2020-9

Discover Influential Mental Workload Attributes Impacting Learners Performance in Third-Level Education

Amisha Mehta

Technological University Dublin

Discover influential mental workload attributes impacting learners performance in third-level education

Amisha Mehta

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

31-08-2020

Declaration

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Data Analytics), is entirely my own work and has not been taken from the work of others save and to the extent that such work 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 on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signed:

Date : 31-08-2020

Abstract

Human Mental Workload is an intervening variable and a fundamental concept in the discipline of Ergonomics. It is deduced from variations in performance. High or low mental workload leads to hampering of performance. Mental workload in an educational setting has been extensively researched. It is applied in instructional design but it is obscure as to which factors are majorly driving mental workload in learners. This dissertation investigates the importance of the features used in the the NASA-Task Load Index mental workload assessment instrument and their impact on the performance of learners as assessed by multiple-choice tests conducted in classrooms of an MSc programme in a university. Model training is performed on these attributes using machine learning approaches including decision tree regression and linear regression. Montecarlo sampling was used in the training phase to ensure model stability. The identification of the importance of selected features is carried on using the permutation feature technique since it is adaptable and applicable across a variety of supervised learning methods. Empirical evidence emphasises the absence of more important features over the others tentatively suggesting their applicability in a multi-dimensional model.

Keywords: Mental Workload, Cognitive Load Theory, Instructional Design, Permutation Feature Importance, Supervised Learning, Unsupervised Learning, Social Constructivism, Collective Working Memory

Acknowledgments

First and foremost, I would like to express my sincere thanks to my supervisor Dr.Luca Longo for always being approachable by providing active guidance, cooperation and encouragement. Thank you for the persistent help and availability through all the phases of the research.

I want to thank all the professors of TU Dublin for their constant guidance and for helping me clear all fundamentals related to Data Science. I would like to extend special gratitude towards college staff and admin department for taking the necessary precaution and ensure smooth function during these COVID times.

I want to thank Dr.Deirdre Lawless, Dr.Brendan Tierney, Prof.David Leonard and Prof.Robert Ross for building a strong foundation of statistics, data mining, machine learning and deep learning which was very helpful in my research.

I want thank my parents Rajesh Mehta and Priti Mehta for having immense faith in me and always being my pillar of support in every walk of my life and for loving me unconditionally. Last but not the least, I would thank by closest friends Kinjal Solanki and Pooja Naik for being my constant support system.

Contents

Declaration **I**

Abstract **II**

Acknowledgments **III**

Contents **IV**

List of Figures **VII**

List of Tables **IX**

List of Acronyms **X**

1 Introduction **1**

1.1 Background 1

1.2 Research Project/problem 3

1.3 Research Objectives 4

1.4 Research Methodologies 5

1.5 Scope and Limitations 6

1.6 Document Outline 7

2 Review of existing literature **8**

2.1 Cognitive Load theory 8

2.1.1 Types of cognitive load theory 10

2.1.2 Social Constructivisim 11

2.2 Mental Workload 12
2.2.1 Types of Measurement Method 15
2.3 Gaps in Research 20

3 Experiment design and methodology 22
3.1 Data Understanding 23
3.1.1 Data Gathering 23
3.1.2 Data Description 24
3.1.3 Data Exploration 25
3.2 Data Preparation 28
3.2.1 Data Selection 28
3.2.2 Data Processing 28
3.3 Modelling 29
3.4 Model Evaluation 32
3.5 Strenghts and Limitations 34

4 Results and discussion 36
4.1 Data Description 36
4.1.1 Multiple Choice Questionnaire (MCQ Score) 38
4.1.2 Pre-Knowledge and Motivation 40
4.1.3 Raw NASA-Task Load Index features 41
4.2 Data Exploration 44
4.2.1 Correlation between MCQ Score and other features 49
4.3 Feature Selection 50
4.4 Model Training 51
4.4.1 Model trained using Regression 52
4.4.2 Classification 64
4.4.3 Clustering 70
4.5 Evaluation of Result 72
4.5.1 Strengths and Limitation: 74

5 Conclusion **76**

5.1 Research Overview 76

5.2 Problem Definition 78

5.3 Design/Experimentation, Evaluation & Results 79

5.4 Contributions and impact 81

5.5 Future Work & recommendations 82

References **83**

List of Figures

2.1 Structure of the literature review 8

3.1 Summary of framework to achieve the research question 23
3.2 Data Gathering Process 25

4.1 Histogram of MCQ Score depicting distribution 37
4.2 Histogram of MCQ Score in Control Group depicts the distribution . . 37
4.3 Histogram of MCQ Score in Experimental Group depicts the distribution 37
4.4 Boxplot of MCQ Score for overall data, control group and experimental group 38
4.5 Histogram depicting the distribution for Knowledge and Motivation . . 40
4.6 Boxplot for Knowledge and Motivation depicting variability 41
4.7 Boxplot showing variance for all mental workload features 43
4.8 Histogram for NASA Task Load Score depicting its distribution 43
4.9 Correlation Matrix of MCQ scores and Mental workload features from NASA-TLX 49
4.10 Linear Regression: Feature Importance - Feature Set 1 (Knowledge and Motivation) 54
4.11 Linear Regression: Feature Importance Score - Feature Set 2 (NASA-TLX features) 55
4.12 Linear Regression: Feature Importance Score - Feature Set 3 (NASA-TLX including Knowledge and Motivation) 56

4.13 Decision Tree Regression: Feature Importance - Feature Set 1 (Knowledge and Motivation) . . . 58
4.14 Decision Tree: Feature Importance - Feature Set 2(NASA-TLX Features) 59
4.15 Decision Tree: Feature Importance - Feature Set 3 (Knowledge and Motivation with NASA-TLX features) . . . 59
4.16 Box plot for Linear Regression (RMSE Score) using Interpolated data for all 3 feature set . . . 61
4.17 Box plot for RMSE - Decision Tree Regression using Interpolated data before hyperparameter tuning . . . 61
4.18 Boxplot for Decision Tree RMSE Score using interpolated data after hyperparameter setting . . . 62
4.19 Decision Tree Regression with Interpolation - Featue Set 1(Knowledge and Motivation) . . . 63
4.20 Decision Tree Regression with Interpolation - Featue Set 2(NASA-TLX Features) . . . 63
4.21 Decision Tree Regression with Interpolation - Featue Set 3(NASA-TLX Features with Knowledge and Motivation) . . . 63
4.22 Accuracy for all feature set across 10 iteration . . . 66
4.23 Decision Tree: Permutation Feature Importance - Feature Set 1(Knowledge and Motivation) . . . 67
4.24 Decision Tree: Permutation Feature Importance - Feature Set 2(NASA-TLX features) . . . 67
4.25 Decision Tree: Permutation Feature Importance - Feature Set 3(NASA-TLX along with Knowledge and Motivation) . . . 68
4.26 Cluster 1 built using all six NASA-TLX Features along with Knowledge and Motivation . . . 71
4.27 Cluster 2 using all six NASA-TLX Features along with Knowledge and Motivation . . . 72
4.28 Cluster 3 using all six NASA-TLX Features along with Knowledge and Motivation . . . 72

List of Tables

3.1 Raw NASA-TLX feature definition 27

4.1 Skewness test of MCQ score 38
4.2 Standardised Skewness Score of Mental Workload Features 42
4.3 Summary statistics 44
4.4 Mean and Standard Deviations of MCQ Score and NASA-TLX grouped by control and experimental group 46
4.5 P-value at significance level=0.05 of T-Test(T) or the Mann Whitney test(M) of the MCQ(Multiple choice question) and NASA-Task Load Index 48
4.6 RMSE Score for Control Group 53
4.7 RMSE Score for Experimental Group 54
4.8 Hyperparameter Setting for Decision Tree Regression 56
4.9 RMSE Score for Control Group - Decision Tree Regression 57
4.10 RMSE Score for Experimental Group - Decision Tree Regression 57
4.11 Hyperparameter setting for Decision Tree Regression with Interpolation 62
4.12 Hyperparameter setting using Grid Search 65
4.13 Train and Test Accuracy of Decision Tree Classifier 65
4.14 Average Performance of Logistic Regression 69
4.15 Accuracy of other learning algorithm 70

List of Acronyms

NASA-TLX	National Aeronautics and Space Administration Task Load Index
WP	Workload Profile
CLT	Cognitive Load Theory
MWL	Mental Workload
RMSE	Root Mean Square Error
SCT	Social Constructivist Theory

Chapter 1

Introduction

1.1 Background

Every person comes across a point when processing and consuming a new set of information becomes difficult for the working memory to concoct. Multiple studies have proven that even the brightest person encounter this issue. Any additional details beyond his/her capacity can result in a reduced performance level; this is because working memory, also known as short term memory stores information temporarily unless re-enacted or actively repeated. Otherwise, the information in working memory usually stays for a short duration of 10-15seconds (Goldstein, 2011).

Mental workload comes into the picture when higher cognitive resources will be required by an individual to accomplish a particular task or to absorb additional information. This demand for extra resources will end up reducing the performance and efficiency of an individual. All these issues arise when he/she is facing high Mental Workload. Mental workload is an interaction between the mental physical demand to perform a task and the cognitive resources required to accomplish them. The relationship between the different demands required to complete a task, performance and human capacity appeared to be a concern for more than thirty years (da Silva, 2014) across fields.

The study of the mental workload falls in the domain of psychology, human factors and ergonomics primarily for safe-critical applications such as aviation, air traffic

control, space and defence. More recently, the study of workload spread across various other domains such as media, medical, behavioural economics, finance and students.

MWL is closely associated with psychological issues such as stress, anxiety, depression, lack of confidence, evoked from cognitive aspects of the task in hand. Past shreds of evidence show that students experience a considerable amount of stress and workload (Aherne, 2001),their physical and psychological behaviour, a shortfall in cognitive ability, examination anxiety are few signs. Students in third-level education are prone to these symptoms, as they are at the peak of their learning curve utilising their cognitive resources to the fullest (Fredricks, Blumenfeld, & Paris, 2004)

Third Level Education in Ireland includes education after second-level education. It comprises of higher education in universities and colleges. A quarter-million students have enrolled for studying in a third - level course since 2018. The Higher Education Authority (HEA,2004) states that from 1965 to 2000, the number of students enrolling in third-level education is growing from 18,200 to 1,20,000 [1]. A quarter-million students have enrolled for studying in a third - level course since 2018 [2]. The total number of students pursuing higher education in Ireland is reaching a record high. However, with this, there is a rapid increase in the number of students seeking help with anxiety, increased stress levels and depression.

A 'Report on Student Mental Health in Third Level Education' compiled by the Union of Students in Ireland(USI) states that up to two in five third-level students are suffering from severe anxiety during the examination, and these numbers are rising at an unparalleled level [3]. According to a survey conducted in 2016, 61.6 percent of students experienced burnout while attending the third level, and 27.6 percent of students dropped out due to anxiety and stress. Mental workload has a great deal of importance in identifying significant academic stress because it has a direct influence on student's performance, anxiety and fatigue levels. Hence, it is essential to measure the vital factors driving the mental workload by collecting written feedback of students. A self-assessment test conducted using the NASA-Task Load Index, which

[1] https://en.wikipedia.org/wiki/Tertiary$_e$*ducation*

[2] https://www.education.ie/en/

[3] https://usi.ie/mentalhealthreport/

is a multidimensional assessment tool used to measure mental workload of learners in a masters classroom before and after giving the test. This study aims to ascertain the factors contributing to the mental workload of students.

1.2 Research Project/problem

Students in third level education battle mental workload because of stress, anxiety, cognitive inability, unable to cope up with the workload of third-level education because of inundating information to consume, ending up with poor performance. Hence it becomes crucial to find out the essential mental workload attributes responsible for this degrade in performance. Cognitive Load Theory tracks how much information does the working memory holds at any given time. Sweller (2011) states that since the working memory is limited in capacity, the direct and explicit instructional method should avoid overloading by incorporating additional activities that do not directly contribute to learning. This inadequacy of working memory capacity gives rise to Mental Workload in learners.

Any set of information after active rehearsing only gets shifted from working memory to long term memory. Therefore, to avoid cognitive overwhelm, and for smooth information grasping among learners and to find out the contribution of mental workload attributes, an inquiry-based technique based on collaborative learning is incorporated. This way, the working memory resources expand as it gets divided amongst many learners. However, even the interaction among learners generates high cognitive cost hampering the learning process keeping the task complexity the same (Kirschner, Paas, & Kirschner, 2009). On the contrary, Jonassen (2009) states that this assumption does not consider all characteristics of the context and learners. Hence, it is hard to find definite experimental evidence for the most reliable way of learning and information transfer into long term memory is achieved interactively or individually.

Research Question : *What is the most influential* ***mental workload*** *attributes that can contribute to explaining the* ***performance*** *of learners in a typical university* ***classroom*** *at the postgraduate level?*

1.3 Research Objectives

To answer the research question as stated above the following research objectives are set:

The initial research objective is to conduct a literature review to understand the current state of the art techniques surrounding mental workload which includes cognitive workload which tracks the usage of working memory, types of cognitive load, the various assessment techniques used to measure mental workload, social constructivism theory, how is cognitive load theory related to working memory and short term memory. The subset of this objective is also to find the research gaps found in the existing and previous research performed on mental workload.

The second research objective is to focus on design and primary research by setting empirical experiments by building an understanding of the data, conducting an exploratory data analysis, performing the pre-processing task and finally work towards choosing the appropriate machine learning approach to take a step ahead to solve the research question.

The third research objective is to implement all the experiments formulated in the previous step to check which experiments are adding value to take the research further on the right path. The fourth objective is to find an appropriate approach to compute the feature importance score of all the mental workload features, which trains the model to predict the learner's performance in the MCQ test.

The final research objective is to evaluate the results to select the best-fit output using different evaluation metric based on the Machine learning approach used to solve the problem. This step also comprise of finding the most critical feature which will provide help in predicting the MCQ score of the learners.

1.4 Research Methodologies

A mixed research methodology is adopted. Firstly, there is a literature review to identify the theoretical knowledge surrounding various concepts related to mental workload. The output of the review led to the formulation of the research question, framework design and identifying the gaps in the research. According to the existing literature review there exist a conflict between direct instructional teaching method and inquiry-based activities. One of the gaps identified in the literature review was that it lacks a decent comparison between both these teaching approaches.

The second research objective was met by conducting summary statistics of all the the feature and target variable in the data. The distribution of all the variables were checked to avoid skewed and imbalanced data. Skewness was checked using standardardise skewness and kurtosis test. A basic exploration will be performed to check if there exists any statistical difference between the control and the experimental group. Correlation test is performed to see if there exist any relationship between the features and target variable. Missing values is treated by imputing the data and outliers detection using interquartile range will be removed from the data. The machine learning algorithm as planned in the framework are Linear Regression to take care of all the linear data, decision tree regression to look into the complex and non-linear data.

The third research objective aims at implementing the experiments formulated previously. These experiments is implemented using machine learning algorithm. The random sampling of existing data is performed ten times to compare all the iteration to determine the consistency of the model. Before the model building the data is split into train and test set at 70:30 ratio respectively.

Finding the feature importance score of all the feature which is nothing but the attribute of the NASA-TLX is one of the primary objective of the research. To achieve this permutation feature importance approach is used as it is applicable for all sorts of supervised machine learning algorithm.

Result evaluation is the final research objective which is conducted using the RMSE score which determines the variation in the residual of the trained model.

As the research involves the use of data which belongs to the existing research; this research will fall under secondary research. The target variables is student's performance which can be measured using their MCQ result. Hence, the objective of this research is quantitative. It is an empirical research because the study is based on actual experiences wherein different statistical and predictive model are used to test the stated hypothesis. It follows a deductive approach because this experiment is concerned with constructing a hypothesis built on a existing theory followed by testing that hypothesis.

1.5 Scope and Limitations

The goal of the research is to apply concepts of mental workload in an educational setting. Hence we can say that domain is limited to learners studying in third-level education. The study is only applicable for learners attending physical college within the university premises. NASA-Task Load Index is the subjective assessment tool used with two additional attributes, namely knowledge and motivation to measure the mental workload within the learners. The primary goal is to search essential features within both the control group and the experimental group, which primarily impacts the performance of the learners. The data consist of 20 classrooms with approximate class strength of 20-30 learners per class.

There is no way to determine the mental workload of the learners taking the virtual class or while they are solving the assignments. The NASA-TLX assessment strategy is used to measure MWL, which is very simple and handy to fill. However, the process of filling the test becomes time-consuming and dull with high chances that the learners build a relationship between their workload ratings and the task performance. While experimenting in the case of the experimental group, the groups are created randomly. Due to which there are high chances that an average learner accidentally goes into the group of bright learners. This way, there are high chances that the average student performs well based collaborative group activity. Hence, this process fails to capture the actual mental workload of the average learner.

1.6 Document Outline

Chapter 2: Literature Review: This chapter covers relevant literature related to the concept of mental workload, cognitive load theory, social constructivism, types of cognitive load and various way to measure mental workload. The chapter starts by covering theoretical framework using Cognitive Load Theory and how to measure the mental workload in the educational set up—further extending the research by discussing collective working memory under social constructivism theory. Subsequently, this research works on finding the relevant research gaps in the previous and existing literature.

Chapter 3: Experiment design and methodology: This section describes the design and implementation which was created after having a detailed literature review. This chapter starts by explaining the design flow along with the steps involved in data collection. It presents a detailed plan which consists of all sorts of possibilities with justified explanation borrowed from literature.

Chapter 4 : Implementation and Results: This section describes the design and implementation which was created after having a detailed literature review. This chapter starts by explaining the design flow along with the steps involved in data collection. It presents a detailed plan which consists of all sorts of possibilities with justified explanation borrowed from literature.

Chapter 5 : Conclusion: This final chapter provides a summary of the results in this study concerning the objectives defined previously. A consideration of things that went well and things that went bad along with something that could have done better was compiled together. Towards the end, the contribution and impact associated with this study were addressed along with recommendation and future work of the study.

Chapter 2

Review of existing literature

This section aims to bring basic notions of cognitive load theory, mental workload, social constructivism, collaborative learning, collective working memory across the readers. The intention behind the review is to identify the existing state of the art concepts and assessment related to mental workload. A critical discussion on the gaps in the existing research is conducted towards the end, which highlights the limitation in the current state of the art research in mental workload. Below is the structure of literature review

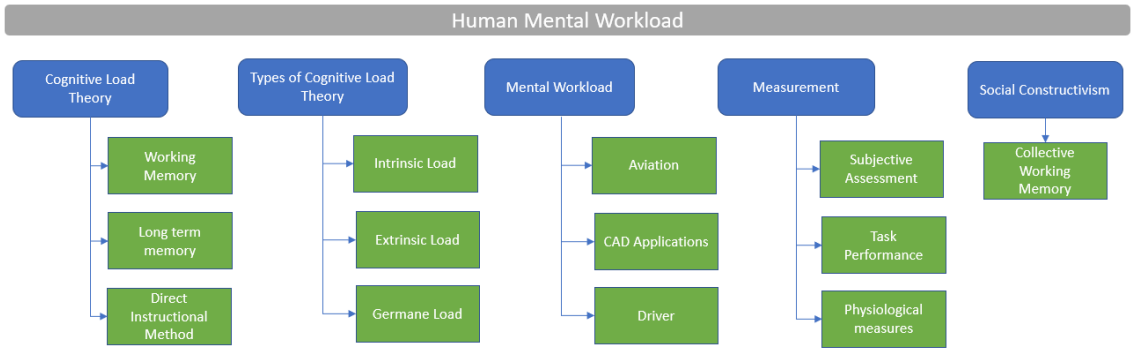

Figure 2.1: Structure of the literature review

2.1 Cognitive Load theory

Cognitive Load Theory(CLT) in cognitive psychology refers to the usage of working memory resources. In other words, it is designed to provide meaningful guidelines

intended to aid in presenting information in a way that helps in better optimisation of the intellectual performance of learners during an ongoing task or activity (Sweller, 1988). CLT is widely recognised in the field of educational psychology to enhance the learning phase by applying instructional teaching techniques based on the knowledge of human cognitive architecture. Human Cognitive architecture is a generic framework in charge of information processing within leaners such as encoding, storing and modifying information for reasoning and decision making purpose (Atkinson & Shiffrin, 1971). Long term memory and short term memory is also known as working memory are the two dimensions of human cognitive architecture. Working memory can be described as temporal decay and the chunk capacity to take up information is limited. In other words, all control processes take place within the short term memory to make any decision and speed up the regulation of information flow thus constraining learning and disremembering shortly (Atkinson & Shiffrin, 1971). As the learning involves new information, the working memory capacity restricts most of the learners to grasp more than four to five pieces of knowledge concurrently. Hence, (Sweller, 1988) suggests avoiding any alteration with instructional techniques to bypass overload with additional activities within learners. Whereas, information can be stored in long term memory after being visited and treated by working memory. As the name suggests, long term memory stays for an extended period. It both stores and recalls details for later use (Goldstein, 2011).

