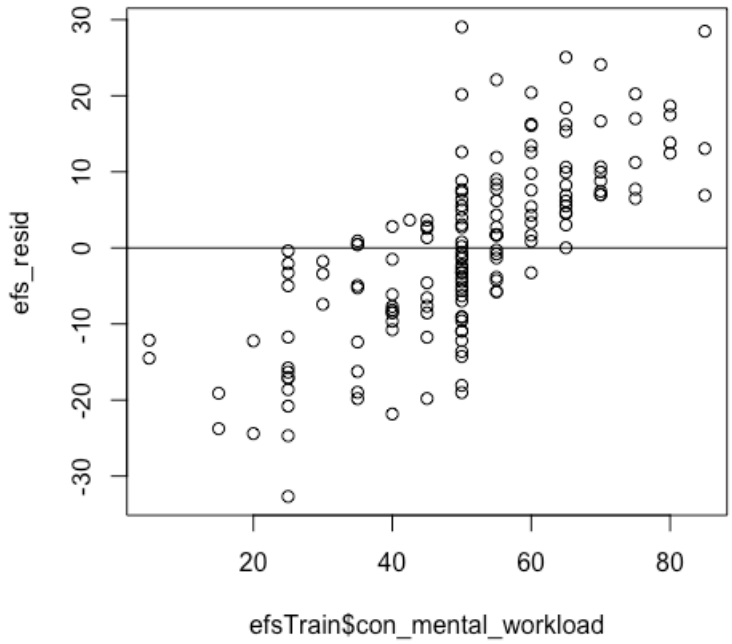

Figure 4.35: Training of EFS in Residual plot

Looking at figure 4.35, there were three main points. First of all, there was a difference between the expected value and the observed. Secondly, there were only some points the residuals centered on zero (median=50) in the range of 15 to 65 (larger than NASA residual plot, figure 4.22). Lastly, the residuals were not normally

distributed and not constantly spread throughout the range (seen histogram in the view of the vertical axis).

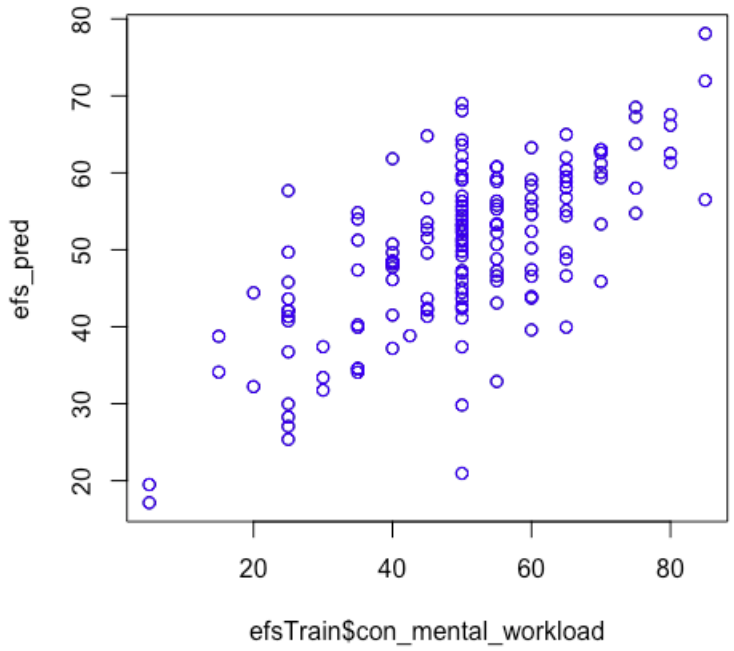

Figure 4.36: Training of EFS in comparison of Actual & Predicted values

The model above had an average R-squared (0.50) which meant 50 percent of the points would be close to the diagonal line. The model had two subsections of performance. The first one was where Actuals <30. Within this zone, the model did not look as same as the Actuals. It greatly overestimated the Actual values. The second one was when Actuals between 30 and 80, the model mostly concentrated within this zone whereas some points were still random. The model looked more correctly than NASA (figure 4.23, 4.30).

Decision Tree Information Gain

Table 4.29: Summary of EFS training result of 10 samples cross-validation (10 times, 10 folds)

Sample	Accuracy	Tuning parameters		
		trials	model	winnow
#1	0.454	10	rules	T
#2	0.428	10	tree	T
#3	0.440	20	tree	F
#4	0.447	20	tree	F
#5	0.428	20	rules	T
#6	0.444	1	tree	T
#7	0.446	20	tree	F
#8	0.421	10	rules	T
#9	0.459	1	tree	T
#10	0.485	20	tree	T

The best tuning parameters among the ten samples was sample 10 (trials=20, model=tree, winnow=T) having the highest accuracy = 0.485.

Table 4.30: Summary of EFS training result of 10 Up-sampling cross-validation (10 times, 10 folds)

Sample	AUC	Accuracy	F1	Precision	Recall	True Pos Rate	True Neg Rate
#1	0.90	0.733	0.72	0.73	0.73	0.73	0.92
#2	0.88	0.721	0.71	0.71	0.72	0.71	0.91
#3	0.91	0.751	0.74	0.75	0.75	0.75	0.92
#4	0.92	0.75	0.74	0.74	0.75	0.74	0.92
#5	0.90	0.735	0.72	0.73	0.74	0.73	0.92
#6	0.87	0.725	0.71	0.72	0.73	0.72	0.92
#7	0.92	0.765	0.75	0.76	0.77	0.76	0.93
#8	0.90	0.734	0.72	0.73	0.73	0.73	0.92
#9	0.86	0.714	0.70	0.71	0.71	0.71	0.91
#10	0.88	0.734	0.72	0.73	0.73	0.73	0.92

AUC: Area Under the Curve,

True Pos Rate: True Positive Rate, True Neg Rate: True Negative Rate

In Decision Tree Information Gain, sample train 7 was the most optimal (Accuracy = 0.765, Precision = 0.764, Recall = 0.765, table 4.30).

Table 4.31: Summary of EFS training result of 10 Up-sampling cross-validation on each class (10 times, 10 folds)

Sample	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Underload										
Accuracy	0.81	0.79	0.81	0.82	0.80	0.79	0.82	0.80	0.77	0.79
Precision	0.80	0.76	0.78	0.81	0.79	0.77	0.80	0.77	0.75	0.77
Recall	0.94	0.94	0.98	0.95	0.94	0.94	0.97	0.96	0.92	0.94
Optimal load 1										
Accuracy	0.60	0.59	0.61	0.60	0.61	0.59	0.61	0.60	0.58	0.59
Precision	0.57	0.55	0.60	0.58	0.57	0.57	0.61	0.56	0.55	0.55
Recall	0.45	0.42	0.46	0.44	0.47	0.41	0.47	0.44	0.38	0.40
Optimal load 2										
Accuracy	0.62	0.62	0.64	0.64	0.62	0.63	0.65	0.63	0.62	0.65
Precision	0.58	0.60	0.63	0.62	0.61	0.59	0.66	0.62	0.57	0.65
Recall	0.54	0.52	0.57	0.61	0.53	0.55	0.62	0.55	0.55	0.60
Overload										
Accuracy	0.92	0.90	0.93	0.93	0.91	0.90	0.94	0.91	0.91	0.90
Precision	0.91	0.89	0.92	0.92	0.90	0.88	0.93	0.91	0.91	0.89
Recall	1	1	1	1	1	1	1	1	1	1

The Accuracy, Precision, and Recall in underload and overload levels were quite high, but the samples in these levels were up-sampled. So, the results may be affected. The optimal load 2 had the three indicators (accuracy, precision, recall) were generally higher than optimal load 1.

Decision Tree Gini Regression

Table 4.32: Summary of EFS Gini Regression training result of 10 samples cross-validation (10 times, 10 folds)

Sample	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
cp	0.05	0.06	0.06	0.07	0.07	0.05	0.07	0.04	0.06	0.04
R-squared	0.34	0.32	0.26	0.35	0.27	0.36	0.27	0.33	0.31	0.31
RMSE	13.48	14.09	14.83	13.84	14.12	14.31	14.37	13.81	14.33	14.35
MAE	10.73	11.26	11.87	11.31	11.10	11.78	11.75	11.02	11.59	11.47

cp: complexity parameter (taken exactly 7 decimal points to prune the tree)

In Decision Tree Gini Regression sample 1 was the optimal model (R-squared = 0.05, RMSE = 13.48, table 4.32).

Decision Tree Gini Classification

Table 4.33: Summary of EFS Gini Classification training result of 10 samples cross-validation (10 times, 10 folds)

Sample	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
cp	0.04	0.05	0.07	0.05	0.06	0.02	0.03	0.04	0.05	0.01
Accuracy	0.44	0.41	0.43	0.38	0.40	0.40	0.36	0.39	0.48	0.51
Kappa	0.08	-0.04	-0.004	-0.05	-0.03	0.04	-0.10	-0.03	0.15	0.20

cp: complexity parameter (taken exactly 7 decimal points to prune the tree)

The best tuning parameters among the ten samples was sample 9 (cp=0.047) having the highest accuracy = 0.484.

Table 4.34: Summary of EFS Gini Classification training result of 10 Up-sampling cross-validation (10 times, 10 folds)

Sample	AUC	Accuracy	F1	Precision	Recall	True Pos Rate	True Neg Rate
#1	0.81	0.599	0.52	0.49	0.60	0.49	0.88
#2	0.80	0.579	0.57	0.59	0.58	0.59	0.87
#3	0.76	0.548	//	//	0.55	//	0.86
#4	0.81	0.601	0.58	0.53	0.60	0.53	0.88
#5	0.79	0.565	0.50	0.49	0.57	0.49	0.87
#6	0.83	0.656	0.63	0.64	0.66	0.64	0.90
#7	0.81	0.619	0.57	0.54	0.62	0.54	0.89
#8	0.81	0.588	0.58	0.56	0.59	0.56	0.88
#9	0.80	0.626	0.43	0.49	0.63	0.49	0.89
#10	0.85	0.681	0.66	0.68	0.68	0.68	0.90

AUC: Area Under the Curve,

True Pos Rate: True Positive Rate, True Neg Rate: True Negative Rate

In Decision Tree Gini Classification was the optimal sample 10 (Accuracy = 0.681, Precision = 0.675, Recall = 0.681, table 4.34).

Table 4.35: Summary of EFS Gini Classification training result of 10 Up-sampling cross-validation on each class

Sample	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Underload										
Accuracy	0.72	0.67	0.69	0.74	0.69	0.72	0.71	0.71	0.72	0.71
Precision	0.68	0.63	0.66	0.69	0.65	0.67	0.62	0.66	0.66	0.66
Recall	0.84	0.74	0.78	0.90	0.78	0.88	0.92	0.86	0.90	0.83
Optimal load 1										
Accuracy	0.57	0.56	0.50	0.50	0.56	0.53	0.51	0.55	0.50	0.57
Precision	0.41	0.42	0.17	0.30	0.36	0.45	0.27	0.38	0.07	0.48
Recall	0.34	0.31	0	0.01	0.37	0.15	0.05	0.24	0	0.34
Optimal load 2										
Accuracy	0.54	0.56	0.60	0.62	0.54	0.63	0.61	0.55	0.61	0.64
Precision	0.40	0.43	0.39	0.45	0.33	0.57	0.51	0.40	0.49	0.69
Recall	0.21	0.28	0.67	0.70	0.23	0.59	0.53	0.27	0.60	0.56
Overload										
Accuracy	0.78	0.76	0.69	0.72	0.79	0.81	0.79	0.77	0.79	0.84
Precision	0.71	0.68	0.70	0.71	0.82	0.77	0.75	0.71	0.73	0.81
Recall	1	1	0.73	0.80	0.88	1	0.98	0.98	1	1

The Accuracy, Precision, and Recall of underload and overload levels were quite high, but the samples in these levels were up-sampled. So, the results may be affected. The Accuracy of the optimal load 2 was similar to optimal load 1, but the Precision and Recall of optimal load 2 were higher than optimal load 1 in most cases.

Data correlation of variables indicates the impact on models. Consequently, in the Linear Regression model the higher weights and more important variables are strong or moderate relationships in the scatter plot matrix (table 4.9, 4.18, 4.27). Also, they are the factors to split branches in the Decision Tree (figure 4.24, 4.25). The additional factors of EFS set had no relationship with Mental Workload score (table 4.8); in spite the impact showing on Linear Regression model (table 4.27).

"The ultimate goal of machine learning is to make a machine system that can automatically build models from data without requiring tedious and time consuming human involvement." One of the difficulties is that learning algorithms (e.g., decision trees, random forests, clustering techniques, etc.) require parameters to be set in the model. The tuning parameters allow to tune the optimal values of a learning task in the best way possible. Thus, tuning an algorithm or machine learning technique can be simply thought of as a process which one goes through in which they optimize the parameters that impact the model to enable the algorithm to perform the best.

In Decision Tree Information Gain, C5.0 is the package applied to train the model which is short of Decision Trees and Rule-Based Models. There are three tuning parameters in this application as in [1] trials specifying the number of boosting iterations, how many the model used, [2] models as rules indicating should the tree be decomposed into a rule-based model, and [3] winnow indicating whether predictor winnowing (i.e feature selection) should be used. In Decision Tree Gini, CART is the package applied to train interval and numeric variables as its name Classification and Regression Trees. Likewise, the tuning parameter cp is the fundamental driver for over/under-fitting. In other words, it determines depth of the tree and number of terminal nodes.

