



Technological University Dublin  
ARROW@TU Dublin

---

Dissertations

School of Computing

---

2017-1

## A Comparison of Supervised Machine Learning Classification Techniques and Theory-Driven Approaches for the Prediction of Subjective Mental Workload

Dmitrii Gmyzin  
*Technological University Dublin*

Follow this and additional works at: <https://arrow.tudublin.ie/scschcomdis>

 Part of the [Computer Engineering Commons](#)

---

### Recommended Citation

Gmyzin, D. (2017) *A Comparison of Supervised Machine Learning Classification Techniques and Theory-Driven Approaches for the Prediction of Subjective Mental Workload*. Masters dissertation, Technological University Dublin, 2017. doi:10.21427/D7533X

This Dissertation is brought to you for free and open access by the School of Computing at ARROW@TU Dublin. It has been accepted for inclusion in Dissertations by an authorized administrator of ARROW@TU Dublin. For more information, please contact [yvonne.desmond@tudublin.ie](mailto:yvonne.desmond@tudublin.ie), [arrow.admin@tudublin.ie](mailto:arrow.admin@tudublin.ie), [brian.widdis@tudublin.ie](mailto:brian.widdis@tudublin.ie).



This work is licensed under a [Creative Commons Attribution-Noncommercial-Share Alike 3.0 License](#)



**A comparison of supervised  
machine learning classification  
techniques and theory-driven  
approaches for the prediction of  
subjective mental workload**



**Dmitrii Gmyzin**

A dissertation submitted in partial fulfilment of the requirements of

Dublin Institute of Technology for the degree of

M.Sc. in Computing (Stream)

**January 2016**

# Declaration

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Stream), 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 Institutes guidelines for ethics in research.

*Signed:*

*Date: January, 2017*

# Abstract

In the modern world of technological progress, systems and interfaces are becoming more and more complex. As a consequence, it is a crucial to design the human-computer interaction in the most optimal way to improve the user experience. The construct of Mental Workload is a valid metric that can be used for such a goal. Among the different ways of measuring Mental Workload, self-reporting procedures are the most adopted for their ease of use and application.

This research is focused on the application of Machine Learning as an alternative to theory-driven approaches for Mental Workload measurement. In particular, the study is aimed at comparing the classification accuracy of a set of induced models, from an existing dataset, to the mental workload indexes generated by well-known subjective mental workload assessment techniques - namely the Nasa Task Load Index and the Workload profile instruments.

**Keywords:** Subjective Mental Workload, Supervised Machine Learning, NASA Task Load Index, Workload Profile, Validity

# Acknowledgments

Firstly, I would like to thank the Professors of Dublin Institute of Technology which gave me invaluable knowledge by providing well structured and state-of-the-art material in Data Analytics.

Specifically, I would like to highlight the support of Dr. Luca Longo. Without his guide it would be impossible to perform this study. He was able to confidentially lead me through all the issues that I have encountered during this dissertation. Additionally, I would like to thank Luca for providing the data used in this research.

# Contents

<b>Declaration</b>	<b>1</b>
<b>Abstract</b>	<b>2</b>
<b>Acknowledgments</b>	<b>3</b>
<b>Contents</b>	<b>4</b>
<b>List of Figures</b>	<b>8</b>
<b>List of Tables</b>	<b>10</b>
<b>1 Introduction</b>	<b>11</b>
1.1 Background . . . . .	11
1.2 Research Question and Research Hypotheses . . . . .	12
1.3 Research Methodologies . . . . .	13
1.4 Scope and Limitations . . . . .	13
1.5 Document Outline . . . . .	13
<b>2 Literature review</b>	<b>15</b>
2.1 Introduction . . . . .	15
2.2 The concept of Mental Workload . . . . .	17
2.2.1 Workload as a multidimensional notion . . . . .	17
2.2.2 Measurements of Mental Workload . . . . .	18
Performance measures . . . . .	18

	Physiological measures . . . . .	19
	Subjective measures . . . . .	19
2.3	Mental Workload subjective assessment techniques . . . . .	19
2.3.1	Introduction . . . . .	19
2.3.2	NASA-TLX overview . . . . .	20
2.3.3	NASA-TLX scale . . . . .	21
2.3.4	Workload profile overview . . . . .	22
2.3.5	Workload profile scale . . . . .	22
2.3.6	Criteria for the evaluation of Mental Workload models . . . . .	23
2.3.7	Face Validity . . . . .	25
2.4	Supervise Machine Learning . . . . .	26
2.4.1	KDD Process . . . . .	26
2.4.2	Data mining approach . . . . .	27
2.4.3	Predictive models . . . . .	30
2.4.4	Supervised Machine Learning techniques . . . . .	31
	Artificial Neural Networks . . . . .	32
	Support Vector Machines . . . . .	34
	Gradient boosting . . . . .	35
	Naive Bayes classifiers . . . . .	36
	K-nearest neighbor algorithm . . . . .	36
2.4.5	Summary and Implications . . . . .	37
2.4.6	Gaps, motivation and limitation . . . . .	38
	Gaps . . . . .	38
	Motivation . . . . .	39
	Limitations . . . . .	39
<b>3</b>	<b>Experiment design</b>	<b>40</b>
3.1	Introduction . . . . .	40
3.2	Software . . . . .	42
3.3	Data Quality investigation . . . . .	43

## CONTENTS

---

3.3.1	Database description . . . . .	43
3.3.2	Data Selection . . . . .	45
3.4	Data Preparation . . . . .	47
3.4.1	Introduction . . . . .	47
3.4.2	Data Transformation . . . . .	48
3.4.3	Missing Values Handling . . . . .	48
3.4.4	Outliers and incompatible values handling . . . . .	49
3.5	Descriptive Statistics . . . . .	51
3.6	Supervise Machine Learning Model Training . . . . .	52
3.6.1	Modelling techniques selection . . . . .	52
3.6.2	Test Design solutions and Model assessment . . . . .	52
3.7	Evaluation and model Adjustment . . . . .	53
3.7.1	Measurement of Face Validity . . . . .	53
3.7.2	Accepting / Rejecting hypothesis . . . . .	53
3.7.3	Strength and limitation of approach taken . . . . .	54
<b>4</b>	<b>Experiment implementation</b>	<b>55</b>
4.1	Introduction . . . . .	55
4.2	Data Investigation . . . . .	58
4.2.1	Data Selection . . . . .	58
4.2.2	Data quality estimation . . . . .	61
4.3	Data Preparation . . . . .	63
4.3.1	Outliers and incompatible values handling . . . . .	63
4.3.2	Missing values handling . . . . .	64
4.3.3	Data Construction . . . . .	68
	Region . . . . .	69
	Daytime . . . . .	69
	NASA-TLX weight . . . . .	71
	Test completion duration . . . . .	71
4.4	SML Modeling Training, Evaluation & Adjustment . . . . .	71



## CONTENTS

---

4.4.1	Selection of SML Classifiers . . . . .	71
4.4.2	Model Building, adjustment & assesment . . . . .	73
4.4.3	Model Building with additional input attributes . . . . .	76
<b>5</b>	<b>Evaluation</b>	<b>78</b>
5.1	Introduction . . . . .	78
5.2	Calculation of Workload NASA-TLX and WP equations . . . . .	79
5.2.1	Strength and limitation of the experiment . . . . .	82
<b>6</b>	<b>Conclusion</b>	<b>83</b>
6.1	Research Overview . . . . .	83
6.2	Problem Definition . . . . .	84
6.3	Design/Experimentation, Evaluation & Results . . . . .	84
6.4	Contributions and impact . . . . .	85
6.5	Future Work & recommendations . . . . .	85
	<b>References</b>	<b>86</b>
<b>A</b>	<b>NASA-TLX questionnaire</b>	<b>92</b>
<b>B</b>	<b>NASA-TLX SQL query</b>	<b>93</b>
<b>C</b>	<b>WP SQL query</b>	<b>97</b>

# List of Figures

1.1	Disadvantages associated with low/high MWL and advantages of optimal workload . . . . .	12
2.1	Outline of the Literature Review Chapter . . . . .	16
2.2	Disadvantages associated with low/high MWL and advantages of optimal workload . . . . .	23
2.3	Overview of the Steps That Compose the KDD Process . . . . .	27
2.4	Overview of CRISP-DM . . . . .	28
2.5	Phases of the CRISP-DM reference model . . . . .	29
2.6	Comparison of Supervised Machine Learning techniques . . . . .	32
2.7	Neural network with three levels of layers . . . . .	34
2.8	Approaches to define the distance between instances . . . . .	37
3.1	Outline of the Research Design Chapter . . . . .	41
3.2	Database Schema . . . . .	45
4.1	Graphical representation of an experiment . . . . .	56
4.2	Outline of the Research Implementation Chapter . . . . .	58
4.3	Deviation by Country of origin Pie Chart . . . . .	61
4.4	Deviation by Country of origin Bar Chart . . . . .	62
4.5	Overall Summary of Missing Values for NASA-TLX dataset . . . . .	65
4.6	Overall Summary of Missing Values for WP dataset . . . . .	67
4.7	SAS Workflow for NASA-TLX dataset without additional features . . . . .	74
4.8	Model Performance for NASA-TLX dataset without additional features . . . . .	75

## LIST OF FIGURES

---

4.9	Model Performance for WP dataset without additional features . . . . .	75
4.10	SAS Workflow for NASA-TLX dataset with additional features . . . . .	76
4.11	Model Performance for NASA-TLX dataset with additional features . . . . .	76
4.12	Model Performance for WP dataset with additional features . . . . .	77
5.1	Outline of the Evaluation Chapter . . . . .	79
A.1	Recent version of NASA-TLX assessment technique . . . . .	92

# List of Tables

2.1	Types of Supervised Machine Learning techniques . . . . .	33
3.1	Database description . . . . .	46
3.2	Value range for NASA-TLX . . . . .	50
3.3	Value range for Worload Profile . . . . .	51
3.4	Confusion Matrix . . . . .	53
4.1	Primary, Secondary Keys and flags . . . . .	59
4.2	Removed features . . . . .	60
4.3	Variable type and level information for NASA-TLX . . . . .	63
4.4	Consistency of data . . . . .	64
4.5	Missing values statitics for NASA-TLX dataset . . . . .	66
4.6	EM Means NASA-TLX . . . . .	67
4.7	EM Means WP . . . . .	68
4.8	Transformation of variable "Region" . . . . .	69
4.9	Transformation of variable "Datetime" . . . . .	70
4.10	Eventual data types . . . . .	72
4.11	Summary of contineous attributes modifications . . . . .	74
5.1	Average square error comparison . . . . .	80
5.2	Correlation between outputs of ML classifier and NASA-TLX equation . . . . .	80
5.3	Correlation between outputs of ML classifier and WP equation . . . . .	81
5.4	Correlation comparison . . . . .	82

# Chapter 1

## Introduction

### 1.1 Background

Nowadays, in the world of fast technological progress, interfaces and systems are becoming more complex because of the influence of different factors such as human behavioural traits, trust and time (Longo, Dondio, & Barrett, 2010). As a consequence, there is a tangible need to design optimal interactions of humans with these systems and interfaces. To support existing design procedures, beside the concept of trust (Dondio & Longo, 2011), the concept of Mental Workload (MWL) has been adopted in areas such as aviation, for instance, for the design of airplane cockpits. However, its applicability is vast, and MWL can be adopted for a wider range of human-computer interaction (HCI) application such as web-based systems, interfaces for medical devices etc. MWL is a key concept for designing interactions that maximise user satisfaction and productivity. For instance, in (Longo, Rusconi, Noce, & Barrett, 2012) it has been showed how the increase in imposed MWL on end-users by a set of tasks executed on two popular web-sites - Google and Wikipedia - is correlated with the perception of usability of the same interfaces.

It is already known that mental workload influences productivity of humans (Xie &

Salvendy, 2000). Both underload and overload negatively affect human productivity. In cases of low levels of mental workload people might often experience annoyance and frustration (Longo, 2015a). On the other hand, high levels may be destructive for a person and negatively influence performance (Rubio, Díaz, Martín, & Puente, 2004). Consequently, the best performance can be achieved with optimal workload, which Longo (2015a) associated with a high user satisfaction, a high system success, a low error rate and a high productivity (figure 1.1).

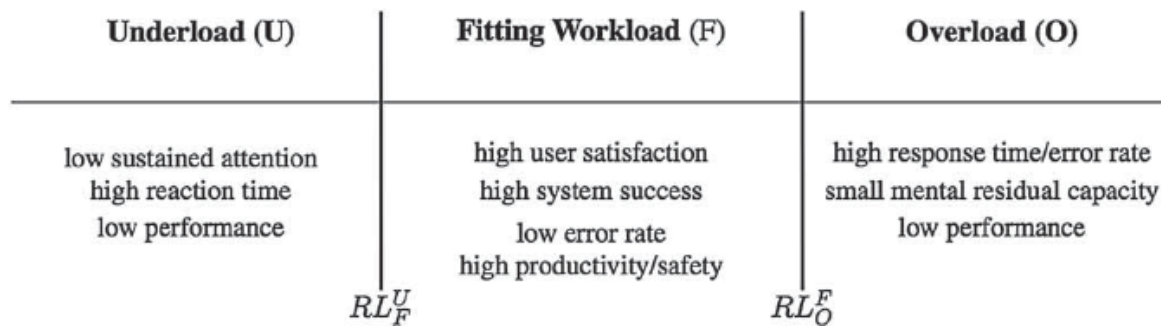


Figure 1.1: Disadvantages associated with low/high MWL and advantages of optimal workload

Source: (Longo, 2015a)

## 1.2 Research Question and Research Hypotheses

The main research question of this research study is:

Can Supervised Machine Learning classifiers outperform two theory-driven approaches, namely the NASA Task Load Index and the Workload Profile, in the prediction of Subjective Mental Workload according to face validity and correlation coefficient.

Following this question, two research hypotheses are proposed:

- H01: Supervised Machine Learning classifiers do not outperform the NASA-TLX in the prediction of Subjective Mental Workload in terms of face validity and

correlation coefficient.

- H02: Supervised Machine Learning classifiers do not outperform the WP in the prediction of Subjective Mental Workload in terms of face validity and correlation coefficient.

### 1.3 Research Methodologies

In order to accept or reject the research hypotheses an existing dataset is used which contains quantitative data. In detail, secondary, empirical, quantitative , deductive research will be done during this study.

### 1.4 Scope and Limitations

This research is using data gathered from students in a university, despite of the fact that some geographical, age and task diversity is provided, the subset is relatively narrow and could not be treated as a representative sample of a population. Another concern is the size of the sample. The dataset size is around 300 hundred records and this could be not enough for achieving the best possible generalisation by machine learning models considering that part of the data is used for validation and testing purposes.

### 1.5 Document Outline

This research consists of five chapters and a brief overview of its content is provided below.

**Chapter 2:** it focuses on a literature divided into two parts. The first part

describes the concept of Mental Workload from a general perspective, highlighting which factors influence it and how it can be measured. Additionally an overview of the state-of-the-art subjective mental workload measurement techniques, used in the comparative analysis, namely the NASA-TLX and the Workload Profile, is provided. The second part touches a different edge of current research particularly it introduces supervised Machine Learning. It then summarises process of knowledge discovery in Databases and the CRISP-DM and it introduces a number of supervised Machine Learning classifiers useful for the subsequent experimental analysis.

**Chapter 3:** it describe the design of the experiment highlighting its sub-activities as well as the software and tools used for data analysis.

**Chapter 4:** it describes the implementation of the experiment and all the activities mentioned in chapter 3. This chapters contains the actual observations and results of the experiment, including the preparation of the dataset, models adjustment and the description of the results.

**Chapter 5:** it demonstrates and evaluate experimental results. In details, the acceptance or rejection of the research hypotheses will be made here. The findings achieved by the Machine Learning classifiers will be compared against the findings achieved by the NASA-TLX and WP instruments.

**Chapter 6:** it provides a conclusion of this thesis and summaries the work performed in the experiment. Additionally, it discusses possible improvements and future work.



# Chapter 2

## Literature review

### 2.1 Introduction

Mental workload is a complex construct borrowed from psychology with several application in aviation and automobile industries. Literature suggests that it is hard to define precisely (Longo, 2014) (Longo, 2015a) (Rizzo, Dondio, Delany, & Longo, 2016). Beside transportation, application of the concept of mental workload are several. (Longo, 2011), proposed to adopt the concept mental workload to contribute to the assessment of cognitive engagement in the World Wide Web, of for the design of adaptive and personalised web-systems (Longo, 2012). He also investigated its relation with the construct of usability (Longo et al., 2012) (Longo & Dondio, 2015) and adopted it in the context of medicine and health-care (Longo, 2015b) (Longo, 2016).

This chapter provides a literature review of different aspects the combination of which is able to find an answer on research problem. Three areas are discussed here: Concept of Mental workload, Assessment of Mental Workload and Supervised Machine Learning.