Optimisation of working memory is a task of utmost importance for the current research work. The aim behind optimisation is promoting the smooth knowledge transfer to long-term-memory and expanding the learning phase. According to Chi, Glaser, and Rees (1981), a schemata of information which consolidates chunks of data from low to high level of complexity which can be perceived into a single chunk of information. In the due course, schemata creation in the working memory required explicit instructional technique.

Roots of the cognitive load if traced back begins from 1982, ever since then the different variation of the theory was updated. Twenty years later, many modifications were observed to the concept of Cognitive Load theory. Firstly, the theoretical basis

of human cognitive architecture lays a stronger foundation, a four-component instructional design which focuses on designing an educational program for a longer time duration. One of the most recent features includes the self-management effect; this feature is based on the assumption that students should be taught to practice CLT on their own. Preferably, these students should only access materials that is designed with consideration of cognitive load. However, due to the internet and other factors, they most likely come under quality learning material. Hence, the learners well versed with a variety of learning material are better equipped than the ones who are exposed to only the material provided by the educational system (Sweller, van Merriënboer, & Paas, 2019).

2.1.1 Types of cognitive load theory

Nearly three decades of research later, three types of load have been defined by (Sweller, 2011): intrinsic, extraneous and germane load. There was a lot of evolution observed in these loads over three decades (Orru & Longo, 2018).

Intrinsic Cognitive Load: It is a term first used in the early 1990s by (Chandler & Sweller, 1991). Intrinsic load indicates the complexity of the information under process. It refers to the notion of element interactivity. It is strenuous to determine the complexity of the information while humans are processing it; this is due to the characteristics of human cognitive architecture. The attributes of information while storing it in the long term memory for learners widely differs before the information storage. According to (Sweller et al., 2019), the complexity or element interactivity depends on two factors. 1. essence of information and 2. knowledge level of individual learner who will process the information. Hence, the intrinsic load can be altered only by changing the requirements to learn or by changing the expertise of the learner.

Extrinsic Cognitive Load: Extrinsic load does not delimit to the intrinsic complexity of the data. Its primary focus is the presentation of information to the learners and how do learners deal with the instructional procedure. The extrinsic load can change by changing the instructional process, which is not the case with the intrinsic load. Efficient instructional method defeats element interactivity to a greater extent

while inefficient way increases it (Sweller et al., 2019).

Germane Cognitive Load: Germane load is associated with the cognitive load needed to learn; this means it shares connections with the working memory just like an intrinsic load. Therefore, the higher the resources are busy dealing with germane load, the less it will be available for intrinsic load, which leads to less learning. Hence, we can say that intrinsic and germane load are closely entwined.(Sweller, 2011),(Sweller et al., 2019)

According to the (Sweller, Van Merrienboer, & Paas, 1998) paper, the germane load was considered to be the total cognitive load replacing the extraneous load. But current research on CLT by (Sweller et al., 2019) has assumed that germane load instead of contributing to the entire cognitive load it can reallocate the working memory with extraneous load to filter the relevant activities for learning.

2.1.2 Social Constructivisim

The Social Constructivism Theory(SCT) is based on the ideas of (Vygotsky, 1980), which states that the learner's engagement in the learning process will lead to better results. The development of human intelligence is socially situated, and the construction of knowledge done through such social interaction can lead to smooth information capture. Dawes, Cresswell, and Pardo (2009) states that social constructivist is useful because it allows tracking and performing qualitative analysis to explore people interact with each other. The SCT affirms that people's ideas harmonise with their experiences in life. The main focus is given to learning taking place due to the interactions within the groups. The difference between cognitive load theory and constructivism is that the former has its basis on human mental architecture, and it strongly supports direct instructional teaching method. Whereas, the later is in support of constructing information with a focus towards collaborative learning employing social interaction. However, (Sweller, 2009) directed that the constructivism theory neglects human mental architecture.

According to research by Reznitskaya, Anderson, and Kuo (2007),Corden (2001),Weber, Maher, Powell, and Lee (2008), increase in learners opportunity to communicate with

one another opens their mind along with letting the students speculate and transfer their knowledge grasped in the class; this not only makes learning light but also helps them integrate others ideas and build a more in-depth perception of what they are learning.

The SCT is based on the collective working memory approach, where a group of learners can share their working memory on a similar task. The underlying assumption is by using working memory of multiple people can reduce the cognitive cost of a job. However, the complexity of the task remains the same, also the capacity of working memory increases because of the collaboration (Sweller, Ayres, & Kalyuga, 2011). According to the educational psychology of (Geary, 2012), concerning the assumption of limited human mental architecture, classified two types of knowledge: biological primary and biological secondary. Humans can develop primary knowledge without any effort because they have it in their genomes. Whereas biologically secondary knowledge requires a lot of effort. In collective working memory, it is assumed that communication is a part of biologically primary. Hence, it does not require any additional effort. On the contrary to this theory (Paas & Sweller, 2012) states that cognitive load increases with task-specific communication. Hence, different literature has a different say on SCT.

2.2 Mental Workload

Mental workload is a study in ergonomics which as started gaining popularity since the 1980s. At the start of 1980, the concept of mental workload was used to study CAD applications. The main focus was to track the strain related to designing a printed circuit board, along with other CAD tasks (Järvenpää, 1986). Similar research was conducted in 1987, which examined mental workload in a software programming team (Young, Brookhuis, Wickens, & Hancock, 2015). The main focus of these of such studies to understand different variation in mental workload. According to a research by Longo (2016) the construct of Mental workload has also been applied to various medical specialist by using hybrid of one or more measurement technique. The main

aim is to study how workload differs between clinical demand and the performance of the health care staff. Further, the focus shifted to application with aviation and driving theme (Prabaswari, Hamid, & Purnomo, 2020), (Wu & Liu, 2006). Mayer, Heiser, and Lonn (2001) tracks cognitive load when students have to deal with more than one multimedia aid in the learning method. But at the same time according to (H. Xie et al., 2017), (De Koning, Tabbers, Rikers, & Paas, 2007) adding some cues such as visual aid can reduce the cognitive load by a decrease in extraneous load of the pupil.

As cognitive resources are limited, which leads to a demand-supply problem when an individual tries to perform more than one tasks that require the same resources. A plethora of workload, caused by the task utilising the same resources can create issues along with a plunge observed in the performance of a task in hand with an increase in error. Increase in workload is not the only reason behind the decline in performance. The drop in performance is caused due to both high as well as low mental workload (Nachreiner, 1995). The high mental workload can be described as a task performed with a high amount of attention, whereas the low mental workload can be described as a task with a low or no amount of attention. The optimal amount of mental workload helps improve the efficiency and performance of a learning task (Orru & Longo, 2019).

A recent study evaluated that mental workload exponentially increases with the increase in fatigue and stress level (Alsuraykh, Wilson, Tennent, & Sharples, 2019), (Gingerich & Yeates, 2019) . Fan and Smith (2018) on the contrary brought up a different argument where people enjoyed being in high workload because that way, they were able to focus more. Hence, we can say MWL definition is subjective depending on the field and the research you are working with no definite definition (Cain, 2007). Therefore, (Gopher & Donchin, 1986) debated classifying MWL as a hypothetical construct instead of intervening variable. The intervening variable in this present scenario is nothing but a theoretical concept which is obtained after manipulation of the values (Gopher & Kimchi, 1989).

One cannot detect Mental workload directly, but it is possible through the measurement of other variables which can highly correlate with it, this includes subjective

rating or some physiological data (B. Xie & Salvendy, 2000). The mental workload consists of both static and dynamic attributes. By static attribute, it means MWL can be determined within an interval of time whereas it can also be determined at a single moment which falls under dynamic attribute.

Mental Workload has been used in collaboration with the field of Artificial Intelligence by using augmentation theory and fuzzy reasoning. The study conducted by L. Rizzo and Longo (n.d.) is a comparison between augmentation theory and fuzzy reasoning model. Based on the convergent and face validity analysis of both the models higher level of inferential capacity was observed for augmentation based models over fuzzy reasoning. Further, the construct of Mental Workload has also been invoked in field of HCI (Longo, 2018a),(Longo, 2017), (Longo & Dondio, 2015),(Longo, 2012). One of the application of Mental Workload in HCI was also applied to assess usability of interactive system under medical domain. In other words user's interaction with medical system (Longo, 2015b).

Mental workload is a multi-dimensional and non-linear concept (Longo, 2015a), (L. Rizzo, Dondio, Delany, & Longo, 2016). Reid and Nygren (1988) classified MWL in three dimensional, namely time load, mental effort load and psychological stress load using Subjective Workload Assessment Technique (SWAT). In 1988, Hart and Staveland (1988) in their National Aeronautics and Space Administration considered mental workload from six prominent aspects: mental, physical, temporal, effort, performance, and frustration. Hence, we can obtain mental workload through various dimensions, although the weights will keep on changing. To design the measurement in an educational setting, the major part of the research incorporated Mental workload in ergonomics as an alternative approach (Longo & Barrett, 2010). In other words, Mental workload is altogether a unique experience which varies from one individual to another by distinct cognitive style, upbringing and separate level of education.

There are numerous research related to measuring and evaluating MWL. However, the effect on instructional teaching technique on the performance measure when associated with the workload is quite unclear (Hancock, 2017).

2.2.1 Types of Measurement Method

There are several measurement methods to measure Mental workload. The advantages and disadvantage are subject to thorough investigation (Gopher & Donchin, 1986); (Hancock & Meshkati, 1988); (Hancock, Meshkati, & Robertson, 1985); (Hart & Staveland, 1988); (Meshkati, Hancock, Rahimi, & Dawes, 1995); (Moray, 2013); (Wilson & O'Donnell, 1988). These measurement methods segregated into three groups:

Task performance measures: Predicting workload solely based on the output efficiency of individuals concerning the task in hand; would provide most of the information this can be classified as primary task performance. But, it is also essential to predict when and how will an individual encounter situation that exceeds their cognitive capability. Here, more than the primary task measure is required. However, this does not justify that primary task performance is only limited instantaneous load levels. One of the ideal examples of primary task measure is that in aviation where we see high workload most likely during taking offs, landing or emergencies. Therefore, in other words, we can say primary task measures can directly record performance which is highly accurate for measuring mental workload in a long task. However, secondary task measure usually wants the individual to perform two tasks concurrently; the first task is primary task whereas the second task is the secondary task which helps evaluate the MWL imposed by the primary task. The intention is supposed an individual has his/her full cognitive capacity designated to a primary task; their performance will hinder during the secondary task even if possibly the secondary task is easier (Cain, 2007). Wästlund, Norlander, and Archer (2008) suggests that the reaction time, which comes during a secondary task, can be used to measure the mental workload. In other words, the more mental demand invested in primary task lesser reaction time is witnessed in a secondary task (Verwey & Veltman, 1996).

Subjective Assessment: These measures aim to measure mental workload by asking to rate themselves within a specific scale about various aspects of the set of tasks. Since these ratings are contemplation after the job and the difficulty level of each task is dependent on the individual, this method is regarded as subjective (Cain, 2007); (Wierwille & Eggemeier, 1993); (B. Xie & Salvendy, 2000). The subjective method

usually evaluates multiple dimensions such as effort and performance, but there exist ways which only have a single dimension (Wierwille & Eggemeier, 1993). Below are a few subjective assessment methods: **N**ASA-Task Load Index: It is a widely used, multidimensional subjective measure (Hart & Staveland, 1988). This method measures explicitly mental workload with applications like communications stations, cockpits in aircraft, control systems and also used in laboratory tests (Tracy & Albers, 2006). The ideal use of NASA-TLX is predicting severe levels of mental workload, which can cause a significant impact to the underlying task. It is not employed widely in the education domain; however, there exist numerous studies which authenticates its legality and sensitivity (Gerjets, Scheiter, & Catrambone, 2004);(Kester, Lehnen, Van Gerven, & Kirschner, 2006). NASA-TLX divides the total workload into six parts:-

- Mental Demand
- Physical Demand
- Temporal Demand
- Performance
- Effort
- Frustration.

The TLX part, on the other hand, plans to create a weighting of each subscale to enable pairwise comparison based on their perceived importance; this makes it easy to pick which measurement is more suitable to workload. A lighter version of NASA-TLX is the RAW NASA-TLX, here the weighting process is eliminated. Many types of research use RAW-NASA-TLX to remove the pairwise comparison (Hart, 2006). There has been proof where the shortened version is evaluated with the full version, and the shortened version received more support since it might increase experimental validity (Bustamante & Spain, 2008). If any individual subscale is less relevant are being dropped in the case of raw-NASA-TLX (Colligan, Potts, Finn, & Sinkin, 2015).

Subjective Workload Assessment Technique: This is one of the most common subjective methods which has been reported in many works of literature (Cain,

2007). Similar to NASA-TLX, it is also a multidimensional measurement. In this approach, subjects rate the workload of a task, and the dimensions used are mental effort, time load and psychological stress load. The definition of cognitive workload influences these dimensions. SWAT works on conjugating measurement and scaling technique to merge assessments at the ordinal level into a separate workload score which is nothing but a value on an interval scale. The dimension time load focuses on the amount of extra time set aside for planning, executing and monitoring activities; mental effort load estimate how much mental effort is consciously allocated for planning and executing; psychological stress load concentrates on measuring the risk, anxiety, frustration and confusion linked to particular task performance (Reid & Nygren, 1988).

Workload Profile: Workload Profile continuously estimates the workload of the subject without interruption with unique values for each point in time (Rusnock & Borghetti, 2018). This method is innately based on multiple resource theory, as a result of which it's dimensions are also directly linked with the dimension which is proposed by the theory (Romero, 2017). The dimensions are as follows (Council et al., 1993):

- Task and Space
- solving and deciding
- auditory attention
- speech response
- visual attention
- response selection
- manual activity

This method is identified to be very reliable as it evaluates different task (Tsang & Vidulich, 2006).

Rating Scale Mental Effort: This one is a unidimensional instrument. It is more related to the Limited Capacity Model. The main task is only to self rate the amount of mental effort the subject had to put into performing a task. RSME consists of 150mm(length) lines comprising nine anchor points, and each has a descriptive label which indicates the level of effort (Widyanti, Johnson, & de Waard, 2013). The rating is distributed as follows:

- close to 0 - "Absolutely no effort."
- about 57 - "a rather much effort"
- about 112 - "the extreme effort"

other labels were, "a little effort", "considerable effort", "great effort", "very great effort" (da Silva, 2014). The subject marks these responses by marking a point on the line corresponding to the amount of effort put into completing a task. (da Silva, 2014) reviewed various studies and identified that this method has a reasonable degree of sensibility despite its simplicity.

In a research conducted by L. M. Rizzo and Longo (2017) it was found that the inferences of NASA-TLX and Workload Profile generated using defeasible reasoning produces decent information even with less information. The inferences are more self-explanatory compared to the results generated using the original measures.

Physiological Measures: This measure performs the analysis of physiological pointers of the human body such as EEG, eye tracking and heartbeat using ECG at the time of completion of the task in hand. Due to current technological advancements, the use of physiological measurement technique has stimulated to measure and predict an individual's mental workload. In recent times, MWL has been distinguished using multiple sensor data. Physiological indicators are associated with humans mental activities such as cognitive load, emotions and frustration (Romero, 2017).Ward and Marsden (2003) reviewed previous studies on these indicators and suggested that the use of these indicators is not as straightforward as it seems. He states that there are a lot of inconsistencies between individual and occasions. Hence, there exist discrepancies in the reading leading to difficulty in interpreting and standardising the signals.

Further, it also becomes hard to quantify and correlate physiological responses with MWL. Cain (2007) studied the main physiological measures studied and evaluated in the MWL context:

- Electroencephalography
- Eye Movement
- Heart Rate
- Respiration

The main benefit of using physiological measures is its capability to measure the operator continuously (Wästlund et al., 2008). Hence we can say this measurement technique is more dynamic in nature. Cain (2007) described that there exists multiple studies where physiological measures are used in collaboration with SWAT and NASA-TLX. The output of that study showed a clear contrast between the results of subjective measures and physiological mesaures. Physiological measures such as eye movement, eye blink, blood pressure, heart rate seems inconsiderate to workload diversity.

One of the most recent work by (Longo & Orru, 2018), (Longo, 2018b) related to education field which was also, conducted in typical third level education class and the self reporting instruments used were NASA-TLX, Workload Profile and Rating Scale Mental Effort. However, in this experiment three instructional design method were used. The first includes the traditional teaching method, the second consists of use of Multimedia Learning and the third involves an extension of second design along with inquiry activity. Based on these three method the self reporting measures are evaluated on the basis of validity, sensitivity and reliability. The experiment points out that these measures are highly reliable but they have moderate moderate face validity and very poor sensitivity indicating almost similar mental workload on learners in all three design methods.

2.3 Gaps in Research

After exploring Mental Workload and Cognitive Load Theory along with all the concepts surrounding them, there were few loopholes that still required more clarity and support.

Multiple models that predict Mental workload exist for numerous domains. According to research by Moustafa and Longo (2018), the current mental workload models are very complex. These models ignore the in-depth evaluation of each feature leading to intricate models. The models are less generalized to employ across multiple fields, discipline and experiments.

In educational psychology, one of the most widely used theory is the Cognitive Load Theory. CLT is aimed at providing guidelines to design instructional material and aims at reducing the cognitive load of learners by expanding their working memory (Orru, Gobbo, O'Sullivan, Longo, et al., 2018). The majority of the models in the research predicts Cognitive Load score through the total cognitive load by multi-criteria or combination of various measurement method (Jung, Kim, & Na, 2016). There is no direct measurement of cognitive load; it is derived from the output of knowledge achieved post-test (De Jong, 2010). In a typical classroom setup student having low knowledge post the test are assumed to have high cognitive load. As we do not have any direct measure, we have to compromise using indirect measures like previous test performance (Mayer, 2005). These measures are not sensitive to variations over time (De Jong, 2010).

Germane Load permanently stores knowledge in the form of the schema (Sweller, 2010). An assumption was made that it becomes easy to store knowledge permanently if there exists some prior knowledge (Paas, Renkl, & Sweller, 2003). However, this was disapproved by (Cheon & Grant, 2012) as there was no correlation seen between the germane load of a student and the prior knowledge.

An evaluation states that increase in MWL will exponentially increase the stress level (Alsuraykh et al., 2019), (Prabaswari et al., 2020). However, this theory had a twist. Gingerich and Yeates (2019) states that there are people who enjoy high

workload due to which the relationship between MWL and stress gains complexity. A very similar conflict observed by (Fan & Smith, 2018) was between MWL and fatigue levels; however there exist scarcity of the research to measure different fatigue levels. Hence, it becomes difficult to build a relationship between fatigue level and MWL.