4.4 Model comparison

4.4.1 NASA Task Load Index

Multiple Linear Regression

Table 4.36: NASA multiple linear regression test results of 10 samples

Sample	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
R-squared	0.18	0.26	0.31	0.23	0.27	0.31	0.29	0.33	0.41	0.13
RMSE	13.28	12.14	12.66	13.24	13.22	11.87	11.69	11.44	11.62	14.35
MAE	9.38	9.23	9.59	10.19	10.33	9.07	8.94	8.11	8.85	10.12

Decision Tree Information Gain

Table 4.37: NASA decision tree Information Gain test results of 10 samples

Sample	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Overall accuracy	0.59	0.51	0.48	0.49	0.51	0.48	0.44	0.49	0.41	0.48
p-value$_{Acc>NIR}$	0.88	**0.01**	0.81	0.74	0.95	0.65	0.55	0.81	0.65	0.55
Underload										
Accuracy	0.56	0.53	0.52	0.54	0.54	0.52	0.55	0.54	0.53	0.52
Precision	0.25	0.50	0.25	0.50	0.50	0.25	0.25	0.50	0.50	0.25
Recall	0.50	0.125	0.11	0.167	0.18	0.09	0.33	0.15	0.12	0.11
Optimal load 1										
Accuracy	0.57	0.60	0.58	0.58	0.54	0.56	0.57	0.56	0.54	0.58
Precision	0.38	0.48	0.43	0.38	0.19	0.33	0.43	0.29	0.24	0.43
Recall	0.44	0.50	0.45	0.50	0.36	0.41	0.38	0.43	0.28	0.47
Optimal load 2										
Accuracy	0.68	0.65	0.63	0.64	0.67	0.66	0.61	0.66	0.63	0.64
Precision	0.73	0.51	0.54	0.57	0.68	0.59	0.46	0.62	0.51	0.54
Recall	0.66	0.83	0.61	0.64	0.66	0.71	0.61	0.68	0.70	0.67
Overload										
Accuracy	0.33	0.57	0.50	0.60	0.60	0.50	0.53	0.50	0.50	0.50
Precision	1	1	0	0	1	0	1	0	0	0
Recall	0.50	0.25	0	0	0.33	0	0.13	0	0	0

Acc: Accuracy, NIR: No Inforamtion Rate

Decision Tree Gini Regression

Table 4.38: NASA Gini Regression test results of 10 samples

Sample	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
R-squared	0.12	0.16	0.26	0.11	0.34	0.08	0.26	0.20	0.26	0.17
RMSE	14.36	13.19	13.07	14.65	12.60	14.76	12.12	13.12	13.02	13.02
MAE	11.40	10.34	10.04	11.60	9.68	11.05	9.52	10.15	10.40	9.91

Decision Tree Gini Classification

Table 4.39: NASA Gini Classification test results of 10 samples

Sample	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Overall accuracy	0.48	0.44	0.48	0.43	0.06	0.30	0.41	0.41	0.44	0.46
p-value$_{Acc>NIR}$	0.97	0.99	0.97	1	1	1	1	1	0.99	0.98
Underload										
Accuracy	0.54	0.53	0.50	0.53	0.52	0.53	0.53	0.53	0.53	0.52
Precision	0.25	0.13	0	0.13	0.08	0.14	0.14	0.12	0.09	0.09
Recall	0.25	0.50	0	0.50	0.75	0.75	0.50	0.50	0.50	0.25
F1	0.25	0.20	//	0.21	0.14	0.23	0.22	0.19	0.15	0.13
Optimal load 1										
Accuracy	0.59	0.58	0.62	0.61	0.50	0.52	0.58	0.57	0.57	0.58
Precision	0.36	0.41	0.48	0.46	//	0.22	0.41	0.38	0.44	0.50
Recall	0.62	0.43	0.67	0.62	0	0.10	0.43	0.38	0.38	0.38
F1	0.46	0.42	0.56	0.53	//	0.13	0.42	0.38	0.41	0.63
Optimal load 2										
Accuracy	0.61	0.62	0.61	0.59	0.50	0.58	0.60	0.61	0.61	0.64
Precision	0.75	0.80	0.73	0.75	//	0.62	0.78	0.73	0.86	0.69
Recall	0.41	0.43	0.43	0.32	0	0.35	0.38	0.43	0.49	0.54
F1	0.53	0.56	0.54	0.45	//	0.45	0.51	0.54	0.62	0.61
Overload										
Accuracy	0.60	0.56	0.50	0.50	0.51	0.52	0.53	0.50	0.50	0.50
Precision	0.33	0.20	0	0	0.04	0.09	0.11	0	0	0
Recall	1	1	0	0	1	1	1	0	0	0
F1	0.50	0.33	//	//	0.08	0.17	0.20	//	//	//

Acc: Accuracy, NIR: No Information Rate

For NASA models, when performing on test sets, sample 8 (RMSE = 11.44) and sample 9 (R-squared = 0.41) were the best one in Linear Regression (table 4.36);

sample 7 (RMSE = 12.12) and sample 5 (R-squared = 0.34) as in Gini Regression (table 4.38); sample 1 (overall Accuracy = 0.59) as in Information Gain (table 4.37) and also sample 1 (overall Accuracy = 0.48) as in Gini Classification (table 4.39).

4.4.2 Workload Profile

Multiple Linear Regression

Table 4.40: WP multiple linear regression in 10 test results

Sample	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
R-squared	0.37	0.17	0.18	0.23	0.36	0.22	0.12	0.24	0.19	0.24
RMSE	14.11	15.45	15.01	15.79	13.50	14.46	16.22	14.56	15.47	14.82
MAE	10.91	12.38	11.65	12.55	10.88	11.28	12.7	11.34	11.64	12.01

Decision Tree Information Gain

Table 4.41: WP decision tree information gain test results of 10 samples

Sample	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Overall accuracy	0.41	0.47	0.41	0.33	0.36	0.36	0.34	0.39	0.33	0.44
$p_{value Acc>NIR}$	0.18	0.65	0.81	0.97	0.44	0.06	0.95	0.99	0.88	0.55
Underload										
Accuracy	0.55	0.52	0.52	0.54	0.55	0.56	0.51	0.57	0.51	0.57
Precision	0.67	0.17	0.17	0.33	0.5	0.67	0.17	0.5	0.17	0.5
Recall	0.2	0.17	0.1	0.2	0.2	0.24	0.08	0.3	0.08	0.33
Optimal load 1										
Accuracy	0.58	0.63	0.56	0.57	0.56	0.59	0.57	0.54	0.56	0.59
Precision	0.39	0.61	0.32	0.39	0.32	0.36	0.29	0.21	0.29	0.39
Recall	0.5	0.55	0.43	0.39	0.41	0.63	0.44	0.38	0.42	0.58
Optimal load 2										
Accuracy	0.59	0.59	0.62	0.56	0.58	0.57	0.60	0.61	0.59	0.60
Precision	0.39	0.43	0.57	0.29	0.39	0.32	0.46	0.57	0.43	0.50
Recall	0.61	0.5	0.55	0.4	0.5	0.53	0.46	0.47	0.48	0.50
Overload										
Accuracy	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
Precision	0	0	0	0	0	0	0	0	0	0
Recall	0	0	0	0	0	0	0	0	0	0

Acc: Accuracy, NIR: No Inforamtion Rate

Decision Tree Gini Regression

Table 4.42: WP Gini Regression test results of 10 samples

Sample	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
R-squared	0.31	0.24	0.19	0.29	0.22	0.18	0.21	0.30	0.28	0.24
RMSE	14.87	14.75	14.86	15.39	14.73	14.98	14.92	14.02	14.84	14.81
MAE	11.65	11.38	11.77	12.59	11.77	12.28	11.58	10.62	11.70	11.57

Decision Tree Gini Classification

Table 4.43: WP Gini Classification test results of 10 samples

Sample	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Overall accuracy	0.42	0.41	0.39	0.31	0.34	0.39	0.27	0.41	0.38	0.39
p-value$_{Acc>NIR}$	0.65	0.73	0.81	0.99	0.95	0.81	1	0.73	0.87	0.81
Underload										
Accuracy	0.56	0.52	0.54	0.55	0.54	0.50	0.53	0.52	0.52	0.55
Precision	0.25	0.11	0.17	0.23	0.17	0	0.12	0.13	0.11	0.21
Recall	0.50	0.17	0.50	0.50	0.67	0	0.50	0.17	0.17	0.50
F1	0.33	0.13	0.25	0.32	0.27	//	0.19	0.14	0.13	0.30
Optimal load 1										
Accuracy	0.61	0.64	0.60	0.62	0.58	0.56	0.56	0.57	0.60	0.61
Precision	0.46	0.61	0.50	0.53	0.59	0.50	0.40	0.53	0.52	0.54
Recall	0.61	0.61	0.50	0.61	0.36	0.25	0.29	0.32	0.46	0.50
F1	0.52	0.61	0.50	0.57	0.44	0.33	0.33	0.40	0.49	0.52
Optimal load 2										
Accuracy	0.57	0.56	0.57	0.50	0.57	0.63	0.55	0.61	0.57	0.57
Precision	0.78	0.58	0.67	//	0.57	0.51	0.67	0.48	0.48	0.62
Recall	0.25	0.25	0.29	0	0.29	0.64	0.21	0.57	0.36	0.29
F1	0.38	0.35	0.40	//	0.38	0.57	0.32	0.52	0.41	0.39
Overload										
Accuracy	0.50	0.52	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
Precision	0	0.07	0	0	0	0	0	0	0	0
Recall	0	0.50	0	0	0	0	0	0	0	0
F1	//	0.12	//	//	//	//	//	//	//	//

Acc: Accuracy, NIR: No Inforamtion Rate

For WP models, when performing on test sets, sample 5 (RMSE = 13.50, R-squared = 0.36) were the best one in Linear Regression (table 4.40); sample 8 (RMSE = 14.02)

and sample 1 (R-squared = 0.31) as in Gini Regression (table 4.42); sample 2 (overall Accuracy = 0.47) as in Information Gain (table 4.41) and sample 1 (overall Accuracy = 0.42) as in Gini Classification (table 4.43).

4.4.3 Extended Feature Sets

Multiple Linear Regression

Table 4.44: EFS multiple linear regression in 10 test results

Sample	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
R-squared	0.48	0.44	0.25	0.33	0.37	0.18	0.35	0.36	0.5	0.43
RMSE	12.81	12.37	14.68	13.25	14.48	14.14	13.17	13.15	12.17	12.79
MAE	10.06	10.02	11.18	10.57	11.47	11.32	10.02	9.9	9.4	9.67

Decision Tree Information Gain

Table 4.45: EFS decision tree information gain test results of 10 samples

Sample	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Overall accuracy	0.39	0.36	0.39	0.37	0.44	0.42	0.44	0.42	0.45	0.44
p-value$_{Acc>NIR}$	0.95	0.65	0.44	0.92	0.87	0.34	0.81	0.34	0.92	0.95
Underload										
Accuracy	0.58	0.56	0.59	0.51	0.52	0.55	0.50	0.50	0.57	0.50
Precision	0.57	0.43	0.71	0.14	0.14	0.43	0	0	0.43	0
Recall	0.33	0.27	0.36	0.08	0.2	0.21	0	0	0.33	0
Optimal load 1										
Accuracy	0.57	0.56	0.58	0.61	0.59	0.57	0.60	0.60	0.56	0.60
Precision	0.41	0.31	0.38	0.52	0.45	0.34	0.48	0.45	0.28	0.41
Recall	0.4	0.39	0.48	0.54	0.5	0.48	0.54	0.54	0.44	0.6
Optimal load 2										
Accuracy	0.56	0.57	0.55	0.55	0.59	0.61	0.58	0.60	0.62	0.61
Precision	0.33	0.38	0.29	0.25	0.54	0.54	0.5	0.46	0.67	0.63
Recall	0.42	0.39	0.3	0.35	0.42	0.51	0.4	0.52	0.48	0.45
Overload										
Accuracy	0.50	0.55	0.60	0.55	0.50	0.50	0.60	0.58	0.60	0.50
Precision	0	0.5	0.5	0.5	//	0	0.5	1	0.5	0
Recall	0	0.2	0.5	0.2	//	0	0.5	0.29	0.5	0

Acc: Accuracy, NIR: No Information Rate

Decision Tree Gini Regression

Table 4.46: EFS Gini Regression test results of 10 samples

Sample	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
R-squared	0.28	0.39	0.26	0.19	0.30	0.15	0.31	0.23	0.16	0.31
RMSE	15.17	12.88	14.32	14.65	14.92	14.67	14.31	14.97	15.26	13.42
MAE	12.01	10.37	11.37	12.42	11.98	11.40	11.91	10.74	11.60	11.10

Decision Tree Gini Classification

Table 4.47: EFS Gini Classification test results of 10 samples

Sample	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Overall accuracy	0.29	0.39	0.39	0.34	0.32	0.39	0.37	0.45	0.37	0.31
p-value$_{Acc>NIR}$	1	0.92	0.92	0.99	0.99	0.92	0.95	0.65	0.95	1
Underload										
Accuracy	0.56	0.61	0.59	0.56	0.58	0.56	0.59	0.58	0.56	0.55
Precision	0.25	0.37	0.32	0.24	0.30	0.27	0.32	0.31	0.27	0.33
Recall	0.71	1	0.86	0.71	0.86	0.57	0.86	0.71	0.57	0.29
F1	0.37	0.54	0.46	0.36	0.44	0.36	0.46	0.43	0.36	0.31
Optimal load 1										
Accuracy	0.58	0.52	0.50	0.50	0.58	0.50	0.50	0.54	0.50	0.55
Precision	0.46	0.40	//	//	0.40	//	//	0.63	//	0.50
Recall	0.38	0.07	0	0	0.48	0	0	0.17	0	0.21
F1	0.42	0.12	//	//	0.44	//	//	0.27	//	0.29
Optimal load 2										
Accuracy	0.50	0.61	0.62	0.61	0.50	0.63	0.61	0.65	0.63	0.57
Precision	//	0.54	0.45	0.44	//	0.47	0.44	0.59	0.50	0.37
Recall	0	0.54	0.71	0.67	0	0.75	0.67	0.67	0.71	0.46
F1	//	0.54	0.55	0.53	//	0.58	0.53	0.63	0.59	0.41
Overload										
Accuracy	0.53	0.54	0.55	0.50	0.50	0.56	0.53	0.55	0.56	0.50
Precision	0.11	0.14	0.20	0	0	0.22	0.14	0.18	0.15	0
Recall	1	1	0.50	0	0	1	0.50	1	1	0
F1	0.20	0.25	0.29	//	//	0.36	0.22	0.31	0.27	//

Acc: Accuracy, NIR: No Inforamtion Rate, //: NaN

For EFS models, When performing on test sets, sample 9 (RMSE = 12.17, R-squared = 0.50) were the best one in Linear Regression (table 4.44); sample 2 (RMSE = 12.88,

R-squared = 0.39) as in Gini Regression (table 4.46); sample 9 (overall Accuracy = 0.45) as in Information Gain (table 4.45) and also sample 8 (overall Accuracy = 0.45) as in Gini Classification (table 4.47).