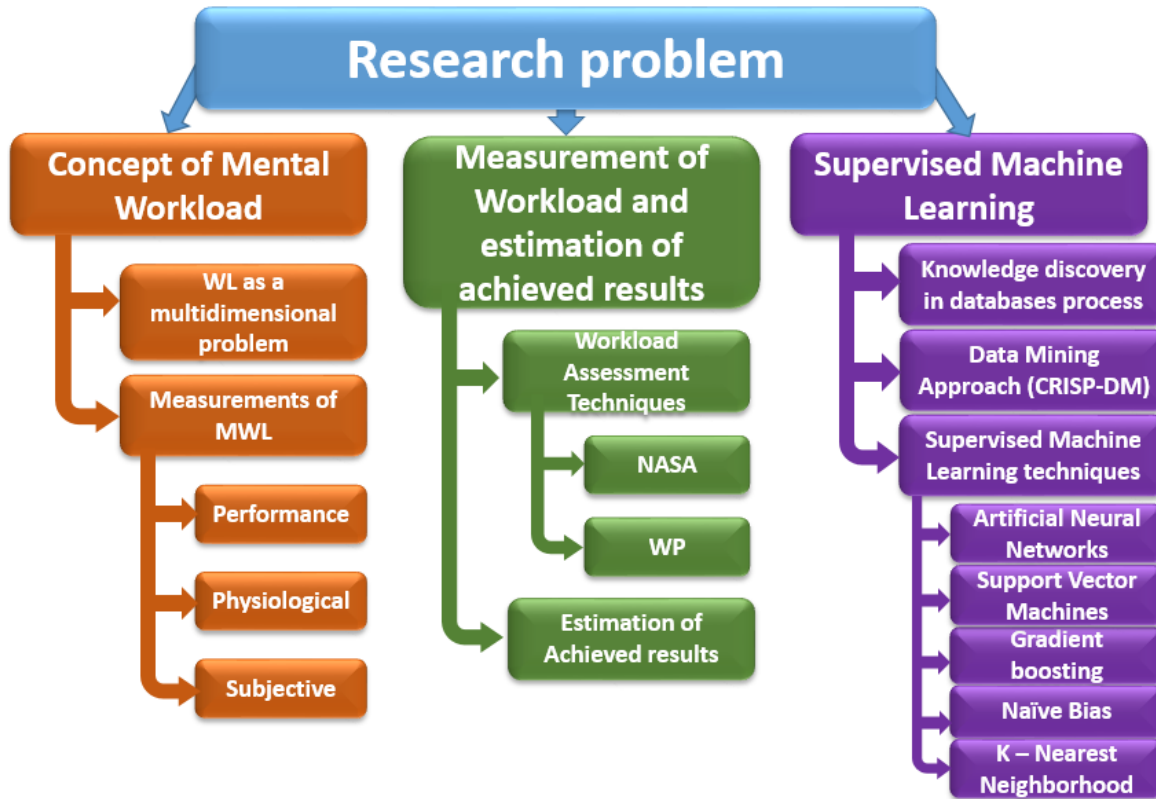


Figure 2.1: Outline of the Literature Review Chapter

Initially, the brief discussion about the importance of MWL for human productivity and its relevance of research provided. Then, it is highlighted that MWL is a complex multidimensional problem which could be influenced by many dimensions and factors. Finally, it is mentioned that there are several approaches for measuring of MWL and their summary gives details about them and showing some advantages and disadvantages of these approaches.

Second part is quite similar to previous one. However, it is focused on practical ways for MWL estimation, and evaluation of robustness of gathered results. Particularly, there is a discussion regarding specific assessment techniques for getting data for analysis of MWL for performed task and the way of calculating subjective estimated workload which suggested by them. Then, the overview of requirements for robustness estimation is provided.

The third part describe particular instruments which are using in comparison with mathematical driven approaches for measuring MWL. The process of knowledge discovery in databases (KDD) is discussed as a foundation for Data Mining approach. In particular, the framework called CRISP-DM is using as a guideline for performing the whole variety of activities applying to dataset for achieving the best results by Predictive Machine Learning. Finally, the comparison of many Supervise Machine Learning according to the number of criterions is given with a brief explanation of the way how they work.

## 2.2 The concept of Mental Workload

### 2.2.1 Workload as a multidimensional notion

Workload, in general, represents the cost in order to accomplish a task and satisfy its requirements (Hart, 2006). Mental Workload focused at measuring the thinking activities of a human. Many researchers agreed that HMWL is a wide and multidimensional problem (Hancock & Meshkati, 1988; Reid & Nygren, 1988). Summarizing information we could formulate three dimensions of MWL problem.

1. Situations. It is the occasions where MWL could take place. It may be as continuous tasks, like aircraft flying, as well as in fast changing systems, for instance video games.
2. Time. The duration of MWL is varying significantly from second to many hours. This factor is influencing perception of MWL in a major way and should be consider properly during evaluation process.
3. Influenced factors. This dimension considers factors which could have an influence at MWL, such as training, practice, motivation etc. In spite of the fact, that some factors are presented in most sets, the final number of such factors is

different for every task.

## 2.2.2 Measurements of Mental Workload

The problem in measuring MWL is that there are not uniform criteria for defining it (Hancock & Meshkati, 1988). It is mostly subjective measure, but consequence factors could reflect it. They could be divided into three main groups: Subjective measures (Self estimated measures) and Objective measures (Performance measures and Physiological measures).

However, Muckler and Seven (1992) in his paper come up with conclusion that the distinction between subjective and objective measurements in human performance studies is meaningless for the reason that subjective elements include in all types of measurement like selecting measures, collecting or interpreting data. Consequently, all three groups should be considered as different approaches for Workload Estimation without superb one over the other.

### Performance measures

In objective measurement, MWL could be estimated by task demand evaluation, performance evaluation, for instance, amount of correct answers in a test.

Performance is indicating how much MWL was required for completing the task. However, it is important to highlight that, according to Longo (2015a) research, not only overload, but also underload could cause a reduce in performance. How stable performance is during time characterized by concentration. There is a variety of research which are showing reducing in concentration while participants were engaged in high demand mental task. This criterion of MWL is significant enough for instance in driving activities (Recarte & Nunes, 2003).

### **Physiological measures**

Physiological measures rely on dependence human psychology and MWL, they are measuring the changes in the body such as heart rate, electro dermal responses and so on, but dont bother about performance. This measurement could be performed continuously during the task, but it requires special equipment for measuring these data.

### **Subjective measures**

In case of subjective evaluation, MWL usually estimated through post-event evaluations, like rating scales or questionnaires. Questionnaires in the same time could ask a participant directly about subjective estimated MWL, or evaluated it by applying a mathematical equation at a number of predefined questions (Hart, 2006; Tsang & Velazquez, 1996).

Overall level of MWL could be influenced by many factors. Additionally, in some cases one reason could defease other(Longo, 2015a), this made such estimation even more complex and sensitive to a variety of factors.

## **2.3 Mental Workload subjective assessment techniques**

### **2.3.1 Introduction**

There is a number of assessment techniques for subjective estimation of subjective workload. One of the most popular and widely used are NASA Task Load Index (Hart, 2006) and Workload Profile (Tsang & Velazquez, 1996) techniques. Both of these procedures relying on the assumption, that workload is a multidimensional and

consequently, should be measured in different dimensions. Additionally, multidimensional approach could not only provide information about levels of demand, but also give insights on the origin of it.

NASA-TLX and Workload profile techniques are showing good results in identifying the origin of workload by measuring many dimensions of workload. They have been used for gathering initial dataset for current research.

Noticeably, that some researchers made an attempt to compare different techniques of workload estimation. For instance, Rubio et al. (2004) made evaluated sensitivity, validity, diagnosticity, and intensiveness of NASA, WP and Subjective Workload Assessment Technique (SWAT). ANOVA test demonstrated that there are no significant differences between them in intrusiveness. WP had the greater sensitivity, and its diagnostic power definitely superior NASA and SWAT.

### **2.3.2 NASA-TLX overview**

Nasa is post-event assessment techniques which was developed by the NASA Human Performance Group (Hart & Staveland, 1988) for investigation of factors influencing subjective perception of mental workload. This technique is one of the most widely used technique in many domains, including healthcare, education, aviation and a variety of other social technical domains (Colligan, Potts, Finn, & Sinkin, 2015).

Through a multi-year research process, scientists have identified six factors which are influencing the subjective estimated MWL. These factors are reflecting the origin clusters which are defining the level of workload for most tasks. Subjective estimation is widely using approach in contrast with performance or theoretical approaches. However, one of the main drawbacks of it is subjective variability. NASA rating technique allows to reduce the influence of this factor.

Another problem of subjective estimation of MWL is a big variation of its sources

across tasks. Nasa partially solves this problem by implementing a rating technique which is multidimensional and allows to find the most relevant sources of workload for a particular task (Hart & Staveland, 1988).

### 2.3.3 NASA-TLX scale

NASA-TLX questionnaire consists of two main parts. First part contains six question which are estimating workload in different dimensions. There is a number of different variations of NASA-TLX, but despite of the words difference the concepts beyond are still the same. The sample questionnaire could be seen from Appendix A,

The overview of scales is listed below:

- **Mental Demand** The amount of calculating, thinking, deciding, remembering and other mental skills was required.
- **Physical Demand** The amount of physical activities such as pulling, pushing, activating, etc.
- **Temporal Demand (Time pressure)** The amount of felt pressure during the task.
- **Performance** The self-estimated performance and satisfaction of the performed work.
- **Effort** How hard the participant work to accomplish achieved level of performance according to his estimation.
- **Frustration level** How irritated or annoying task was.

Second part of NASA-TLX is a rating system included binary choice in pairs among all combination of Factor-to-factor pars. Then, according to the gained weight the most important factor for particular task is getting higher weight and by this is has more influence at overall MWL. The final formula is represented at figure below:

$$\frac{\sum_{i=1}^6 n_i * w_i}{15} \quad (2.1)$$

### 2.3.4 Workload profile overview

Workload profile (WP) is another used multidimensional assessment procedure for the estimation of subjective workload. WP has demonstrated a better performance in contrast to other uni-dimensional procedures in terms of reliability, validity and sensitivity to task demand (O'Donnell & Eggemeier, 1986).

### 2.3.5 Workload profile scale

Tsang and Velazquez (1996) identified four (stages, responses, codes and modalities) dimension which could be required for completion a task. Each dimension consists of two ways of processing are illustrated on figure 2.2. WP technique based on assumption that demand resources for task completion are formulating the workload dimensions. WP identifies similar dimensions: stages of processing, processing codes, input modalities and output modalities. Each dimension estimated with range from zero to one individually by a participant after completion of a test, where zero could be treated as no demand on the particular dimension, whereas one proposes maximum attention. In contrast with NASA-TLX WP does not have rating system and overall workload estimated through average figure of all eight questions.

Stages are divided into perceptual and response processing. In first case involved activities which are requiring attention for problem-solving or remembering activities. In second case, attention required for execution or selection. The typical response processing is for example choosing the right pedal for car drivers.

Processing codes mostly could be separated between spatial and verbal processing. The deviation depends of the nature of the task, for instance driving is a special process



whereas reading is a representative of verbal processing.

There are two main ways for input modality (receiving information) visual processing and auditory processing in depend on which sense was involved into receiving process.

Output modalities. In most cases output modalities consist of manual like typing and speech responses like participating into debates or conversation.

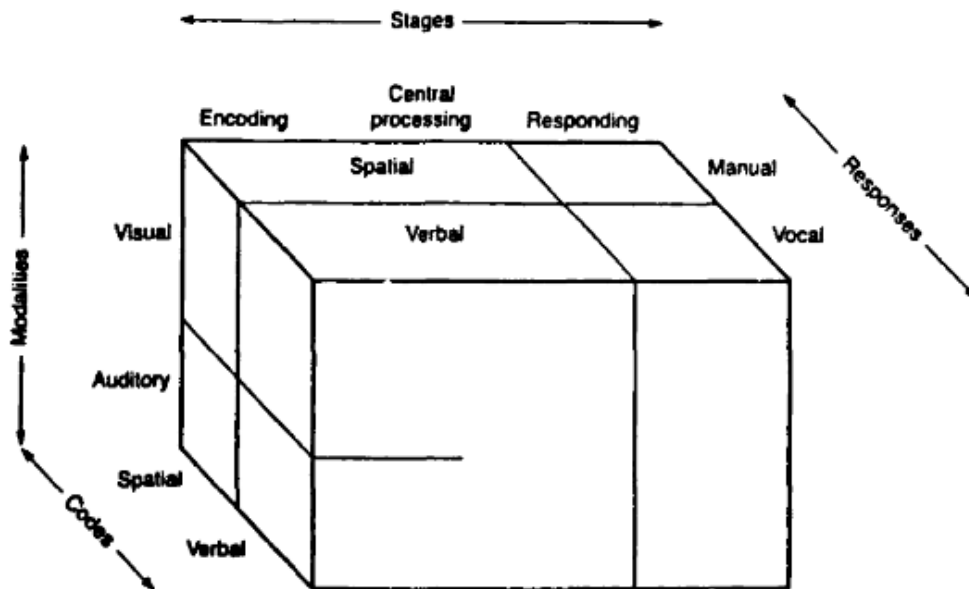


Figure 2.2: Disadvantages associated with low/high MWL and advantages of optimal workload

Source: (O'Donnell & Eggemeier, 1986)

### 2.3.6 Criteria for the evaluation of Mental Workload models

Current research is trying to compare Theoretical Driven Approaches and Machine Learning Classifiers in quantitative manner. In order to perform this, it is necessary to choose the way how they could be compared. Boff, Kaufman, and Thomas (1986)

highlighted a number of criteria for estimation of technique robustness, they could be listed as following:

- Sensitivity. The ability of technique should show the reflection of changes in task difficulty or demand
- Diagnosticity. How the measurement techniques are able to find reasons for changes in MWL.
- Validity. Technique should measure what is expected to measure. Changes in amount of stress or physical demand should not effect of evaluated figure.
- Intrusiveness. Where techniques are causing degradations at continuous task performance.
- Requirements for implementation. How easy or difficult to implement a technique. What are operator or equipment requirements.
- Operator Acceptance. The level of operators willingness to strictly follow requirements, how much a technique results could be affected by operator.

Langdridge and Hagger-Johnson (2009) described assessment measures including sensitivity and predictive validity as well as their suitability for the research. They also discussed many aspects of Data Analysis applying particularly in field of Psychology.

Eventually, it has been dived to use Face Validity as criterion for comparison of Theoretical driven and Machine Learning approaches because they could be represented in qualitative terms and because of this easily compared. Moreover, the dataset has been gathered in the same way for both approaches, which means that, for instance, Implementation Requirements and Operator Acceptance measures are not significantly different for them.

### 2.3.7 Face Validity

Test validity is widely-used characteristic during psychological test. From point of view of test respondents this measure is a degree of relevance of test content and the purpose for which test was performed.

Face validity is based upon three main principles which were highlighted by Weiner and Craighead (2010):

1. Face validity is more based on the opinion of people who take part in an experiment and their restricted knowledge in this domain rather than opinion and judgments of professionals and psychologists.
2. Face validity imply that the measuring content is obvious for test takers.
3. The environment and situation where the test took part are influencing face validity. The test defensiveness could appear in case in participants are not fully opened for a test, a result it could be developed into incorrect results (Bornstein, Rossner, Hill, & Stepanian, 1994).

The combination of these three aspects is important for achieving the high level of face validity. The good example for distinguishing low and high levels of face validity relying under the obviousness for participants and environment is the asking the same question into two different situations. The question Does trying something new is always scary? could be traded by potential employers for a manufacturing position is absolutely different rather than the patients during a mitting with a psychotherapist.

There are debates about the helpfulness of Face validity in psychological tests. On one hand, Downing (2006) beliefs that the importance of face validity is overrated in scientific and in particular in Medical education. It his research 67 papers were investigated in terms of usage Face validity. He concludes that about 19% of the papers incorrectly discussed face validity and two out of 16 papers mislabeled validity

evidence as face validity. He claims that the usage of this measure in scientific papers is more associated with marketing to consistency rather than real evidence of validity.

On the other hand, many scientists find valuable usage of face validity in the research and the positive correlation between face validity and test item accuracy (Holden & Jackson, 1985). For instance, Holden and Jackson (1979) conduct an experiment where he discussed a distinction between face validity and item subtlety. He concludes that then higher face validity and lower subtlety level; then higher item validity was observed.

## **2.4 Supervise Machine Learning**

### **2.4.1 KDD Process**

The current study makes use of an existing dataset for the creation of predictive models. This operation requires many preparation steps. Consequently, for better results a widely-used process called Knowledge Discovery in Databases (KDD) should be implemented here.

In general, a KDD framework allows to find a new knowledge in already existed data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). De Martino et al. (2002) describes KDD as an integration of multiple technologies for data management such as database management and data warehousing, statistic machine learning, decision support, and others such as visualisation and parallel computing. This is quite old process, but it is using a standard approach in most of the tasks in many areas (Han, Altman, Kumar, Mannila, & Pregibon, 2002; Yang & Wang, 2012). It's overview is represented below.