Iqbal, Zheng, and Bailey (2004), Tungare and Pérez-Quiñones (2009) performed a very similar experiment of correlating Mental Workload with the pupillary response by mounting an eye tracker on the computer. The user's had to perform different tasks. The completion time and subjective ratings to measure task difficulty were used to evaluate mental workload. However, this technique faced problems with the hierarchical task. This experiment was unable to reflect changes in MWL that user experiences throughout the task. Hence, despite using both the physiological and subjective measurement, the output still lacked inconsistency.

NASA-Task Load Index which is a multidimensional subjective method to measure the mental workload is easy to apply and understand. However, at the same time it is very time consuming and strenuous. Many times participant tend to forget various detail of the task which makes NASA-TLX less ideal approach. The participants perception on their performance can differ heavily. Hence using subjective assessment test can be labourious and intrusive to the participants (Rubio, Díaz, Martín, & Puente, 2004).

Chapter 3

Experiment design and methodology

In order to answer the empirical research question, a hypothesis along with the comparative study, has been outlined. This chapter is devoted to the design of a framework with the aim to solve the research question.

Research Hypothesis

The research aims to investigate the influential mental workload features which contributes to the performance of the learners which is measured using the MCQ test. The alternative hypothesis is as follows:

H1 : A **higher number** of statistical significant differences in the feature importance coefficients of the mental workload attributes, used to train models of mental workload (with decision trees regression multiple linear regressions), is expected to be found in the experimental group (direct instructions + constructivism learning) than in the control group (direct instruction learning).

The implementation of the investigation takes place in four parts. The first phase is data understanding which includes data gathering, exploratory data analysis. The second phase comprises data preparation which describes data cleaning and preprocessing to proceed ahead with the study. Thirdly, the data modelling phase which describes different machine learning algorithm which is to be incorporated, how is the data split into train and test set, assessing the feature importance score. Lastly,

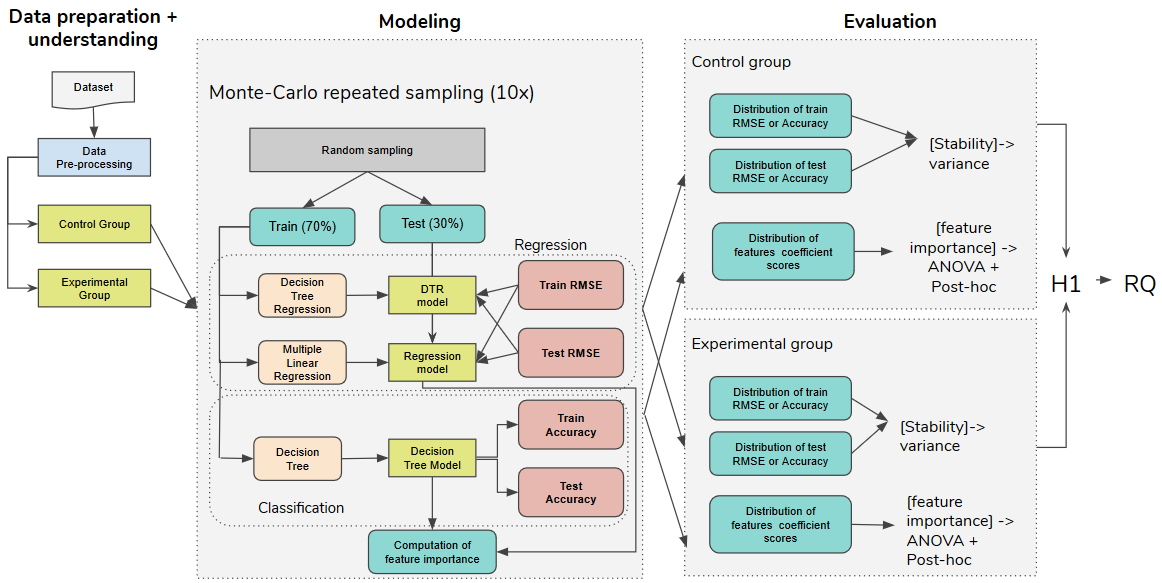

Figure 3.1: Summary of framework to achieve the research question

the evaluation phase which explains the stability of the model and helps understand the essential mental workload feature which impacts the student's performance in the MCQ test. Figure 1. shows the flow of the research. The data division takes place in two parts: the control and the experimental group. Mental workload features are measured using NASA-Task Load Index; they are used to compute learners mental workload which impacts their performance in MCQ test. The target variable is learners MCQ score which can be both continuous or categorical feature. Hence, the data will be trained and tested using Multiple Linear Regression and Decision Tree Regression if the target variable is continuous and Decision Tree classifier in case it is categorical.

3.1 Data Understanding

3.1.1 Data Gathering

Data gathered from ongoing classes in a master's classroom for 19 modules such as Research Methods, Operating systems, Machine Learning, Statistics and many more. A total of 455 records captured in the dataset. Initially, a consent form, along with

task information, was circulated to the learners to maintain transparency about data usage. The classroom division was done in two parts: the control and the experimental group. Both the group received direct instructions while only the experimental group underwent with the collaborative group activity which involves discussing the cognitive trigger questions associated to the topic being guided about in the class before; this is nothing but the social constructivism theory. Social constructivism theory states that knowledge grows faster with shared interaction with each other.

The learners in the control group received a NASA-TLX questionnaire that contains questions about the subjective effort and mental workload followed by the Multiple Choice question on the topic which was being taught at the beginning of the class. The experimental group, on the other hand, was divided into a group of 3-4 learners for inquiry-based group activity. Students in each group should be discussing the answers to different questions on the topic discussed initially and jotting down the discussed answers individually. This step was essential to make the information transfer and processing in working memory. Learners in the experimental group also received the questionnaire similar to the control group. The learners part of the experimental group were given an added advantage to use the written answers they had agreed upon as a group while giving the MCQ test. This helped in maintaining clarity between the output of the constructivism approach and the knowledge achieved at the end. The questions asked in the MCQ test were related to the trigger questions which the learners in the experimental group worked on in the group activity; they had to fill the Raw NASA-TLX questionnaire even after the MCQ test. The main aim is to perceive which mental workload feature derived from NASA-TLX contributes to the growth or decline in learners performance which is mesaured using MCQ score.

3.1.2 Data Description

Raw NASA-TLX is the shorter version of NASA-Task load Index. The only difference is that Raw-NASA-TLX features does not have weightage. Both of these are multi-dimensional measures for mental workload. They consist of six sub-scales which consists of independent bunch of variables: mental, physical, temporal, performance,

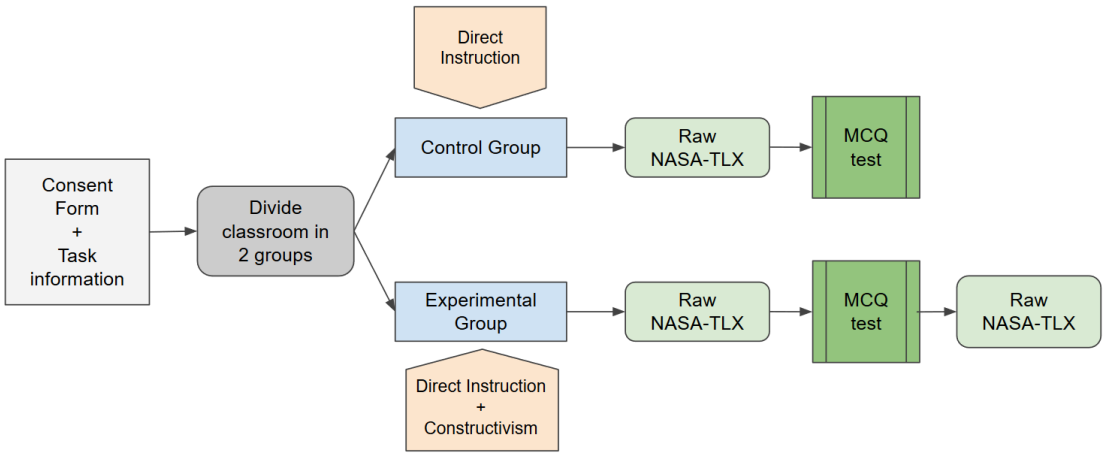

Figure 3.2: Data Gathering Process

effort, frustration. The other additional measures taken into consideration are knowledge and motivation as shown in table 3.1.

3.1.3 Data Exploration

The primary objective of running an exploratory data analysis is to investigate each feature within the data set and analyse their relationship concerning other variables.

1. The initial step towards the EDA would be to look for a quick statistical summary of the data which will include the number of missing values, minimum, median, max, mean, standard deviation, interquartile range and skewness for all features in the data. A comparison between mean and median for all features will help to determine the distribution. If the mean and median is same, the variable is normally distributed whereas if there exist difference we can say that the distribution is not normal. The standard deviation of each variable, which will help recognise the spread and how far is the observed data point away from the mean. In an ideal scenario, about 95 percent of the data will be within two standard deviations if the distribution is normal. A coefficient of variance (CV) will be calculated (standard deviation/mean), if CV is less than 1 we can consider standard deviation to be low, while if CV greater than equal to 1 it

indicates high variation.

2. A univariate analysis will be conducted to check the distribution of all the variables. This analysis will use both histogram for graphical representation and Skewness test. Both these methods will help us determine if the data is normally distributed or skewed. The histogram will additionally help how many times each value occurred in the dataset. The standardised score for skewness between -2 to +2 are considered acceptable to prove normal univariate distribution [1]. If the distribution is skewed, we check whether at 0.05 level if 95 percent of our data falls within +/-1.96, we can still treat the data as normal. Skewness test will be used over Shapiro-Wilkomen test as the latter is senstive to samples greater than 200.

3. The data consist of two groups control and experimental group, and according to the design above, a separate model will be created for both the groups. Hence, the distribution of both groups will be examined to ensure balanced data and sufficient samples to train and test model for both the groups. This will be checked simply by counting the number of samples in each group.

4. After looking into basic summary statistics and distribution of the data, a preliminary analysis to check a significant statistical difference between the following:

   - The control group vs the experimental group for both MCQ Score and NASA-TLX score computed using the six features.
   - The control group vs the experimental group for both MCQ Score and NASA-TLX score computed using the six features for all the topics covered in the classroom.

5. After verifying the data to check the normality of the distribution using Skewness Test an independent t-test for normal distribution (p-value>0.05) and a Mann Whitney test for not normal distribution(p-value<0.05). The principal reason

[1] george2010spss

behind this difference test is to compare the means of both the groups. The same experiment was replicated for various topic.

6. Check the distribution of the categorical variable using histogram by using count as the aggregation method. This way, we can have a more in-depth look towards understanding the data.

Feature	**Description**
Mental Demand	The amount of mental and perceptual activity required while working on a task
Physical Demand	The amount of physical activity required while working on a task
Temporal Demand	The amount of time and pressure felt while performing a task in hands
Performance	The success of the task in reaching towards its goal
Effort	The amount of hard-work required to accomplish the task
Level of Frustration	The amount of emotional drainage and irritated vs. rewarded and satisfying feeling was felt while performing the task
Knowledge	The amount of knowledge an individual or group has pertaining to the task in hand
Motivation	How much the group or individual is motivated to perform the task

Table 3.1: Raw NASA-TLX feature definition

3.2 Data Preparation

3.2.1 Data Selection

Data will be divided into the control and the experimental group, as stated above. Each group will be split into 70:30 ratio, train and test set respectively. The random sample of the data will be produced from the existing data. The sampling is performed using repeated random sampling that is Monte Carlo sampling. The target variable is MCQ Score which measures the learner's performance. In contrast, the independent variables are 6 Raw NASA-TLX features (mental demand, physical demand, temporal demand, performance, frustration and effort) along with the two additional features motivation and knowledge. The model will be trained and tested on the sampled data, and this process of sampling will repeat ten times to note the model results and to evaluate the consistency of different models.

3.2.2 Data Processing

Initially, the target variable will be tested for normality both graphically by histogram and numerically by Skewness Test. A Pearson correlation test(interval scaled descriptive data) will be performed to check the relationship between MCQ Scores and other independent variables. Missing values can be easily be found in the summary statistics. These missing values will be imputated using arithmatic mean. Imputation is useful because it helps improve precision and ensures robust statistics with more resistance towards outlier. Dong and Peng (2013) asserts missing values below 5 %, or lower is inconsequential in such cases, missing values will be dropped. If the total amount of missing values crosses 5 percent, it will be imputed by computing arithmetic mean as discussed above.

The detection of outliers and anomalies in the data is done using the interquartile range. Box plot of each feature will help to visualise the outlier quickly. Any point above upper whisker and below lower whisker in the box plot is assumed to be the outlier. If outlier(s) are present in the data, it will be removed. The reason behind

dropping them is they increase the variability in the data, which decreases the statistical power. Therefore, to obtain statistically significant results, it is better to exclude outliers.

The bivariate relationship checks the correlation between different variables and target variables with the independent variables. If the target variable that is MCQ Score is numeric and parametric i.e. normally distributed, a Pearson correlation test can be used to check if there exists a linear relationship among the variables. If the MCQ Scores is non-parametric, Kruskal-Wallis (for nominal data) and Spearmen (for numeric data) will be used. The correlation will stand as a fair representation for the critical variables in models of Multiple Linear Regression and Decision Tree Regression.

After having a detailed look in the data, if the independent variables - six independent features of NASA-TLX plus knowledge and motivation are not in the same range then normalisation technique such as Min-Max technique will be considered. However, as raw NASA-TLX assessment test is being used, which means the elements have no weightage, and they might lie within the same range, which might be typically between (0-20). Hence, there is a strong chance that normalisation might not be considered.

Just before the data is ready to enter the modelling phase it is randomly sampled using Monte Carlo sampling method. Here, the same dataset will undergo testing under different condition. In other words, each sample of data extracted by random chance and each data point of a dataset has an equal probability of getting selected. Sampling randomly shuffles the data; hence each time a new set of data is observed in train and test set after splitting the data. This sampling method allows calculation of sampling error, and it works on reducing the selection bias. This method of sampling is known as Monte Carlo random sampling is the most straightforward approach to sampling.

3.3 Modelling

The principal aim of this stage is to create models using Machine Learning approaches. The goal is to create mathematical models which can predict the values of the target

variable, which is the MCQ scores with the help of the values of independent variables. The intention is to build a model which helps in finding out the essential mental workload attribute influencing the learner's performance in a class test.

The initial step consists of dividing the data into two, where one model will be trained for the control group, and another another model will be trained for the experimental group.

The beginning of the modelling phase involves splitting the data into 70:30 ratio into train and test set respectively. The model will be trained on ten random sample generated from the same data. This is achieved by Monte Carlo sampling which is nothing but a form of repeated random sampling. This process will help us find out the consistency and stability of the models. With the training set, data will be trained using Decision Tree Regression, Decision Tree classifier and Linear Regression. If the outcome variable MCQ score is continuous which is quantitative a Multiple Linear Regression and Decision Tree Regression will be used to train the model. In contrast, Decision Tree classification will be used when the target variable is ordinal such as Grades(A, B, C). Therefore, based on the type of the MCQ Score, the appropriate machine learning algorithm will be applied for learning.

The reason behind selecting two learning algorithms Multivariate Linear Regression and Decision Tree Regression if the MCQ score is a continuous variable are:

- Decision Tree Regression will better be able to capture any non-linear relationship within the data.
- Linear Regression will capture linear relationship in the data points.

In Machine learning usually using simple algorithm at the beginning and later shifting to complex one is found out to be a fitting approach. While comparing linear and non-linear algorithm, the linear algorithm is better because it has a less computational cost and higher interpretability. However, non-linear can capture unusual and complex relations.

In total, six models will be created. The first three models belong to the control group part of the data and the remaining three to the experimental group. All six

model will have MCQ score as the target variable. The feature set for the first model is Knowledge and Motivation. Whereas the second model will be created by using six features of NASA-TLX and the third model comprises of both Motivation, Knowledge along with NASA-TLX features combined. Hence, altogether the last model will comprise of 8 feature with the same target variable. The same process of model building will be replicated for the experimental group as well.

Hyperparameter tuning is an essential step to know the right parameter setting for the model while training. The best hyperparameter selection manually can be a tedious task as there exist multiple permutation combination to give a shot. Hence, to make this task manageable, a grid search algorithm will be used to get the best value for each hyperparameter. This process will internally try executing various combinations to ensure the improvement of the model performance by reducing the prediction error and boosting accuracy.

The primary purpose behind creating these model is to find out the essential mental workload features which majorly influences the learner's performance in the class test. Feature importance computation is implemented using an algorithm called Permutation feature importance. It measures the importance of the feature by tracking the increment in prediction error after the permuting the feature. Here, permuting is nothing but randomly shuffling the values of a particular feature. The feature is allowed to be shuffled as many time as per requirement. A feature is considered to be important if the prediction error after shuffling increases. On the contrary, a feature is said to be unimportant if there is no change observed in the prediction error even after shuffling because, in this scenario, the model does not consider the feature for prediction. This concept of feature importance was introduced by (Breiman, 2001). Based on this idea (Fisher, Rudin, & Dominici, 2019) made various modification to propose a model agnostic version of feature importance. This feature importance algorithm will compute the importance score for all ten iterations. Hence, we can say each feature will have ten feature importance score; this will help determine the endurance of each feature and make the process of selecting variables straightforward. Feature importance score can be computed for both the train and test data. If the score is

computed for the train data, it shows that the model relies on each feature for making the prediction. In contrast, if it is computed for test data, it shows how much does each feature contribute to the overall performance of the model on unseen data.

Last part of the modelling phase will be the model evaluation. In this part, the model will be evaluated using Root Mean square error in case of the continuous dependent variable. Whereas, accuracy will be used for evaluation if the target variable is ordinal.

3.4 Model Evaluation

In the model evaluation part, important issues such as consistency and stability of the results will be considered. In the data cleaning step, missing value and outlier treatment was successfully applied to the dataset. The training samples are randomly created by using the Monte Carlo sampling, which is a form of repeated random sampling. From a dataset of records, 70 percent of instances are selected in random, which is nothing but the train set. This process is repeated for ten iterations to receive ten different sets of data on which training can be done. The correlation between MCQ Score and other relevant feature in the NASA-TLX subjective test will be tested using Pearson or Kruskal-Wallis, Spearman by p-value. If p-value¡0.05, there exist a connection. The magnitude of how strong the correlation will be is determined as follows:

- +/-.1 = small/weak correlation
- +/-.3 = medium/moderate correlation
- +/-.5 = large/strong correlation

As discussed above, the result of the six models will undergo testing. Test sets are 30 percent of the whole data, which consist of different instances then ten training sets. Various metrics such as RMSE in case of regression and accuracy for classification will be counted on ten results of the ten iterations through hypothesis test and visualisations. For the evaluation of the optimal model, there are two ways to evaluate.

- Testing the difference between actual and predicted for each model through RMSE and Accuracy.

- ANOVA test to be performed for the hypothesis testing to identify if one model is statistically significant than the other by using the ten RMSE score and accuracy captured through ten iterations.

A feature importance score will also be computed to recognise which features are contributing to predict the MCQ score, which in this case is the target variable. The feature importance score is also being computed for ten iterations. After which an ANOVA test followed by Post-Tukey is used to find out:

- Whether there exists a significant difference between Mental Workload feature of both the control group and the experimental group, a post-Tukey test will tell how much difference is present between two feature. This test will be executed individually for both groups.

From a visualisation stand-point, a box plot will be used for all the models to compare the RMSE, accuracy and feature importance score. The box plot will help explain the variation in the results and stability by the spread and size of the whiskers. All the test mentioned above will be repeated for both the training and test data. The threshold of significant difference between both the models is decided using p-value¡0.05 with 95 percent confidence interval. The intent behind the evaluation is to determine the following:

- the suitable model in both the groups

- measure the performance of the model

- Ensure that by using these models, we will get close to achieving the final goal, which is to find out the mental workload feature, which impacts the performance of learners in an MCQ test.