4.5 Model selection

4.5.1 Within three subjective rating scales

NASA models

The Mental Workload score of NASA models (Linear Regression, Decision Tree Gini Regression, Actual values) did not significantly change over models (chi-squared = 3.53, p = 0.17).

The chi-squared test for independence indicated no significant difference between Information Gain NASA models and Actual values (chi-squared = 14.70, p = 0.1). Similarly, a chi-squared test for independence indicated no significant difference between Decision Tree Gini Classification NASA models and Actual values (chi-squared = 13.32, p = 0.15).

On the whole, in testing the difference between models with NASA sets, there was no statistically significant difference between predicted value and actual values in four NASA models.

WP models

The Mental Workload score of WP models (Linear Regression, Decision Tree Gini Regression, Actual values) did not significantly change over models (chi-squared = 2.47, p = 0.29).

A chi-squared test for independence indicated no significant difference between Information Gain WP models and Actual values (chi-squared = 2.75, p = 0.97). Similarly, a chi-squared test for independence indicated no significant difference between Decision Tree Gini Classification WP models and Actual values (chi-squared = 12.05, p = 0.21).

On the whole, in testing the difference between models with WP sets, there was no statistically significant difference between predicted value and actual values in four WP models.

EFS models

The Mental Workload score of EFS models (Linear Regression, Decision Tree Gini Regression, Actual values) did not significantly change over models (chi-squared = 1.68, p = 0.43).

A chi-squared test for independence indicated no significant difference between Information Gain EFS models and Actual values (chi-squared = 10.01, p = 0.35). Similarly, a chi-squared test for independence indicated no significant difference between Decision Tree Gini Classification EFS models and Actual values (chi-squared = 15.14, p = 0.09).

On the whole, in testing the difference between models with EFS sets, there was no statistically significant difference between predicted value and actual values in four EFS models.

In each training sets of three instrumental scales, there were four models applied to see the difference of Mental Workload among ten sample sets. In the event of Mental Workload as a continuous feature, Linear Regression and Decision Tree Gini Regression were chosen; and in the event of Mental Workload as a categorical feature, Decision Tree Information Gain and Decision Tree Gini Categorical were chosen ones.

4.5.2 Between subjective rating scales

Models of Mental Workload as continuous feature

Training results

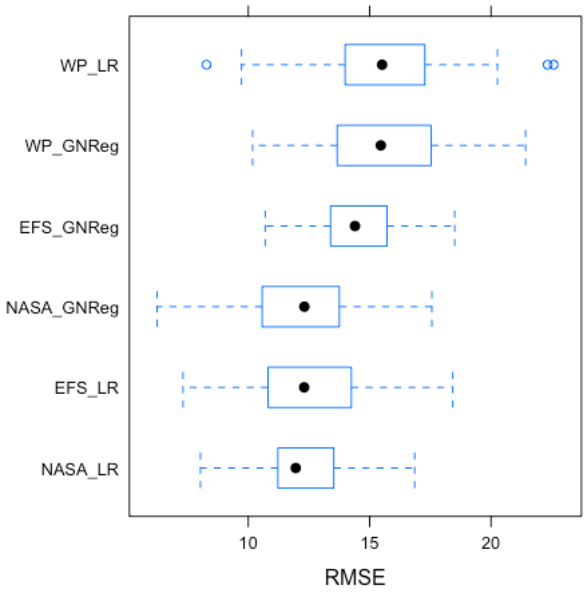

Figure 4.37: Training results of RMSE of Mental Workload score boxplots

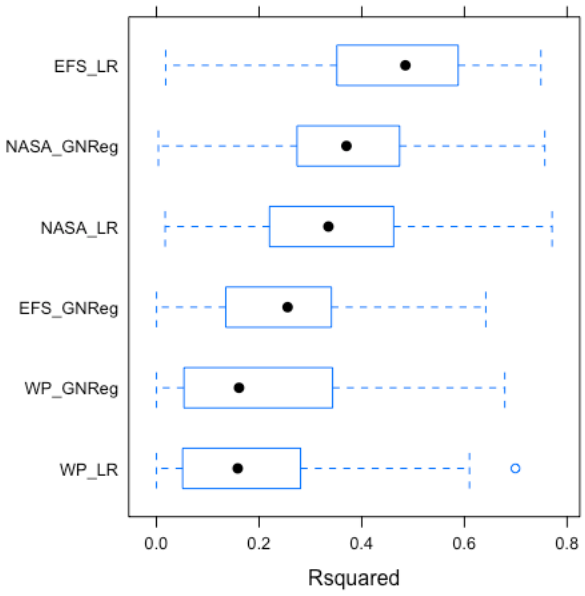

Figure 4.38: Training results of R-squared of Mental Workload score boxplots

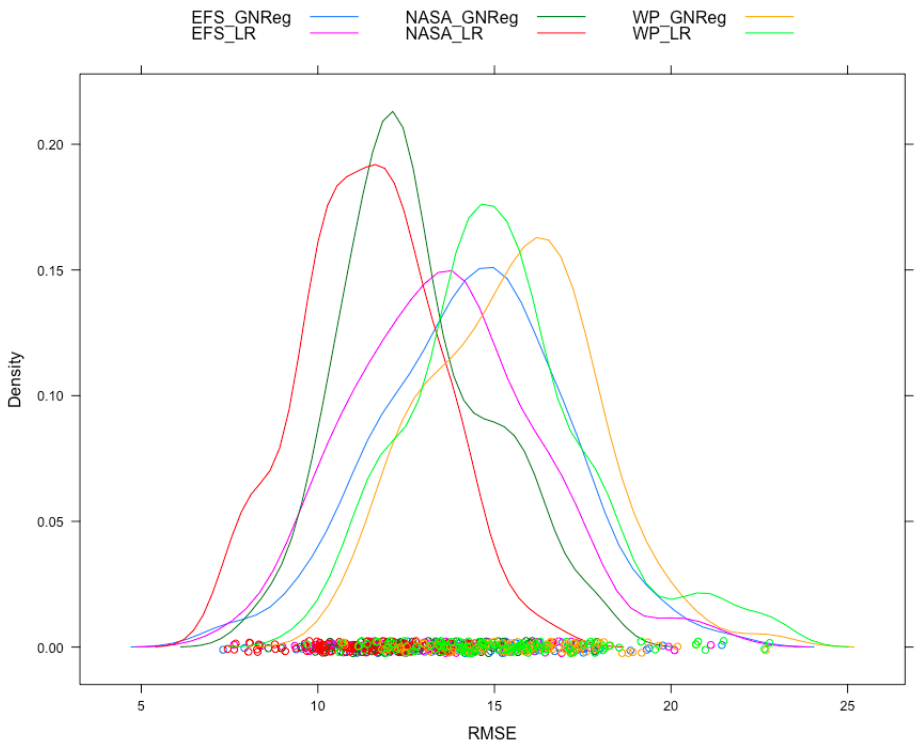

Figure 4.39: Training results of RMSE of Mental Workload score density plots

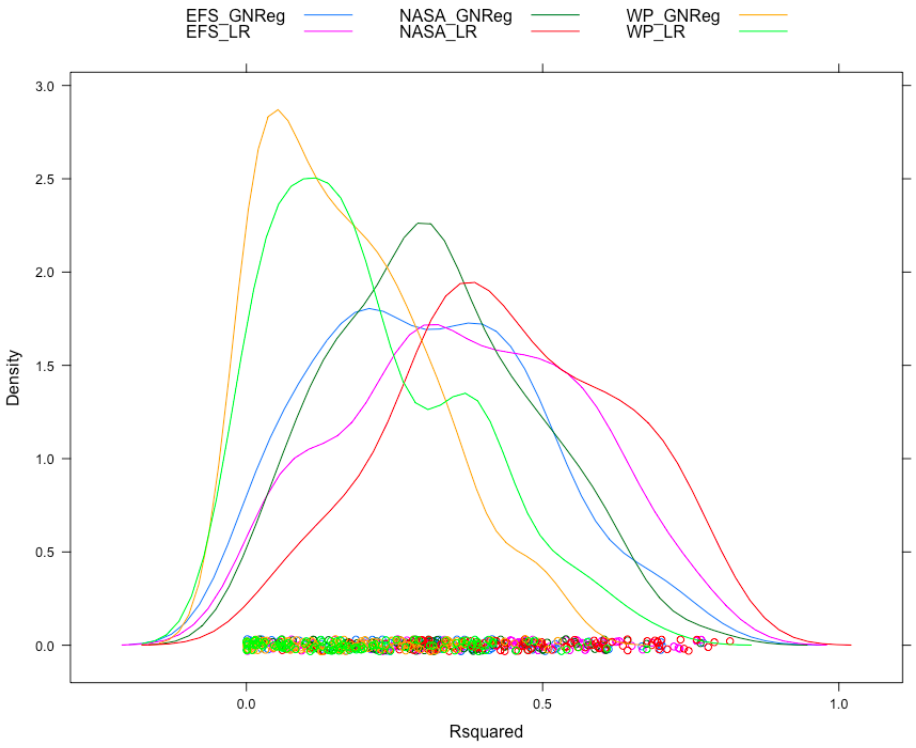

Figure 4.40: Training results of R-squared of Mental Workload score density plots

```
RMSE
           NASA_LR   NASA_GNReg WP_LR     WP_GNReg  EFS_LR    EFS_GNReg
NASA_LR               -1.3329    -3.7140   -4.0477   -2.1110   -2.9086
NASA_GNReg 3.822e-05             -2.3811   -2.7149   -0.7782   -1.5758
WP_LR      < 2.2e-16 6.289e-08             -0.3337    1.6030    0.8054
WP_GNReg   < 2.2e-16 1.491e-12  1.0000000             1.9367    1.1391
EFS_LR     2.711e-08 0.2904889  0.0001158 1.034e-06            -0.7976
EFS_GNReg  1.166e-13 2.661e-05  0.3462026 0.0202402 0.5936176

Rsquared
           NASA_LR   NASA_GNReg WP_LR     WP_GNReg  EFS_LR    EFS_GNReg
NASA_LR                0.122282   0.230342  0.267462  0.068366  0.131806
NASA_GNReg 2.165e-05              0.108060  0.145180 -0.053917  0.009524
WP_LR      1.855e-14 0.0001794              0.037120 -0.161976 -0.098535
WP_GNReg   < 2.2e-16 2.137e-09  1.0000000            -0.199096 -0.135655
EFS_LR     0.0899582 0.4423077  1.493e-07 3.974e-13             0.063441
EFS_GNReg  3.694e-05 1.0000000  0.0005071 1.094e-06 0.3441172
```

Figure 4.41: Significance test of difference (lower) & estimates of the difference of RMSE, R-squared as in Mental Workload score

In the box plots and density plots of RMSE and R-squared (figure 4.37, 4.38, figure 4.39, 4.40), the higher in box plots, the more right-inclined in density plots. With the training-result that NASA_{LR} of RMSE Mental Workload had the lowest error and EFS_{LR} had the highest R-squared. There were difference between models but how significantly different was shown in figure 4.41.

As the threshold of significant test 0.05, in terms of **RMSE**, the model NASA_{LR} had a statistically significant difference ($p_{value} < 0.001$) with NASA_{GNReg} (-1.33), WP_{LR} (-3.71), WP_{GNReg} (-4.05), EFS_{LR} (-2.11), EFS_{GNReg} (-2.91). The model NASA_{GNReg} had a statistically significant difference with WP_{LR} (-2.38), WP_{GNReg} (-2.71), EFS_{GNReg} (-1.58). The model WP_{LR} had a statistically significant difference with EFS_{LR} (1.60). The model WP_{GNReg} had a statistically significant difference with EFS_{LR} (1.94) and EFS_{GNReg} (1.14). In terms of **R-squared**, the model NASA_{LR} had a statistically significant difference ($p_{value} < 0.001$) with NASA_{GNReg} (0.12), WP_{LR} (0.23), WP_{GNReg} (0.27), EFS_{GNReg} (0.13). The model NASA_{GNReg} had a statistically significant difference with WP_{LR} (0.11), WP_{GNReg} (0.15). The model WP_{LR} had a statistically significant difference with EFS_{LR} (-0.16), EFS_{GNReg} (-0.10). The model WP_{GNReg} had a statistically significant difference with EFS_{LR} (-0.20) and EFS_{GNReg} (-0.14).