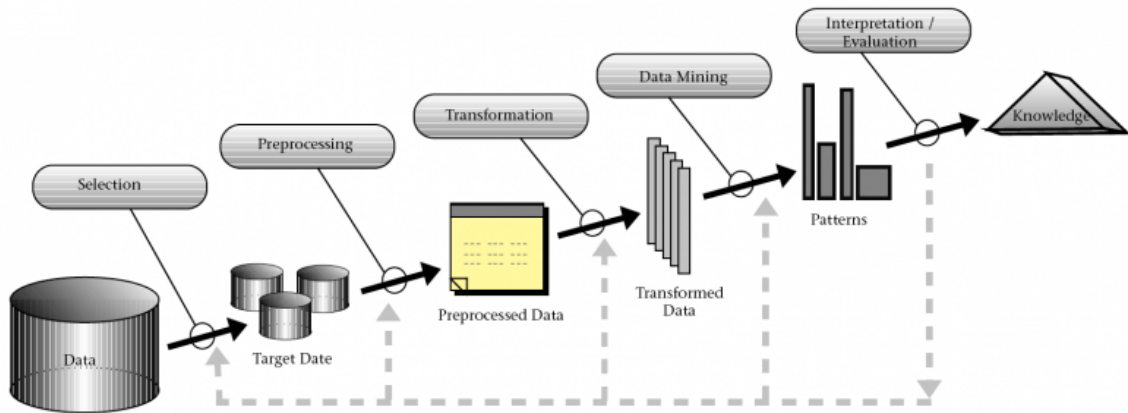


Figure 2.3: Overview of the Steps That Compose the KDD Process

Source: (Fayyad et al., 1996)

## 2.4.2 Data mining approach

Data mining is only one part of KDD process. However, it is requiring a plenty of preparation steps for achieving the best results. Fortunately, in early 2000s have been designed One of the most well-known, standard and well-describe approaches for achieving so is Cross Industry Standard Process for Data Mining (CRISP-DM). In spite of the fact that this approach was developed mostly for industry purposes based on practical experience of experts in Data Mining market, it very reliable for scientific purposes as well. (Chapman et al., 2000) This approach fully describes steps and particular activities which should be done for gaining knowledge from row dataset. The figure 2.4 demonstrates the overview of CRISP-DM activities.

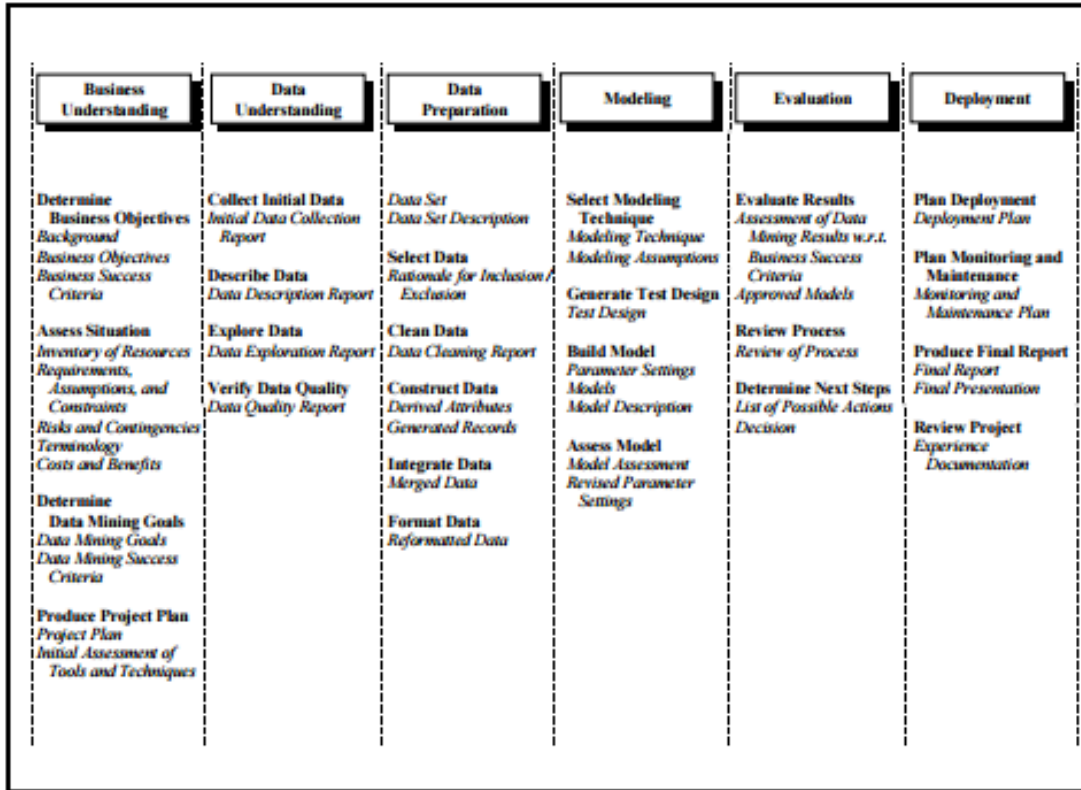


Figure 2.4: Overview of CRISP-DM

Source: (Chapman et al., 2000)

Regarding the current study, a gathered dataset has a variety of issues which have to be solved in purpose of designing quality models and by doing this, robust results.

This approach conducts researches through the main stages of DM process with detailed practical recommendations for each particular stage and sub stages. CRISP-DM contains next 6 parts with many back-forward paths between them: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment 2.5.

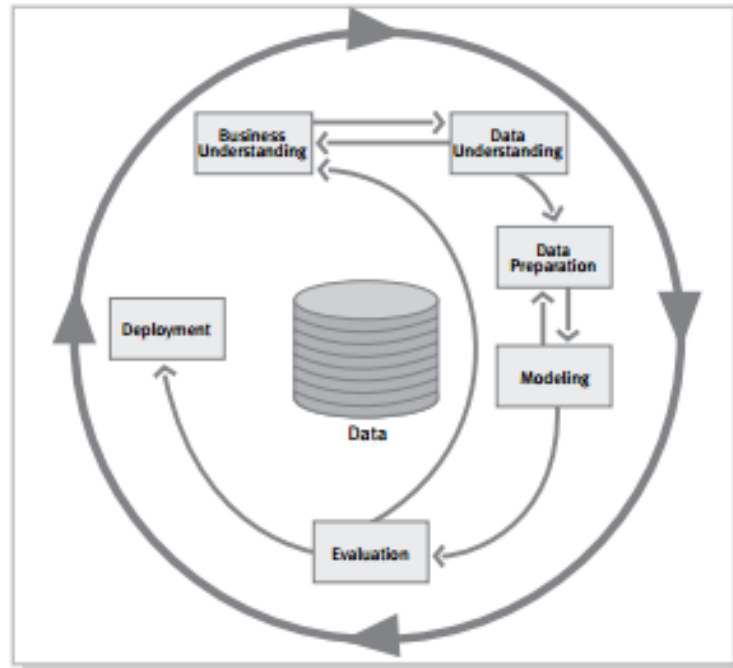


Figure 2.5: Phases of the CRISP-DM reference model

Source: (Chapman et al., 2000)

All of these stages are quite generic and could be adjusted to satisfy a specific problem. The overview of each stage is represented below:

- **Data Understanding.** In order to choose the most appropriate mathematical models it is necessary to make a solution relying on the following: the absence or presence of outliers in a dataset, noise as well as the type of the data and type and range of a target variable, because model performance is dependent on them. For example, Support-Vector Machine and Neural Networks do not allow discrete figures as target variables (Kotsiantis, 2007). Data understanding also includes visualization activities, for the reason that it becomes easier to see patterns and features of data.
- **Data Preparation.** This phase consists of all activities for creating the final dataset for modeling. It is quite possible that this phase will be performed many

times and also the output might have more than one prepared dataset to meet specific requirements of each model.

- **Modelling.** Usually more than one model applying with a wide range of parametric sets. Modeling and Data preparation stages are closely connected with each other and performs one after another multiple times.
- **Evaluation of the models.** After creation of the models it is important to understand whether they met predefined business requirements in relation to desirable evaluation criteria. Some models are easy to understand from business point of view, whereas other could be better in their prediction power but being a black box for stakeholders (Kotsiantis, 2007).
- **Deployment.** This phase is carrying about implementation of finding knowledge for business. This phase is not applicable for current research and will be bypassed.

### 2.4.3 Predictive models

Nyce and CPCU (2007) interpret predictive analytics as a variety of statistical techniques, including predictive modelling, machine learning and data mining, which analyses current and historical facts to make predictions about future, or otherwise unknown, events. One of the key component in predictive analytics is a target variable. Predictive analytics uncovers relationships and patterns within large volumes of data that can be used to predict behavior and events, and it is widespread nowadays. Predictive analytics is widely used in business areas, where it helps to reach conclusions about customer behavior and helps to understand purchasing patterns to create new sales and reduce churn to the competition (Linoff & Berry, 2011).

Predictive analytics uses supervise Machine Learning in attempt to calculate a target variable according to a labeled dataset.



#### 2.4.4 Supervised Machine Learning techniques

During the past 20 years, alongside the growing amount of data available for analysis, a variety of different Machine Learning techniques have been developed. All of them could be divided into Supervised and Unsupervised according to whether they have labeled instances or not. If they have, then these techniques are called supervised, if not they are referred to as unsupervised or in another words, clustering techniques (Kotsiantis, 2007) (Jain, Murty, & Flynn, 1999). In the current study only the first type is relevant and will be taken into consideration.

Machine learning have been used in many fields of application, some of them including adaptive web-sites (Aslan & Inceoglu, 2007), natural language processing (Collobert & Weston, 2008), healthcare (Longo & Hederman, 2013), software engineering (Srinivasan & Fisher, 1995).

A detailed review of the most popular Supervised Machine Learning techniques like Decision Trees, Neural Networks Nave Bayes, k-Nearest Neighbor, Support vector Machine and Rule-learners techniques was made by (Kotsiantis, 2007). The assumptions for using each of ML classifier were listed. Particularly, type of the accepted dependent and independent input variables was mentioned here. A quantitative analysis ranked classifiers in 13 properties 2.6.

	Decision Trees	Neural Networks	Naïve Bayes	kNN	SVM	Rule-learners
Accuracy in general	**	***	*	**	****	**
Speed of learning with respect to number of attributes and the number of instances	***	*	****	****	*	**
Speed of classification	****	****	****	*	****	****
Tolerance to missing values	***	*	****	*	**	**
Tolerance to irrelevant attributes	***	*	**	**	****	**
Tolerance to redundant attributes	**	**	*	**	***	**
Tolerance to highly interdependent attributes (e.g. parity problems)	**	***	*	*	***	**
Dealing with discrete/binary/continuous attributes	****	***(not discrete)	***(not continuous)	***(not directly discrete)	** (not discrete)	***(not directly continuous)
Tolerance to noise	**	**	***	*	**	*
Dealing with danger of overfitting	**	*	***	***	**	**
Attempts for incremental learning	**	***	****	****	**	*
Explanation ability/transparency of knowledge/classifications	****	*	****	**	*	****
Model parameter handling	***	*	****	***	*	***

Figure 2.6: Comparison of Supervised Machine Learning techniques

Source: (Chapman et al., 2000)

The majority of algorithms could be divided into six groups:

### Artificial Neural Networks

This algorithm consists of a big number of neurons which are connected together as at presented in figure 2.5. ANN usually achieve a good rate of accuracy, however it requires a lot of computation power to create a model and it is almost impossible to understand why some decision was made, i.e. it is a black box for researchers. Noticeably, that ANN not allow discrete variables, they have to be transformed in binary ones during data preparation steps. Also, ANN characterized by high risk of overfitting. Haykin, Haykin, Haykin, and Haykin (2009) describes the Artificial Neural Network (ANN) as a machine that is designed to model the way in which the brain

Table 2.1: Types of Supervised Machine Learning techniques

<b>Algorithm type</b>	<b>Example</b>
Logic based algorithms	C4.5
Perceptron-based techniques	Artificial Neural Network
Statistical learning algorithms	Naive Bayes classifiers
Instance-based learning	K-Nearest Neighbor algorithm
Support Vector Machines	Support Vector Machines
Regressions	Logistic regression

performs a particular task or function of interest. He indicates that a neural network is a massively parallel distributed processor composed of simple processing units that have a natural propensity for storing experiential knowledge and making it available for use.

The example structure of an artificial neural network is provided below (Haykin et al., 2009). The first six source nodes ( $x_1 \dots x_6$ ) comprise the receptive field for hidden neuron 1, and so on for the other hidden neurons in the network. The receptive field of a neuron is defined as that region of the input field over which the incoming stimuli can influence the output signal produced by the neuron. The mapping of the receptive field is a powerful and shorthand description of the neurons behavior, and therefore of its output.

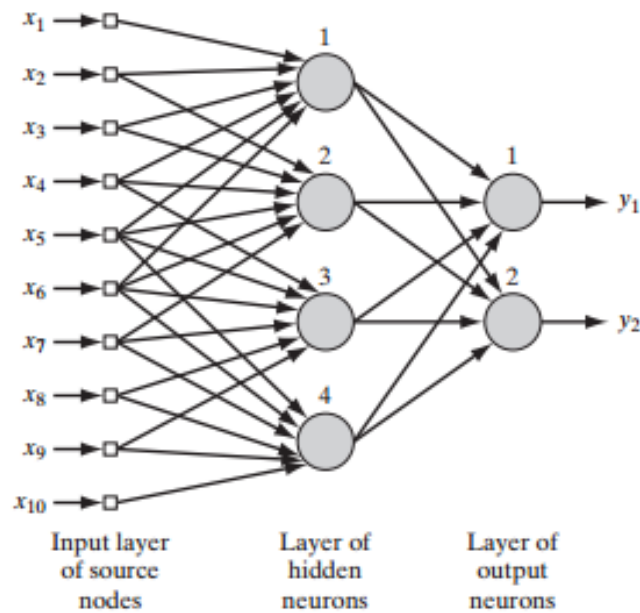


Figure 2.7: Neural network with three levels of layers

Source: (Haykin et al., 2009)

Artificial neural networks have shown strong performance in many areas; for instance, (Kara, Boyacioglu, & Baykan, 2011) reported significant performance in predicting the direction of stock price movement. The average prediction performance for his research was 75.74%.

### Support Vector Machines

Another machine learning technique is the Support Vector Machine (SVM). This technique uses associated learning algorithms for highlighting patterns and understanding data in order to use them for classification and regression analysis (Cortes & Vapnik, 1995). It is calculating the maximal margin between all dimensions i.e. it is creating the largest distance between instances, which is reducing the generalization Error. As well as ANN, SVM have very high level of accuracy but has a need in computation power. Also, this technique is not allowing using discrete variables.

The SVM is widely used in systems whereby classification is needed. (Meier et al., 2012) used the SVM truing to classify younger adult brains, distinguishing them from older ones, according to resting-state functional connectivity. He concluded that this technique can successfully solve this issue. The use of the SVM allowed him to find three general patterns in age-related brain changes.

### **Gradient boosting**

Gradient boosting is also Machine Learning algorithm which is using an ensembles of decision trees for solving classification problems. Gradient boosting work similarly to the desertion trees, e.g. it is searching for an optimal deviation into two categories in respect to a target variable. Noticeably, that it is less sensitive to overfitting and more robust solution in contrast with a single decision tree because relying on many decision trees models (Maldonado, Dean, Czika, & Haller, 2014).

It consists of two parts Gradient descent or Gradient ascent and Boosting. In general, Gradient descent and Gradient asking are optimization algorithms. They are looking for a minimum or maximum of the function by adjusting its arguments. On each step the function become slightly closer to a minimum or maximum. The step size usually determinate by a precision which want to be achieve by algorithm.

Boosting is meta algorithm which is allowing to convert an ensemble of week learners (Gradient algorithms) to one strong learner. However, there are some restrictions of boosting. It is very sensitive to noise data because in case of misclassification it is trying hardly to correct prediction. Long and Servedio (2010) demonstrated this features of this meta algorithm and conclude that this causes a significant restriction of using it in a real word scenario with noisy and misclassified datasets. Gradient boosting is also could be very expensive in terms of computation power and time, especially close to the function minimum Yuan (2008).

### **Naive Bayes classifiers**

This approach based on Bayes theorem with strict assumption of independence between features (Wu et al., 2008). This classifier is widely using in text estimation. For instance, many spam filters are using it in order to divide acceptable content from unacceptable. Usually, the accuracy of this method is relatively low in contrast with other approaches 2.6. However, an advantage of this technique is very high speed of classification and also very good level of tolerance to missing values. Additionally, NB algorithm characterized by low tolerance to redundant attributes. Continuous features are not permitted here.

### **K-nearest neighbor algorithm**

This algorithm is representative of lazy algorithms. It based on assumption that records within a dataset are generally having the same properties. Consequently, in labeled data, when new instances are coming for labeling, the model is finding the closest k neighbors in many dimensional spaces and classifies new one according to the mode in case of categorical label or the average in case of continuous label. K-NN algorithm is relatively slow in classification of new instances coming into model, but fast during training process. Also, this algorithm is very sensitive to noise in dataset. There are different metrics for calculation neighbor distance: The most common of them are presented in figure 2.1.

Minkowsky: $D(x,y)=\left(\sum_{i=1}^m  x_i - y_i ^r\right)^{1/r}$
Manhattan: $D(x,y)=\sum_{i=1}^m  x_i - y_i $
Chebychev: $D(x,y)=\max_{i=1}^m  x_i - y_i $
Euclidean: $D(x,y)=\left(\sum_{i=1}^m  x_i - y_i ^2\right)^{1/2}$
Camberra: $D(x,y)=\sum_{i=1}^m \frac{ x_i - y_i }{ x_i + y_i }$
Kendall's Rank Correlation: $D(x,y)=1 - \frac{2}{m(m-1)} \sum_{i=j}^m \sum_{j=1}^{i-1} \text{sign}(x_i - x_j) \text{sign}(y_i - y_j)$

Figure 2.8: Approaches to define the distance between instances

Source: (Kotsiantis, 2007)

### 2.4.5 Summary and Implications

The list below is outlining key points highlighted during literature:

- Workload is multidimensional problem with next main dimensions: Situations, time and influenced factors. Each of these areas
- could influence workload in different proportion depend on particular situation.
- There are three main groups for Workload measurement such as Performance measures (estimate results of performed task),
- Psychological measures (estimate body behaviour such as heart rate, electrodermal responses) and Subjective measures (rating scales and questionnaires)
- There is a number of specifically designed for measuring workload questionnaires. One of the most widespread are NASA-TLX and WP

- Results gathered by questionnaires could be estimated in seven criterias.
- Supervise Machine Learning could be applied to solve many different problems including Prediction of Subjective estimated Workload, in case of providing proper data as an input.
- The most suitable Classifier for a dataset highly depends of it's features.