For the categorical target variable, our indicators will be accuracy which states the number of correct labels classified out of the total number of names which will

reflect the optimal predictive model. Precision can also be used which will capture when the model predicts the positive values correctly. For a continuous target variable, the RMSE score can be used for evaluation. RMSE score is a standard deviation of residuals where residuals are the difference between the actual value and predicted value. In other words, the RMSE score measures the spread of these residuals. RMSE is better than other error metrics because:

- It can present the variance on the same scale as the target variable.
- RMSE works on measuring error's average magnitude. The difference between the actual and predicted value is squared and then averaged over the sample. Later the square root of the sample is taken. Since the square of the residual(error) is computed first and later averaged, the RMSE score will heavily penalise the large errors. Hence, RMSE can be useful when more large errors are not desirable. In the case of MAE, it is more of a linear score, which means it will give equal weights to all the errors.

3.5 Strenghts and Limitations

The framework in the design chapter is accomplished and ready to accept features of any type(nominal, ordinal, interval). The key take away from the design chapter is that it has the facility to handle both linear and non-linear data. The current design consists of regression algorithms such as linear regression and decision tree regression. Linear regression takes care of the linear data with meagre computational cost and high interpretability. On the other hand, decision tree regression is responsible for handling the non-linear relationship and also uncover complex relations within the data. Hence, the use of both these learning approaches makes the framework more robust.

Very few mental workload research based on machine learning focuses on optimizing the hyperparameters. Tuning hyperparameter can control the training behaviour along with improving the performance of the model significantly. Hyperparameter tuning

is conducted using a grid search. It also helps us find out which parameter of the framework is crucial.

Training and testing the model using on different random sample every iteration helps keep track of the sturdiness of the model. By using multiple samples for multiple iterations, one can determine the stability of the model. Permutation feature importance technique is one of the best picks for computation of feature importance because this approach can be used across all supervised learning algorithm. Hence, making the model creation more approachable and flexible. Overall, the design showcases an end to end machine learning framework, which is accessible to more set of data.

Limitations: The model training and evaluation part is given the utmost importance in the design as the data received after the collection was clean. Hence, the framework does not invest much behind data cleaning. The use of subjective assessment results in vague data points which leads to high bias in the data; there is no way of handling this problem in the design framework. The data is limited to educational setting specific to learners in the third level education; however, the design can be extended to accommodate feature of the different domain by using the test set.

Chapter 4

Results and discussion

This chapter is organised to discuss and describe the evaluation and relevant study in-dept:

- The data description of the mental workload features along with the outcome variable
- A quick exploration of data which looks into the summary statistics, distribution and correlation between each other
- The result of all the experiments performed using various supervised and unsupervised learning approaches.
- Evaluating each model output and choosing the best fit model
- The final section discusses the strengths and weakness related to the findings

4.1 Data Description

The variable MCQ score is a continuous variable. Three histograms in figure 4.1, 4.2 and 4.3 shows MCQ score (N=406), MCQ Score - control group (N=209) and MCQ Score - experimental group (N=197) and it depicts that the histogram does not show discreteness and normality in the data. The left tail is shorter than the right tail. The histogram shows more students scoring between 80-100 in the class test.

Standard score(skewness values/standard error) for skewness between -2 and +2 are considered acceptable to prove normal univariate distribution. Whereas, for normal skewness score if the skewness is less than -1 or greater than +1, the distribution is highly skewed.

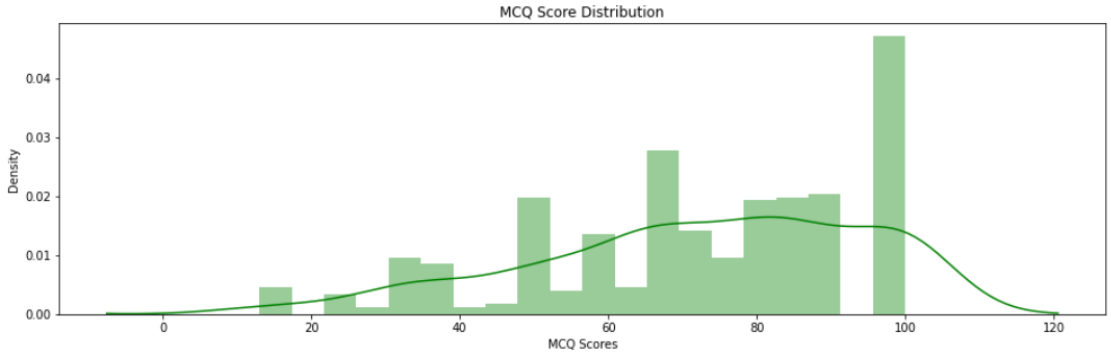

Figure 4.1: Histogram of MCQ Score depicting distribution

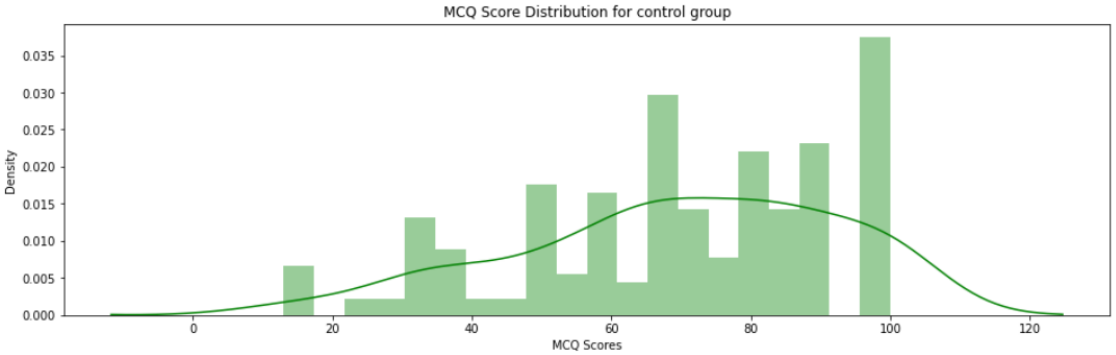

Figure 4.2: Histogram of MCQ Score in Control Group depicts the distribution

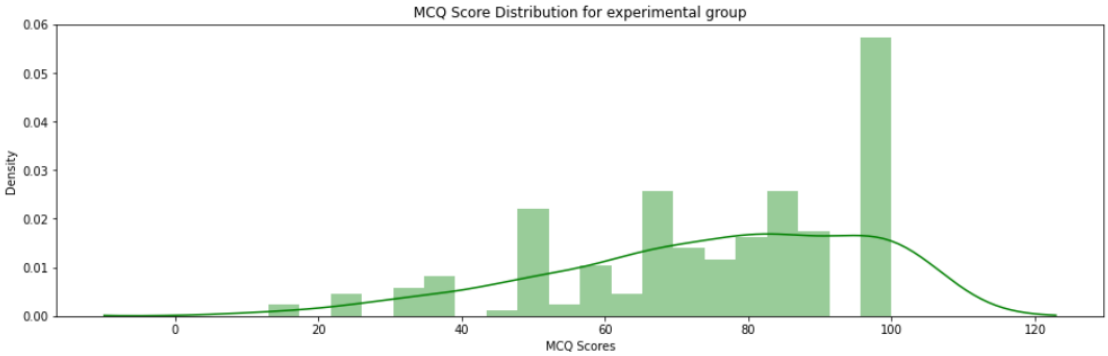

Figure 4.3: Histogram of MCQ Score in Experimental Group depicts the distribution

MCQ Score	Standardised Skewness score	Standardised Kurtosis Score
MCQ Score(overall)	-4.7	-1.6
MCQ Score(control group)	-2.9	-1.44
MCQ Score(experimental group)	-3.8	-0.67

Table 4.1: Skewness test of MCQ score

The non-standardised skew value is (Skewness: -0.51,-0.49,-0.52) for MCQ score overall, control and experimental group respectively which is between -1 and -0.5, which indicates that there exist moderate skewness in the data. However, after looking at standardised skewness and kurtosis score, it was observed that the scores go beyond -2 to +2, which is not acceptable. Hence, a further look into the data was given to check if at 0.05 level 95 percent of our data falls within +/-1.96(rounded as 2) the data can safely be treated as normal. Since the sample size of the data is beyond 80, we can take into account this criterion. After sorting the data, it was observed that around 15 values fall outside +/- 1.96, which is only 4 percent of the total data. Hence, it is safe to treat this data as normal. [1].

4.1.1 Multiple Choice Questionnaire (MCQ Score)

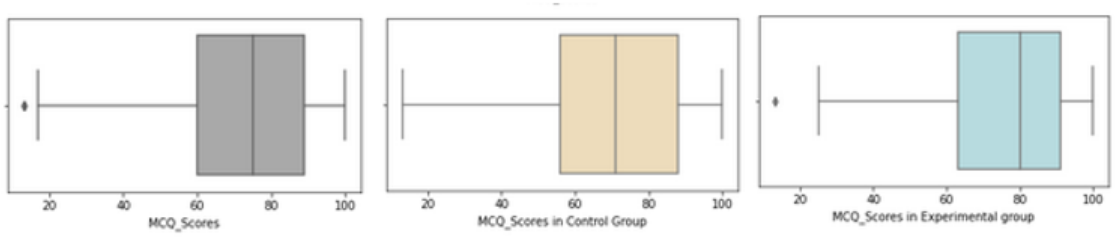

Figure 4.4: Boxplot of MCQ Score for overall data, control group and experimental group

[1] george2010spss

The box plot above 4.4 shows the shape of distribution of all the three MCQ Score. The central value of the control group is somewhere between 60-80, whereas, for the experimental group, the median value is 80 itself. As the size of the box plot is not spread out, and between small to medium size, it can be said that there is not much of the variability. The box start point is from 55 to 85 approximately, which states 50 percent of the data is between this range. A Shapiro Wilk test was performed to confirm MCQ score distribution explicitly. If p-value<=0.05 the test rejects the hypothesis of normality within the data. However, for all three MCQ Scores the Shapiro Wilk test had a p-value¡0.05. However, this test tends to be very sensitive for sample size larger than 100-200. It will tell you the data is not normal even if that is not the case. Hence, we look into other tests to be extra sure about the distribution. Therefore, skewness and kurtosis tests were computed. Skewness and Kurtosis standardised scores is shown in Table 4.1.

The kurtosis test measures the tailedness of the probability distribution. If the kurtosis value is positive, it states that the distribution is peaked and has a thick tail whereas if it is negative means you have light tails. The standardised score for kurtosis between -2 and +2 are considered acceptable. In this case 95 percent of the data is within this range. Hence, MCQ score for both the control group and experimental group is acceptable and proves normal univariate distribution.

Looking into descriptive statistics, we can see the measure of central tendency of MCQ scores indicates the number of samples (n=406) with average 72(SD=22.2) making the coefficient of variation which is the ratio of standard deviation and mean to be 0.3 which is less than 1. As a result, it indicates a relatively low standard deviation. Similarly, for the control group the total number of records (n=209), MCQ score ranged from 13 to 100 (M=69.5, SD=22.6) and the experimental group with records (n=197) MCQ Score having (M=75, SD=21.43). Both the control and the experimental group has low standard deviation.

4.1.2 Pre-Knowledge and Motivation

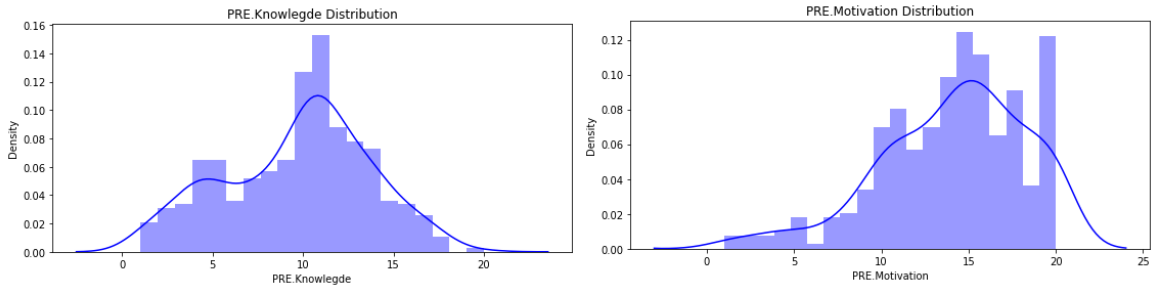

Figure 4.5: Histogram depicting the distribution for Knowledge and Motivation

Knowledge and Motivation, specifically, do not belong to NASA-TLX. It is not one of the Mental Workload attributes. Still, it can be beneficial to have a characteristic such as prior knowledge and prior motivation to better know learners state of mind. Knowledge and Motivation, just like other features, ranging from (0-20). The distribution of knowledge see be in the figure is entirely symmetrical with standardised skewness score falling within +/- 1.96. Whereas kurtosis is slightly falling outside the range with a negative value, this phenomenon is called Platykurtic, which signifies the tails are lighter than a normal distribution. However, from a sample of (n=406) which is only 1.9 percent of the total data falling outside the range. Hence, motivation can be treated as normal.

The box plot in figure 4.6 shows the variability of both the variables, which is almost similar. The standard deviation, range and interquartile range of both the variable are close when compared with one another. Examining the box plot reveals that knowledge distribution is close to appearing symmetrical, whereas motivation is less clear from the box plot.

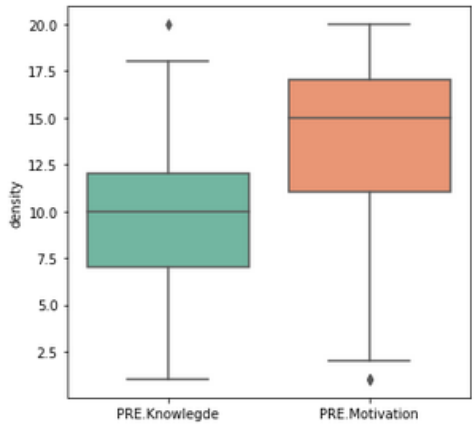

Figure 4.6: Boxplot for Knowledge and Motivation depicting variability

4.1.3 Raw NASA-Task Load Index features

The Raw NASA-Task Load Index consists of six features, namely Mental, Physical, Temporal, Performance, Effort and Frustration. The NASA-Task Load Index has weights assigned to each feature whereas this is not the case for Raw NASA-TLX. The range of all the features is from 0-20, just like Motivation and Knowledge. Having a glance over the skewness of these six features, we see that the features such as physical and frustration has the highest skewness of 6.7 and 5.0 respectively. It falls outside the limit of +/-2, as shown in the table below. Whereas, features like Mental, Temporal, Performance and Effort look perfectly normal in distribution. The skewness of physical and frustration is not acceptable. A detailed investigation was further carried out to see wheather 95 percent of the data at 0.05 level falls within +/-2 range. After scaling the data, it was observed that 3.6 percent for physical and 3.4 percent for frustration was falling outside the limit of +/-2. Since the dataset is large than 80 samples, it is safe to accept the data to be normal.

MWL Feature	Standardised Skew Score
Mental	-1.65
Physical	6.7
Temporal	1.08
Performance	-0.37
Effort	-1.25
Frustration	5.0

Table 4.2: Standardised Skewness Score of Mental Workload Features

The box plot for the NASA-TLX features in figure 4.7 showcases variability of all the six features. The variability of these features is very similar to each other. The standard deviation, mean and interquartile range are similar for Mental, Temporal, Performance and Effort. In the case of physical and frustration, the standard deviation, mean and interquartile range falls in the same range. Examining the figure 4.7 it can be seen that the mental, effort, performance are close to symmetrical shape. In contrast, physical and frustration have long box signifying more variance in the data compared to other models. The box with a long tail from the top of the box would be consistent and be considered as a positive skew. But having median at the top of the box is generally regarded as negative skew.

The NASA-TLX score was also calculated using the six features without incorporating weights of the features. The skewness(1.91) and kurtosis(2.0) is entirely within the range of +/-2. Hence, NASATLX is normal in terms of distribution. This measure is computed by summing all the features and dividing the sum by 15. The distribution of NASATLX is shown below in figure 4.8:

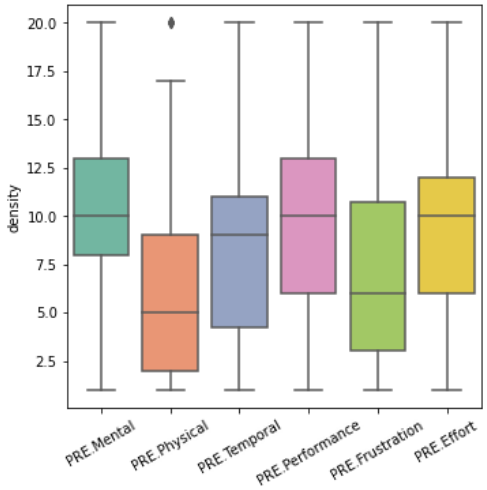

Figure 4.7: Boxplot showing variance for all mental workload features

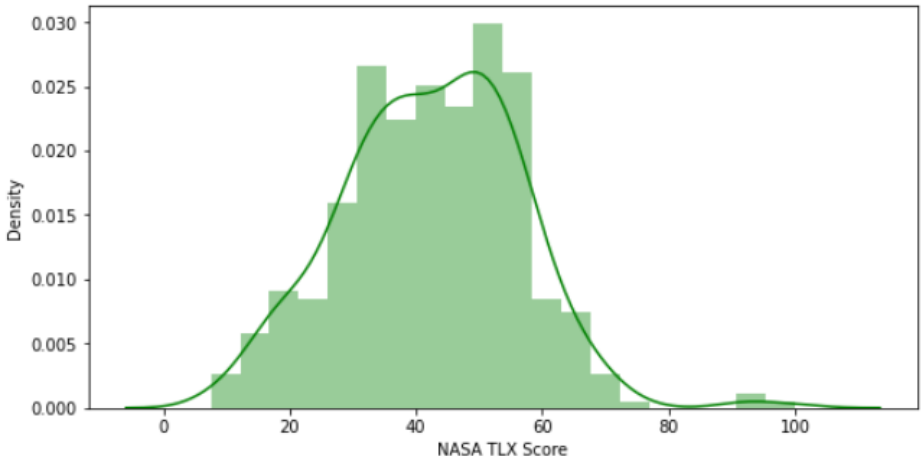

Figure 4.8: Histogram for NASA Task Load Score depicting its distribution

4.2 Data Exploration

Feature	Missing	N	min	median	max	mean	sd
MCQ Score	0	455	13	75	100	72.4	21.4
Mental	0	455	1	10	20	10.26	3.76
Physical	0	455	1	5	20	5.7	4.22
Temporal	0	455	1	9	20	8.33	4.4
Performance	0	455	1	10	20	9.67	4.7
Level of Frustration	0	455	1	6	20	7.07	4.69
Effort	0	455	1	10	20	9.53	4.30
Knowledge	46	409	1	10	20	9.64	4.05
Motivation	23	432	1	15	20	14.23	4.15

Table 4.3: Summary statistics

There are 455 records collected from 19 lectures conducted on 19 different modules by 11 different lecturers in a master's classroom of Technological University Dublin. The class strength is roughly between 20-40, and each class divided into two groups control and experimental. The range of MCQ Score is between 0-100 with (Mean 72.4, Median 75). The range of other features is between 1-20. Physical demand has the lowest score (mean = 5.7, Median=5) and Mental Demand has the highest score (Mean=10.26, Median=10) among all features influencing the performance of learners in an MCQ test. Level of Frustration has the next lowest score of (Mean=7.07, Median=6). Apart from Raw NASA-TLX features, other features such as motivation has the highest score with (Mean=14.23 and Median=15). Motivation and Knowledge both these variables have missing values. At an individual level, the sample size distribution of the control group (N=211) and the experimental group(N=198) considerably equally divided.

The target variable is MCQ Scores which will evaluate the performance of the learners. The distribution of MCQ scores is moderately skewed with approximate skewness = -0.57. Features like mental demand, motivation, temporal demand, performance

demand, effort, knowledge has a normal distribution with skewness score between -0.5 to +0.5. Whereas, variables like physical, frustration are again moderately skewed.