Test results

Table 4.48: Test results of Mental Workload score in comparison of RMSE, R-squared

	$NASA_{LR}$	$NASA_{GNReg}$	WP_{LR}	WP_{GNReg}	EFS_{LR}	EFS_{GNReg}
$_{Mean}$R-squared	0.272	0.195	0.232	0.246	0.369	0.257
$_{Mean}$RMSE	12.55	13.39	14.94	14.82	13.30	14.46

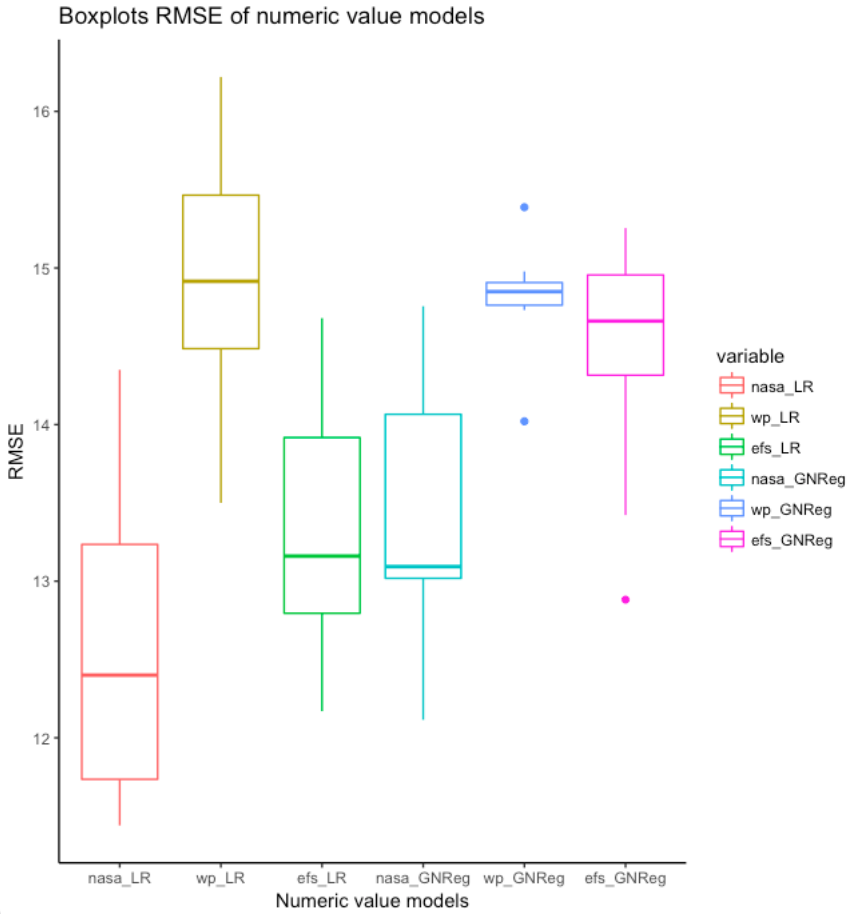

Figure 4.42: Test results of RMSE of Mental Workload score boxplots

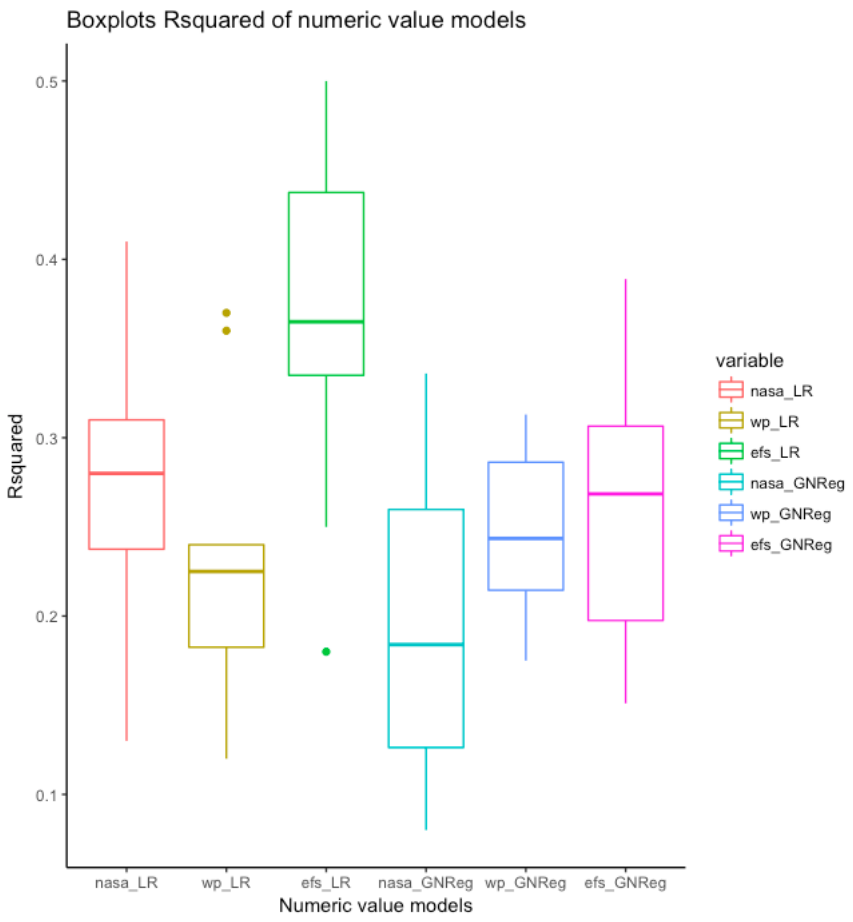

Figure 4.43: Test results of R-squared of Mental Workload score boxplots

A Bartlett test was conducted to evaluate the difference of RMSE in 6 models when treating Mental Workload as a continuous feature. The data of variance in RMSE did not significantly differ between models of Mental Workload as a continuous feature ($p = 0.12$). A one-way between groups analysis of variance was conducted to explore the impact of models of Mental Workload as a continuous feature on RMSE. There was a statistically significant difference at the $p<0.05$ level in RMSE for six models: $F = 14.665$, $p<0.001$.

As same for R-squared, the hypothesis test steps proceeded. A Bartlett test was conducted to evaluate the difference of R-squared in 6 models when treating Mental Workload as a continuous feature. The data of variance in R-squared did not significantly differ between models of Mental Workload as a continuous feature ($p = 0.47$). A one-way between groups analysis of variance was conducted to explore the impact

of models of Mental Workload as a continuous feature on R-squared. There was a statistically significant difference at the $p<0.05$ level in R-squared for six models: $F = 5.553$, $p<0.001$.

As in table 4.48 and figure 4.42, figure 4.43, in comparison to Linear Regression and Gini Regression within rating scales, Linear Regression was better with a higher R-squared and lower RMSE. In comparison to Mental Workload score between rating scales, on the one hand, EFS_{LR} had the highest R-squared (=0.37) which indicated how close the data are to the fitted regression line or the model fits data; next to was $NASA_{LR}$ (=0.27). On the other hand, $NASA_{LR}$ had the lowest RMSE (=12.55), which emphasizes large individual errors.

Models of Mental Workload as categorical feature

Training results

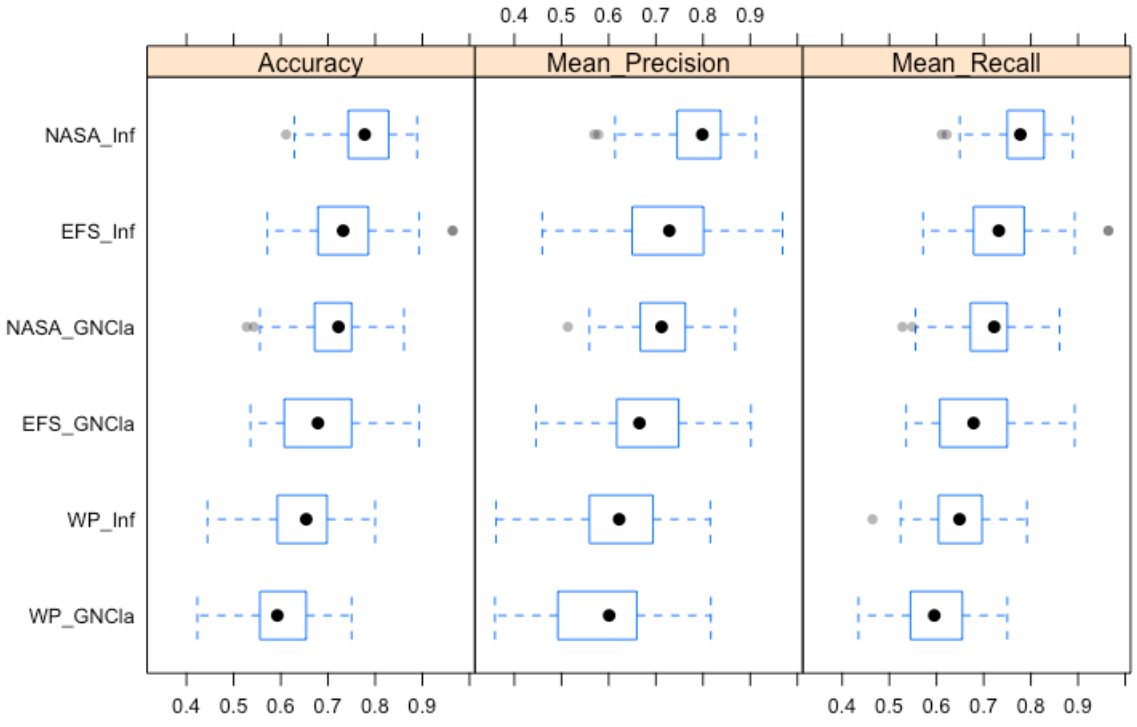

Figure 4.44: Training results of Accuracy, Precision, Recall of Mental Workload classes boxplots

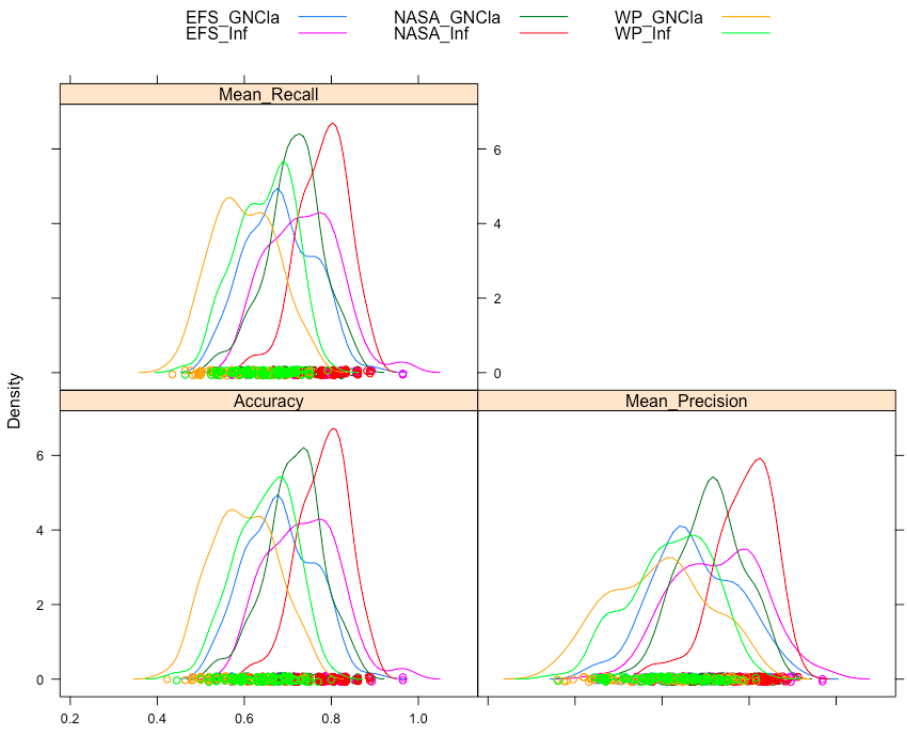

Figure 4.45: Training results of Accuracy, Precision, Recall of Mental Workload classes density plots

```
> summary(difValues2)

Call:
summary.diff.resamples(object = difValues2)

p-value adjustment: bonferroni
Upper diagonal: estimates of the difference
Lower diagonal: p-value for H0: difference = 0

Accuracy
           NASA_Inf  NASA_GNCla WP_Inf    WP_GNCla  EFS_Inf   EFS_GNCla
NASA_Inf              0.06829    0.13512   0.18215   0.04773   0.10130
NASA_GNCla 2.338e-11             0.06683   0.11385  -0.02056   0.03301
WP_Inf     < 2.2e-16 3.035e-10             0.04702  -0.08739  -0.03382
WP_GNCla   < 2.2e-16 < 2.2e-16  4.641e-05           -0.13442  -0.08084
EFS_Inf    0.0001431 0.8421085  1.241e-10 < 2.2e-16            0.05357
EFS_GNCla  < 2.2e-16 0.0261984  0.0136907 4.304e-10 0.0001696
```

Figure 4.46: Significance test of difference (lower) & estimates of the difference of Accuracy in Mental Workload level classes

```
Mean_Precision
           NASA_Inf  NASA_GNCla WP_Inf    WP_GNCla  EFS_Inf   EFS_GNCla
NASA_Inf              0.07300    0.16688   0.19391   0.05823   0.11187
NASA_GNCla 4.655e-10             0.09404   0.12288  -0.01280   0.03990
WP_Inf     < 2.2e-16 2.154e-12             0.02721  -0.10769  -0.05501
WP_GNCla   < 2.2e-16 1.164e-12  0.9494615           -0.13634  -0.08411
EFS_Inf    0.0001799 1.0000000  2.789e-10 1.556e-13             0.05030
EFS_GNCla  9.459e-15 0.0225858  0.0010225 1.204e-05 0.0136550

Mean_Recall
           NASA_Inf  NASA_GNCla WP_Inf    WP_GNCla  EFS_Inf   EFS_GNCla
NASA_Inf              0.06792    0.13506   0.18101   0.04738   0.10095
NASA_GNCla 2.377e-11             0.06714   0.11310  -0.02054   0.03304
WP_Inf     < 2.2e-16 1.982e-10             0.04595  -0.08768  -0.03411
WP_GNCla   < 2.2e-16 < 2.2e-16  4.955e-05           -0.13363  -0.08006
EFS_Inf    0.0001753 0.8222708  5.176e-11 < 2.2e-16             0.05357
EFS_GNCla  < 2.2e-16 0.0262922  0.0102511 3.033e-10 0.0001696
```

Figure 4.47: Significance test of difference (lower) & estimates of the difference of Mental Workload in Precision & Recall

In the box plots and density plots of Accuracy, Precision and Recall (figure 4.44, figure 4.45, the higher in box plots, the more right-inclined in density plots and conversely. With the training-result that NASA_{Inf} of Accuracy, Precision and Recall of Mental Workload classification had the best, next to EFS_{Inf}. There was a difference between models but how significantly different was shown in figure 4.46, 4.47.