Overall, literature review was focussing in three different areas: concept of Workload, ways if its measuring using Subjective Approach and Tools wich could perform such activity in particular - Supervise Machine Learning.

### 2.4.6 Gaps, motivation and limitation

#### Gaps

After 18 years Hart (2006) conducted new implementation of the NASA multidimensional scale by applying it on new generation of users. It was highlighted that there is no clear understanding in this process and necessity of provide new insights into area of MWL estimation and reflection of results.

Although, the phenomena of Mental Workload is under consideration of many researches, there was not attempts to apply Machine Learning in attempt to predict Subjective Estimated Workload instead of using mathematical Equations. Current research will try to bring new knowledge in are of MWL by using relatively new capabilities of Machine Learning algorithms.

New opportunities of modern software products are able to give additional insights into MWL by revealing hidden patents and they could be more precise in prediction of subjective estimated workload.



### **Motivation**

The deeper understanding of MWL will provide to humanity many ways to apply this knowledge into something more practical starting from web-site creation and end by designing interfaces for Space Ships. It could help reduce amount of workload which is influencing people productiveness, and not only this, but also reduce the chance to make a critical fault.

### **Limitations**

Current research as well as all questionnaires are using Subjective way to measure Workload which is only one of the three possible ways to do it. Consequently, the better results could be achieved in combination of all three approaches, for example by measuring physiological indicators during the task and estimation results and filling questionnaires afterwards.

# Chapter 3

## Experiment design

### 3.1 Introduction

This chapter is describing experiment design which will be performed during an experiment. The variety of software products will be used, in order to achieve robust results. The summary of all software participated in research is provided.

The Cross Industry Standard Process for Data Mining (CRISP - DM) was taken as a basis for designing current experimental phase. The sections order and content were changed in respect to specific workflow of an experiment and in purpose of keeping waterfall way of presenting information, because CRISP-DM supposing the movement back and forward during developing process, which is not desirable in a thesis paper.

In addition to Supervise Machine Learning Models it is necessary to design a solution which is calculating Mental Workload using NASA-TLX and WP mathematical formulas in order to evaluate and compare results of both approaches into Evaluation Chapter.

Additionally, design experiment will be performed for two types of datasets. The fact is that, gathered dataset contains more features such as age, participator nation

or professor name, relegated to each questionnaire in addition to 21 NASA-TLX or 8 WP questions. In order to make a fair comparison, the input (set of input features) for both Theoretical and ML approaches have to be the same. However, in order to extend the knowledge about current area and find implementation of other gathered features, it is decided to perform two types of experiment including and excluding additional features. It will give an understanding of impact and value of other data at Mental Workload Demand. For instance, whether daytime when an experiment took place or nation of participant has an impact on results.

It is a critical point to highlight the research questions and by doing so research objectives as well, because with their clear identifying, research overall losing a main goal. The main research question for current research is: Do Supervised Machine Learning classifiers outstrip Theory Driven approaches NASA and WP in estimation of Subjective Mental Workload in terms of face validity and correlation coefficient? Also, it will be trying to find an additional knowledge beside main goal, by investigation next an additional research question: Does participants age, country of origin, date time of experiment have a significant impact on Subjective Mental Workload?

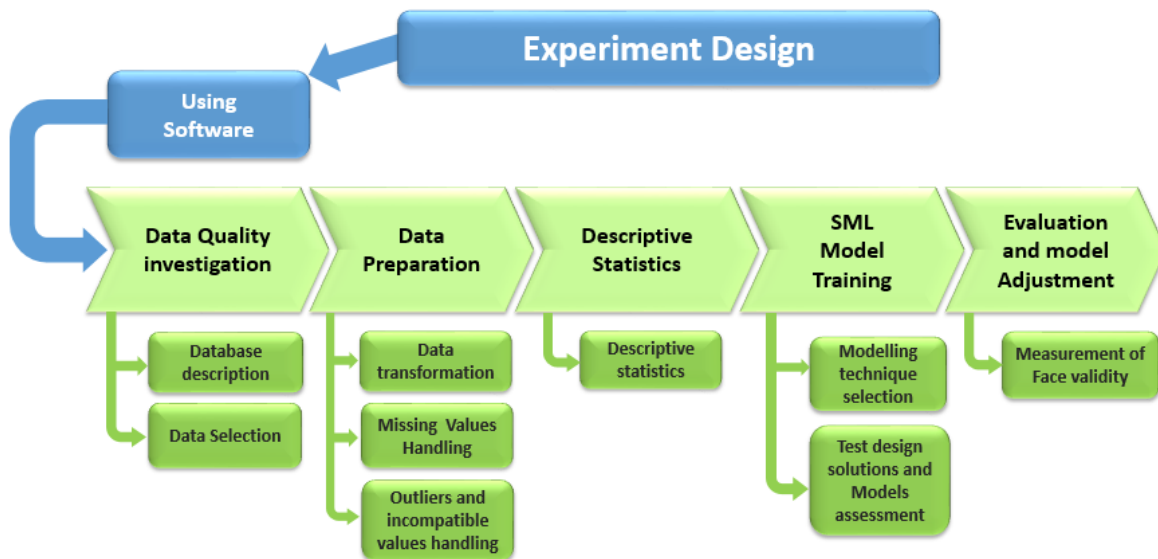


Figure 3.1: Outline of the Research Design Chapter

## 3.2 Software

Nowadays, programmers could use a huge amount of programming languages as well as software for achieving the same goals. For getting robust results initial row dataset have to go through entire cycle of CRSIP-DM meteorology. Consequently, the wide range of specific issues should be solved from investigation of data to evaluation of generated models. There is not union tool which could perform all these activities at the same time with the same level of efficiency. As a result, it has been decided that each part of the experiment will be done by the most appropriate tool for it.

Database schema was originally design for MySQL database, because of this Microsoft Work Bench was chosen for manipulating data and its querying. This software product is providing the Database Administration tools for Database management, administration and creation. It allowed to set up the Local SQL server and populated dataset with manually inserted data. This tool gives capability to manipulate data using SQL queries and create a flat SCV file as an output for further data processing.

In order to solve all issues with data, preparing and transform and construct data was chosen C# Programming Language. This very clear and powerful language is able to manipulate and change data easily with high level of control of each manipulation. The most powerful and nature environment for C# is Microsoft Visual Studio 2015. This application gives a full range of developing tools.

Next, output flat CSV files from Microsoft Visual Studio went to Data Mining tool. SAS Enterprise Miner was chosen for this task. It allows to process data in a streamline process with many descriptive and data investigation capabilities. One of the most important features of Enterprise Miner is that All models have predefined set of setting according to the provided dataset. This feature allows significantly speed up the Data Mining investigation. Moreover, all these settings could be adjusted for creating better models. Enterprise Miner could also be used for data preparation steps, but it is not so powerful for this stage in contrast with Models creation stage.

For measuring statistical figures IBM SPSS Statistics was chosen as the most suitable tool. This tool is able to give extensive information about dataset in general and a desirable feature in particular. IBM SPSS could be used for getting initial insights in data standard deviation, normalization or other data quality properties.

Finally, the visualisation of the data is another separated area during the process of creation high quality models. In the data investigation and data preparation stages the visualisation of the data is very important part, because it is improving the understanding of data which is almost impossible to achieve without visualizing it. The most suitable tool for it is Tableau 9.1. This tool allows to create advance and interactive graphs, charts and ability to customize their appearance widely according to user requirements.

### 3.3 Data Quality investigation

#### 3.3.1 Database description

The initial dataset was gathering from 2014 to 2016 years the following way: at the end of a class, the printed copy of NASA-TLX or WP questionnaire were randomly given to students for filling. The experiment was performed among university students in couple countries, by several professors and for different classes. All of it are providing the diversity in demand for task completion. Then, these papers were manually processed and populated into a MySQL database.

The database consists of six tables; each table represent a different object.

- `Students` gives student age and also connected to `Nationality` table. It should be mentioned, that one student could participate in more than one task.
- `Nationalities` contains only information about student nationality.

- Lectures gives information about lecturer First and Second Names in one field.
- Courses - this table contains class names where questionnaires took places. It has a connection to lecturer Name by a Foreign Key.
- Tasks includes information about a particular task, such as short text description with some details, date and time of performed task, and task duration. The field which is representing time has a text data type, which is not suitable for such measure. This issue should be overcome into Data Transformation stage.
- The main table for analysis is questionnaire which is containing primary and secondary keys, question about subjective perception of Workload, fields of both NASA-TLX and WP techniques as well as subjective task difficulty (RMSE). Each question asking about Subjective Mental Workload scaled to 20 pieces from the lowest level to the highest. The gradation of RMSE is lying between 0 and 120. Both NASA-TLX and WP question designed in the same way as described into (NASA) (WP) papers.

Database schema is represented below 3.2

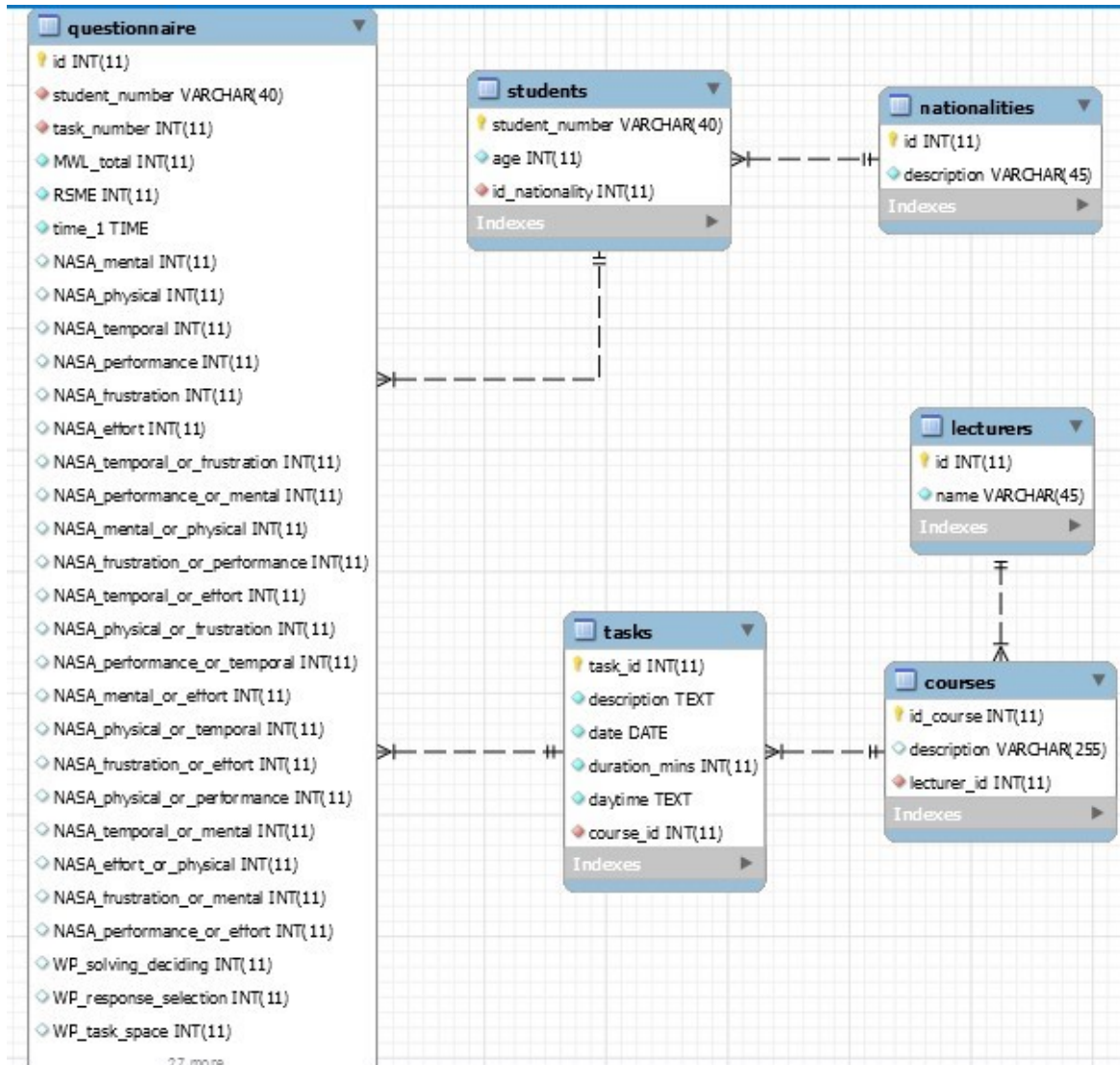


Figure 3.2: Database Schema

Before data cleaning steps, current dataset has next properties:

### 3.3.2 Data Selection

Initial Database Schema for storing data from questionnaires consists of many tables. It is typical structure of any relational database with is satisfying the low of third normal form (Date, 1999). However, for model creation all data have to be presented into a one flat table. These requirements come from data modelling tools such SAS

Table 3.1: Database description

Table name	Size	Field	Type	Size	Description
Students	244	student_number	VARCHAR	40	PK
		age	INT	11	Age of a student
		id_nationality	INT	11	FK
Nationalities	27	id	INT	11	PK
		description	VARCHAR	45	Nationality name
Lectures	3	id	INT	11	PK
		name	VARCHAR	45	Lectures name and surname
Courses	5	id_course	INT	11	PK
		description	VARCHAR	255	Course description
		lecturer_id	INT	11	FK
Tasks	50	task_id	INT	11	PK
		description	TEXT	-	Task description
		date	DATE	-	Date of the class
		duration_mins	INT	11	Duration of the class in minutes
		daytime	TEXT	-	Time of start and finishing
		course_id	INT	11	FK
Questionnaire	619	id	INT	11	PK
		student_number	VARCHAR	40	FK
		task_number	INT	11	FK
		<b>MWL_total</b>	<b>INT</b>	<b>11</b>	<b>Subjective estimation of Mental Workload</b>
		RSME	INT	11	Subjective Task difficulty
		time.id	TIME	-	Start questionnaire time
		<i>Groups of 21 Fields of NASA survey</i>	-	-	-
		<i>Groups of 8 Fields of WP survey</i>	-	-	-
		time.2	TIME	-	Finish questionnaire time
		intensiveness	INT	11	Subjective test intensiveness
not_valid	INT	11	Flag for validation		



Enterprise Miner which are consuming data into flat file format. The query which as merge all tables together have to be designed in Implementation Chapter.

As it could be seen from Questionnaire table from 3.1, both NASA and WP questionnaires data are storing in the same table. The Machine Learning Classifiers have to be applied for each model separately, and also data investigation process will give more details and insights in case of separated investigation of NASA-TLX and WP data. Consequently, the additional requirement is to separate merged dataset into two parts.

Finally, as a result of merging data there will be many Foreign and Primary keys columns where are useless for any further objectives. They have to excluded from dataset on the very initial stage before performing any type of analysis.

## 3.4 Data Preparation

### 3.4.1 Introduction

As it was described before, the initial dataset was gathered through manual filling of NASA-TLX and WP paper questionnaires. Consequently, it is very likely that during the manual process of filling the form or during a transformation question were left unfilled or filled by inappropriate information. Particularly, it has been noticed that:

1. Many students were unwilling to give information about their student identification number, nation or age, consequently many questionnaires were incomplete. Additionally, some question regarding Workload were also leaved as blank.
2. Particularly in NASA-TLX questionnaire, in few cases in section of binary choice between two types of workload there were ticked both option, which is inappropriate for a technique and such questionnaire should be treated as not valid.

3. It happened in a quite few cases that, some students highlight a range of Workload instead of choosing the most suitable level. Such questionnaires were also rejected for further investigation or transformed to satisfy requirements.
4. Duration of a task has some unacceptable values.

### 3.4.2 Data Transformation

As it was mentioned before, during Experiment implementation stage, the Supervised Machine Learning Classifiers will be also applied for dataset which includes additional set of features. This means, that these features have low information gain have to be transform in more reliable ones.

In some cases, the format of inserted data has to be changed into union format or some of the features should be grouped together in case of a big number of distinct values.

### 3.4.3 Missing Values Handling

Missing values is a significant issue for some of Machine Learning Classifiers. As it could be seen from Literature review, Neural Networks and K-Nearest Neighborhood algorithm are absolutely intolerant to missing values. It has been decided to fill missing values on initial stage or if it is not possible, remove them from a dataset in order to avoid their influence on database statistics and models. In advance, mathematical equations used in Theoretical Driven Approached are not allow to have missing values.

There is an important assumption regarding how random are missing values. If values are missing randomly, then it is possible to replace them using synthetic values. Expectation maximization estimation techniques is able to answer this question.