A raw NASA-TLX score has been calculated for both questionnaires filled before the MCQ test (NASA-TLX Pre) which applies to both the groups and after the MCQ Test (NASA-TLX Post) which applies only to the experimental group. This score derived by taking the summation of features of mental workload and dividing by 15 - which is nothing but the total number of paired comparisons. Table 3 shows the mean and standard deviation of NASA-TLX pre, NASA-TLX post and MCQ scores associated to each topic for individual group.

Table 4.4 shows the mean and standard deviation of NASAT-TLX and MCQ, which is associated with each topic and related group. According to the table above, on an average, the experimental group experienced more cognitive load (NASA) than the control group. Therefore, instinctively this can be attributed as extra mental load and cognitive cost required in collaboration activity. However, the learners in the experimental group perform (MCQ Score) better than the control group. Hence, we can say that even though there is more cognitive load associated with the collaborative activity, but it did increase the overall performance level of the learners belonging to the experimental group. Further to confirm the normality of the data topics a skewness test was conducted, which was followed by T-test if the distribution is normal else a Mann Whitney test for not normal distribution (p-value$<$0.05). These tests are performed to compare the means of the control and experimental group. Difference test on the below combination was conducted:

- Difference between both the groups that is the control group and the experimental group for the entire class for MCQ Score and NASA-TLX score.
- Difference between both the groups that is the control group and the experimental group for every individual topic for MCQ Score and NASA-TLX score.

	MCQ Mean(SD)		**NASA-Pre Mean(SD)**	
Topic	Control	Exp	Control	Exp
Data Mining	37.8(14.4)	32.2(11.04)	42.9(8.37)	47.9(21.2)
IT Forensic	52.6(17.8)	56.3(17.4)	34.34(14.32)	40.06(16.8)
Image Processing	69.2(15.3)	82.5(7.5)	35.4(13.06)	54.5(2.59)
Lit Comprehension	73.3(16.3)	75.5(16.6)	52.2(14.79)	43.8(9.84)
Literature Review	69.4(19.5)	68.5(15.2)	47.29(10.79)	45.5(10.9)
Machine Learning	77(8.21)	77.4(8.0)	33.1(5.5)	37.38(5.57)
O.System	65.6(22)	84.1(14.5)	37.5(12.5)	42.36(13.7)
Operating Systems	80(14.14)	84.6(14.5)	39.35(10.43)	42.3(13.8)
P.Solving	76.2(24.17)	54.8(25.2)	39.9(14.03)	46.2(10.63)
Program Design	85.3(19.2)	88(16.56)	35.3(15.8)	39.6(17.8)
R.Hypothesis	82.3(17.37)	89(13.9)	45.79(17.6)	45.15(10.93)
R.Methods	71.5(22.12)	75.3(14.36)	47.14(12.03)	47.3(17.4)
Res Hypothesis	82.5(17.7))	98.46(5.54)	46.8(11.8)	34.74(12.9)
Research Methods	69.2(22.3)	87.3(11.3)	43.8(19.34)	41.74(16.8)
Statistics	46.8(29.6)	64(22.27)	58.1(7.39)	54.7(7.4)
Strings	57.3(23.8)	74.9(21.89)	44.8(16.9)	30.34(8.35)
V.Geo.Data	45.4(20.8)	58.85(17.3)	35.8(13.16)	43.69(11.7)
Virtual Mem	75.12(12.3)	73.4(8.7)	42.29(8.8)	47.8(9.9)

Table 4.4: Mean and Standard Deviations of MCQ Score and NASA-TLX grouped by control and experimental group

Taking into account the first combination, we see no significant difference in the score for NASA-TLXPre (M=41.9, SD=14 for the control group and M=42.7, SD=14.10 for the experimental group), (t(406)=-0.59, p=0.55) but, we see a substantial difference in the score for MCQ Score (M=69.52, SD=22.6 for the control group and M=75, SD=21.43 for the experimental group), (t(406)=1.29, p=0.01). Hence, we further deep down at the second combination, which is the classwise approach for all topics.

Despite the experimental group performing better than the control group, the result of T-Test or Mann Whitney test, shown in table 4.5, a stastically significant difference in the MCQ scores was found between both the groups for the following topics *O.Systems ($U=87.5, p<0.05$)*, *P.Solving($U=40.5, p<0.05$), Research Methods($U=14.5, p<0.05$)*, *Res. Hypothesis($U=51, p<0.05$)* and relevant difference was also found in the NASA-TLX scores for the following topics Image Processing($U=2.5, p<0.05$) and Res. Hypothesis($U=56, p<0.05$) only. Unfortunately, in the case of other topics, after both conducting independent sample t-test or Mann Whitney test it was witnessed there is no significant difference between the control and experimental group for the MCQ scores or NASA-TLX scores where the P-value is greater than the significance level (P-value>0.05 with 95 % confidence interval).

Orru and Longo (2019) performed a similar experiment in his paper. According to the article, no significant difference was observed for various topic between both the groups in the MCQ and NASA-TLX scores. One of the core reason behind it was also scarcity of data. The sample size of every class was nearly 20-30 students which is low sample size. Hence, this motivated a new angle to the research question and instead analyzed the impact of each Mental Workload feature on student's performance.

Topic	**MCQ**	**Nasa-TLX**
R.Methods	0.5(M)	0.9(T)
R.Hypothesis	0.09(M)	0.43(M)
V.Geo.Data	0.46(T)	0.3(T)
O. Systems	***0.003(M)***	0.26(T)
P.Solving	***0.02(M)***	0.23(T)
Data Mining	0.36(T)	0.41(M)
Literature Review	0.88(T)	0.65(T)
Research Hypothesis	0.09(M)	***0.01(T)***
Strings	0.09(T)	***0.02(M)***
Prog.Design	0.37(M)	0.49(T)
Mac.Learning	0.9(T)	0.22(T)
Image Processing	0.07(M)	***0.018(M)***
Research Methods	***0.03(M)***	0.82(T)
Statistics	0.29(T)	0.46(T)
IT Forensics	0.56(T)	0.31(T)
Lit.Compreh.	0.8(T)	0.21(T)
Virtual Mem	0.79(M)	0.31(T)
Res. Hypothesis	***0.03(M)***	***0.013(T)***
Operating Systems	0.4(T)	0.54(T)

Table 4.5: P-value at significance level=0.05 of T-Test(T) or the Mann Whitney test(M) of the MCQ(Multiple choice question) and NASA-Task Load Index

4.2.1 Correlation between MCQ Score and other features

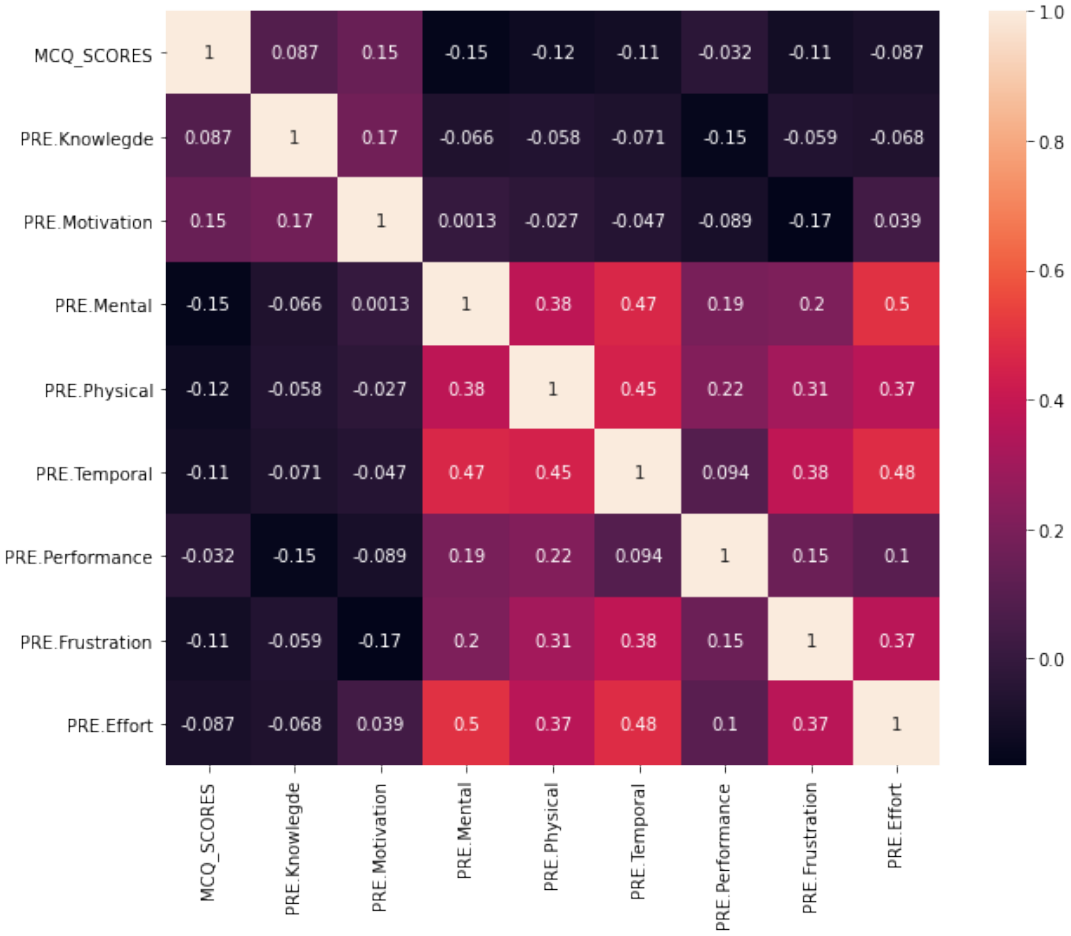

Figure 4.9: Correlation Matrix of MCQ scores and Mental workload features from NASA-TLX

As both the MCQ squares and NASA-TLX features are quantitative, a Pearson correlation test was performed on all the features. Figure 4.9 shows the correlation matrix of MCQ scores with features of NASA-TLX set. It is essential to determine if there exists a linear relationship between two variables and its strength. The Cohen's effect size close to +/- 1 the stronger is between two variables. If Cohen's effect size is larger than [-0.5,0.5], there exists a strong relationship. If it's higher than [-0.3,0.3], there exists a moderate relationship, whereas if it's higher than [-0.1,0.1], it's said to have a weak relationship. Figure 4.11 shows that there is a positive correlation between mental demand and effort, r=0.5,p<=.001. Apart from effort, mental demand is pos-

itively and moderately correlated with temporal demand r=0.47, p<=0.01 followed by moderate correlation with physical demand r=0.38, p<=0.01. Similarly, there was a moderate relationship between temporal demand and physical demand r=0.45 and frustration with r=0.38 with p<=0.01. Frustration is moderately correlated with effort with r=0.37 with p<0.01. There was a weak correlation existing between MCQ and other NASA-TLX features along with Knowledge and Motivation.

4.3 Feature Selection

The foremost research objective is to find for Mental workload features that have the most impact on learners which in turn hampers their performance. Initially, Decision Tree Regression and Linear Regression was only going to be part of the analysis. Decision Tree in scikit learns it is own Feature importance function which computes the importance of each feature in the model. Whereas, for the linear regression coefficient of the features is sometimes used as feature importance. However, the coefficient determines the direction of the relationship between a dependent and independent variable. However, it will not answer as to which variable was most important to predict the target variable. To compare the feature importance across different learning approaches, it was necessary to find something versatile which can be applied to all the machine learning algorithm.

Permutation Feature Importance: A feature importance method called permutation feature importance holds a speciality where we can modify it to work with any machine learning algorithm. Initially, the idea was introduced for the random forest algorithm, which later went ahead and got scaled up to multiple learning algorithms. In simple terms, permutation feature importance works behind checking the increase in prediction error after the feature is being permuted (permute here means shuffling of a particular column) which breaks the relationship between the output variable and feature. The logic behind this concept is very straightforward. The importance of the feature is computed by increasing the prediction error of the model after permuting

the feature. In simple words, every feature is shuffled, and it is claimed to be important if the model error increases after the shuffling of the feature because technically the model relies on the feature for prediction. If the model error remains unchanged after shuffling the feature, then that feature is claimed to be unimportant.

There are multiple evaluation metrics used based on the type of algorithm being applied. For instance, for classification problem accuracy, precision is measured to see if any change in model accuracy was noticed after shuffling. Similarly, for the regression problem metric like rmse, mae, mse are used.

As previously, both regression and classification algorithms were used. Therefore rmse and accuracy, respectively, are used to compute the feature importance score. The original error from the model is computed. Later, the feature importance score is calculated by taking the difference between model error after permutation and the original model error, which was computed initially. Feature Importance score for all three feature set.

4.4 Model Training

Repeated random sampling was used to evaluate machine learning models on a limited sample of data. A single parameter "random state" can shuffle data into a different combination. In other words, the sample is shuffled before each repetition which results in a different split of the same sample. This process of random sampling is known as Monte Carlo Sampling. A value of 10 is observed to be accepted in the machine learning field, which is found via various experiments. It also results in low bias and medium variance (Kuhn, Johnson, et al., 2013). With the help of the ten iterations, it will be handy to have a look at the consistency of the model.Model training will require splitting the data into train and test set. Here, 70 percent train and 30 percent test set is allocated to the model. The model is trained on three feature set, which is as follows:

- The feature set with only six NASA-TLX attributes
- The feature set with Motivation and Knowledge

- The feature set with a combination of first and second features.

4.4.1 Model trained using Regression

Multiple Linear Regression:

Multiple Linear Regression model was build for both control and experimental. The model was trained using MCQ Score as the dependent variable and NASA-TLX feature along with knowledge and motivation as independent variables. The main aim was to find out the most important characteristics affecting learners performance in a class test. Initially, the data is split into 70:30 ratio for train and test set, which is standard across all learning algorithms. The model was trained on three features set as discussed above.

The model was trained using Multiple Linear Regression on ten different sets of randomly sampled data. It computes ten iterations of the model through which we can derive the stability and consistency of the model. Model evaluation was performed using the RMSE(Root mean square error) as it is better at penalising error with high weightage. RMSE score was computed both for train and test to determine how much the error deteriorates from training to test. This will help to examine the model better. As the model has been trained on ten different random samples, we have ten RMSE values per model. As shown in the table below we can see that the RMSE score is between 20 to 25. In the case of RMSE, it is generally said that there is no specific range that determines a good or bad score. However, the dependent variable, which is the MCQ Score ranges from 0-100 based on which we can very well determine whether the score is reliable or not. Hence, looking at the MCQ Score range, having an RMSE score in the 20's, state that the model has high error. If we look into the individual group as shown in table 4.6 and 4.7, we can see both the group has the same range of RMSE score representing consistency of the model. The RMSE score for both the train and test set is close to each other. Hence, we can say that the model is not overfitting. However, while training the model does not tests well for the data inside and outside the sample.

	RMSE Score					
Iteration	Feature Set 1		Feature Set 2		Feature Set 3	
No of Iteration	Train	Test	Train	Test	Train	Test
Iteration 1	21.8	24.3	21.3	24.7	21.2	24.8
Iteration 2	22.8	22.03	22.4	22.4	22.3	22.5
Iteration 3	23.2	20.8	22.3	23.5	22.1	23.8
Iteration 4	22.9	21.5	22.8	21.5	22.7	21.3
Iteration 5	22.9	21.7	22.6	21.9	22.5	22.3
Iteration 6	22.8	21.8	22.6	22.03	22.5	21.9
Iteration 7	22.5	22.6	21.9	23.7	21.9	23.7
Iteration 8	22.5	22.7	22.3	22.7	22.15	22.9
Iteration 9	23.2	21.0	22.9	21.57	22.8	21.9
Iteration 10	22.8	21.8	22.7	21.6	22.6	21.8

Table 4.6: RMSE Score for Control Group

Feature Importance for Linear Regression: In the case of Linear Regression, rmse score was used to measure the permutation feature importance score. Ten iterations were executed to determine the consistency of the feature importance score. The number of repeats was set to 10, which is nothing but the number of times a particular feature value should get shuffled. The feature importance score projected below is for comprehensive data.

Feature set 1: As we can see in figure 4.12, the feature set 1 comprises of knowledge and motivation. Knowledge has high importance for both train and test set compared to motivation in control as well as the experimental group. Motivation in both the groups is negative, which indicates its shuffling had no impact on the model's prediction error. Hence motivation is an unimportant feature. However, if we look into both the group's test set even though knowledge has a high feature importance score, it has high variance. The coefficient of variation, which is a ratio of standard deviation and mean is greater than 1 in case of knowledge—looking into both the control and

the experimental group it is observed that learners having previous knowledge about the topic can perform better in class test.

	RMSE Score					
Iteration	Feature Set 1		Feature Set 2		Feature Set 3	
No of Iteration	Train	Test	Train	Test	Train	Test
Iteration 1	21.45	22.3	22.3	22.12	21.14	24.67
Iteration 2	22.67	22.12	22.6	23.4	24.3	25.5
Iteration 3	23.4	20.12	21.3	22.4	22.07	24.3
Iteration 4	23.1	22.4	22.7	21.2	22.42	23.6
Iteration 5	22.5	21.4	21.8	22.3	22.5	24.4
Iteration 6	22.12	21.6	22.4	22.03	22.3	21.4
Iteration 7	22.3	21.5	21.7	23.4	21.9	23.9
Iteration 8	22.12	22.6	22.12	22.5	22.43	22.6
Iteration 9	23.5	21.3	22.7	21.6	22.12	21.56
Iteration 10	22.8	21.5	22.12	21.7	23.2	23.3

Table 4.7: RMSE Score for Experimental Group

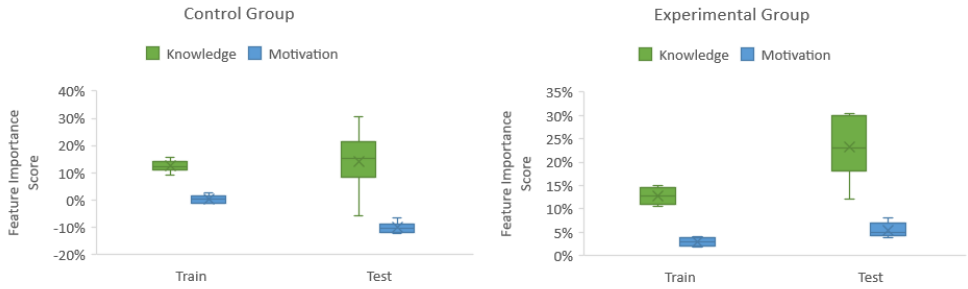

Figure 4.10: Linear Regression: Feature Importance - Feature Set 1 (Knowledge and Motivation)

Feature Set 2: This set contains only NASA-TLX features. Figure 4.13 represents that Mental demand followed by Physical demand and Effort is the essential feature in both train and test having positive and high feature importance score in the control

group. Whereas, in the experimental group we see mental demand, physical demand and frustration has higher feature importance. Features such as temporal is unimportant in both the group. However, frustration and performance in control group and performance in experimental group have small but positive importance while training and testing, which states that the model moderately relies on these features for better model prediction. Mental demand in test data of experimental has high variability, which indicates a high standard deviation. Comparing both the control and experimental group we see that learners in experimental group faces more frustration then in the control group. As learners in experimental group are involved in communicating with other learners which can lead to frustration.