As the threshold of significant test 0.05, in terms of **Accuracy**, the model NASA_{Inf} had a statistically significant difference ($p_{value} < 0.001$) with the difference estimation of NASA_{GNClas} (0.07), WP_{Inf} (0.14), WP_{GNClas} (0.18), EFS_{Inf} (0.05), EFS_{GNClas} (0.10). The model NASA_{GNClas} had a statistically significant difference with WP_{Inf} (0.07), WP_{GNClas} (0.11), EFS_{GNClas} (0.03). The model WP_{Inf} had a statistically significant difference with WP_{GNClas} (0.05), EFS_{Inf} (-0.09), EFS_{GNClas} (-0.03). The model WP_{GNClas} had a statistically significant difference with EFS_{Inf} (-0.13) and EFS_{GNClas} (-0.08). The model EFS_{Inf} had a statistically significant difference with EFS_{GNClas} (0.05). In terms of **Precision**, the model NASA_{Inf} had a statistically significant difference ($p_{value} < 0.001$) with NASA_{GNClas} (0.07), WP_{Inf} (0.17), WP_{GNClas} (0.19), EFS_{Inf} (0.06), EFS_{GNClas} (0.11). The model NASA_{GNClas} had a statistically significant difference with WP_{Inf} (0.09), WP_{GNClas} (0.12), EFS_{GNClas} (0.04). The model WP_{Inf} had a statistically significant difference with EFS_{Inf} (-0.11), EFS_{GNClas}

(-0.06). The model WP_{GNClas} had a statistically significant difference with EFS_{Inf} (-0.14) and EFS_{GNClas} (-0.08). The model EFS_{Inf} had a statistically significant difference with EFS_{GNClas} (0.05). In terms of **Recall**, the model NASA_{Inf} had a statistically significant difference ($p_{value} < 0.001$) with NASA_{GNClas} (0.07), WP_{Inf} (0.14), WP_{GNClas} (0.18), EFS_{Inf} (0.05), EFS_{GNClas} (0.10). The model NASA_{GNClas} had a statistically significant difference with WP_{Inf} (0.07), WP_{GNClas} (0.11), EFS_{GNClas} (0.03). The model WP_{Inf} had a statistically significant difference with WP_{GNClas} (0.05), EFS_{Inf} (-0.09), EFS_{GNClas} (-0.03). The model WP_{GNClas} had a statistically significant difference with EFS_{Inf} (-0.13) and EFS_{GNClas} (-0.08). The model EFS_{Inf} had a statistically significant difference with EFS_{GNClas} (0.05).

Test results

Table 4.49: Test results of Mental Workload classes in comparison of Accuracy, Precision, Recall

	NASA_{Inf}	NASA_{GNClas}	WP_{Inf}	WP_{GNClas}	EFS_{Inf}	EFS_{GNClas}
$_{Mean}$Accuracy	0.782	0.714	0.647	0.600	0.734	0.681
$_{Mean}$Precision	0.787	0.714	0.620	0.592	0.728	0.675
$_{Mean}$Recall	0.782	0.714	0.647	0.601	0.734	0.681

A Bartlett test was conducted to evaluate the difference of Accuracy in 6 models when treating Mental Workload as a categorical feature. There was a statistically significant difference in data of variance in Accuracy ($p < 0.001$, $df = 23$). So, Kruskal-Wallis test was applied to find some statistically significant differences between models of the Accuracy of Mental Workload as a categorical feature ($p<0.001$, Kruskal-Wallis chi-squared = 141.61, df = 23).

A Bartlett test was conducted to evaluate the difference of Precision in 6 models when treating Mental Workload as a categorical feature. There was a statistically significant difference in data of variance in Precision ($p < 0.001$, $df = 23$). Kruskal-Wallis test was applied to find some statistically significant differences between models of the Precision of Mental Workload as a categorical feature ($p<0.001$, Kruskal-Wallis chi-squared = 125.88, df = 23).

A Bartlett test was conducted to evaluate the difference of Recall in 6 models when treating Mental Workload as a categorical feature. There was a statistically significant difference in data of variance in Recall ($p < 0.001$, $df = 23$). Kruskal-Wallis test was applied to find some statistically significant differences between models of the Recall of Mental Workload as a categorical feature ($p<0.001$, Kruskal-Wallis chi-squared = 118.41, $df = 23$).

Table 4.50: Legendary for boxplots of 6 models of Accuracy, Precision, Recall

V1	NASA Information Gain underload	V13	NASA Gini Classification underload
V2	NASA Information Gain optimal load 1	V14	NASA Gini Classification optimal load 1
V3	NASA Information Gain optimal load 2	V15	NASA Gini Classification optimal load 2
V4	NASA Information Gain overload	V16	NASA Gini Classification overload
V5	WP Information Gain underload	V17	WP Gini Classification underload
V6	WP Information Gain optimal load 1	V18	WP Gini Classification optimal load 1
V7	WP Information Gain optimal load 2	V19	WP Gini Classification optimal load 2
V8	WP Information Gain overload	V20	WP Gini Classification overload
V9	EFS Information Gain underload	V21	EFS Gini Classification underload
V10	EFS Information Gain optimal load 1	V22	EFS Gini Classification optimal load 1
V11	EFS Information Gain optimal load 2	V23	EFS Gini Classification optimal load 2
V12	EFS Information Gain overload	V24	EFS Gini Classification overload

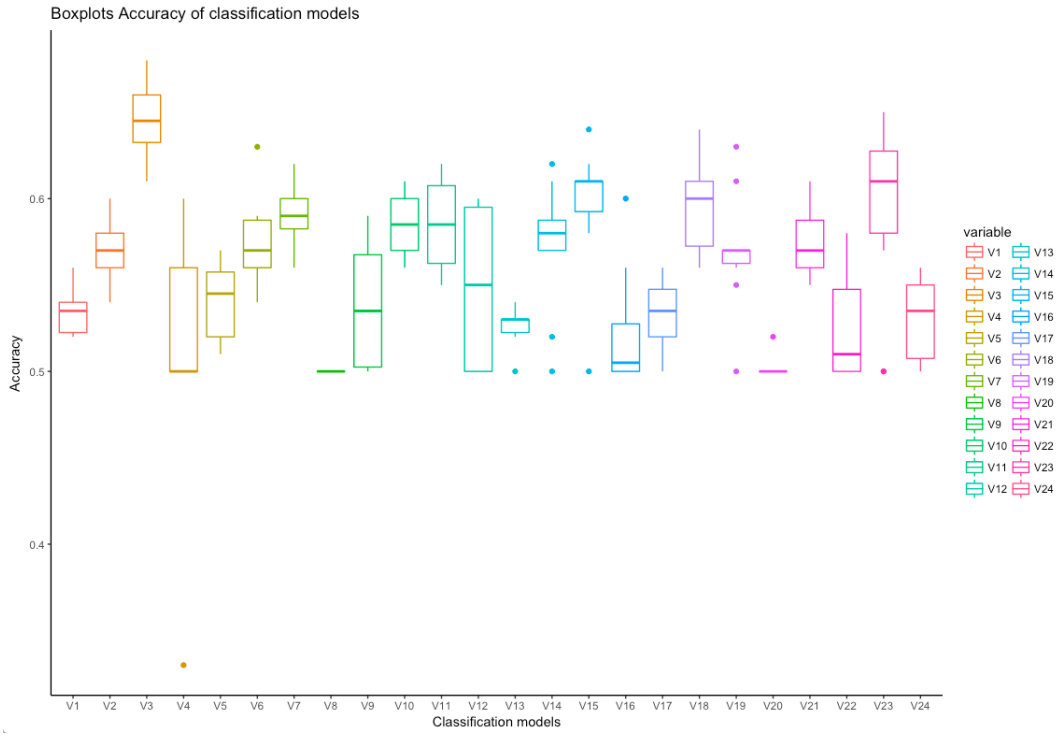

Figure 4.48: Test results of Accuracy of Mental Workload classes boxplots

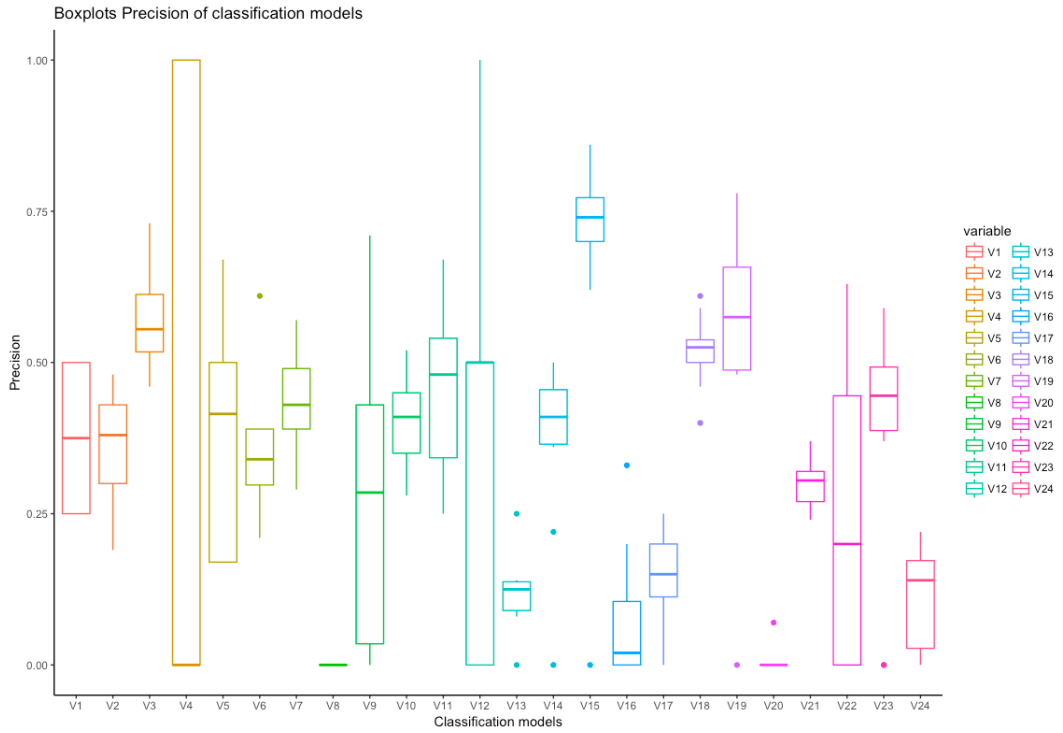

Figure 4.49: Test results of Precision of Mental Workload classes boxplots

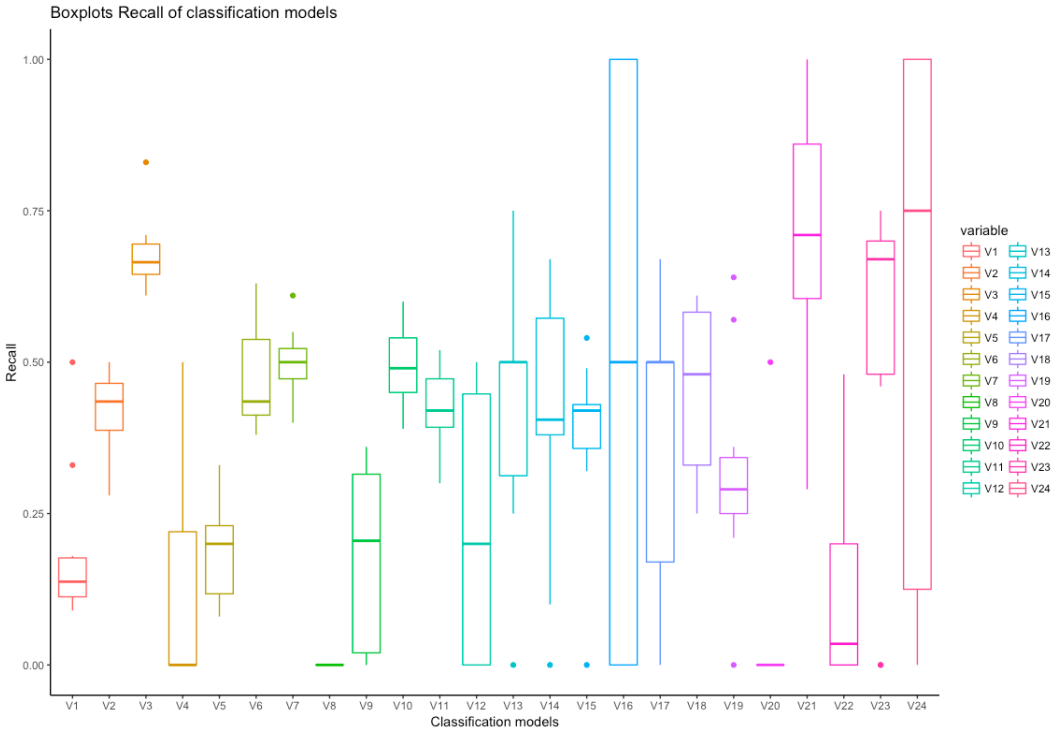

Figure 4.50: Test results of Recall of Mental Workload classes boxplots

Looking at table 4.49, in comparison to Information Gain and Gini Classification within each rating scales, Information Gain was better with a higher Accuracy, Precision, and Recall. In comparison to Mental Workload score between rating scales, NASA_{Inf} was the best having the highest Accuracy (=0.782), Precision (=0.787), Recall (=0.782) which indicated how optimal the predictive models are; next to was EFS_{Inf} with Accuracy (=0.734), Precision (=0.728), Recall (=0.734).

In figure 4.48, Accuracy or classification accuracy is the number of correct classifications in the total number on each class within and between models, which was the highest in NASA_{Inf} optimal load 2; and most cases with optimal load 2 in (WP_{Inf}, NASA_{GNClas}, EFS_{GNClas}), as same optimal load 1 as optimal load 2 in (EFS_{Inf}), optimal load 1 in (WP_{GNClas}). In figure 4.49, Precision captures when a model makes correct prediction, which was the highest in NASA_{GNClas} optimal load 2; also the rest of all with optimal load 2 in (NASA_{Inf}, WP_{Inf}, EFS_{Inf}, WP_{GNClas}, EFS_{GNClas}). In figure 4.50, Recall defines how confident that all the instances with a positive target level have been found, which was the highest in NASA_{Inf} optimal load 2; and two

cases with optimal load 2 in ($NASA_{Inf}$, WP_{Inf}, two cases with optimal load 1 in (EFS_{Inf}, WP_{GNClas}), two cases with overload in ($NASA_{GNClas}$, EFS_{GNClas}).

As the result above, the best model regarding Mental Workload score was Linear Regression, and regarding classes of Mental Workload was Decision Tree Information Gain. NASA (G Hart, 2006) showed the best measure between WP and EFS when the validated indicators had a statistically significant difference among the other models.