If values are missing completely at random (MCAR), then there is number of

different techniques of its replacement:

- First way is to replace missing values, it the fully conditional specification method, which is represented into IBM SPSS Statistics, is able to insert missing values relying on other values incomplete record. This method is suitable for data with a likelihood of existing pattern in missing values and it gives relatively accurate results. However, it is not working when there are too many empty values for an individual record.
- In case if previous method not able to give results, the replacement with the most frequent values (mode) in case of categorical feature or with average value for quantitative feature should be used as a second choice. Such replacement is allowing to reduce the impact of such value on dataset, but the dataset also becomes less representative of real data in contrast with first method.

### 3.4.4 Outliers and incompatible values handling

The database schema doesnt have any restrictions on input values, as a consequence it may have incompatible values. Outliers could have as significant impact at all models especially at Neural Networks.

The verification of values existence in expected range have to be done. The table of such range for NASA-TLX and WP features is provided below.

The removing outliers could be performed by calculating F-value and removing all records with value more than 3.29 in absolute. IBM SPSS Statistic software is having necessary tools for performing such tasks.

Table 3.2: Value range for NASA-TLX

Feature name	Acceptable Range	
	Minimal Value	Maximal Value
NASA_mental	0	20
NASA_physical	0	20
NASA_temporal	0	20
NASA_performance	0	20
NASA_frustration	0	20
NASA_effort	0	20
NASA_temporal_or_frustration	0	2
NASA_performance_or_mental	0	2
NASA_mental_or_physical	0	2
NASA_frustration_or_performance	0	2
NASA_temporal_or_effort	0	2
NASA_physical_or_frustration	0	2
NASA_performance_or_temporal	0	2
NASA_mental_or_effort	0	2
NASA_physical_or_temporal	0	2
NASA_frustration_or_effort	0	2
NASA_physical_or_performance	0	2
NASA_temporal_or_mental	0	2
NASA_effort_or_physical	0	2
NASA_frustration_or_mental	0	2
NASA_performance_or_effort	0	2

Note: "0" for empty values

Table 3.3: Value range for Worload Profile

Feature name	Acceptable Range	
	Minimal Value	Maximal Value
WP_solving_deciding	0	20
WP_response_selection	0	20
WP_task_space	0	20
WP_verbal_material	0	20
WP_visual_resources	0	20
WP_auditory_resources	0	20
WP_manual_response	0	20
WP_speech_response	0	20

### 3.5 Descriptive Statistics

After all preliminary steps of data preparation and manipulation, alongside with achieving improvement in data quality and dataset readiness for applying Machine Learning classifiers, the activities regarding data investigation should be done. Such activities are able to give helpful insights in trends and patterns of gathered data. This information will be used as foundation for choosing the most suitable Machine Learning techniques in respect of founded dataset features and also it might help in designing partitioning for Manual Decision Trees. This steps should be done via data visualisation through the suitable tool for it as Tableau 9.1.

It is also necessary to capture descriptive statistic such us mean, skewness and kurtosis to check the assumption of normally distributed data which will be required for measuring sensitivity of obtained results.

## **3.6 Supervise Machine Learning Model Training**

### **3.6.1 Modelling techniques selection**

There are many Machine Learning techniques, many of them have been described into Literature review chapter, and it is necessary to not test all of them, but choose the most suitable once according to existed information about data and focus more on their adjustment and tuning for increasing performance. On this stage, it should be enough information for making such choice.

### **3.6.2 Test Design solutions and Model assessment**

In order estimate an effectiveness of created models the Misclassification rate will be used as a criterion for this. For estimation of such criteria the entire dataset has to be divided into Training and Validation sets. SAS notation is recommended to 80%/20% deviation in Training and Validation Sets Respectively. In additional, the removing of insignificant attributes, which could be treated as nosy data, from input of some of the Machine Learning Techniques such as Neural Networks, could potentially increase techniques performance.

This phase has to be done in SAS Enterprise Miner because during the process of adjustment parameters for each particular chosen Machine Learning Techniques they could vary from one technique to another.

## 3.7 Evaluation and model Adjustment

### 3.7.1 Measurement of Face Validity

In order find an answer of research question and compare Theoretical Driven Approaches and Machine Learning Approaches it is necessary to measure Face Validity which, in this case, is represented by Accuracy. Metz (1978) describe accuracy as the proportion of true positives and true negatives among the total amount of observed cases and it could be calculated using Confusion matrix 3.4 and by the following equation:

$$\frac{TP + TN}{TP + FP + FN + TN} \quad (3.1)$$

Table 3.4: Confusion Matrix

	Confusion matrix	Predicted condition positive	Predicted condition negative
True condition	Condition Positive	True positive (TP)	False negative (FN) (Type II error)
	Condition negative	False Positive (FP) (Type I error)	True negative (TN)

### 3.7.2 Accepting / Rejecting hypothesis

Acceptation or rejection of main and secondary hypotheses will be based on Face validity and correlation coefficient between values predicted by the best Supervise Machine Learning Classifier and values caclulated by NASA-TLX and WP equations. Comparison will be done in pairs in a such way: ML classifier(NASA-TLX) and NASA-TLX; ML classifier(WP) and WP. Hypothesie will be accepted if it ML classifier is overperform Theoretically driven approach in both face validity and Correlation coeficient

### **3.7.3 Strength and limitation of approach taken**

A detailed experiment design highlighting many issues and ways to solve them in order to achieve robust results. However, the main potential issue is amount of data for SML input. After all preparation and deleting all inappropriate questionnaires, the size of the dataset could be still not big enough in order to facilitate representative sample, especially considering the fact, that dataset will be splitted into NASA-TLX and WP which divides sample size at a half.



# Chapter 4

## Experiment implementation

### 4.1 Introduction

This Chapter perform an experiment which was designed in a previous chapter. Expectedly, the workflow is quite similar as it was for design phase. However, all steps propose the practical implementation of phases described before.

The core of an experiment is provided at figure 4.1. The initial dataset could be roughly divided into three parts: WP/NASA-TLX questionnaires, additional criteria gathered through questionnaire and the level of workload estimated by participant.

Questions from Q1 to Qn are only input for WP/NASA-TLX mathematical equations. On the other side, for ML classifiers Subjective estimated workload is also using for model creation. Second experiment will also include set of additional features as an input attributes for model generation.

Next, values predicted by ML classifiers with lowest average square error will be compare with values generated by mathematical equation in terms of face validity and correlation coefficient which will give information to prove or reject hypothesis.

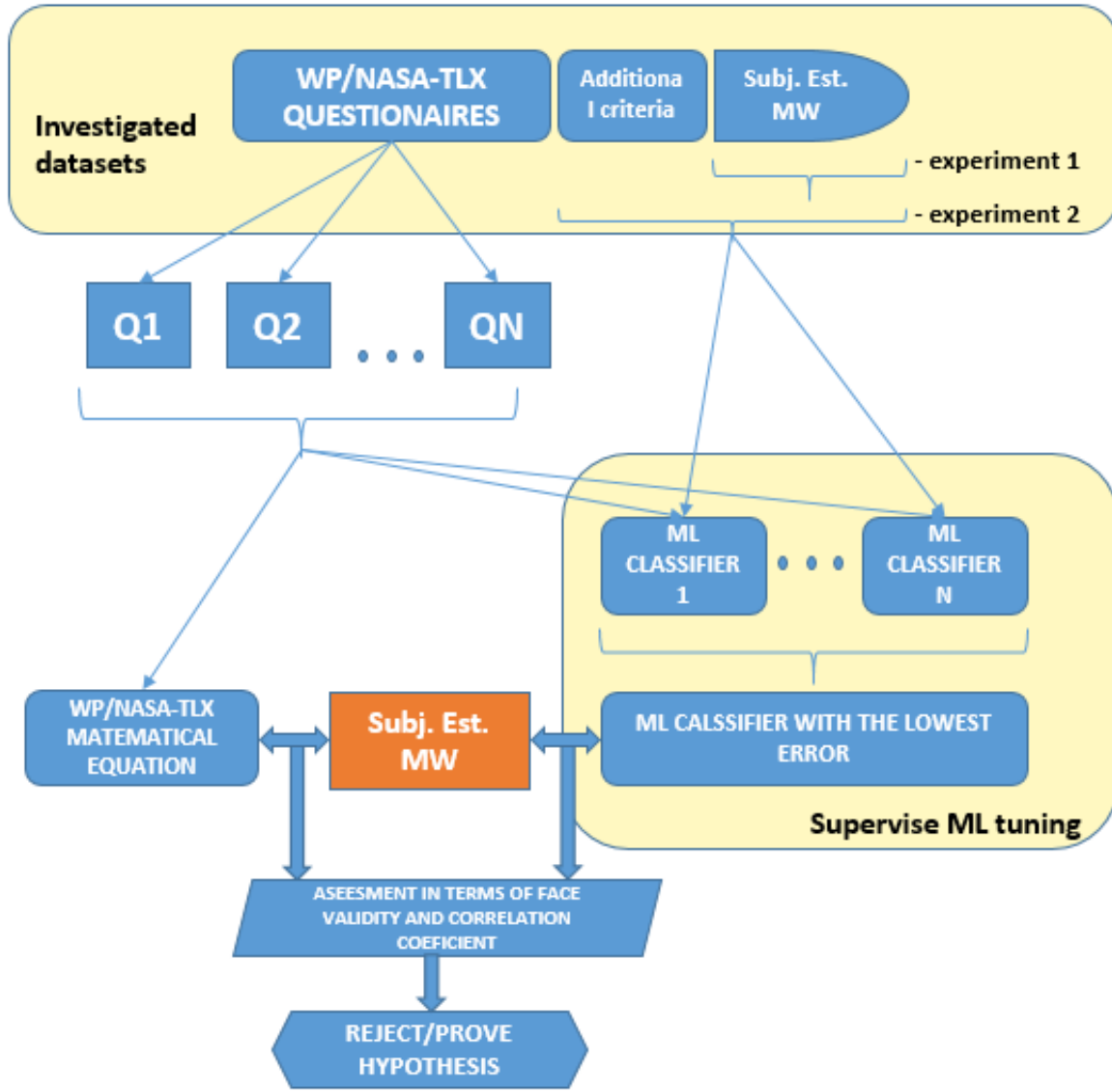


Figure 4.1: Graphical representation of an experiment

Also, Pseudocode was developed to make an experiment design more structured:

```

//data selection
SELECT only records of NASA-TLX and WP questionnaires
DIVIDE initial dataset into NASA-TLX and WP subsets
REMOVE useless features for current research (Primary and
    Secondary keys, flags)
//data transformation
FOR each record IF there are no too many missing values in a
    record THEN replace using algorithm
  
```

```
        ELSE delete such record
FOR each record TRANSFORM start and finish time INTO duration of a
    task in time format
FOR each record GROUP daytime of experiment INTO morning,
    afternoon and evening groups
FOR NASA-TLX records TRANSFORM weights into binary variables
//models generation
PERFORM data investigation in order to choose a number of suitable
    ML classifiers
DIVIDE NASA-TLX and WP datasets into five training and validation
    subsets (80% and 20% respectively)
APPLY each chosen ML classifier to each subset
CALCULATE mental workload for validation subsets using NASA-TLX
    and WP mathematical equations
//correlation coefficient for ML classifiers
CALCUCATE correlation coefficient for five NASA-TLX subsets and
    five WP subsets
EVALUATE how output from two approaches correlates with Subjective
    estimated workload from questionnaire and make a decision
    about Convergent Validity
```

In addition to Supervise Machine Learning Models it is necessary to create a solution which is calculating Mental Workload using NASA-TLX and WP mathematical formulas in order to evaluate and compare results of both approaches into Evaluation Chapter.

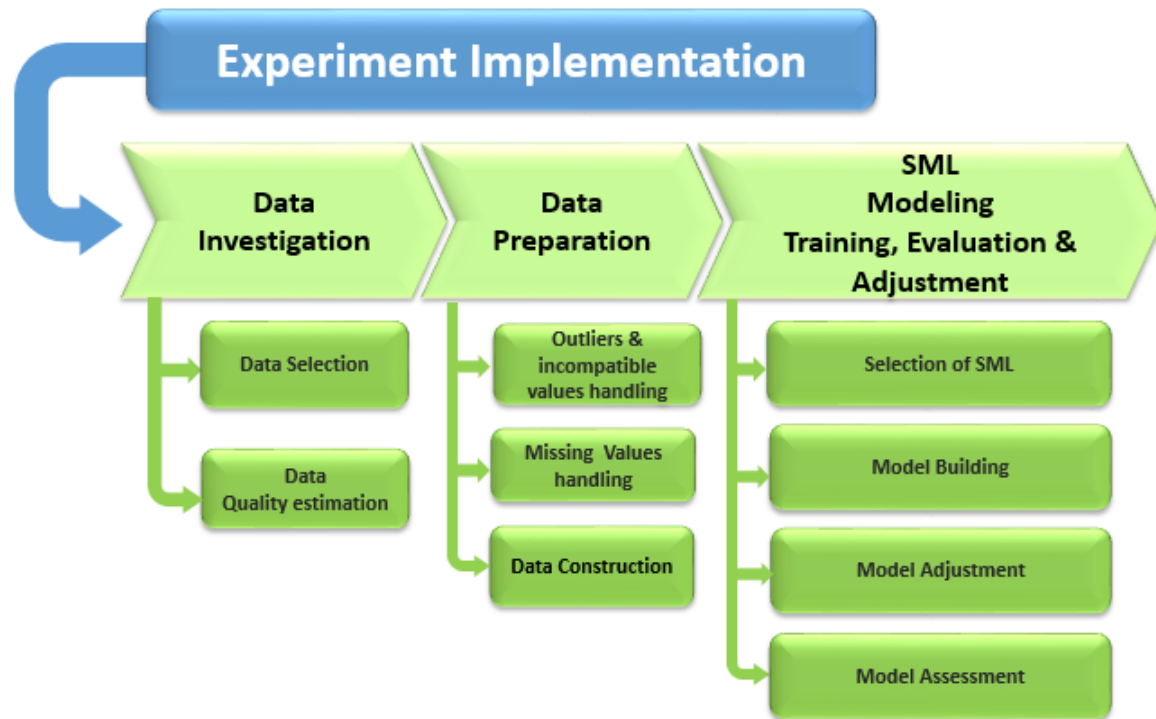


Figure 4.2: Outline of the Research Implementation Chapter

## 4.2 Data Investigation

### 4.2.1 Data Selection

As it was mentioned before, comparison between Machine Learning Classifiers and Theory driven approaches will be performing for NASA-TLX and WP separately. For this reason, datasets should be divided into two groups. It gives better understanding and clarification of datasets distinctions.

Additionally, it is compulsory to construct one flat file from all six tables for chosen data analytics software. However, this table consists IDs, Primary Secondary keys, flags, which are useless for analysis because unique and have zero information gain, consequently should be excluded from dataset at the beginning.

As a result, the initial set is represented by 46 columns and about three hundred records for NASA-TLX and WP, after executing a query from Appendix B and Appendix C. Table 4.1 and table 4.2 separating all features by type and show whether they have been removed or not.

Table 4.1: Primary, Secondary Keys and flags

Table Name	Field Name	Status
Students	student number	Removed
Students	id nationality	Removed
Nationalities	id	Removed
Lectures	id	Removed
Courses	id course	Removed
Courses	lecturer id	Removed
Tasks	task id	Removed
Tasks	course id	Removed
Questionnaire	id	Removed
Questionnaire	student number	Removed
Questionnaire	task number	Removed
Questionnaire	task id	Removed
Questionnaire	not valid	Removed

Next, NASA-TLX and WP subsets of questionnaires will be processed separately, for deviation of records into two different datasets an SQL query from appendix B has been applied to initial dataset. As a results the number of unique records for NASA-TLX and WP questionnaires about 300 and 330 respectively.

After this operation two datasets in SCV format are available for its transformation

Table 4.2: Removed features

<b>Table Name</b>	<b>Field Name</b>	<b>Status</b>
Students	age	Not removed
Nationalities	description	Not removed
Lectures	name	Not removed
Courses	description	Not removed
Tasks	description	Not removed
Tasks	date	Removed
Tasks	duration mins	Not removed
Tasks	daytime	Not removed
Questionnaire	MWL total	Not removed
Questionnaire	RSME	Not removed
Questionnaire	time 1	Not removed
Questionnaire	Groups of 21 Fields of NASA survey	Not removed
Questionnaire	Groups of 8 Fields of WP survey	Not removed
Questionnaire	time 2	Not removed
Questionnaire	intensiveness	Not removed
Questionnaire	Time 3	Added

during next step.

### 4.2.2 Data quality estimation

The feature deviation shows that country of origin for more than 60% (375) of participants. Next tree the most frequent countries are Poland, Thanzania and China with the 33, 27 and 24 participants respectively.

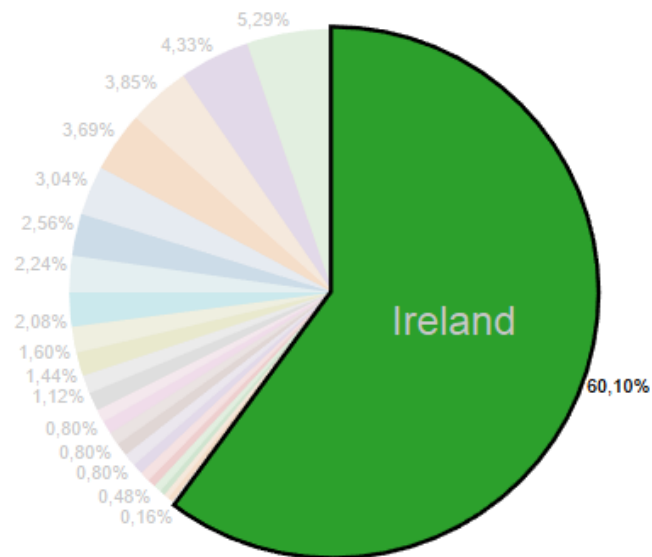


Figure 4.3: Deviation by Country of origin Pie Chart

The bar chart 4.4 showing that is only 19 participants did not fill the nation field into questionnaire.