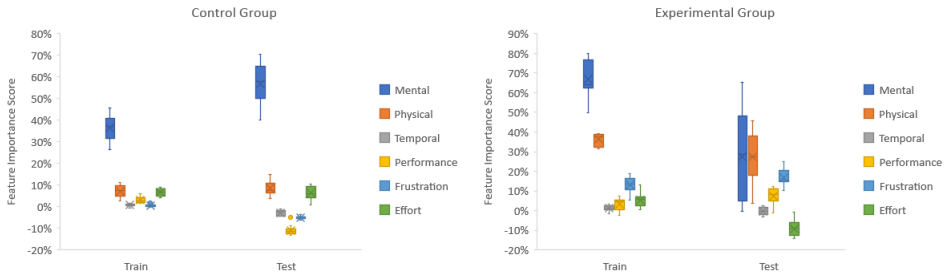

Figure 4.11: Linear Regression: Feature Importance Score - Feature Set 2 (NASA-TLX features)

Feature Set 3: The feature set 3 is a union of feature set 1 and 2. The boxplot in figure 4.14 shows feature importance score for the feature set 3 for both the groups. According to figure 4.14, it was observed that features such as mental demand, knowledge, effort and physical are essential in the control group. Whereas, in the experimental group motivation, mental demand, physical and performance contribute more towards predicting the MCQ score. Knowledge in case of the experimental group was significant only in the train set, but it had no contribution in the test set. It was observed that learners in the experimental group have motivation as one of the critical factors in predicting their performance which was not the case in the control group.

The boxplot in figure 4.14 shows that Mental Demand followed by Knowledge and Physical Demand is having high feature importance score. The effort has a minimal

contribution towards model prediction and improvising model performance for better prediction on unseen data. However, during the test set, it was observed that there exist a high standard deviation, which indicates the importance spread over an extensive range of values. Hence, we cay say the stability of these variables is low.

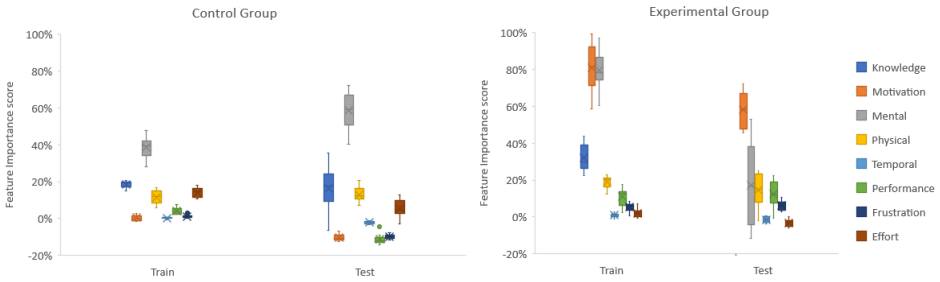

Figure 4.12: Linear Regression: Feature Importance Score - Feature Set 3 (NASA-TLX including Knowledge and Motivation)

Decision Tree Regression: A Decision Tree Regression was used to capture unusual and complex relations which Linear Regression might miss on capturing. The model is trained using the same feature set as Linear Regression. Initially the hyperparameters were set to default values. However, setting to default caused overfitting of the model. Therefore, grid search was used to tune the hyparameters set them with relevant values. The settings were as follows:

Parameter	Feat Set 1	Feat Set 2	Feat Set 3
criteria	mse	mse	mse
maximum depth	10	10	2
maximum leaf nodes	100	40	60
minimum sample leaf	20	40	40
minimum sample split	5	10	20
splitter	random	random	random

Table 4.8: Hyperparameter Setting for Decision Tree Regression

The RMSE score for train and test after setting the hyperparameter is as follows:

	RMSE Score					
Iteration	Feature Set 1		Feature Set 2		Feature Set 3	
No of Iteration	Train	Test	Train	Test	Train	Test
Iteration 1	21.6	23.4	21.3	23.7	21.8	24.5
Iteration 2	22.8	22.9	22.7	22.3	23.7	23.9
Iteration 3	21.01	21.1	21.4	22.7	23.0	24.1
Iteration 4	22.4	22.1	22.8	21.9	22.8	23.1
Iteration 5	22.4	22.7	22.6	21.3	22.5	21.8
Iteration 6	22.8	20.9	21.9	21.4	22.2	21.9
Iteration 7	23.3	22.4	22.6	21.6	22.5	23.7
Iteration 8	22.1	22.0	22.5	22.6	22.5	22.7
Iteration 9	23.5	20.7	23.3	20.5	23.1	21.5
Iteration10	22.9	21.8	22.6	22.0	23.7	24.0

Table 4.9: RMSE Score for Control Group - Decision Tree Regression

	RMSE Score					
Iteration	Feature Set 1		Feature Set 2		Feature Set 3	
No of Iteration	Train	Test	Train	Test	Train	Test
Iteration 1	20.3	22.3	19.9	23.7	20.2	23.7
Iteration 2	20.5	22.6	20.8	22.0	20.7	21.6
Iteration 3	22.0	19.3	21.5	20.2	21.3	20.1
Iteration 4	20.6	23.1	20.6	23.1	20.5	23.6
Iteration 5	22.3	19.2	21.8	18.6	21.8	19.7
Iteration 6	22.4	18.4	22.6	18.4	22.5	18.0
Iteration 7	21.5	21.1	20.7	21.8	21.1	20.6
Iteration 8	21.3	20.5	21.6	19.8	21.7	19.8
Iteration 9	21.4	20.2	21.8	19.7	21.7	19.7
Iteration 10	21.6	21.1	21.3	20.8	21.7	20.8

Table 4.10: RMSE Score for Experimental Group - Decision Tree Regression

Feature Importance : Decision Tree Regression Feature Importance in Decision Tree regression is also calculated using RMSE score just like Linear Regression. Decision tree itself has a feature importance function which calculates the significance of each variable. However, an algorithm common to all learning algorithm was the primary motive. In the case of the Decision Tree, if a particular feature is not considered for splitting the feature importance score of that feature will be set to 0. The feature importance score is calculated for both the control and the experimental group. The importance is calculated on 10 random sample of train and test data.

Feature Set 1: Figure 4.15 shows the feature importance score for the control and experimental group of feature set 1, which consist of knowledge and motivation. The critical feature found in the control group is knowledge, whereas, in the experimental group, it is the motivation for both train and test set. Feature importance of knowledge and motivation is very close to each other in the test set of the control group. Hence, looking at the test set of a control group, it can be said that the model relies moderately on knowledge and motivation.

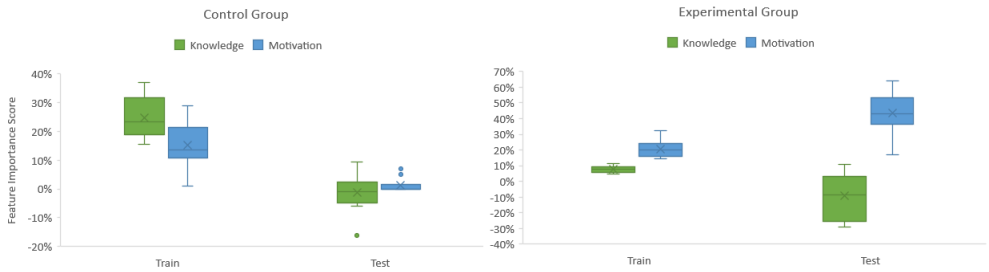

Figure 4.13: Decision Tree Regression: Feature Importance - Feature Set 1 (Knowledge and Motivation)

Feature Set 2: The feature importance score for feature set 2 can be seen in figure 4.16. It can be observed that mental demand is an essential feature in both the control group and the experimental group. However, in the experimental group, apart from mental demand features such as performance, frustration, and temporal moderate are of high importance. Out of all the features having high importance in the train set of the control group except mental, every other variable was unimportant. In other words, these features have no contribution to the performance of the model on

unseen data. Frustration level, mental demand, performance and temporal demand are essential features influencing the performance of the students in the experimental group.

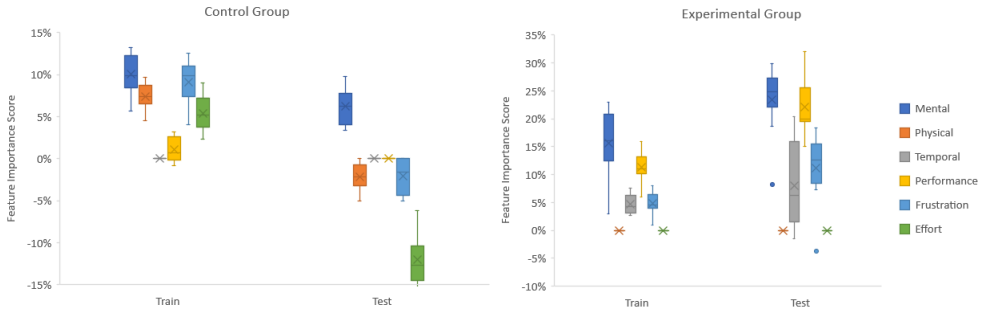

Figure 4.14: Decision Tree: Feature Importance - Feature Set 2(NASA-TLX Features)

Feature Set 3: Figure 4.17 shows the feature importance of feature set 3, which is a blend of knowledge plus motivation and six feature of NASA-TLX. There are, in total, eight features used in this feature set to train the model. The critical feature in the control group is frustration, mental demand, effort and knowledge for both train and test set. The critical feature in the experimental group is motivation, mental and frustration in both the train and test set. However, the variance of mental demand and knowledge in the test set of the control group is too high; this leads to high variance and high bias issue where bias is determined from the RMSE score table which is very high.

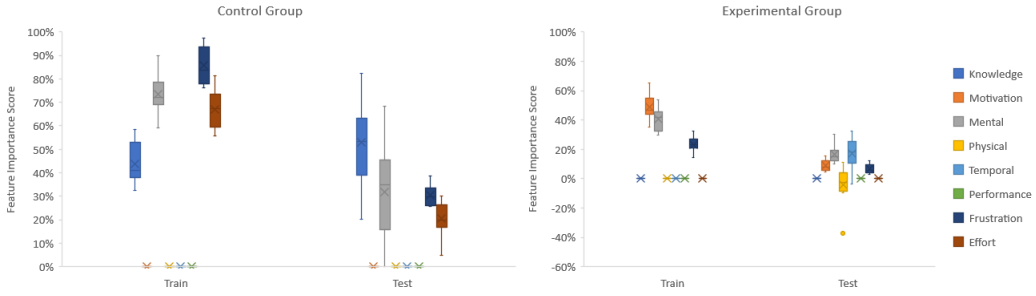

Figure 4.15: Decision Tree: Feature Importance - Feature Set 3 (Knowledge and Motivation with NASA-TLX features)

The feature importance for both Decision Tree and Linear Regression is for both

the control and experimental group. The features important in the control group common across the regression model includes knowledge, mental and effort. Whereas for the experimental group features such as motivation, frustration, temporal demand and mental demand are critical. However, both the regression model had a high variance in the feature importance, especially in the test set and high bias in the RMSE score for both train and test set. This was observed in both the control and the experimental group. The the model RMSE score is high due to which we cannot obtain a generalised model.

Regression with Interpolation: After training the model with all three feature set to predict the MCQ scores, both the decision tree regressor and linear regressor had high RMSE scores which states that the model is unable to generalise well to predict the target variable (MCQ score) accurately. As there is no acceptable range for RMSE score, it was determined by seeing how the dependent variable is scaled. Therefore, compared to the MCQ score range, RMSE score range was very high. One of the possible reason for this high prediction error is the lack of unique points in MCQ score. In other words, the data points in MCQ are very concentrated to a specific range. Hence, the research was further extended by incorporating interpolated data points to the actual data. Interpolation will construct new data points inside the range of a discrete set of known data. It helps to create more unique data points to increase the sample size to increase the training performance of the model. Two types of interpolation technique were used 1. Linear 1D interpolation and 2. Spline interpolation.

- Linear Interpolation is used when we are looking for values within set of values. Basically, it does the job of filling the gaps in the data. Often Linear Interpolation is not the best idea for non-linear data.

- Spline Interpolation is treated as polynomial interpolation and it is often used to smoothen the error. It is one of the most supple ways to interpolate.

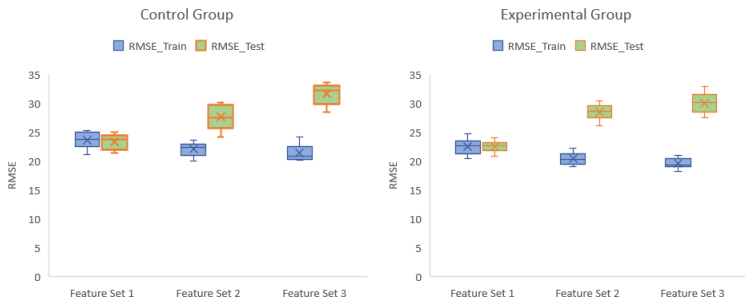

Figure 4.16: Box plot for Linear Regression (RMSE Score) using Interpolated data for all 3 feature set

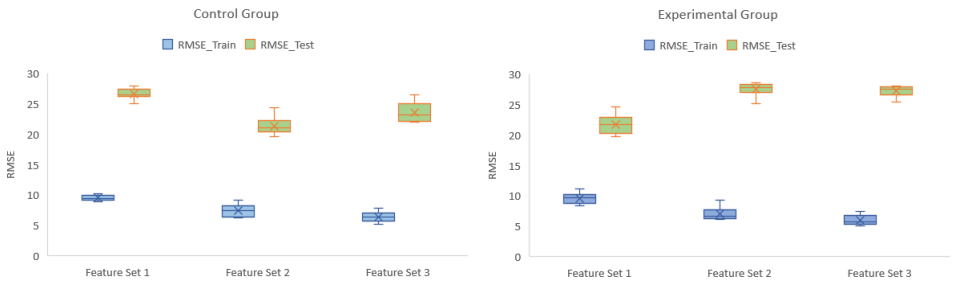

Figure 4.17: Box plot for RMSE - Decision Tree Regression using Interpolated data before hyperparameter tuning

However, it was observed in figure 4.18 and 4.19 the RMSE score after interpolation is still in the same range as it was before interpolation. The only difference which can be observed is that the error for the train set had reduced, but the test error increased a lot. The reason being the model was trained on interpolated data, whereas the testing was performed on the original data, which was without interpolation. In the case of Decision Tree Regression, overfitting was observed, as seen in figure 4.19. However, the overfitting issue was resolved after the hyperparameter tuning was performed. This tuning was conducted using the grid search-relevant parameter value is retrieved out of the pool of values given to the algorithm. The parameter set for each feature set are as follows:

Parameter Name	Feature Set 1	Feature Set 2	Feature Set 3
Criterion	mse	mse	mse
Maximum Depth	8	10	15
Maximum Leaf Node	60	60	100
Minimum Sample Leaf	11	5	3
Splitter	random	random	random

Table 4.11: Hyperparameter setting for Decision Tree Regression with Interpolation

After setting the model parameter overfitting was reduced in Decision Tree Regression at the cost of an increase in training error, as shown below in figure 4.20. The training error and test error has reached a range between 20-30. Hence, we can say there was no improvement observed. Interpolation helped reduce the training error, but the test error remained the same. The problem of generalisation remains even after applying interpolation.

Figure 4.18: Boxplot for Decision Tree RMSE Score using interpolated data after hyperparameter setting

Feature Importance Score:Regression with Interpolation The feature importance score for every feature after interpolation for both Linear regression and Decision Tree regression remains similar to the previous score. Feature Importance of Decision Tree:

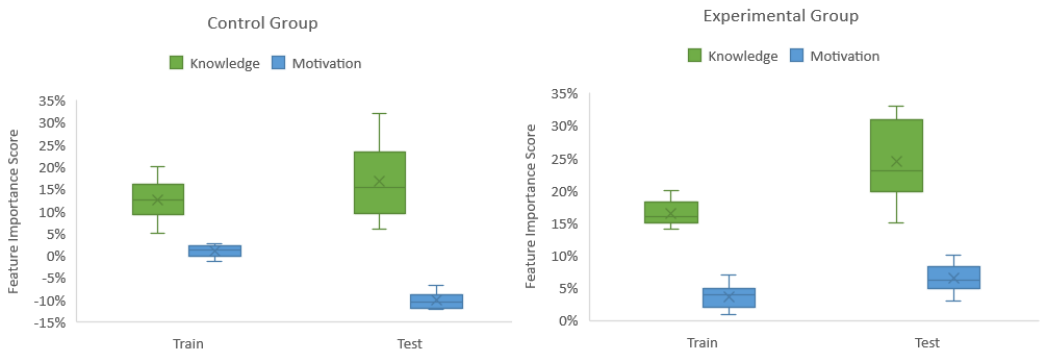

Figure 4.19: Decision Tree Regression with Interpolation - Featue Set 1(Knowledge and Motivation)

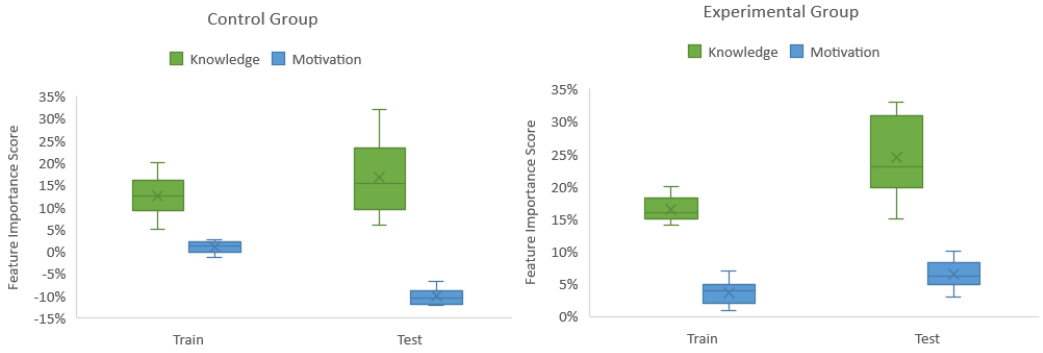

Figure 4.20: Decision Tree Regression with Interpolation - Featue Set 2(NASA-TLX Features)

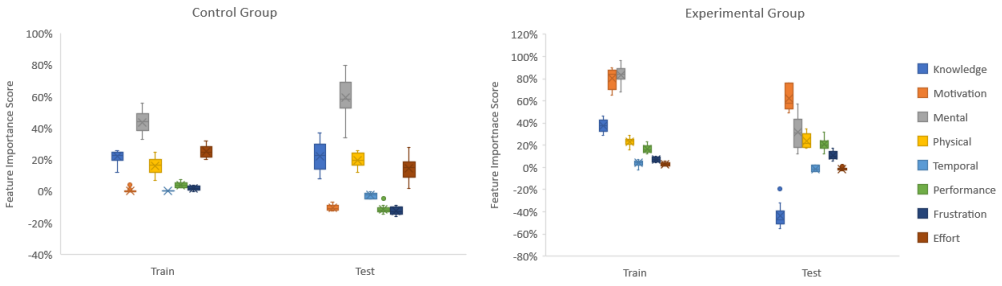

Figure 4.21: Decision Tree Regression with Interpolation - Featue Set 3(NASA-TLX Features with Knowledge and Motivation)

Interpolation is only used to train the model; the model testing is performed in the same set without interpolation. Hence in all three feature set, not much of a

difference in feature importance score is visible for the test set. The three feature importance score shown across ten iteration in figure 4.19,4.20 and 4.21 for all three feature set is similar to the previous approach, which was trained using decision tree without interpolation. The same was the case with linear regression. Essential set feature found in linear regression with interpolation was identical to elements in the linear regression model without interpolation.

4.4.2 Classification

After trying multiple combinations to train all three model using regression, another approach used was to convert between classification and regression problem. This approach is known as discretization, where the resulting target variable is a classification where the labels have an ordinal relationship. In the current data, the range of MCQ score was between 0-100, which can alternatively be used for the classification task. The scores of learners were grouped into five sets:

- 0-40 - Extremely Low
- 41-60- Moderate
- 61-80- Good
- 81-100-Optimum

The distribution of these classes was imbalanced with maximum data points in Optimum and Good group. Hence, SMOTE(Synthetic minority oversampling) was used to overcome imbalance. It oversamples the minority instances and makes it equal to the majority classes. The classification model was built on two learning approaches: Decision Tree Classification and Logistic Regression. Similar to regression, three feature set were trained:

- Feature Set 1: Knowledge and Motivation
- Feature Set 2: NASA-TLX feature set (six features)

- Feature set 3: Combination of Feature set 1 and Feature set 2

Decision Tree Classifier:

Decision Tree is used because it is easier to understand, and it can use different subset and decision rule at various stages which helps improve the predictability of the model. The data was split into 70:30 ratio like regression. A repeated random sample similar to ten-fold cross-validation was also used to generate ten iterations on a various sample of the same dataset. Again, grid search was performed to choose the most suitable hyperparameter setting for all three feature set.