4.6 Strengths and limitations of the results

4.6.1 Strengths of the results

The primary strength of the results is the establishment of an interpretable model of Mental Workload in the third level of education that would be easy to consider the Mental Workload and student interaction. Consequently, the purpose is detecting and improving the student performance in the early stage.

The secondary strength is discovering the main factors in third level education, which affect student Mental Workload. This discovery will be the evidence-based design for the research relevant to the Cognitive Theory of Multimedia Learning and active learning in the future.

The final strength of the findings is support of machine learning in feature selection, which aims at auto-training and testing for more features or more extensive data sets.

4.6.2 Limitations of the results

In the aspect of modeling, there were more domain representations (20 independent features) in EFS than NASA (6 independent features) and WP (8 independent features), the more instances would be needed (larger sample size) in analyzing EFS in Decision Tree models.

In the aspect of subjective rating scales, WP was used for the overloaded environment such as working conditions, multi-task jobs. However, the design of Research

Design & Proposal Writing module did not cause overload state on a student. So, there may create restrictions of WP subjective assessment. This restriction was also in the findings in (Luximon & Goonetilleke, 2001) with SWAT assessment techniques. Also, in Tsang and L. Velazquez (1996) it was suggested that the individual workload profiles would only have limited predictive value on performance.

Last but not least, the target feature in three subjective rating scales were the perception of Mental Workload in students, which was uni-dimensional measures. It was sensitive but not specific to differentiate the difference in a small data set. For this limitation, the comparison of two target features: the perception of MWL and the MWL calculated based on its internal factors would be the better solution.

Chapter 5

Conclusion

5.1 Research Overview

As mentioned in the background, Mental Workload can affect mental state which then has an impact on task performance or the learners capacity to absorb knowledge. This study evaluated the Mental Workload on three subjective rating scales in third level education. From then, some findings are as:

- The Extended Feature Sets which is the mixture of NASA and WP factors with additional factors showed the potential rating scale in the multi-dimensional measurement in education;
- With the easy-to-interpret and training, Multiple Linear Regression and Decision Tree Information Gain will be the driver models for more research in this area.

5.2 Problem Definition

The model trained with Extended Feature Sets was significantly more accurate and had less errors in predicting perception of Mental Workload than the model trained with Workload Profile; but not in the circumstance of NASA - Task Load Index. The reason may be the small sample size of EFS, despite having more features than the others.

5.3 Design, Evaluation & Results

Two learning methods applied to the study are error-based and information-based as Multiple Linear Regression and Decision Tree Information Gain, Decision Tree GINI Regression, Decision Tree GINI Classification. This study has drawn one point of attention that having two different models testing the two types of output feature simultaneously: Linear Regression and Decision Tree GINI Regression trained and tested on Mental Workload as a continuous feature; Decision Tree Information Gain and Decision Tree GINI Classification trained and tested on Mental Workload as an ordinal feature. RMSE and R-squared are useful for assessing the Mental Workload as continuous feature models; while the Accuracy, Precision, and Recall are for the Mental Workload as ordinal feature models. The results highlight some points below:

- Mental demand, Temporal demand, Frustration, Effort, Central Processing, Visual attention and Parallelism were significant factors;
- Decision Tree Information Gain and Linear Regression were the optimal model to predict Mental Workload as categorical target or continuous target;
- Seeing Mental Workload as a categorical feature, the Accuracy, Precision and Recall in optimal load 2 were the highest; next to optimal load 1, underload and overload. However, there were some exceptions in the case of WP Gini Classification in that precision and recall of optimal load 2 were lower than optimal load 1. The same situation happened for EFS Information Gain.

5.4 Contributions and impact

The study is neither the novice research nor is it to highlight machine learning methods. However, the results and findings of this research will contribute to the ground of Mental Workload research in education as well as suggesting the measurable and feasible models for machine learning.

In consideration of prediction speed and capacity for retraining, Multiple Linear Regression and Decision Tree are standards and the ideal suggestion.

5.5 Future Work & recommendations

The study needs to have more experiments on the other technique, such as Support Vector Machine or ensemble methods when considering prediction accuracy. Otherwise, the same method can apply to enlarge sample size for EFS rating scales and putting the weighted Mental Workload measured by internal factors in models.

In addition the study could be useful for Human-centered computing by exploring Interaction design theory, concepts & paradigms for Interaction design applications. The most applicable idea is to create a predictive model including factors to reduce student's risk of failure from early detection of increasing Mental workload.

References

Adar, T., & Delice, E. K. (2017). Evaluating mental work load using multi-criteria hesitant fuzzy linguistic term set (hflts). *Turkish Journal of Fuzzy Systems (TJFS)*, *8*(2), 90-101. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=127528652&site=ehost-live

Agostinho, S., Tindall-Ford, S., Ginns, P., Howard, S., Leahy, W., & Paas, F. (2015). Giving learning a helping hand: Finger tracing of temperature graphs on an ipad. *Educational Psychology Review*, *27*(3), 427-443. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=108930574&site=ehost-live

Balfe, N., Crowley, K., Smith, B., & Longo, L. (2017). Estimation of train driver workload: Extracting taskload measures from on-train-data-recorders. *International Symposium on Human Mental Workload: Models and Applications*, 106-119.

Blayney, P., Kalyuga, S., & Sweller, J. (2015). Using cognitive load theory to tailor instruction to levels of accounting students' expertise. *Journal of Educational Technology Society*, *18*(4), 199-210. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=110247514&site=ehost-live

Byrne, A., Tweed, N., & Halligan, C. (2014). A pilot study of the mental workload of objective structured clinical examination examiners. *Medical Education*, *48*(3), 262-267. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=94449019&site=ehost-live

Cain, B. (2007). A review of the mental workload literature. *Defence Research and Development Canada, Toronto*.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction*, *8*(4), 293-332.

Cheon, J., & Grant, M. (2012). The effects of metaphorical interface on germane cognitive load in web-based instruction. *Educational Technology Research Development*, *60*(3), 399-420. Retrieved from `http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=75177485&site=ehost-live`

Colligan, L., Potts, H. W., Finn, C. T., & Sinkin, R. A. (2015). Cognitive workload changes for nurses transitioning from a legacy system with paper documentation to a commercial electronic health record. *International Journal of Medical Informatics*, *84*(7), 469-476. Retrieved from `http://www.sciencedirect.com/science/article/pii/S1386505615000635` doi: https://doi.org/10.1016/j.ijmedinf.2015.03.003

Colombi, J. M., Miller, M. E., Schneider, M., McGrogan, M. J., Long, C. D. S., & Plaga, J. (2012). Predictive mental workload modeling for semiautonomous system design: Implications for systems of systems. *Systems Engineering*, *15*(4), 448-460. Retrieved from `http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=82370471&site=ehost-live`

Davids, M. R., Halperin, M. L., & Chikte, U. M. E. (2015). Optimising cognitive load and usability to improve the impact of e-learning in medical education. *African Journal of Health Professions Education*, *7*(2), 147-152. Retrieved from `http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=111920213&site=ehost-live`

D.Kelleher, J., Namee, B. M., & Darcy, A. (2015). *Fundamentals of machine learning for predictive data analytics*. MIT Press.

Foo, J.-L., Martinez-Escobar, M., Juhnke, B., Cassidy, K., Hisley, K., Lobe, T., & Winer, E. (2013). Evaluating mental workload of two-dimensional and three-dimensional visualization for anatomical structure localization. *Journal of Laparoendoscopic Advanced Surgical Techniques*, *23*(1), 65-70. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=84766253&site=ehost-live

Fredricks, J., Blumenfeld, P., & Paris, A. (2004). School engagement: Potential of the concept, state of the evidence. *Review of educational research*, *74*(1), 59-109.

Galy, E., & Mlan, C. (2015). Effects of cognitive appraisal and mental workload factors on performance in an arithmetic task. *Applied Psychophysiology Biofeedback*, *40*(4), 313-325. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=111003263&site=ehost-live

Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2013). *An introduction to statistical learning with applications in r*. Springer.

G Hart, S. (2006). Nasa-task load index (nasa-tlx); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*, 904-908. doi: 10.1177/154193120605000909

G. Hart, S., & E. Stavenland, L. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in Psychology*, *52*, 139-183. doi: 10.1016/S0166-4115(08)62386-9

Gough Young, B., Wodehouse, A., & Sheridan, M. (2015). Questioning conventions: Are product conventions trading off the usability of products for short term user satisfaction. *International Journal of Cognitive Research in Science, Engineering Education (IJCRSEE)*, *3*(2), 47-58. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=111816456&site=ehost-live

Guastello, S. J., Marra, D. E., Correro, A. N., Michels, M., & Schimmel, H. (2017). Elasticity and rigidity constructs and ratings of subjective workload for individuals

and groups. *International Symposium on Human Mental Workload: Models and Applications*, 51-76.

Haji, F. A., Cheung, J. J. H., Woods, N., Regehr, G., Ribaupierre, S., & Dubrowski, A. (2016). Thrive or overload? the effect of task complexity on novices' simulation-based learning. *Medical Education*, *50*(9), 955-968. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=117672505&site=ehost-live

Haji, F. A., Rojas, D., Childs, R., Ribaupierre, S., & Dubrowski, A. (2015). Measuring cognitive load: performance, mental effort and simulation task complexity. *Medical Education*, *49*(8), 815-827. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=103668919&site=ehost-live

Hancock, P. A. (2017). Whither workload? mapping a path for its future development. in international symposium on human mental workload: Models and applications. *Conference: International Symposium on Human Mental Workload: Models and Applications*, 3-17.

Jaewon, J., Dongsik, K., & Chungsoo, N. (2016). Effects of woe presentation types used in pre-training on the cognitive load and comprehension of content in animation-based learning environments. *Journal of Educational Technology Society*, *19*(4), 75-86. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=120696888&site=ehost-live

Jeng-Chung, W. (2014). Digital game-based learning supports student motivation, cognitive success, and performance outcomes. *Journal of Educational Technology Society*, *17*(3), 291-307. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=98543306&site=ehost-live

Jihyun, S., Dongsik, K., & Chungsoo, N. (2014). Adaptive instruction to learner expertise with bimodal process-oriented worked-out examples. *Journal of Educational Technology Society*, *17*(1), 259-271. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=94937816&site=ehost-live

Jimenez-Molina, A., Retamal, C., & Lira, H. (2018). Using psychophysiological sensors to assess mental workload during web browsing. *Sensors (Basel)*, *18*(2), 458. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=128260693&site=ehost-live

Joshi, A., Kale, S., Chandel, S., & Pal, D. (2015). Likert scale: Explored and explained. *British Journal of Applied Science Technology*, *7*, 396-403. doi: 10.9734/BJAST/2015/14975

Kalyuga, S., Chandler, P., & Sweller, J. (1998). Levels of expertise and instructional design. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *40*(1), 1-17.

Kalyuga, S., & Singh, A.-M. (2016). Rethinking the boundaries of cognitive load theory in complex learning. *Educational Psychology Review*, *28*(4), 831-852. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=119539238&site=ehost-live

Kim, B. (2001). Social constructivism. *Emerging perspectives on learning, teaching, and technology*, *1*(1), 16.

Kuo-Kuang, F., Chung-Ho, S., Shuh-Yeuan, D., & Wei-Jhung, W. (2013). An achievement prediction model of meaningful learning, motivation, and cognitive on spani: Partial least square analysis. *Mathematical Problems in Engineering*, 1-11. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=94814272&site=ehost-live

Leahy, W., Hanham, J., & Sweller, J. (2015). High element interactivity information during problem solving may lead to failure to obtain the testing effect. *Educational Psychology Review*, *27*(2), 291-304. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=102748390&site=ehost-live

Liew, T. W., & Tan, S.-M. (2016). The effects of positive and negative mood on cognition and motivation in multimedia learning environment. *Journal of Educational*

Technology Society, *19*(2), 104-115. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=114601237&site=ehost-live

Lin, L., Lee, C. H., Kalyuga, S., Wang, Y., Guan, S., & Wu, H. (2017). The effect of learner-generated drawing and imagination in comprehending a science text. *Journal of Experimental Education*, *85*(1), 142-154. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=119150851&site=ehost-live

Liu, T.-C., Lin, Y.-C., & Paas, F. (2013). Effects of cues and real objects on learning in a mobile device supported environment. *British Journal of Educational Technology*, *44*(3), 386-399. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=86864411&site=ehost-live

Longo, L. (2015a). A defeasible reasoning framework for human mental workload representation and assessment. *Behaviour Information Technology*, *34*(8), 758-786. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=103122869&site=ehost-live

Longo, L. (2015b, June). Designing medical interactive systems via assessment of human mental workload. *Computer-Based Medical Systems (CBMS)*, 364-365.

Longo, L. (2016, June). Mental workload in medicine: foundations, applications, open problems, challenges and future perspectives. *Computer-Based Medical Systems (CBMS)*, 106-111.

Longo, L. (2017, September). Subjective usability, mental workload assessments and their impact on objective human performance. *IFIP Conference on Human-Computer Interaction*, 202-223.

Longo, L. (2018). Experienced mental workload, perception of usability, their interaction and impact on task performance. *PLoS ONE*, *13*(8), 1-36. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=131014115&site=ehost-live

Longo, L. (2018, March). On the reliability, validity and sensitivity of three mental workload assessment techniques for the evaluation of instructional designs: a case study in a third-level course. *CSEDU 2018*, *2*, 166-178.

Longo, L., & Barrett, S. (2010, August). Cognitive effort for multi-agent systems. *Yao Y., Sun R., Poggio T., Liu J., Zhong N., Huang J. (eds) Brain Informatics. BI 2010. Lecture Notes in Computer Science*, *6334*, 55-66.

Longo, L., & Dondio, P. (2015, December). On the relationship between perception of usability and subjective mental workload of web interfaces. *Web Intelligence and Intelligent Agent Technology (WI-IAT)*, *1*, 345-352.

Longo, L., & Leva, M. C. E. (2017). Human mental workload: Models and applications: First international symposium. *H-WORKLOAD 2017, Dublin, Ireland, June 28-30, 2017, Revised Selected Papers*, *726*.