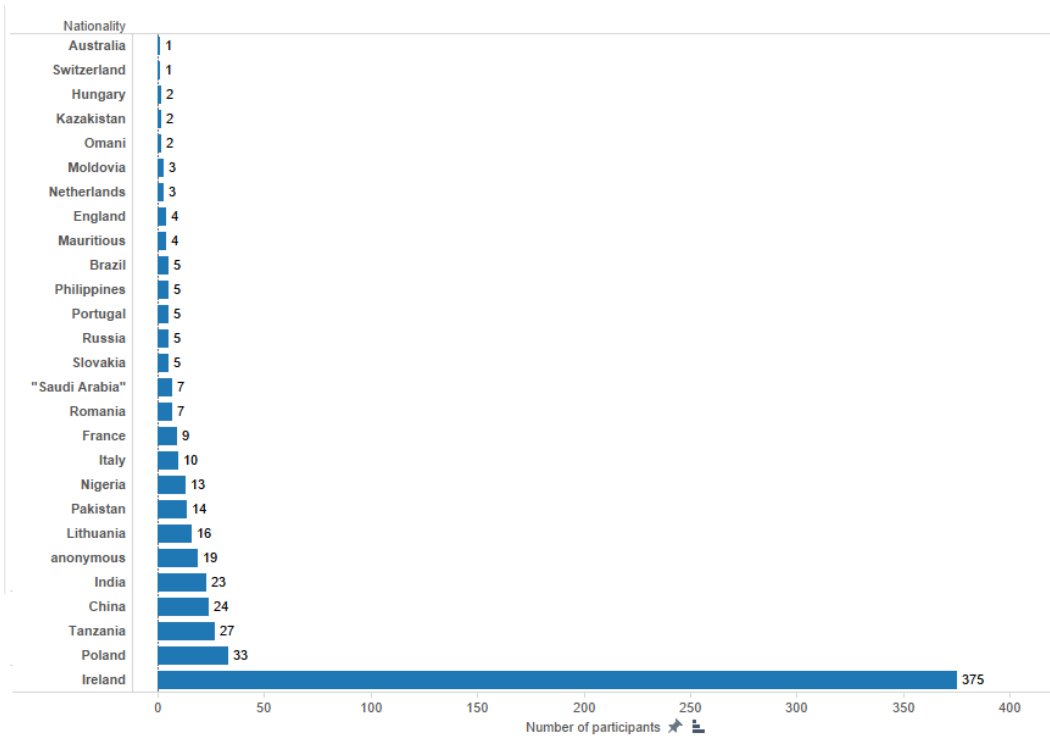


Figure 4.4: Deviation by Country of origin Bar Chart

Table 4.3 showing that weighs for NASA-TLX questionnaire have four levels. Particularly, NULL and 0 values point empty values, and 1 or 2 show whether first or second criteria have been chosen. This incontinency of labeling empty values could reduce robustness of data.



Table 4.3: Variable type and level information for NASA-TLX

Variable Label	Type	Levels
NASA_effort_or_physical	C	4
NASA_frustration_or_effort	C	4
NASA_frustration_or_mental	C	4
NASA_frustration_or_performance	C	4
NASA_mental_or_effort	C	4
NASA_mental_or_physical	C	4
NASA_performance_or_effort	C	4
NASA_performance_or_mental	C	4
NASA_performance_or_temporal	C	4
NASA_physical_or_frustration	C	4
NASA_physical_or_performance	C	4
NASA_physical_or_temporal	C	4
NASA_temporal_or_effort	C	4
NASA_temporal_or_frustration	C	3
NASA_temporal_or_mental	C	4

## 4.3 Data Preparation

### 4.3.1 Outliers and incompatible values handling

The acceptable range of values for NASA-TLX and WP questionnaires is clearly defined and it is necessary to check whether this restriction is true for current dataset.

It has been founded that 30 and 10 values were increasing the maximum threshold mentioned in table 4.4 for NASA-TLX and WP questionnaires respectively. The same issue was detected for 26 values of WP\_auditory\_resources variable; 10 incompatible values for NASA-TLX MWL and two for NASA\_mental. These values labeled as missing.

Table 4.4: Consistency of data

Variable name	Number of unacceptable values	Levels
MWL (NASA-TLX dataset)	10	4
NASA_mental	10	4
NASA_temporal_or_frustration	2	4
MWL (WP dataset)	30	4
WP_auditory_resources	26	4

After removing all incompatible values f value did not appear with values more than 3.29, which means that additional modification regarding this issue are not required.

### 4.3.2 Missing values handling

NASA-TLX missing values analysis demonstrated that all 20 questions have at least one missing value. The overall amount of cases with incomplete data is approximately 18%. The percentage of missing values is about 9%. The detailed description is provided in table 4.5 and figure 4.5

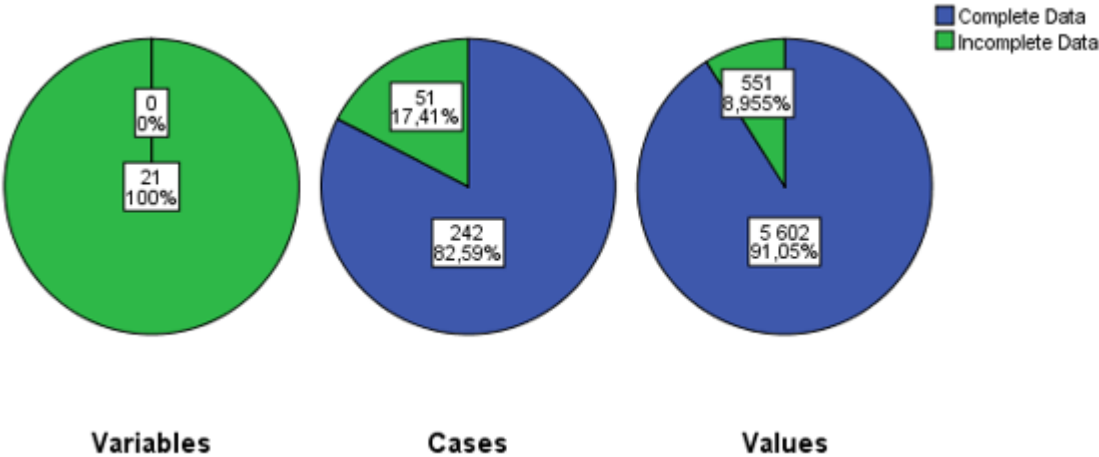


Figure 4.5: Overall Summary of Missing Values for NASA-TLX dataset

Table 4.5: Missing values statistics for NASA-TLX dataset

	Missing		Valid N
	N	Percent	
NASA_temporal_or_frustration	39	13,30%	254
NASA_effort_or_physical	38	13,00%	255
NASA_performance_or_effort	37	12,60%	256
NASA_frustration_or_mental	37	12,60%	256
NASA_temporal_or_mental	36	12,30%	257
NASA_frustration_or_effort	36	12,30%	257
NASA_physical_or_temporal	36	12,30%	257
NASA_physical_or_performance	35	11,90%	258
NASA_mental_or_effort	34	11,60%	259
NASA_performance_or_temporal	34	11,60%	259
NASA_physical_or_frustration	34	11,60%	259
NASA_temporal_or_effort	34	11,60%	259
NASA_frustration_or_performance	33	11,30%	260
NASA_mental_or_physical	33	11,30%	260
NASA_performance_or_mental	33	11,30%	260

If values are missing randomly, then It is possible to replace them using synthetic values. Expectation maximization estimation techniques is able to tell about how randomly data missed (table 4.6).

Table 4.6: EM Means NASA-TLX

NASA_effort	NASA_mental	NASA_physical	NASA_temporal	NASA_performance	NASA_frustration	NASA_performance_or_mental	NASA_mental_or_physical	NASA_frustration_or_performance	NASA_temporal_or_effort	NASA_physical_or_frustration	NASA_performance_or_temporal	NASA_mental_or_effort	NASA_physical_or_temporal	NASA_frustration_or_effort	NASA_physical_or_performance	NASA_temporal_or_mental	NASA_effort_or_physical	NASA_frustration_or_mental	NASA_performance_or_effort
9,11	11,41	6,68	8,27	7,85	6,88	0,52	0,2	0,79	0,6	0,43	0,38	0,43	0,69	0,78	0,8	0,65	0,25	0,76	0,47

a. Little's MCAR test: Chi-Square = 155,408, DF = 184, Sig. = ,938

In table 4.6 the null hypothesis for Little's MCAR test is that the data are missing completely at random (MCAR). The level of significance for NASA-TLX is equal 0,938. It is much more than 0,05 consequently the null hypothesis could not be rejected and missing values can be replaced with synthetic values.

In contrast with NASA-TLX, WP dataset has just few missing values (figure 4.6). The level of significance for Littles MCAR test is higher than threshold value as well (table 4.7).

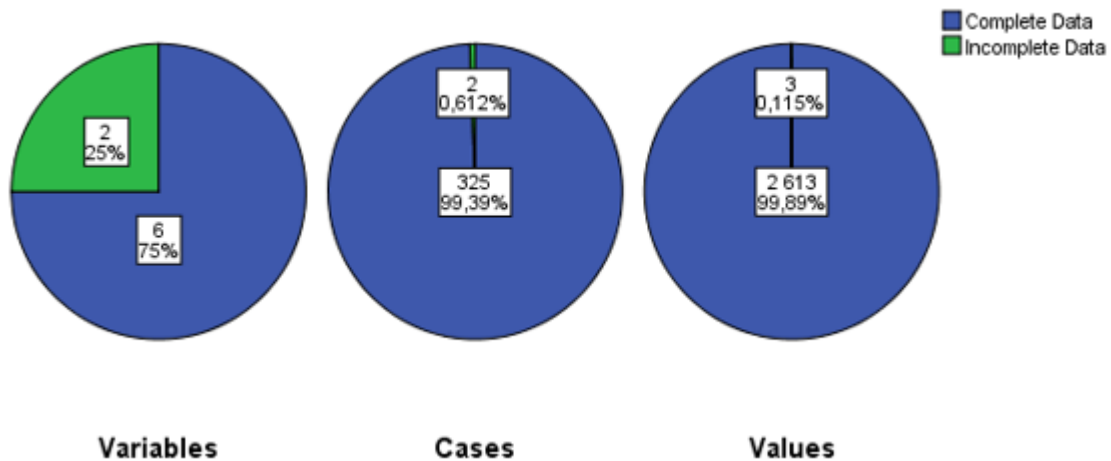


Figure 4.6: Overall Summary of Missing Values for WP dataset

Table 4.7: EM Means WP

WP_auditory_resources	WP_manual_response	WP_response_selection	WP_solving_deciding	WP_speech_response	WP_task_space	WP_verbal_material	WP_visual_resources
14,27	9,19	9,39	10,43	8,58	8,61	10,8	11,38

a. Little's MCAR test: Chi-Square = 14,542, DF = 13, Sig. = ,337

As a result, the it can be concluding that there is not pattern in how values are missing, and because of these missing values will be replacing with generated ones.

### 4.3.3 Data Construction

Fully conditional specification method was used for inserting the predictive value instead of missing for both NASA-TLX and WP questionnaires. This method suitable for data with a likelihood of existing relation between values in a dataset.

For better accuracy of method not only specific for questionnaires factors were used for predictions, but all features of dataset. It gives algorithm more data and potentially increases the accuracy of inserted values.

The new generated datasets will be used in further steps.

## Region

There are two attributes where grouping could be applied in order to improve model accuracy in terms of face validity. Both of them are categorical variables and have to many dimensions for current size of the dataset. Particularly, the country of origin has about 24 different dimensions, but as it was described before the vast majority belong to one category (Ireland) (table 4.8). It was decided to divide this category into two groups, Ireland and not Ireland. However, the dataset of both original and grouped columns will be tested as an input for a model. This operation will perform for both WP and NASA-TLX datasets.

Table 4.8: Transformation of variable "Region"

Country of origin (original)			Country of origin (grouped)
Ireland			Ireland
Saudi Arabia	Anonymous	Australia	Not Ireland
Brazil	China	England	
France	Hungary	India	
Italy	Kazakistan	Lithuania	
Mauritious	Moldovia	Netherlands	
Nigeria	Omani	Pakistan	
Poland	Portugal	Romania	
Russia	Slovakia	Switzerland	
Tanzania			

## Daytime

The second feature for consideration is daytime. This feature presented in a database into a string format, consequently the value for model design is very low. Day time

columns will be transformed into three groups according to the time when the task started. The result table of such transformation is represented bellow (table ??).

Table 4.9: Transformation of variable "Datetime"

Day time (original)			Day time (grouped)
09:42-10:50	09:43-11:19	11.	Morning
11:00 -14:00	11:10-13:15	11:10-14:00	
11:15-11:45	11:15-12:10	11:15-13:50	
11am-14pm	12:00-13:00	9-10am	
9-11am	9:10-10:30	9:10-9:57	
9:10:10:05	9:15-10:05	9:15-9:45	
9:15-9:50	9:20-10:30	9am-10am	
14:10-14:40	15:00-16:13	16-16:35	Afternoon
16-17	16:05-17:05	16:05:17:05	
16:10-17:10	16:12:16:52	16:15-16:45	
16:15-17:10	16:18-17:18	16:30-17:30	
17:13-18:03	17:15-17:43	3pm-6pm	
18:30-21:30	18:35-14:38	18:35-18:45	Evening
18:35-19-35	18:35-21:16	18:38-20:13	
18:38-21000	18:38-21:08	18:45-19:20	
18:45-19:30	18:45-19:53		



### **NASA-TLX weight**

Next, transforming of data for NASA-TLX questionnaires mentioned in before has been performed via modifying of SQL query. It has been decided to label empty values as NULL and make this attribute as binary variable by transforming 1 to 0 and 2 to 1. The query for this is presented in Appendix B.

### **Test completion duration**

There are two fields which are related to duration to the questionnaire completion. The time'1 indicating the time of starting the test and time'2 indicating the finish time. Both fields have seconds precision. They represent how long it took to compile a questionnaire, consequently for getting more value from a data the subtraction of this fields could be calculated as representation of task completion time in seconds. The resultative field called time'3 and initial two fields have been removed from a query.

## **4.4 SML Modeling Training, Evaluation & Adjustment**

### **4.4.1 Selection of SML Classifiers**

After preliminary preparation it is necessary to narrow the set of ML classifiers up to few most suitable for current experiment. Decision will be based on futures of dataset and rely on performed literature review and table 2.6.

One of the crucial things is a type of a variable because ML classifier have different assumption in relation to both dependent and independent variables. Some of the previously mentioned modifications affected variable types. The eventual types are presented in a table 4.10.

Table 4.10: Eventual data types

Table Name	Field Name	Modified?	Variable type
Students	Age	No	Discrete
Nationalities	description	Yes	Binary
Lectures	Name	Yes	Categorical
Courses	description	Yes	Categorical
Tasks	description	No	Categorical
Tasks	duration mins	No	Continuous
Tasks	daytime	Yes	Ordinal
Questionnaire	MWL total	No	Ordinal
Questionnaire	RSME	No	Ordinal
Questionnaire	time 1	Yes	Continuous
Questionnaire	Groups of 21 Fields of NASA survey	No	Ordinal/Binary
Questionnaire	Groups of 8 Fields of WP survey	No	Ordinal
Questionnaire	time 2	Yes	Continuous
Questionnaire	intensiveness	No	Ordinal
Questionnaire	Time 3	New	Continuous

Finally, considering the set of criteria it could be concluded that such factors as speed of learning and speed of classification do not have any importance for current research, because the size of the dataset is relatively small and there is no need to have fast training models. After data preparation there is no missing, irrelevant attributes or noise as well as absence of redundant features. Consequently, the choice falls into Gradient Boosting and Neural Networks because of their relatively high level of accuracy and Decision trees which will be able to give helpful insights about usefulness of each particular attribute for current research.

#### **4.4.2 Model Building, adjustment & assessment**

For the first experiment which includes only NASA-TLX/WP fields and a target variable MWL total the assumptions about type are already met. However, for the second experiment which has additional features as an input for ML classifiers the categorical features have to be modified into a set of binary dummy variables (one binary variable for each level of categorical variable), for the reason that the presence of such variables as an input for Gradient Boosting and Neural Networks could corrupt the model. The variable representing the class where the experiment took place has been modified into five binary variables; country into three and country of origin variable has been already in previous steps (table 4.11). In contrast, for Decision trees input still be used datasets with categorical variables.

As it was decided Model design and tuning will be performed through SAS Enterprise Miner. The workflow for the first experiment, which includes only features from questionnaires is designed on figure 4.7. The schema is the same for NASA-TLX and WP, the difference is only in the set of attributes.