Parameter	Feature Set 1	Feature Set 2	Feature Set 3
Splitting Criterion	gini	gini	gini
Max Depth	10	10	35
Maximum Leaf nodes	60	60	45
Minimum Sample leaf	15	15	20
Minimum Sample split	5	5	45
Splitter	random	random	random

Table 4.12: Hyperparameter setting using Grid Search

After tuning the model with appropriate hyperparameter train and test accuracy of the model is on average as follows:

Accuracy	Feature Set 1	Feature Set 2	Feature Set 3
Train Accuracy	45	25	29
Test Accuracy	40	18	25

Table 4.13: Train and Test Accuracy of Decision Tree Classifier

From table 4.12, we can see the accuracy average accuracy across ten iterations for the model with feature set one is the highest as the number of features are less in the model. Whereas model 2 has very low train and test accuracy, which means features used for training this model does not correctly predict as to which learner lies

in which class. The model built with feature set 3 has the highest accuracy compared to other models. One of the reason is that it has the highest number of components used during training. Also, test accuracy is close to training accuracy. In the figure, we can see how is the spread of the accuracy across all ten iterations and how stable is the model.

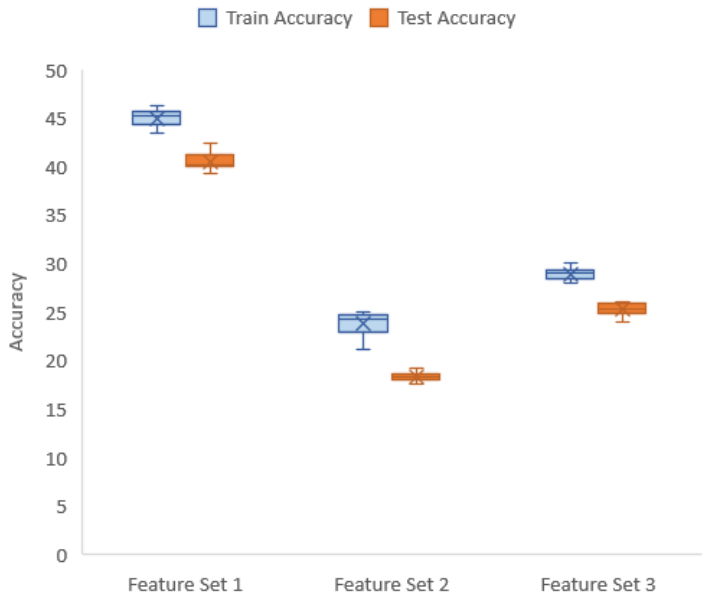

Figure 4.22: Accuracy for all feature set across 10 iteration

Decision Tree Classifier: Feature Importance - Feature Importance score computed using permutation feature importance in Decision Tree uses accuracy to generate the importance score. Feature importance score for all three feature set is shown below:

- **Feature Set 1:** In feature set 1 both for train and test set knowledge is considered to have higher importance which means shuffling which took place in knowledge feature increased the model error making it a significant variable for training the model. A similar pattern was observed in the test data. However, the standard deviation of knowledge is much higher than motivation. Motivation, on the contrary, has low importance compared to knowledge, but it has low variability making it more reliable. But, knowledge is very much consistent

in the test set with standard deviation n proper range. Therefore, in this case, knowledge has more importance over motivation.

Figure 4.23: Decision Tree: Permutation Feature Importance - Feature Set 1(Knowledge and Motivation)

- **Feature Set 2:** Effort, Mental Demand, Frustration are observed to essential features in train and test set. However, the variation in Effort and Frustration increased in the test set. Whereas, for Mental demand, the variation declined compared to the train set.

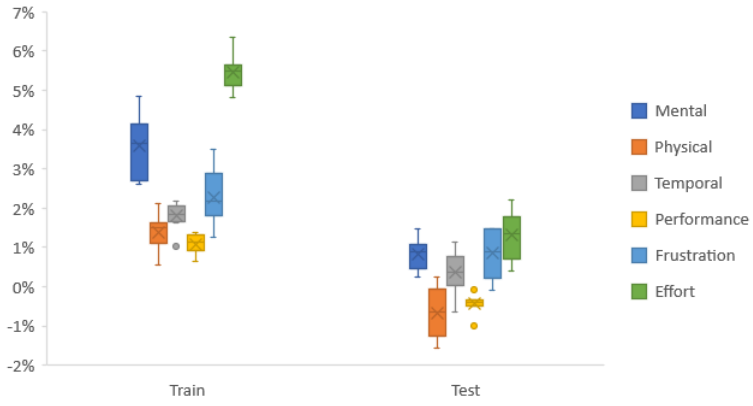

Figure 4.24: Decision Tree: Permutation Feature Importance - Feature Set 2(NASA-TLX features)

- **Feature Set 3:** Mental, Physical, Temporal and Performance had no contribu-

tion to either training the model or help predict the model. These variables were not used to construct the Decision Tree. Hence, the importance of these features is 0. However, Knowledge, Frustration, Effort and Motivation are significant features contributing to both training and test set. In the case of test set, the importance score slightly changed. Motivation became the most critical variable followed by knowledge, and the deviation of effort increase a lot in the test set.

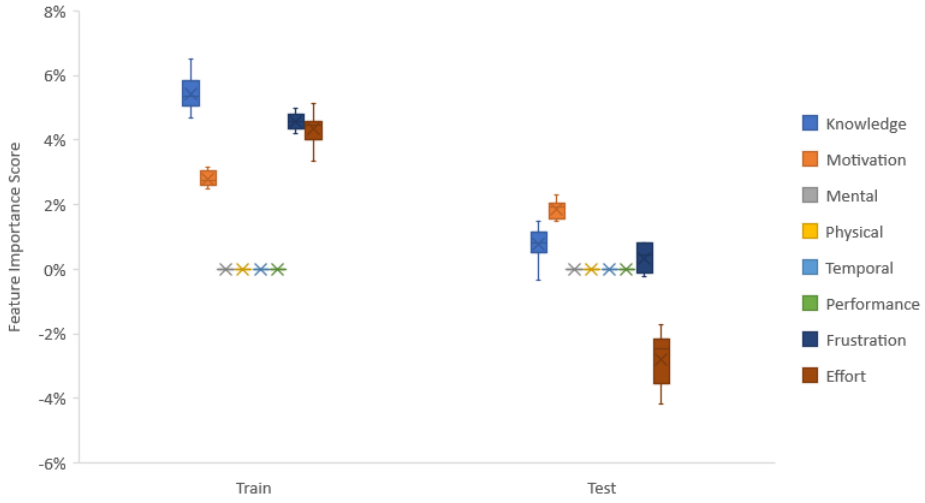

Figure 4.25: Decision Tree: Permutation Feature Importance - Feature Set 3(NASA-TLX along with Knowledge and Motivation)

Logistic Regression: As the accuracy of decision tree classifier was not good enough to ahead with Logistic Regression was also taken into consideration to see if it can capture the complexity of the data. Logistic Regression was used because it is one of the most straightforward learning algorithms which provides excellent learning efficiency. Hence, training with Logistic Regression is not computationally costly. Updating the model to reflect new data becomes convenient, which less likely in decision tree or support vector machine.

As there are multiple classes, the model is trained using multinomial logistic regression. The training algorithm uses the cross-entropy loss function in case of a multinomial problem. To reduce the loss "sag" optimizer is used as it is highly recommended for a multi-class problem (Pedregosa et al., 2011). After training the model

using Logistic Regression it was observed that the accuracy falls in the same range as Decision Tree Classifier and there was no improvement in the performance. The table below shows the average accuracy of ten iteration for all three feature set:

Accuracy	Feature Set 1	Feature Set 2	Feature Set 3
Train Accuracy	43	40	42
Test Accuracy	40	38	36

Table 4.14: Average Performance of Logistic Regression

The test accuracy declines as the number of features in the model increases. The last feature set consist of 8 features and the test accuracy is the lowest of 36 percent. However, the difference between train and test accuracy is higher in Decision Tree compared to Logistic Regression. The reason being non-linear approaches are more prone to overfitting compared to linear approaches. **Feature Importance:Logistic Regression** The important features in case of logistic regression is similar to other model trained and tested above.

Other Learning Algorithm: Apart from Decision Tree and Logistic Regression, other classifier used were Random Forest to incorporate ensemble learning, Neural Network because of it's ability to dynamically solve complex prediction problems and Support Vector Machine as its model has generalisation in practise and it less prone to overfitting. However, the accuracy of all these classification algorithm was very similar and at all time low. However, we can say the performance of SVM was much better than other approaches as SVM are good when it comes to generalising the model. Table 4.15 shows the output of other learning algorithm.

In the case of Random Forest, the trees were split using 'mse' mean square error. The parameter setting was done using a grid search. All three models were built on a different parameter setting. A similar process was performed in SVM as well. But even after hyperparameter tuning the model had low accuracy for both train and test set. The feature importance score for these models were analysed at comprehensive level as the task was to train the data with as many samples as possible to improve

the prediction of the model.

Learning Algo	Accuracy	Feat Set 1	Feat Set 2	Feat Set 3
Random Forest	Train	35	40	39
	Test	27	34	32
Neural Network	Train	37	40	45
	Test	31	32	30
Support Vector Machine	Train	40	45	50
	Test	31	34	42

Table 4.15: Accuracy of other learning algorithm

4.4.3 Clustering

As the main aim was to find the features which highly influences the performance of the learners and as regression or classification is unable to predict the performance correctly. Another approach that can help achieve this is unsupervised learning approach. Clustering comes under unsupervised learning. It assists in clustering data points to groups called clusters. Clustering can help us find out the group of learners having similar score have which mental workload attribute in common. The profiling variable used to form clusters was MCQ Scores. K-means clustering can be used for this process as it is computationally faster, and it goes on producing more robust results than other types of clustering. Kmeans require the number of clusters as one of the parameters. However, getting to know the optimal number of a cluster needs a granular level of clustering information. Nevertheless, this is possible using the elbow method, which finds the ideal number of clusters. Elbow method plots the explained variation as a function to several clusters and choosing the elbow of the curve to finalise the number of clusters. Each observation belongs to the cluster nearest to its mean.

Just like regression and classification, clustering model was also built using three feature set. It was observed that as the number of features increases in the model,

the elbow plot becomes smooth, and it becomes difficult to determine the elbow in the curve. The major problem was that the algorithm is not able to separate the data into clusters. clusters were made using the different feature set which excludes the target variable and consist of Knowledge, Motivation and NASA-TLX features, and profiling was done using the MCQ score. Profiling variable helps in identifying the behaviour of each cluster. However, it was witnessed that for different clusters, there was no pattern observed in the MCQ score. One of the possible reason can be most of the MCQ Score value is scattered around the same range. In other words, there exist slight non-gaussian distribution in MCQ score. K-means clustering is sensitive to a slight imbalance in the data. Hence, K-means clustering was not able to obtain familiar mental workload attributes based on learners performance. The distribution of MCQ score in each cluster is shown below:

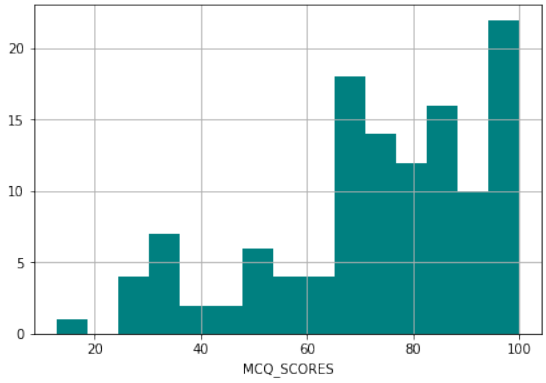

Figure 4.26: Cluster 1 built using all six NASA-TLX Features along with Knowledge and Motivation

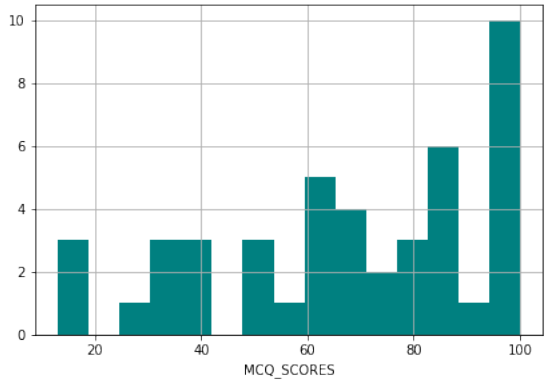

Figure 4.27: Cluster 2 using all six NASA-TLX Features along with Knowledge and Motivation

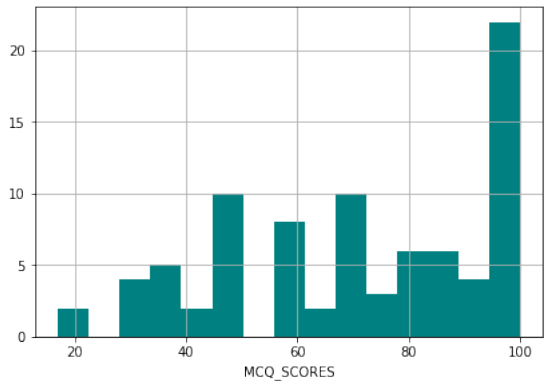

Figure 4.28: Cluster 3 using all six NASA-TLX Features along with Knowledge and Motivation

4.5 Evaluation of Result

The aim of performing these diverse experiments was not only to find out the most critical variables which create an impact on learners class test performance but also find out the characteristic of learners in both the groups by using NASA-TLX attributes. The model initially selected was linear regression and decision tree regression. However, even after tweaking with the hyperparameters settings such as the depth of the tree, the criteria used to split the tree,sample leaf node required,types of splitter etc. there was no improvement found in decision tree regression. The evaluation metric

used was RMSE score which is a difference between the predicted value received by the model and actual value; this is also known as residual. The best way to determine whether a particular RMSE score is acceptable or not; it can get compared with the scale of the target variable. In this case, the scale of the target variable was between 0-100, and the RMSE score was between 19-25. Hence, it can be stated that the model is experiencing a high bias.

The research was further extended to various learning approaches which were not a part of the design framework. However, multiple experiments were performed to find a suitable fit for the data. Regression was later implemented by incorporating linear interpolation. Interpolation creates new data points within the range of the existing data. Linear 1D and Spline interpolation were implemented. However, as the model was only trained on the interpolated data and tested on the original data, a sudden dip in train error and rise in test error was noticed in decision tree regression as it is prone to overfitting. After tuning the hyperparameter, both the train and test error came in the same range with the same error as it was before interpolation.

Switching to a classification problem was considered, to approach more straightforward and more interpretable model. This process of switching from regression to classification is known as discretization. Decision Tree classifier, Logistic Regression, Random Forest, Neural Network and Support Vector machine are the various classifiers applied to the data. However, the accuracy of all these model was even less than 50%. Yet, a linear regression and decision tree regression model was built using a single class. In other words, a separate model was built for learners having extremely high values, moderate values and optimum values. This model had a very low error with an RMSE score between 6 and 7. However, this is possible as the range of prediction is now made limited.

The feature importance score of each model was measured using the permutation feature importance. Every model had a few set of features standard in both the control group and the experimental group. This aided in determining the characteristic of learners in both the group. But, the feature importance score for a few of the features had high variance.

An independent t-test was performed to find the best fit model out of the multiple models trained. However, it was noticed that there is no statistically significant difference between the RMSE score of linear regression(M=22.3, SD=0.83) and decision tree regression(M=22.6, SD=0.92),(t(209)=-0.645,p=0.26) for the control group and the experimental group, linear regression(M=21.3,SD=0.36), decision tree regression(M=21.7, SD=0.4),(t(197)=-1.97,p=0.09). This infers that the performance of all the models is similar and stable without any variance. The variance in feature importance score improves gradually from one learning approach to another.

The essential features for the control group common across different feature set are knowledge, mental demand, physical demand and effort. Important features shared across all learning algorithm for the experimental group are motivation, mental demand, frustration and temporal demand.

Unfortunately, looking into the performance of the model, there is a shred of substantial evidence that these features are not sufficient to carry the prediction of the MCQ score of the learners.

4.5.1 Strengths and Limitation:

Strengths:

The existing framework has been updated a lot during the model training phase making it a compact model which can be reusable with any data falling within the mental workload domain.

The model is capable of understanding various characteristics of learners in both control and the experimental group and the key attributes which impact the mental workload, which further hampers the performance of the learners with a condition of providing the right feature set.

The use of permutation feature importance supports mache learning whose primary aim is automated training and testing results which can be easily scaled across multiple features.

At every evaluation step, the stability and consistency of the model are checked regularly by evaluating ten iterations of the random sample.

The metrics used, which is RMSE score and accuracy for the regression and the classification task, respectively is broadly acceptable and candid about the results enabling the establishment of a good model.

The final strength is the identification of the core factors responsible for the mental workload in the third level of education.

Limitations of the results:

The model building takes place on two sets of data the control group(N=20) and the experimental group(N=197). However, training a model to achieve generalisation is not possible with such a limited set of data.

The features used to train the model NASA-TLX score with previous knowledge and motivation to predict the model are not enough to perform the prediction of MCQ scores.

The feature importance method called the permutation feature importance is compatible with all kinds of the supervised learning algorithm. However, the scale of this method is not fixed, and it can range from positive to negative infinity. Hence, there is no way that any two variables can be compared on the same scale using permutation feature importance.

Chapter 5

Conclusion

This chapter sums the thesis, highlighting the main structure and key findings. It outlines the work which requires to be done along with presenting a fair path towards future work in the research of mental workload within learners.

5.1 Research Overview

This thesis started with an aim to explore the existing state of the art theories about measuring, defining and describing mental workload. The research initially focused on common factors such as stress and anxiety, that cause an increase in the mental workload of learners in third-level education. It further throws light on various concepts surrounding mental workload such as cognitive load theory, working memory, collective working memory, instructional design and ways to measure mental workload. The field of educational psychology lacks ways to measure the cognitive load of the learning task. This encourages the aim of this study which is to investigate various mental workload attribute, and discover which feature influences the performance of the learners in a masters classroom.

The data in question was collected from the university classroom of students pursuing their masters and PhD. The dataset comprises of data from 20 lectures along with learners performance in-class test and the amount of mental workload each student had to undergo to finish the task. This data was initially used in the research, which

compares the learning efficiency between traditional teaching method extended with collaborative group activity and traditional direct instructional teaching technique alone. The collaborative group activity was based on the social constructivist theory, which gives learners a fair opportunity to communicate along with allowing them to open their mind gates to grasp additional information from each other. According to the literature review, Cognitive Load Theory supports the traditional teaching method and assumes any alteration with the direct instructional method will eventually fail. This was supported by a theory where collaboration among learners can also cause MWL due to too much of communication.

During the data collection, the classroom was divided into two groups: the control group and the experimental group. Control group comprises of individual learners giving the test. In contrast, the experimental group consisted of learners who gave test after the inquiry-based group activity where they discussed cognitive trigger questions followed by MCQ test. The performance of the students was tracked using the MCQ test. The MCQ score of each student along with NASA-TLX test output later used to find out the most critical mental workload attribute, which possibly impacts their performance in the MCQ test.