Luximon, A., & Goonetilleke, R. (2001). Simplified subjective workload assessment technique. *Ergonomics*, *44*, 229-243. doi: 10.1080/00140130010000901

Macbeth, G., Razumiejczyk, E., del Carmen Crivello, M., Bolzn, C., Pereyra Girardi, C. I., & Campitelli, G. (2014). Mental models for the negation of conjunctions and disjunctions. *Europe's Journal of Psychology*, *10*(1), 135-149. Retrieved from `http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=94867725&site=ehost-live`

Macken, L., & Ginns, P. (2014). Pointing and tracing gestures may enhance anatomy and physiology learning. *Medical Teacher*, *36*(7), 596-601. Retrieved from `http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=96729515&site=ehost-live`

Max, K., & Kjell, J. (2013). *Applied predictive modelling*. Springer.

Mitropoulos, P., & Memarian, B. (2013). Task demands in masonry work: Sources, performance implications, and management strategies. *Journal of Construction Engi-*

neering Management, *139*(5), 581-590. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=86929016&site=ehost-live

Miyake, S. J. (2001). Multivariate workload evaluation combining physiological and subjective measures. *Psychophysiol*, *40*(3), 233-238.

Moustafa, K., Luz, S., & Longo, L. (2017). Assessment of mental workload: A comparison of machine learning methods and subjective assessment techniques. In *Longo L., Leva M. (eds) Human Mental Workload: Models and Applications. H-WORKLOAD 2017. Communications in Computer and Information Science* (Vol. 726, p. 30-50). Springer, Cham.

Naismith, L. M., Cheung, J. J. H., Ringsted, C., & Cavalcanti, R. B. (2015). Limitations of subjective cognitive load measures in simulation-based procedural training. *Medical Education*, *49*(8), 805-814. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=103668908&site=ehost-live

Ngu, B., & Phan, H. (2016). Unpacking the complexity of linear equations from a cognitive load theory perspective. *Educational Psychology Review*, *28*(1), 95-118. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=112966721&site=ehost-live

ODonnel, R., & Eggemeier, F. (1986). *Workload assessment methodology. handbook of perception and human performance* (Vol. 2). John Wiley and Sons, Inc.

Orru, G., Gobbo, F., O'Sullivan, D., & Longo, L. (2018). An investigation of the impact of a social constructivist teaching approach, based on trigger questions, through measures of mental workload and efficiency. In *Proceedings of the 10th international conference on computer supported education, CSEDU 2018, Funchal, Madeira, Portugal, March 15-17, 2018, volume 2.* (p. 292-302).

Ree, E., Odeen, M., Eriksen, H., Indahl, A., Ihlebk, C., Hetland, J., & Harris, A. (2014). Subjective health complaints and self-rated health: Are expectancies more

important than socioeconomic status and workload? *International Journal of Behavioral Medicine*, *21*(3), 411-420. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=95865667&site=ehost-live

Reid, G. B., & Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (Vol. 52, p. 185-218). North-Holland. Retrieved from http://www.sciencedirect.com/science/article/pii/S0166411508623870 doi: https://doi.org/10.1016/S0166-4115(08)62387-0

Rizzo, L., & Longo, L. (2017). Representing and inferring mental workload via defeasible reasoning: A comparison with the nasa task load index and the workload profile. In *Proceedings of the 1st workshop on Advances in Argumentation in Artificial Intelligence co-located with XVI International conference of the Italian Association for Artificial Intelligence, AI^3@AI*IA 2017, Bari, Italy, November 16-17, 2017* (p. 126-140). Retrieved from http://ceur-ws.org/Vol-2012/AI3-2017_paper_13.pdf

Rubio, S., Daz, E., Martn, J., & Puente, J. M. (2004). Evaluation of subjective mental workload: A comparison of swat, nasa-tlx, and workload profile methods. *Applied Psychology*, *53*(1), 61-86. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1464-0597.2004.00161.x doi: 10.1111/j.1464-0597.2004.00161.x

Schwonke, R. (2015). Metacognitive load - useful, or extraneous concept? metacognitive and self- regulatory demands in computer-based learning. *Journal of Educational Technology Society*, *18*(4), 172-184. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=110247512&site=ehost-live

Seker, A. (2014). Using outputs of nasa-tlx for building a mental workload expert system. *Gazi University Journal of Science*, *27*(4), 1131-1142. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=99765692&site=ehost-live

Sewell, J. L., Boscardin, C. K., Young, J. Q., Cate, O., & O'Sullivan, P. S. (2016). Measuring cognitive load during procedural skills train-

ing with colonoscopy as an exemplar. *Medical Education*, *50*(6), 682-692. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=115268657&site=ehost-live

Smith, K. T. (2017). Observations and issues in the application of cognitive workload modelling for decision making in complex time-critical environments. *International Symposium on Human Mental Workload: Models and Applications*, 77-89.

Stuart, R., & Peter, N. (2016). *Arificall intelligence - a modern approach*. Pearson.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 257-285.

Sweller, J. (1994). Cognitive load theory, learning difficulty and instructional design. *Learning and Instruction*, 295-312.

Sweller, V. M. J., J., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *3*(10), 251-296.

Tsang, P., & L. Velazquez, V. (1996). Diagnosticity and multidimensional subjective workload ratings. *Ergonomics*, *39*, 358-381. doi: 10.1080/00140139608964470

Valdehita, S., Ramiro, E., Garca, J., & M. Puente, J. (2004). Evaluation of subjective mental workload: a comparison of swat, nasa-tlx, and workload profile methods. *Applied Psychology: An International Review*, *53*(1), 61-86.

Wei, C., Chen, N., & Kinshuk. (2012). A model for social presence in online classrooms. *Educational Technology Research Development*, *3*(60), 529-545.

Weigl, M., Mller, A., Angerer, P., & Hoffmann, F. (2014). Workflow interruptions and mental workload in hospital pediatricians: an observational study. *BMC Health Services Research*, *14*(1), 433-440. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=99838194&site=ehost-live

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors*, *50*(3), 449-455. Retrieved from https://doi.org/10.1518/001872008X288394 (PMID: 18689052) doi: 10.1518/001872008X288394

Wickens, C. D. (2017). Mental workload: assessment, prediction and consequences. *International Symposium on Human Mental Workload: Models and Applications*, 18-29.

Wong, M., Castro-Alonso, C., J., Ayres, P., & Paas, F. (2015). Gender effects when learning manipulative tasks from instructional animations and static presentations. *Journal of Educational Technology Society*, *18*(4), 37-52. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=110247502&site=ehost-live

Xiaoru, W., Damin, Z., & Huan, Z. (2014). Improving pilot mental workload evaluation with combined measures. *Bio-Medical Materials Engineering*, *24*(6), 2283-2290. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=108585556&site=ehost-live

Xie, H., Wang, F., Hao, Y., Chen, J., An, J., Wang, Y., & Liu, H. (2017). The more total cognitive load is reduced by cues, the better retention and transfer of multimedia learning: A meta-analysis and two meta-regression analyses. *PLoS ONE*, *12*(8), 1-20. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=124894963&site=ehost-live

Xu, X., Xiaoru, W., & Damin, Z. (2015). Mental workload prediction based on attentional resource allocation and information processing. *Bio-Medical Materials Engineering*, *26*, S871-S879. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=109031174&site=ehost-live

Young, J. Q., Van Merrienboer, J., Durning, S., & Ten Cate, O. (2014). Cognitive load theory: Implications for medical education: Amee guide no. 86. *Medical Teacher*, *36*(5), 371-384. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=95661696&site=ehost-live

Yuling, H., Yuan, G., Tzu-Chien, L., & Sweller, J. (2015). Interactions between levels of instructional detail and expertise when learning with computer simulations. *Journal of Educational Technology Society*, *18*(4), 113-127. Retrieved from `http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=110247508&site=ehost-live`

Yung, H. I., & Paas, F. (2015). Effects of cueing by a pedagogical agent in an instructional animation: A cognitive load approach. *Journal of Educational Technology Society*, *18*(3), 153-160. Retrieved from `http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=109155628&site=ehost-live`

Zijlstra, F. R. (1993). Efficiency in work behavior: A design approach for modern tools. *Delft University Press, Holland*, 177.

Appendix A

Additional content

A.1 NASA Task Load Index

A.1.1 Data description

```
> shapiro.test(nasa$con_mental_workload)

        Shapiro-Wilk normality test

data:  nasa$con_mental_workload
W = 0.96697, p-value = 0.00003476

> shapiro.test(log(nasa$con_mental_workload))

        Shapiro-Wilk normality test

data:  log(nasa$con_mental_workload)
W = 0.87216, p-value = 0.0000000000005639

> shapiro.test(sqrt(nasa$con_mental_workload))

        Shapiro-Wilk normality test

data:  sqrt(nasa$con_mental_workload)
W = 0.93306, p-value = 0.00000000984
```

Figure A.1: Shapiro-wilk test of NASA normality

A.1.2 Model Training

Linear Regression

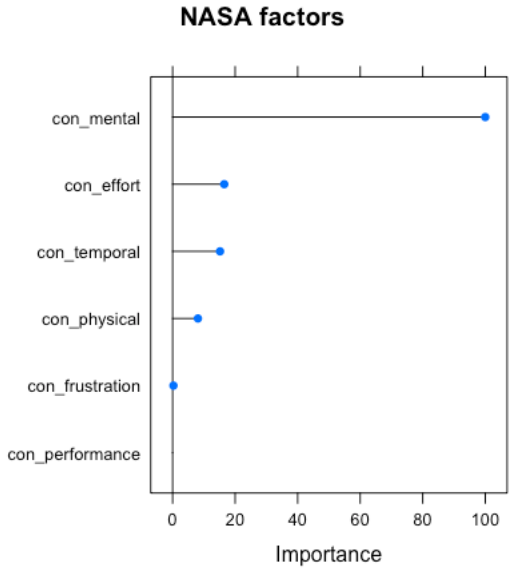

Figure A.2: Variable importance of NASA in model

Decision Tree Information Gain

```
> nasa_dt1
C5.0

153 samples
  6 predictor
  4 classes: 'optimal1', 'optimal2', 'overload', 'underload'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 10 times)
Summary of sample sizes: 138, 138, 139, 138, 137, 139, ...
Resampling results across tuning parameters:

  model  winnow  trials  Accuracy   Kappa
  rules  FALSE    1      0.5422328  0.07582268
  rules  FALSE   10      0.5417924  0.07165775
  rules  FALSE   20      0.5417924  0.07165775
  rules   TRUE    1      0.5318964  0.02916356
  rules   TRUE   10      0.5281625  0.02577855
  rules   TRUE   20      0.5287875  0.02729370
  tree   FALSE    1      0.5370343  0.06266565
  tree   FALSE   10      0.5388309  0.07584198
  tree   FALSE   20      0.5388309  0.07584198
  tree    TRUE    1      0.5291870  0.01912718
  tree    TRUE   10      0.5303943  0.02577488
  tree    TRUE   20      0.5303943  0.02577488

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were trials = 1, model = rules and winnow = FALSE.
```

Figure A.3: NASA decision tree trained by Information Gain (N=154)

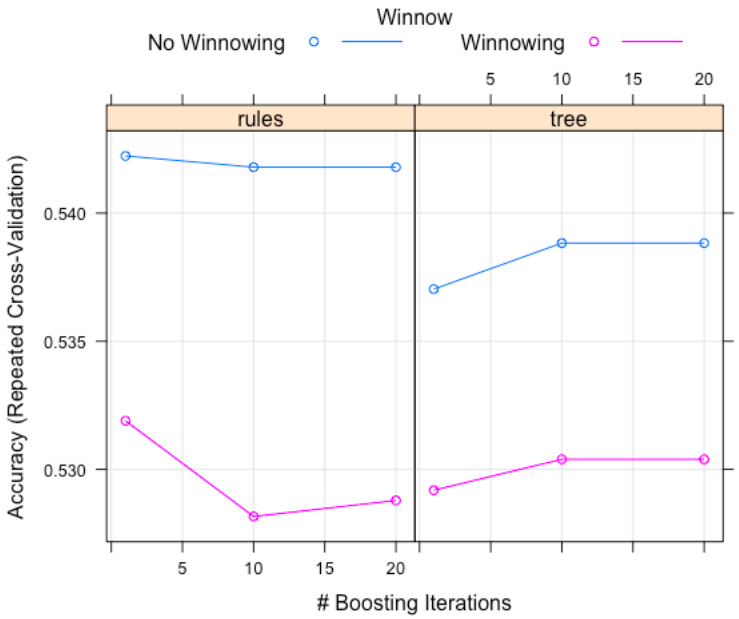

Figure A.4: NASA decision tree Information Gain with cross-validation (N=154)

```
> nasa_dt3  #missing values in resampled performance measures, NaN (precision & f1-measure)
C5.0

153 samples
  6 predictor
  4 classes: 'optimal1', 'optimal2', 'overload', 'underload'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 10 times)
Summary of sample sizes: 138, 138, 139, 138, 137, 139, ...
Resampling results:

  logLoss   AUC        prAUC      Accuracy   Kappa       Mean_F1  Mean_Sensitivity
  13.17685  0.5488464  0.1587154  0.5422328  0.07582268  NaN      0.2657407
  Mean_Specificity  Mean_Pos_Pred_Value  Mean_Neg_Pred_Value  Mean_Precision  Mean_Recall
  0.7694212         NaN                  0.780243             NaN             0.2657407
  Mean_Detection_Rate  Mean_Balanced_Accuracy
  0.1355582            0.5137145

Tuning parameter 'trials' was held constant at a value of 1
Tuning parameter 'model' was
 held constant at a value of rules
Tuning parameter 'winnow' was held constant at a value of FALSE
```

Figure A.5: NASA decision tree Information Gain trained by Grid, tuning parameters and cross-validation on actual sample (N=154)

```
> nasa_dt3
C5.0

352 samples
  6 predictor
  4 classes: 'optimal1', 'optimal2', 'overload', 'underload'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 10 times)
Summary of sample sizes: 316, 317, 318, 316, 317, 316, ...
Resampling results:

  logLoss   AUC        prAUC      Accuracy   Kappa      Mean_F1   Mean_Sensitivity
  8.504323  0.8359143  0.2437929  0.7450874  0.6600237  0.722943  0.7455556
  Mean_Specificity  Mean_Pos_Pred_Value  Mean_Neg_Pred_Value  Mean_Precision  Mean_Recall
  0.9149913         0.7517435            0.9231084            0.7517435       0.7455556
  Mean_Detection_Rate  Mean_Balanced_Accuracy
  0.1862719            0.8302734

Tuning parameter 'trials' was held constant at a value of 1
Tuning parameter 'model' was
 held constant at a value of rules
Tuning parameter 'winnow' was held constant at a value of FALSE
```