Table 4.11: Summary of contineous attributes modifications

Previous continuous variable	New binary variable
<b>Course</b>	Web Application Architectures
	Web development and deployment
	Enterprise application development
	Data management
	Programming paradigms
<b>Daytime</b>	Morning
	Afternoon
	Evening

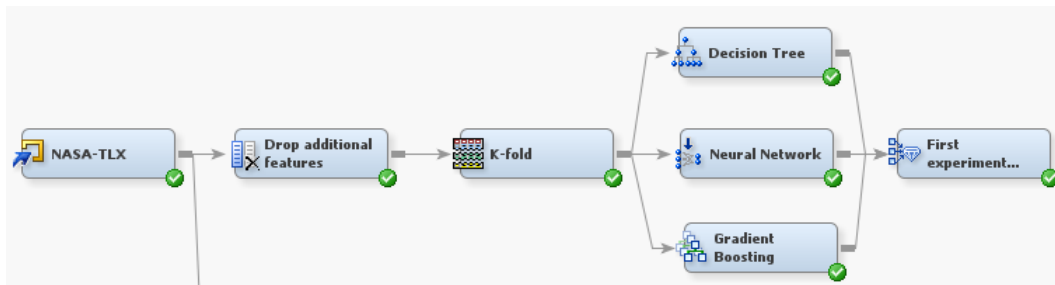


Figure 4.7: SAS Workflow for NASA-TLX dataset without additional features

The overall workflow consists of seven sections. Initially, NASA-TLX dataset prepared for analysis is uploading into application. This step is required to set attributes datatypes into proper ones and specification of target variable. For this experiment only NASA-TLX questions plus target variable are required. Consequently, any other attributes have been removed from the second stage. Before applying the chosen ML models the dataset have to be splitted into two parts. One part for model design and another part for testing purposes. As it was described model creation part is using 60%/20% of data for training and validation purposes respectively, another 20% re-

served for testing. In addition, Stratified Sampling have been chosen as partitioning method in order to prove better representatives of samples. Finally, three models such as Decision Tree, Neural Networks and Gradient Boosting have been applied and their outputs compared between each other into a last step. Average square error was a selection criteria of models success.

The results of models for NASA-TLX dataset are provided at figure 4.8

Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Test: Average Squared Error
Tree2	Tree2	Decision Tr...	MWL_total	MWL_total	7.95004
Boost	Boost	Gradient Bo...	MWL_total	MWL_total	7.975832
AutoNeural2	AutoNeural2	Neural Net...	MWL_total	MWL_total	10.97911

Figure 4.8: Model Performance for NASA-TLX dataset without additional features

Decision Tree algorithm was able to achieve to lowest average square error equal to 7.95 for Test dataset. Gradient Boosting algorithm has almost the same figure of error. Neural Networks were able to achieve only 10.8 of average square error. Results for WP are provided at figure 4.9

Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Test: Average Squared Error
Tree2	Tree2	Decision Tr...	MWL_total	MWL_total	7.95004
Boost	Boost	Gradient Bo...	MWL_total	MWL_total	7.975832
AutoNeural2	AutoNeural2	Neural Net...	MWL_total	MWL_total	10.97911

Figure 4.9: Model Performance for WP dataset without additional features

Results for WP gave less accuracy represented by Error. The most precise model is again Gradient Boosting for test dataset. However, the situation for Decision trees

is exact opposite. Particularly, Decision trees have the lowest error among all three models.

### 4.4.3 Model Building with additional input attributes

Schema for the second experiment is different. As it was mentioned, categorical variables could corrupt Neural Networks and Gradient Boosting models, to avoid these additional dummy variables have been designed. Datasets for current experiment are containing both sets of additional variables. Consequently, after dataset partitioning, only set of necessary attributes used as an input relying on model type. The schema is similar for NASA-TLX and WP and represented at figure 4.10.

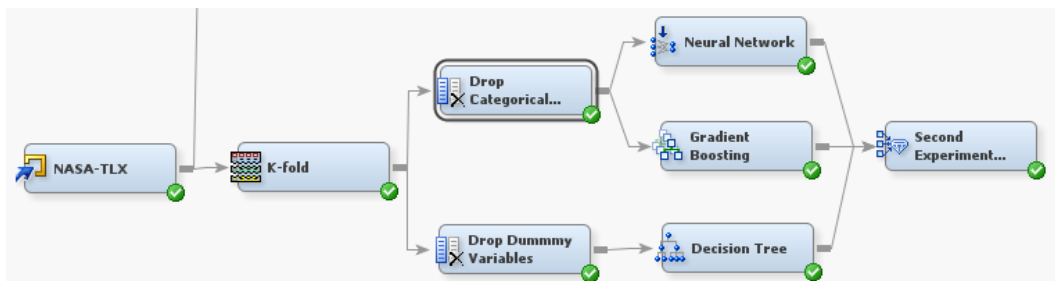


Figure 4.10: SAS Workflow for NASA-TLX dataset with additional features

As is could be seen from figure 4.11. The set of additional features in NASA-TLX dataset was not able only improve results for any algorithm.

Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Test: Average Squared Error
Tree	Tree	Decision Tr...	MWL_total	MWL_total	7.95004
Boost2	Boost2	Gradient Bo...	MWL_total	MWL_total	7.98589
AutoNeural	AutoNeural	Neural Net...	MWL_total	MWL_total	11.06317

Figure 4.11: Model Performance for NASA-TLX dataset with additional features

WP results slightly increased for Gradient boosting algorithms, but Neural Networks and Decision Trees showed almost the same figure of average square error.

Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Test: Average Squared Error
Boost2	Boost2	Gradient Bo...	MWL_total	MWL_total	7.63301
AutoNeural	AutoNeural	Neural Net...	MWL_total	MWL_total	9.180638
Tree	Tree	Decision Tr...	MWL_total	MWL_total	9.303528

Figure 4.12: Model Performance for WP dataset with additional features

Finally, it can be concluded that Gradient Boosting algorithm was able to demonstrate the most stable and precise results for NASA-TLX and WP datasets in both experiments. Consequently, this algorithm could be chosen for comparison with results from Theoretical driven approaches in a next chapter.

# Chapter 5

## Evaluation

### 5.1 Introduction

This chapter is evaluating results which were achieved in a previous chapter. The comparison of results between Theatrically driven approaches and Machine Learning approaches in prediction power of Subjective estimated workload is described here. Gradient Boosting Algorithm has been chosen as the most powerful model with the most stable results and low figure for average square error.

For such comparison it is necessary to calculate Workload using mathematical equations describes during Literature review. Additionally, comparison of correlation between predicted and target variables will be performed for both NASA-TLX and WP datasets.



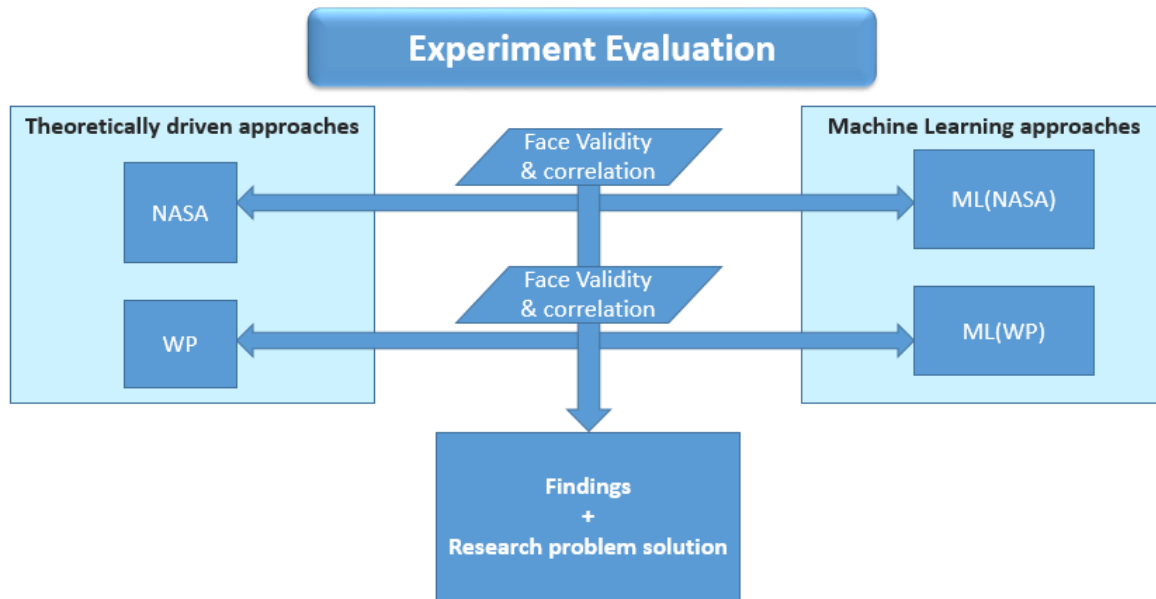


Figure 5.1: Outline of the Evaluation Chapter

## 5.2 Calculation of Workload NASA-TLX and WP equations

Initial datasets we divided into parts for model generation and model testing in proportion of 80%/20% respectively. Consequently, for fair experiment only the same 20% of rows will be used.

SAS Enterprise miner has an opportunity to see which records it has randomly chosen for testing dataset. Such records have been copied into excel sheet and using NASA-TLX and WP equations next average square errors have been calculated:

NASA-TLX average square error - 11,84342

WP average square error - 10,01450033

Table 5.1 gives relative comparison of two approaches:

Table 5.1: Average square error comparison

	ML average square error (Gradient Boosting)	Average square error derived from equations	Improvement percentage
NASA-TLX	7.796	11.84	34%
WP	8.67	10.01	13.40%

As it can be seen from a table, ML algorithm showed much better results for NASA-TLX questionnaires, rather than WP. However, ML model for both questionnaires over perform results achieved by mathematical approaches.

Next, these algorithms will be compared in degree of correlation between predicted and target variable (table 5.2, 5.3).

Table 5.2: Correlation between outputs of ML classifier and NASA-TLX equation

<b>Correlations</b>				
		MWL	ML_Predicted_MWL	EQ_Predicted_MWL
MWL	Pearson Correlation	1	,705**	,629**
	Sig. (2-tailed)		0	0
	N	58	58	58
ML_Predicted_MWL	Pearson Correlation	,705**	1	,668**
	Sig. (2-tailed)	0		0
	N	58	58	58
EQ_Predicted_MWL	Pearson Correlation	,629**	,668**	1
	Sig. (2-tailed)	0	0	
	N	58	58	58

\*\* . Correlation is significant at the 0.01 level (2-tailed).

Table 5.3: Correlation between outputs of ML classifier and WP equation

**Correlations**

		MWL	ML_Pred	EQ_Pred
MWL	Pearson Correlation	1	,434**	,457**
	Sig. (2-tailed)		0	0
	N	66	66	66
ML_Pred	Pearson Correlation	,434**	1	,834**
	Sig. (2-tailed)	0		0
	N	66	66	66
EQ_Pred	Pearson Correlation	,457**	,834**	1
	Sig. (2-tailed)	0	0	
	N	66	66	66

\*\* . Correlation is significant at the 0.01 level (2-tailed).

Results of correlation are provided below. It has been founded that correlation level is significant for all four cases at 0.01 level. For NASA-TLX questionnaire the correlation between values predicted by ML models is equal to 7.05 which is more than 10 percent higher than value calculated by mathematical equation. However, for WP the situation is opposite and ML model is less correlated to target variable then calculated value (table 5.4).

Table 5.4: Correlation comparison

	ML predicted value target value correlation	Calculated from equation target value correlation	Improvement percentage
NASA-TLX	0.705	0.629	10.8%
WP	0.434	0.457	-5.2%

In conclusion, according to available data Supervise Machine learning classifiers outstrip Theory Driven approaches NASA-TLX and WP in terms of Face Validity. SML models for NASA-TLX questionnaires are also have higher Correlation Coefficient, whereas SML models for WP questionnaire less correlated with values calculated by mathematical equations. Hypothesis H01 is accepted, wherease H02 is rejected.

### 5.2.1 Strength and limitation of the experiment

Highly detailed experiment design was developed and performed in previous two chapters. It was considering a lot of issue which could affect an experiment results. Presence of Graphs and Pseudocode did help to perform experiment without losing the right path. However, the awareness is that despite of the fact that hypothesis was clearly designed, it was not mentioned how significant should be raise in degree of face validity and correlation coefficient in order to accept hypotheses. This issue could be addressed during further work.

# Chapter 6

## Conclusion

### 6.1 Research Overview

This research was trying to extend the knowledge about such phenomena as Mental Workload. It is basing on assumption about high level of its importance and wide range of possible implementations for modern technological society.

From a number of different approaches to measure MWL it has been decided to choose Self estimated measures as the Most suitable for current study. It was decided to compare Machine learning models with two theoretical Driven approaches called NASA-TLX and Workload provide.

It has been find out that Machine learning models are able to give more robust results in prediction of Subjective Estimated Workload rather than NASA-TLX and WP Mathematical equations. Such decision based on Average Square Error value and correlation coefficient among ML and Theoretical Driven Approaches. Despite of observing improving it was decided that prediction quality could be still improved in big manner in case of providing bigger sample size for model generation.

Alongside with main research question, it attempts find additional insights into

gathered data was performed by executing second experiment. As it has been founded, additional features as age, daytime of an experiment and language of country of origin were not able to improve model performance in major way to gain new knowledge in such area as Mental Workload.

## 6.2 Problem Definition

The main purpose for current research was developed the most robust machine Learning model and compare Subjective estimated workload and values calculated by NASA-TLX and WP in terms of face validity. In order to find the answer, it was necessary to overcome a sequence of issues. They could be described as follows:

- Investigate methods for Measuring Mental workload
- Identify dataset issues such as missing values, inconsistencies, noisy data.
- Choose statistical methods to solve dataset issues.
- Choose software which could face perform previously mentioned methods
- Find out dataset features and properties
- Identify the most suitable techniques according to the features
- Identify instrument to improve models efficient

## 6.3 Design/Experimentation, Evaluation & Results

In relation to formulated hypothesis. according to tables 5.4 and 5.1 we can accept hypotheses H01 that according to gathered data, SML classifier oustrip Theoretically driven approach for NASA-TLX questionair, but reject hypotheses H02 because correlation coefficient is lower for ML classifier for WP questionnaire.

## 6.4 Contributions and impact

Finally, after getting the results of an experiment, listed below findings are able to contribute for the body of Knowledge:

- Implementing approach of applying Machine Learning models to the subject of Mental Workload.
- Showing that standard approaches of Mental workload calculation could be extended by opportunities of Machine Learning models.
- Raw data were significantly processed. This could everyone who want to continue exploring knowledge from data a fast start.

## 6.5 Future Work & recommendations

Machine Learning classifiers require relatively big dataset for designing quality models, gathering additional data will defiantly give more space for model designing.

The idea of Supervise Machine learning could be developing in a future researches. Mental workload as a natural problem could be influenced by hundred and hundred factors and as a consequence ML models is a good way of processing such amount of information. In current research only self-estimated approach was used, but combining this approach with actual participant performance would be able to give wider picture about experienced Workload.

# References

- Aslan, B. G., & Inceoglu, M. M. (2007). Machine learning based learner modeling for adaptive web-based learning. In O. Gervasi & M. L. Gavrilova (Eds.), *Computational science and its applications – iccsa 2007: International conference, kuala lumpur, malaysia, august 26-29, 2007. proceedings, part i* (pp. 1133–1145). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from [http://dx.doi.org/10.1007/978-3-540-74472-6\\_94](http://dx.doi.org/10.1007/978-3-540-74472-6_94) doi: 10.1007/978-3-540-74472-6\_94
- Boff, K. R., Kaufman, L., & Thomas, J. P. (1986). Handbook of perception and human performance.
- Bornstein, R. F., Rossner, S. C., Hill, E. L., & Stepanian, M. L. (1994). Face validity and fakability of objective and projective measures of dependency. *Journal of Personality Assessment*, 63(2), 363–386.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). Crisp-dm 1.0. *CRISP-DM Consortium*.
- 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.
- Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on machine learning* (pp. 160–167).



## REFERENCES

---

- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273–297.
- Date, C. J. (1999). *An introduction to database systems (introduction to database systems)*. {Addison Wesley Longman}.
- De Martino, M., Bertone, A., Albertoni, R., Hauska, H., Demsar, U., & Dunkars, M. (2002). *Technical report of data mining* (Tech. Rep.). Tech. Rep., European Commission, IST-2000-29640, INVISIP Project Derivable 2.2, Bruxelles.
- Dondio, P., & Longo, L. (2011). Trust-based techniques for collective intelligence in social search systems. In N. Bessis & F. Xhafa (Eds.), *Next generation data technologies for collective computational intelligence* (pp. 113–135). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from [http://dx.doi.org/10.1007/978-3-642-20344-2\\_5](http://dx.doi.org/10.1007/978-3-642-20344-2_5) doi: 10.1007/978-3-642-20344-2\_5
- Downing, S. M. (2006). Face validity of assessments: faith-based interpretations or evidence-based science? *Medical education*.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3), 37.
- Han, J., Altman, R. B., Kumar, V., Mannila, H., & Pregibon, D. (2002). Emerging scientific applications in data mining. *Communications of the ACM*, 45(8), 54–58.
- Hancock, P. A., & Meshkati, N. E. (1988). *Human mental workload*. North-Holland.
- 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. *Advances in psychology*, 52, 139–183.
- Haykin, S. S., Haykin, S. S., Haykin, S. S., & Haykin, S. S. (2009). *Neural networks and learning machines* (Vol. 3). Pearson Upper Saddle River, NJ, USA:.