The initial design had Linear Regression and Decision Tree regression in the plan to train the model to predict the MCQ score of the learner and check which attribute is majorly contributing towards the prediction. Permutation feature importance algorithm was used to compute the feature importance score as it is handy across different learning algorithm. However, the model was not able to generalise the data, and there were not enough number of features required to predict the MCQ score. Various other machine learning approaches such as decision tree classifier, logistic regression, support vector machine, neural network, random forest and k-means clustering were also applied. Hyperparameters were tuned using a grid search to ensure relevant values to the parameter settings. The feature importance score for the different machine learning approaches had several features in common. In the case of the control group features such as knowledge, mental demand, physical demand and effort were observed to be critical features. Whereas in the case of the experimental group motivation, mental,

temporal, performance and frustration were witnessed as essential features contributing towards the performance of the model on both seen and unseen data. However, due to high error and low accuracy experienced in all models which were trained, we can state that the model did not generalise considerably on this combination of features.

5.2 Problem Definition

Consumption of a plethora of information leads to cognitive overwhelm, which in return causes mental workload. This issue is prominent in student's life as they have to overboard themselves with too much direct information which many times requires more cognitive resources. A piece of new information is always stored in the working memory, which is also known as short term memory. However, to ensure this information gets transferred to the long term memory, an active amount of rehearsing should be done with the information.

Cognitive Load Theory(CLT) is in charge of keeping track of the information working memory holds at any given point. CLT assumes that working memory can have a grip on direct, explicit instruction. Any alteration done to this traditional teaching method will hamper the learning process. However, the concept of social constructivism states that collaborative learning improves the learning power in learners. On the contrary, a few pieces of literature also says that communication during these group activities drastically rises the mental workload among learners. Hence, it is essential to construct a machine learning models one for the control group and other for the experimental group. The former received only the direct instruction, whereas the latter also participated in inquiry-based activity associated with collaborative group exercise where the learners discuss cognitive trigger questions. The MCQ test result tracked the performance of both these groups. Mental workload while giving the test was recorded using the self-assessment test, which is the NASA-TLX test. Building a machine learning model will help determine the influence of mental workload attribute on each group; this will also help to define the characteristic of each group. However, there are high chances that there exist high bias in the learner's response while fill-

ing the NASA-TLX questionnaire. Also, subscale weighting is very time consuming, repetitive and arduous this might lead to similar rating applied to each subscale. The mental workload can also be very subjective from individual to individual. In other words, the extreme workload can deteriorate an individuals performance, or some people enjoy experiencing a high workload; this pressure helps them perform even better.

5.3 Design/Experimentation, Evaluation & Results

The initial design consisted of data understanding and the pre-processing phase where imputing missing values, removing outliers and performing fundamental exploratory analysis to check the distribution of all the features was taken into account. The next step was the modelling phase in which the data was randomly sampled for ten iterations using monte Carlo sampling to check the consistency of the model across all iterations. The sample was split into 70:30 ratio into train and test set. As the target variable is MCQ score, and it is continuous hence decision tree regression, and linear regression is used to train the model. The evaluation metrics used was RMSE score which will be computed for both train and test set. The study took place between two groups the control and experimental group and three different feature sets:

- Feature Set 1: Knowledge and Motivation
- Feature Set 2: Mental Demand, Physical Demand, Temporal Demand, Performance, Frustration, Effort
- Feature Set 3: Feature Set 1+Feature Set 2

A total of six models were created one for the control group using these features set and one for the experimental group using the same feature set. The evaluation metric used was RMSE score for regression approach. However, the error in all the six models was very high in comparison to the scaling of the target feature. To improve the performance of the model various other machine learning approaches were also used such as Decision Tree classifier as it easy to understand and it becomes convenient to obtain a better result by tweaking with the parameter settings of Decision

Tree. For the classification problem, the data was divided into five classes: extremely low, low, moderate, optimum and good. Logistic regression was applied to efficiently model the non-linear data in a linear way using log transformation. Support Vector Machine followed by Neural Networks and Random forest was applied. The data was interpolated with training done on this data, and testing was done on the original data. However, none of the learning algorithms was fruitful with all-time low RMSE and accuracy. An unsupervised learning approach such as clustering using K-means was also put to use with clusters created using the features set mentioned above, and profiling was done using MCQ score. Unfortunately, there was no clear pattern visible between each cluster. The optimal number of the cluster was selected using the knee plot method. The feature importance of each model was calculated using permutation feature importance score. The results highlight some points below:

The RMSE score of Decision Tree Regression and Linear Regression had no statistical difference with a p-value greater than 0.05 at 95 percent confidence interval. Hence, we can say RMSE score of both the model is equally high.

For classification model, the model built using the feature set 3, which is a combination of NASA-TLX, Knowledge and Motivation has better train and test accuracy compared to other feature sets. However, it is still shallow. Out of all the classification approaches, Support Vector Machine performed slightly better with 50 percent train accuracy and 42 percent test accuracy though the accuracy of other classifier was not too far.

After applying interpolation, the train RMSE and Accuracy improved, but there was no improvement observed in test RMSE and Accuracy.

After performing various experiments, an observation was made that there exist a typical pattern in the feature importance score across models. However, the high model error and low accuracy suggests that the number of features is not sufficient to predict the MCQ score.

5.4 Contributions and impact

The research has been able to incorporate the concept of mental workload in the educational setting, which is rare in the field of psychology. This approach helps find out the critical elements responsible for mental workload in the learners. A significant contribution was the presentation of methodology, which consist of a framework which can adopt more mental workload feature in future, to represent the mental workload attributes influencing the learner's performance in an MCQ test conducted in the classroom. Once the research is over, the same framework can be replicated with a new and an extended set of features. This study is based on the optimized use of cognitive resources. It gives researchers a new direction to measure the cognitive load of the learning process. The concept of permutation feature importance score is adapted to provides us with the list of features which contributes to the growth or decline of learners performance.

This study majorly contributes to find out the substantial mental workload attribute which in this case is NASA-TLX feature which ultimately impacts the learners performance. This also helps in determining the causes of mental workload in the control group and the experimental group. In addition to the major contribution, this thesis also makes several minor contributions. Firstly, it considers the optimization of cognitive resources through the concept of collective working memory. This means the learners share working memory while working on the same task. The assumption is to use working memory of multiple people, ultimately reducing the cognitive cost. Secondly, apart from the traditional raw NASA-TLX features, it also incorporates elements such as knowledge and motivation before the MCQ test, which gives a more detailed picture while noting the mental workload characteristic of each group. Thirdly, it throws a light on concepts like inquiry based method and instructional design though literature survey explaining how these concept are important part of human cognitive architecture.

5.5 Future Work & recommendations

The solution proposed can be improvised in many different ways. Firstly, we observed based on the evidence in the previous section that a model with only eight features is not sufficient to predict the performance of the learners. Hence, the immediate next steps would be to incorporate more elements to improve the predictability of the model. This can also lead to a better understanding of the complicated and captivating construct of mental workload and go a step closer to the goal of building a highly generalisable model. Secondly, to conduct multiple subjective assessment test together, which will include a mix of the unidimensional and multi-dimensional questionnaire. At the design side finding ways to calculate feature importance which generates feature importance score of all features on the same scale. More experiments can be performed on ensemble models with two different learning algorithm. In other words, the probability of one model can be passed as a feature set to the next model, which can also be said as a mix of two different machine learning model.

A step by step approach can be initiated towards a systematic quality check of the data towards the end of the data collection task. Secondly, to focus on composing a mild collaborative activity to relieve the extra cost overheads due to communication among learners along with measuring the cognitive load, that the learners experience during the collaborative activity. Incorporating visual aids along with direct instructional technique can lead to an increase in the germane load and decline in the extraneous load among learners. Occasionally, switching to real-time mental workload measurement methods which allow tracking mental demands at real-time using eye trackers while learners are working in laboratories. This will help track their workload in an ongoing activity.

References

Aherne, D. (2001). Understanding student stress: A qualitative approach. *The Irish Journal of Psychology*, *22*(3-4), 176–187.

Alsuraykh, N. H., Wilson, M. L., Tennent, P., & Sharples, S. (2019). How stress and mental workload are connected. In *Proceedings of the 13th eai international conference on pervasive computing technologies for healthcare* (pp. 371–376).

Atkinson, R. C., & Shiffrin, R. M. (1971). The control of short-term memory. *Scientific american*, *225*(2), 82–91.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Bustamante, E. A., & Spain, R. D. (2008). Measurement invariance of the nasa tlx. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 52, pp. 1522–1526).

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

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

Cheon, J., & Grant, M. M. (2012). The effects of metaphorical interface on germane cognitive load in web-based instruction. *Educational Technology Research and Development*, *60*(3), 399–420.

Chi, M. T., Glaser, R., & Rees, E. (1981). *Expertise in problem solving.* (Tech. Rep.). Pittsburgh Univ PA Learning Research and Development Center.

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.

Corden, R. (2001). Group discussion and the importance of a shared perspective: Learning from collaborative research. *Qualitative Research*, *1*(3), 347–367.

Council, N. R., et al. (1993). *Workload transition: Implications for individual and team performance*. National Academies Press.

da Silva, F. P. (2014). Mental workload, task demand and driving performance: What relation. *Procedia-Social and Behavioral Sciences*, *162*, 310–319.

Dawes, S. S., Cresswell, A. M., & Pardo, T. A. (2009). From "need to know" to "need to share": Tangled problems, information boundaries, and the building of public sector knowledge networks. *Public Administration Review*, *69*(3), 392–402.

De Jong, T. (2010). Cognitive load theory, educational research, and instructional design: some food for thought. *Instructional science*, *38*(2), 105–134.

De Koning, B. B., Tabbers, H. K., Rikers, R. M., & Paas, F. (2007). Attention cueing as a means to enhance learning from an animation. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *21*(6), 731–746.

Dong, Y., & Peng, C.-Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, *2*(1), 222.

Fan, J., & Smith, A. P. (2018). Mental workload and other causes of different types of fatigue in rail staff. In *International symposium on human mental workload: Models and applications* (pp. 147–159).

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, *20*(177), 1–81.

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

Geary, D. C. (2012). Evolutionary educational psychology. In *Apa educational psychology handbook, vol 1: Theories, constructs, and critical issues.* (pp. 597–621). American Psychological Association.

Gerjets, P., Scheiter, K., & Catrambone, R. (2004). Designing instructional examples to reduce intrinsic cognitive load: Molar versus modular presentation of solution procedures. *Instructional Science*, *32*(1-2), 33–58.

Gingerich, A., & Yeates, P. (2019). The mental workload of conducting research in assessor cognition. *Perspectives on medical education*, *8*(6), 315–316.

Goldstein, E. (2011). *Cognitive psychology: Connecting mind, research, and everyday experience*. Wadsworth Cengage Learning. Retrieved from `https://books.google.ie/books?id=wIbBQwAACAAJ`

Gopher, D., & Donchin, E. (1986). Workload: An examination of the concept.

Gopher, D., & Kimchi, R. (1989). Engineering psychology. *Annual Review of Psychology*, *40*(1), 431–455.

Hancock, P. A. (2017). Whither workload? mapping a path for its future development. In *International symposium on human mental workload: Models and applications* (pp. 3–17).

Hancock, P. A., & Meshkati, N. (1988). *Human mental workload*. North-Holland Amsterdam.

Hancock, P. A., Meshkati, N., & Robertson, M. (1985). Physiological reflections of mental workload. *Aviation, space, and environmental medicine*.

Hart, S. G. (2006). Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 50, pp. 904–908).

Hart, S. G., & Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139–183). Elsevier.

Iqbal, S. T., Zheng, X. S., & Bailey, B. P. (2004). Task-evoked pupillary response to mental workload in human-computer interaction. In *Chi'04 extended abstracts on human factors in computing systems* (pp. 1477–1480).

Järvenpää, E. (1986). Mental workload in cad-work: Computer aided design of printed circuit boards as an example. *Acta Psychologica Fennica*.

Jonassen, D. (2009). Reconciling a human cognitive architecture. In *Constructivist instruction* (pp. 25–45). Routledge.

Jung, J., Kim, D., & Na, C. (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.

Kester, L., Lehnen, C., Van Gerven, P. W., & Kirschner, P. A. (2006). Just-in-time, schematic supportive information presentation during cognitive skill acquisition. *Computers in Human Behavior*, *22*(1), 93–112.

Kirschner, F., Paas, F., & Kirschner, P. A. (2009). A cognitive load approach to collaborative learning: United brains for complex tasks. *Educational psychology review*, *21*(1), 31–42.

Kuhn, M., Johnson, K., et al. (2013). *Applied predictive modeling* (Vol. 26). Springer.

Longo, L. (2012). Formalising human mental workload as non-monotonic concept for adaptive and personalised web-design. In *International conference on user modeling, adaptation, and personalization* (pp. 369–373).

Longo, L. (2015a). A defeasible reasoning framework for human mental workload representation and assessment. *Behaviour & Information Technology*, *34*(8), 758–786.

Longo, L. (2015b). Designing medical interactive systems via assessment of human mental workload. In *Computer-based medical systems (cbms), 2015 ieee 28th international symposium on* (pp. 364–365).

Longo, L. (2016). Mental workload in medicine: foundations, applications, open problems, challenges and future perspectives. In *2016 ieee 29th international symposium on computer-based medical systems (cbms)* (pp. 106–111).

Longo, L. (2017). Subjective usability, mental workload assessments and their impact on objective human performance. In *Ifip conference on human-computer interaction* (pp. 202–223).

Longo, L. (2018a). Experienced mental workload, perception of usability, their interaction and impact on task performance. *PloS one*, *13*(8), e0199661.

Longo, L. (2018b). 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. In *Proceedings of the 10th international conference on computer supported education, CSEDU 2018, funchal, madeira, portugal, march 15-17, 2018, volume 2.* (pp. 166–178). doi: 10.5220/0006801801660178

Longo, L., & Barrett, S. (2010). Cognitive effort for multi-agent systems. In *International conference on brain informatics* (pp. 55–66).

Longo, L., & Dondio, P. (2015). On the relationship between perception of usability and subjective mental workload of web interfaces. In *Web intelligence and intelligent agent technology (wi-iat), 2015 ieee/wic/acm international conference on* (Vol. 1, pp. 345–352).

Longo, L., & Orru, G. (2018). An evaluation of the reliability, validity and sensitivity of three human mental workload measures under different instructional conditions in third-level education. In *International conference on computer supported education* (pp. 384–413).

Mayer, R. E. (2005). Cognitive theory of multimedia learning. *The Cambridge handbook of multimedia learning*, *41*, 31–48.

Mayer, R. E., Heiser, J., & Lonn, S. (2001). Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *Journal of educational psychology*, *93*(1), 187.

Meshkati, N., Hancock, P. A., Rahimi, M., & Dawes, S. M. (1995). Techniques in mental workload assessment.

Moray, N. (2013). *Mental workload: Its theory and measurement* (Vol. 8). Springer Science & Business Media.

Moustafa, K., & Longo, L. (2018). Analysing the impact of machine learning to model subjective mental workload: a case study in third-level education. In *International symposium on human mental workload: Models and applications* (pp. 92–111).

Nachreiner, F. (1995). Standards for ergonomics principles relating to the design of work systems and to mental workload. *Applied Ergonomics*, *26*(4), 259–263.

Orru, G., Gobbo, F., O'Sullivan, D., Longo, L., et al. (2018). An investigation of the impact of a social constructivist teaching approach, based on trigger questions, through measures of mental workload and efficiency. In *Csedu (2)* (pp. 292–302).

Orru, G., & Longo, L. (2018). The evolution of cognitive load theory and the measurement of its intrinsic, extraneous and germane loads: a review. In *International symposium on human mental workload: Models and applications* (pp. 23–48).

Orru, G., & Longo, L. (2019). Direct instruction and its extension with a community of inquiry: A comparison of mental workload, performance and efficiency. In *Csedu (1)* (pp. 436–444).

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational psychologist*, *38*(1), 1–4.

Paas, F., & Sweller, J. (2012). An evolutionary upgrade of cognitive load theory: Using the human motor system and collaboration to support the learning of complex cognitive tasks. *Educational Psychology Review*, *24*(1), 27–45.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Prabaswari, A., Hamid, A., & Purnomo, H. (2020). The mental workload analysis of gojek drivers. In *Iop conference series: Materials science and engineering* (Vol. 722, p. 012008).

Reid, G. B., & Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. In *Advances in psychology* (Vol. 52, pp. 185–218). Elsevier.

Reznitskaya, A., Anderson, R. C., & Kuo, L.-J. (2007). Teaching and learning argumentation. *The Elementary School Journal*, *107*(5), 449–472.

Rizzo, L., Dondio, P., Delany, S. J., & Longo, L. (2016). Modeling mental workload via rule-based expert system: a comparison with nasa-tlx and workload profile. In *Ifip international conference on artificial intelligence applications and innovations* (pp. 215–229).

Rizzo, L., & Longo, L. (n.d.). Inferential models of mental workload with defeasible argumentation and non-monotonic fuzzy reasoning: a comparative study..

Rizzo, L. M., & Longo, L. (2017). Representing and inferring mental workload via defeasible reasoning: a comparison with the nasa task load index and the workload profile.

Romero, J. F. (2017). An investigation of the correlation between mental workload and web user's interaction.

Rubio, S., Díaz, E., Martín, 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.

Rusnock, C. F., & Borghetti, B. J. (2018). Workload profiles: A continuous measure of mental workload. *International Journal of Industrial Ergonomics*, *63*, 49–64.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive science*, *12*(2), 257–285.

Sweller, J. (2009). What human cognitive architecture tells us about constructivism. In *Constructivist instruction* (pp. 139–155). Routledge.

Sweller, J. (2010). Cognitive load theory: Recent theoretical advances.

Sweller, J. (2011). Cognitive load theory. In *Psychology of learning and motivation* (Vol. 55, pp. 37–76). Elsevier.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). Measuring cognitive load. In *Cognitive load theory* (pp. 71–85). Springer.

Sweller, J., van Merriënboer, J. J., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, 1–32.

Sweller, J., Van Merrienboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational psychology review*, *10*(3), 251–296.

Tracy, J. P., & Albers, M. J. (2006). Measuring cognitive load to test the usability of web sites. In *Annual conference-society for technical communication* (Vol. 53, p. 256).

Tsang, P. S., & Vidulich, M. A. (2006). Mental workload and situation awareness.

Tungare, M., & Pérez-Quiñones, M. A. (2009). Mental workload in multi-device personal information management. In *Chi'09 extended abstracts on human factors in computing systems* (pp. 3431–3436).

Verwey, W. B., & Veltman, H. A. (1996). Detecting short periods of elevated workload: A comparison of nine workload assessment techniques. *Journal of experimental psychology: Applied*, *2*(3), 270.

Vygotsky, L. S. (1980). *Mind in society: The development of higher psychological processes*. Harvard university press.

Ward, R. D., & Marsden, P. H. (2003). Physiological responses to different web page designs. *International Journal of Human-Computer Studies*, *59*(1-2), 199–212.

Wästlund, E., Norlander, T., & Archer, T. (2008). The effect of page layout on mental workload: A dual-task experiment. *Computers in human behavior*, *24*(3), 1229–1245.

Weber, K., Maher, C., Powell, A., & Lee, H. S. (2008). Learning opportunities from group discussions: Warrants become the objects of debate. *Educational Studies in Mathematics*, *68*(3), 247–261.

Widyanti, A., Johnson, A., & de Waard, D. (2013). Adaptation of the rating scale mental effort (rsme) for use in indonesia. *International Journal of Industrial Ergonomics*, *43*(1), 70–76.

Wierwille, W. W., & Eggemeier, F. T. (1993). Recommendations for mental workload measurement in a test and evaluation environment. *Human factors*, *35*(2), 263–281.

Wilson, G. F., & O'Donnell, R. D. (1988). Measurement of operator workload with the neuropsychological workload test battery. In *Advances in psychology* (Vol. 52, pp. 63–100). Elsevier.

Wu, C., & Liu, Y. (2006). Queuing network modeling of age differences in driver mental workload and performance. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 50, pp. 190–194).

Xie, B., & Salvendy, G. (2000). Review and reappraisal of modelling and predicting mental workload in single-and multi-task environments. *Work & stress*, *14*(1), 74–99.

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), e0183884.

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). State of science: mental workload in ergonomics. *Ergonomics*, *58*(1), 1–17.