Figure A.6: NASA decision tree Information Gain trained by Grid, tuning parameters and cross-validation on upSampling (N=154)

Decision Tree Gini Regression

```
> nasa_GN1
CART

154 samples
  6 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 10 times)
Summary of sample sizes: 139, 139, 138, 139, 137, 138, ...
Resampling results across tuning parameters:

  cp          RMSE       Rsquared   MAE
  0.04156853  0.2858310  0.2896771  0.2175626
  0.07776363  0.2968590  0.2340753  0.2291056
  0.28328208  0.3168994  0.1325616  0.2455291

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was cp = 0.04156853.
```

Figure A.7: NASA model decision tree Regression trained by Gini Index (N=154)

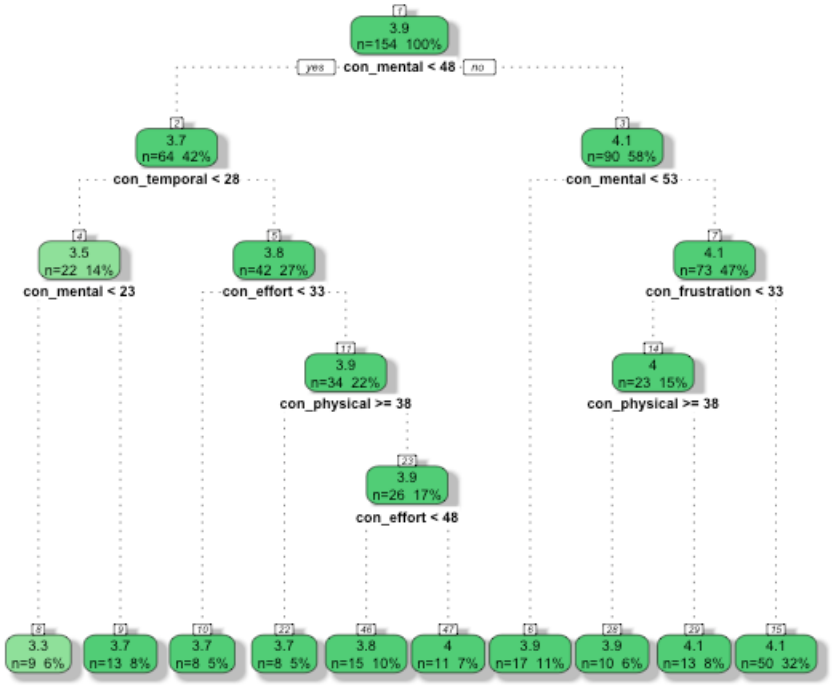

Figure A.8: NASA decision tree Regression trained by Gini Index (N=154)

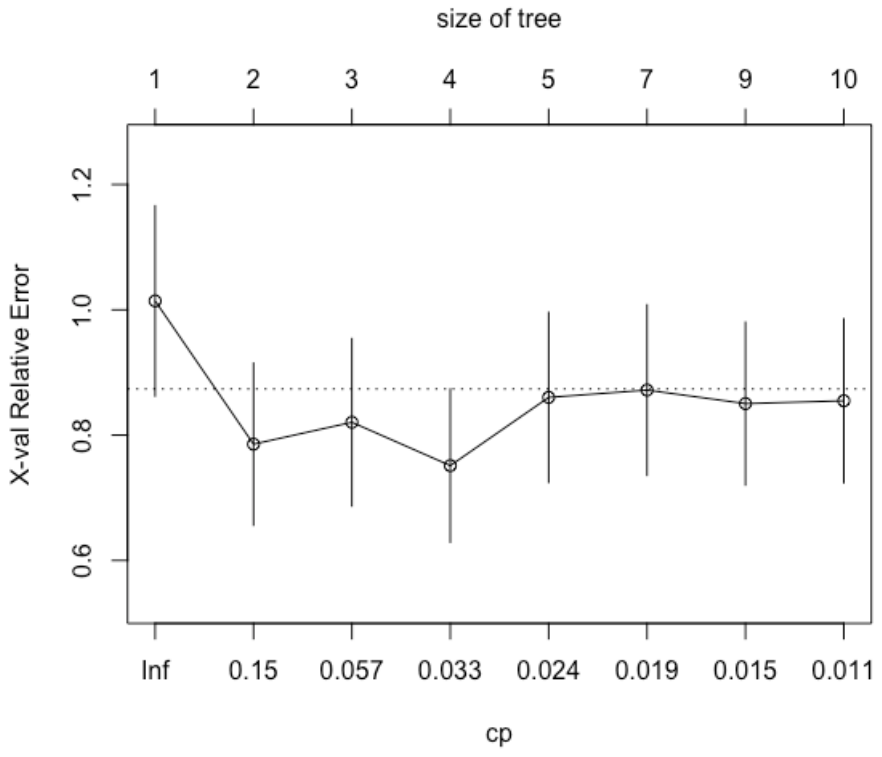

Figure A.9: NASA decision tree Gini Regression with cross-validation (N=154)

Decision Tree Gini Classification

```
> nasa_GN2
CART

154 samples
  6 predictor
  4 classes: 'optimal1', 'optimal2', 'overload', 'underload'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 10 times)
Summary of sample sizes: 139, 140, 139, 139, 138, 139, ...
Resampling results across tuning parameters:

  cp          Accuracy   Kappa
  0.01990050  0.5180994  0.05461500
  0.02985075  0.5381677  0.07424477
  0.11940299  0.5369072  0.01841484

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.02985075.
```

Figure A.10: NASA model decision tree Classification trained by Gini Index (N=154)

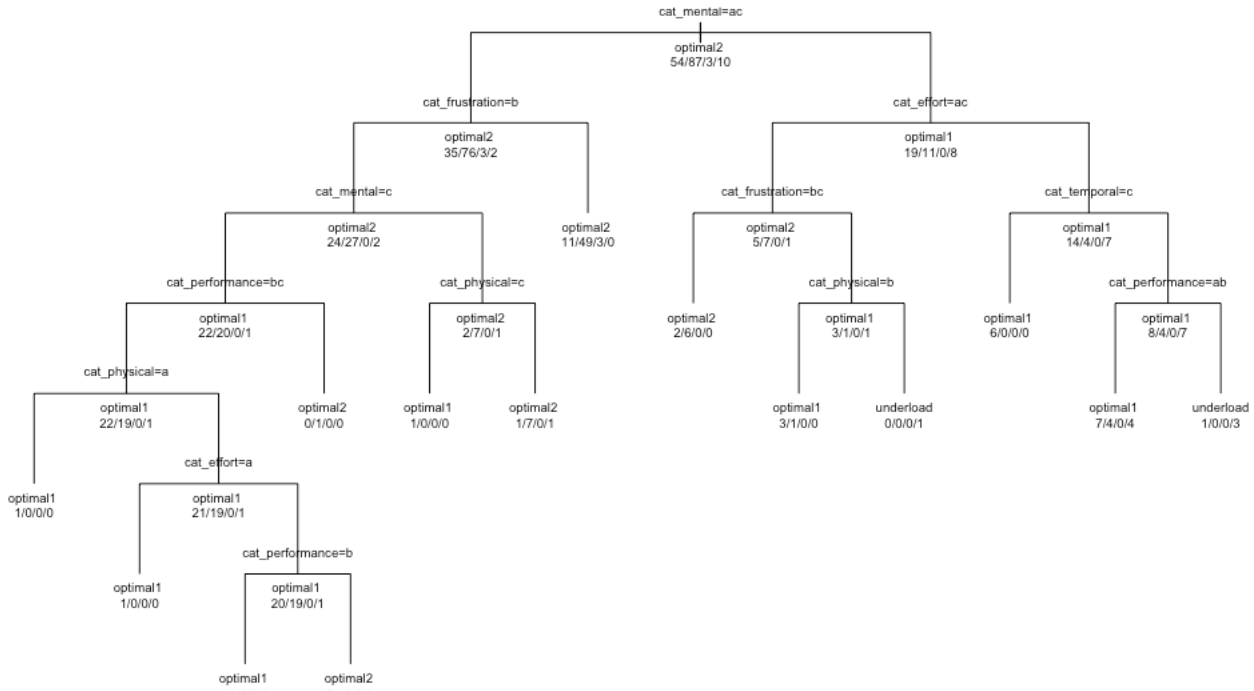

Figure A.11: NASA decision tree Classification trained by Gini Index (N=154)

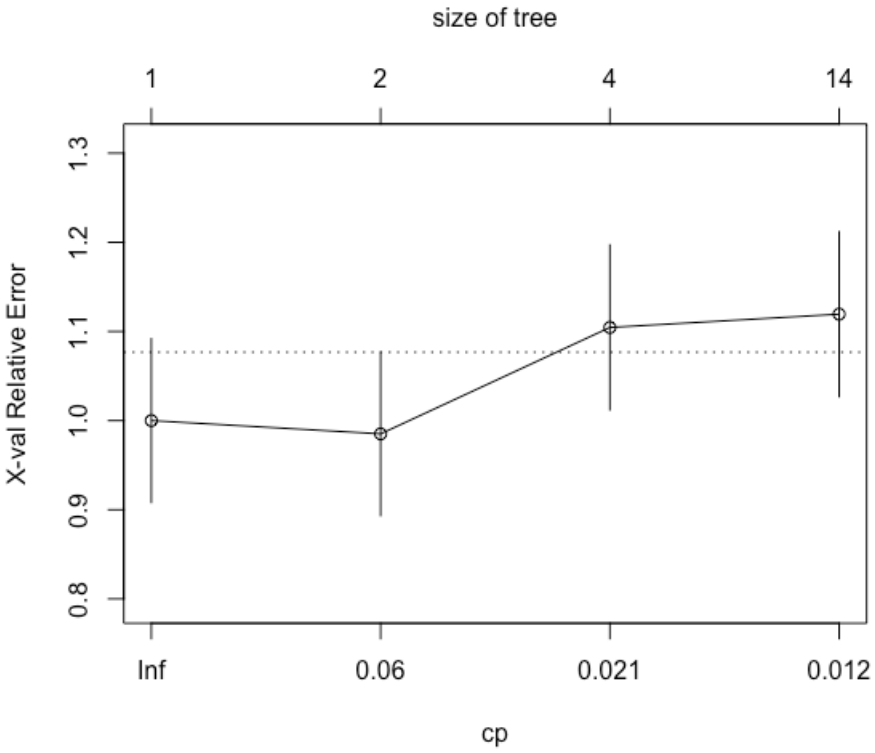

Figure A.12: NASA decision tree Gini Classification with cross-validation (N=154)

A.2 Workload Profile

A.2.1 Data description

```
> shapiro.test(wp$con_mental_workload)

        Shapiro-Wilk normality test

data:  wp$con_mental_workload
W = 0.97799, p-value = 0.001798
```

Figure A.13: Shapiro-wilk test of WP normality

A.2.2 Linear regression

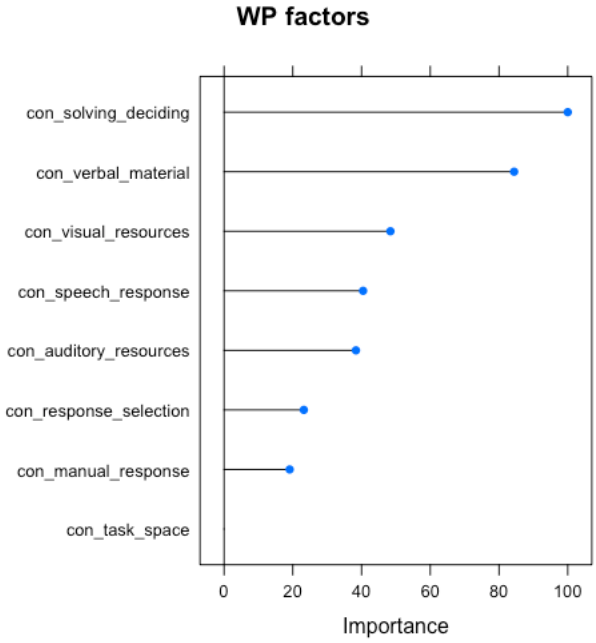

Figure A.14: Variable importance of WP in model

A.3 Extended Feature Sets

A.3.1 Data description

```
> shapiro.test(efs$con_mental_workload)

        Shapiro-Wilk normality test

data:  efs$con_mental_workload
W = 0.97966, p-value = 0.00175
```

Figure A.15: Shapiro-wilk test of EFS normality

A.3.2 Linear regression

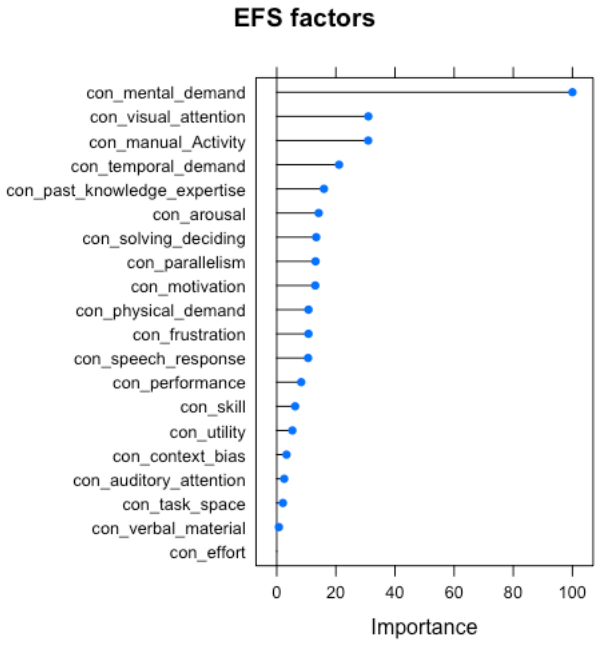

Figure A.16: Variable importance of EFS in model