## REFERENCES

---

- Holden, R. R., & Jackson, D. N. (1979). Item subtlety and face validity in personality assessment. *Journal of Consulting and Clinical Psychology, 47*(3), 459.
- Holden, R. R., & Jackson, D. N. (1985). Disguise and the structured self-report assessment of psychopathology: I. an analogue investigation. *Journal of Consulting and Clinical Psychology, 53*(2), 211.
- Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR), 31*(3), 264–323.
- Kara, Y., Boyacioglu, M. A., & Baykan, Ö. K. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the istanbul stock exchange. *Expert systems with Applications, 38*(5), 5311–5319.
- Kotsiantis, S. (2007). Supervised machine learning: A review of classification techniques. *Informatica, 31*, 249–268.
- Langdrige, D., & Hagger-Johnson, G. (2009). *Introduction to research methods and data analysis in psychology*. Pearson Education.
- Linoff, G. S., & Berry, M. J. (2011). *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons.
- Long, P. M., & Servedio, R. A. (2010). Random classification noise defeats all convex potential boosters. *Machine Learning, 78*(3), 287–304.
- Longo, L. (2011). Human-computer interaction and human mental workload: Assessing cognitive engagement in the world wide web. In *Ifip conference on human-computer interaction* (pp. 402–405).
- 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).

## REFERENCES

---

- Longo, L. (2014). *Formalising human mental workload as a defeasible computational concept. trinity college.* (Unpublished doctoral dissertation). Trinity College.
- Longo, L. (2015a). A defeasible reasoning framework for human mental workload representation and assessment. *Behaviour & Information Technology*, *34*(8), 758.
- 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 *Computer-based medical systems (cbms), 2016 ieee 29th international symposium on* (pp. 106–111).
- 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., Dondio, P., & Barrett, S. (2010). Transactions on computational collective intelligence ii. In N. T. Nguyen & R. Kowalczyk (Eds.), (pp. 46–69). Berlin, Heidelberg: Springer-Verlag. Retrieved from <http://dl.acm.org/citation.cfm?id=1985614.1985617>
- Longo, L., & Hederman, L. (2013). Argumentation theory for decision support in health-care: A comparison with machine learning. In K. Imamura, S. Usui, T. Shirao, T. Kasamatsu, L. Schwabe, & N. Zhong (Eds.), *Brain and health informatics: International conference, bhi 2013, maebashi, japan, october 29-31, 2013. proceedings* (pp. 168–180). Cham: Springer International Publishing. Retrieved from [http://dx.doi.org/10.1007/978-3-319-02753-1\\_17](http://dx.doi.org/10.1007/978-3-319-02753-1_17) doi: 10.1007/978-3-319-02753-1\_17
- Longo, L., Rusconi, F., Noce, L., & Barrett, S. (2012). The importance of human mental workload in web design. In *Webist* (pp. 403–409).

## REFERENCES

---

- Maldonado, M., Dean, J., Czika, W., & Haller, S. (2014). Leveraging ensemble models in sas® enterprise miner. In *Proceedings of the sas global forum 2014 conference*.
- Meier, T. B., Desphande, A. S., Vergun, S., Nair, V. A., Song, J., Biswal, B. B., . . . Prabhakaran, V. (2012). Support vector machine classification and characterization of age-related reorganization of functional brain networks. *Neuroimage*, *60*(1), 601–613.
- Metz, C. E. (1978). Basic principles of roc analysis. In *Seminars in nuclear medicine* (Vol. 8, pp. 283–298).
- Muckler, F. A., & Seven, S. A. (1992). Selecting performance measures:” objective” versus” subjective” measurement. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *34*(4), 441–455.
- Nyce, C., & CPCU, A. (2007). Predictive analytics white paper. *American Institute for CPCU. Insurance Institute of America*, 9–10.
- O’Donnell, R. D., & Eggemeier, F. T. (1986). Workload assessment methodology.
- Recarte, M. A., & Nunes, L. M. (2003). Mental workload while driving: effects on visual search, discrimination, and decision making. *Journal of experimental psychology: Applied*, *9*(2), 119.
- Reid, G. B., & Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. *Advances in psychology*, *52*, 185–218.
- 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).
- 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.

## REFERENCES

---

- Srinivasan, K., & Fisher, D. (1995, Feb). Machine learning approaches to estimating software development effort. *IEEE Transactions on Software Engineering*, *21*(2), 126-137. doi: 10.1109/32.345828
- Tsang, P. S., & Velazquez, V. L. (1996). Diagnosticity and multidimensional subjective workload ratings. *Ergonomics*, *39*(3), 358–381.
- Weiner, I. B., & Craighead, W. E. (2010). *The corsini encyclopedia of psychology* (Vol. 4). John Wiley & Sons.
- Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... others (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, *14*(1), 1–37.
- 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.
- Yang, J., & Wang, Y. L. (2012). A new outlier detection algorithms based on markov chain. In *Advanced materials research* (Vol. 366, pp. 456–459).
- Yuan, Y.-x. (2008). Step-sizes for the gradient method. *AMS IP Studies in Advanced Mathematics*, *42*(2), 785.

# Appendix A

## NASA-TLX questionnaire

Figure 8.6

### **NASA Task Load Index**

*Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.*


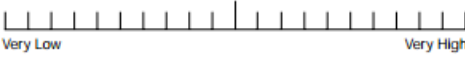

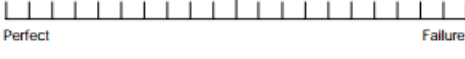
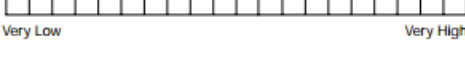
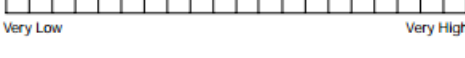
Name	Task	Date
Mental Demand	How mentally demanding was the task?	
Very Low		Very High
Physical Demand	How physically demanding was the task?	
Very Low		Very High
Temporal Demand	How hurried or rushed was the pace of the task?	
Very Low		Very High
Performance	How successful were you in accomplishing what you were asked to do?	
Perfect		Failure
Effort	How hard did you have to work to accomplish your level of performance?	
Very Low		Very High
Frustration	How insecure, discouraged, irritated, stressed, and annoyed were you?	
Very Low		Very High

Figure A.1: Recent version of NASA-TLX assessment technique

# Appendix B

## NASA-TLX SQL query

```
SELECT
s.age,
CASE
WHEN n.description = 'Ireland' THEN 1
ELSE 0
END isEnglishSpeaking,
l.name,
c.description,
t.description,
t.duration_mins,
CASE
WHEN t.daytime = '9am-10am' THEN 'Morning'
WHEN t.daytime = '9:20-10:30' THEN 'Morning'
WHEN t.daytime = '9:15-9:50' THEN 'Morning'
WHEN t.daytime = '9:15-9:45' THEN 'Morning'
WHEN t.daytime = '9:15-10:05' THEN 'Morning'
WHEN t.daytime = '9:10:10:05' THEN 'Morning'
WHEN t.daytime = '9:10-9:57' THEN 'Morning'
WHEN t.daytime = '9:10-10:30' THEN 'Morning'
WHEN t.daytime = '9-11am' THEN 'Morning'
WHEN t.daytime = '18:35-21:16' THEN 'Evening'
WHEN t.daytime = '18:45-19:53' THEN 'Evening'
WHEN t.daytime = '18:45-19:53' THEN 'Evening'
```

## APPENDIX B. NASA-TLX SQL QUERRY

```
WHEN t.daytime = '18:45-19:20' THEN 'Evening'
WHEN t.daytime = '18:38-21:08' THEN 'Evening'
WHEN t.daytime = '18:38-21000' THEN 'Evening'
WHEN t.daytime = '18:38-20:13' THEN 'Evening'
WHEN t.daytime = '18:35-19-35' THEN 'Evening'
WHEN t.daytime = '18:35-18:45 - pre questions 18:45-19:35 - class
19:35-21:25 - lab' THEN 'Evening'
WHEN t.daytime = '18:35-14:38' THEN 'Evening'
WHEN t.daytime = '18:30-21:30' THEN 'Evening'
WHEN t.daytime = '18:45-19:30' THEN 'Evening'
WHEN t.daytime = '9-10am' THEN 'Morning'
WHEN t.daytime = '3pm-6pm' THEN 'Afternoon'
WHEN t.daytime = '17:15-17:43' THEN 'Afternoon'
WHEN t.daytime = '17:13-18:03' THEN 'Afternoon'
WHEN t.daytime = '16:30-17:30' THEN 'Afternoon'
WHEN t.daytime = '16:18-17:18' THEN 'Afternoon'
WHEN t.daytime = '16:15-17:10' THEN 'Afternoon'
WHEN t.daytime = '16:15-16:45' THEN 'Afternoon'
WHEN t.daytime = '16:12:16:52' THEN 'Afternoon'
WHEN t.daytime = '16:10-17:10' THEN 'Afternoon'
WHEN t.daytime = '16:05:17:05' THEN 'Afternoon'
WHEN t.daytime = '16:05-17:05' THEN 'Afternoon'
WHEN t.daytime = '16-17' THEN 'Afternoon'
WHEN t.daytime = '16-16:35' THEN 'Afternoon'
WHEN t.daytime = '15:00-16:13' THEN 'Afternoon'
WHEN t.daytime = '14:10-14:40' THEN 'Afternoon'
WHEN t.daytime = '12:00-13:00' THEN 'Afternoon'
WHEN t.daytime = '11am-14pm' THEN 'Morning'
WHEN t.daytime = '11:15-13:50' THEN 'Morning'
WHEN t.daytime = '11:15-12:10' THEN 'Morning'
WHEN t.daytime = '11:15-11:45' THEN 'Morning'
WHEN t.daytime = '11:10-14:00' THEN 'Morning'
WHEN t.daytime = '11:10-13:15' THEN 'Morning'
WHEN t.daytime = '11:00 -14:00' THEN 'Morning'
WHEN t.daytime = '11-12' THEN 'Morning'
WHEN t.daytime = '09:43-11:19' THEN 'Morning'
```



## APPENDIX B. NASA-TLX SQL QUERRY

---

```
WHEN t.daytime = '09:42-10:50' THEN 'Morning'
END daytime,
CASE
WHEN time_1 = '00:00:00' THEN NULL
WHEN time_2 = '00:00:00' THEN NULL
ELSE time_2 - time_1
END time_3,
q.MWL_total,
q.rsme,
CASE
WHEN q.WP_auditory_resources = 0 THEN NULL
ELSE WP_auditory_resources
END WP_auditory_resources,
CASE
WHEN q.WP_manual_response = 0 THEN NULL
ELSE WP_manual_response
END WP_manual_response,
CASE
WHEN q.WP_response_selection = 0 THEN NULL
ELSE WP_response_selection
END WP_response_selection,
CASE
WHEN q.WP_solving_deciding = 0 THEN NULL
ELSE WP_solving_deciding
END WP_solving_deciding,
CASE
WHEN q.WP_speech_response = 0 THEN NULL
ELSE WP_speech_response
END WP_speech_response,
CASE
WHEN q.WP_task_space = 0 THEN NULL
ELSE WP_task_space
END WP_task_space,
CASE
WHEN q.WP_verbal_material = 0 THEN NULL
ELSE WP_verbal_material
```

## APPENDIX B. NASA-TLX SQL QUERRY

---

```
END WP_verbal_material ,
CASE
WHEN q.WP_visual_resources = 0 THEN NULL
ELSE WP_visual_resources
END WP_visual_resources ,
q.intrusiveness
FROM questionnaire q
INNER JOIN students s
ON s.student_number = q.student_number
INNER JOIN nationalities n
ON n.id = s.id_nationality
INNER JOIN tasks t
ON t.task_id = q.task_number
INNER JOIN courses c
ON c.id_course = t.course_id
INNER JOIN lecturers l
ON l.id = c.lecturer_id
WHERE q.AT_mental IS NULL
AND q.AT_parallelism IS NULL
AND q.WP_auditory_resources IS NOT NULL
OR q.WP_manual_response IS NOT NULL;
```

# Appendix C

## WP SQL query

```
SELECT
s.age,
CASE
WHEN n.description = 'Ireland' THEN 1
ELSE 0
END isEnglishSpeaking,
l.name,
c.description,
t.description,
t.duration_mins,
CASE
WHEN t.daytime = '9am-10am' THEN 'Morning'
WHEN t.daytime = '9:20-10:30' THEN 'Morning'
WHEN t.daytime = '9:15-9:50' THEN 'Morning'
WHEN t.daytime = '9:15-9:45' THEN 'Morning'
WHEN t.daytime = '9:15-10:05' THEN 'Morning'
WHEN t.daytime = '9:10:10:05' THEN 'Morning'
WHEN t.daytime = '9:10-9:57' THEN 'Morning'
WHEN t.daytime = '9:10-10:30' THEN 'Morning'
WHEN t.daytime = '9-11am' THEN 'Morning'
WHEN t.daytime = '18:35-21:16' THEN 'Evening'
WHEN t.daytime = '18:45-19:53' THEN 'Evening'
```

## APPENDIX C. WP SQL QUERY

---

```
WHEN t.daytime = '18:45-19:53' THEN 'Evening'
WHEN t.daytime = '18:45-19:20' THEN 'Evening'
WHEN t.daytime = '18:38-21:08' THEN 'Evening'
WHEN t.daytime = '18:38-21000' THEN 'Evening'
WHEN t.daytime = '18:38-20:13' THEN 'Evening'
WHEN t.daytime = '18:35-19-35' THEN 'Evening'
WHEN t.daytime = '18:35-18:45 - pre questions 18:45-19:35 - class
19:35-21:25 - lab' THEN 'Evening'
WHEN t.daytime = '18:35-14:38' THEN 'Evening'
WHEN t.daytime = '18:30-21:30' THEN 'Evening'
WHEN t.daytime = '18:45-19:30' THEN 'Evening'
WHEN t.daytime = '9-10am' THEN 'Morning'
WHEN t.daytime = '3pm-6pm' THEN 'Afternoon'
WHEN t.daytime = '17:15-17:43' THEN 'Afternoon'
WHEN t.daytime = '17:13-18:03' THEN 'Afternoon'
WHEN t.daytime = '16:30-17:30' THEN 'Afternoon'
WHEN t.daytime = '16:18-17:18' THEN 'Afternoon'
WHEN t.daytime = '16:15-17:10' THEN 'Afternoon'
WHEN t.daytime = '16:15-16:45' THEN 'Afternoon'
WHEN t.daytime = '16:12:16:52' THEN 'Afternoon'
WHEN t.daytime = '16:10-17:10' THEN 'Afternoon'
WHEN t.daytime = '16:05:17:05' THEN 'Afternoon'
WHEN t.daytime = '16:05-17:05' THEN 'Afternoon'
WHEN t.daytime = '16-17' THEN 'Afternoon'
WHEN t.daytime = '16-16:35' THEN 'Afternoon'
WHEN t.daytime = '15:00-16:13' THEN 'Afternoon'
WHEN t.daytime = '14:10-14:40' THEN 'Afternoon'
WHEN t.daytime = '12:00-13:00' THEN 'Afternoon'
WHEN t.daytime = '11am-14pm' THEN 'Morning'
WHEN t.daytime = '11:15-13:50' THEN 'Morning'
WHEN t.daytime = '11:15-12:10' THEN 'Morning'
WHEN t.daytime = '11:15-11:45' THEN 'Morning'
WHEN t.daytime = '11:10-14:00' THEN 'Morning'
WHEN t.daytime = '11:10-13:15' THEN 'Morning'
WHEN t.daytime = '11:00 -14:00' THEN 'Morning'
WHEN t.daytime = '11-12' THEN 'Morning'
```

## APPENDIX C. WP SQL QUERRY

---

```
WHEN t.daytime = '09:43-11:19' THEN 'Morning'
WHEN t.daytime = '09:42-10:50' THEN 'Morning'
END daytime,
CASE
WHEN time_1 = '00:00:00' THEN NULL
WHEN time_2 = '00:00:00' THEN NULL
ELSE time_2 - time_1
END time_3,
q.MWL_total,
q.rsme,
CASE
WHEN q.WP_auditory_resources = 0 THEN NULL
ELSE WP_auditory_resources
END WP_auditory_resources,
CASE
WHEN q.WP_manual_response = 0 THEN NULL
ELSE WP_manual_response
END WP_manual_response,
CASE
WHEN q.WP_response_selection = 0 THEN NULL
ELSE WP_response_selection
END WP_response_selection,
CASE
WHEN q.WP_solving_deciding = 0 THEN NULL
ELSE WP_solving_deciding
END WP_solving_deciding,
CASE
WHEN q.WP_speech_response = 0 THEN NULL
ELSE WP_speech_response
END WP_speech_response,
CASE
WHEN q.WP_task_space = 0 THEN NULL
ELSE WP_task_space
END WP_task_space,
CASE
WHEN q.WP_verbal_material = 0 THEN NULL
```

```
ELSE WP_verbal_material
END WP_verbal_material,
CASE
WHEN q.WP_visual_resources = 0 THEN NULL
ELSE WP_visual_resources
END WP_visual_resources,
q.intrusiveness
FROM questionnaire q
INNER JOIN students s
ON s.student_number = q.student_number
INNER JOIN nationalities n
ON n.id = s.id_nationality
INNER JOIN tasks t
ON t.task_id = q.task_number
INNER JOIN courses c
ON c.id_course = t.course_id
INNER JOIN lecturers l
ON l.id = c.lecturer_id
WHERE q.AT_mental IS NULL
AND q.AT_parallelism IS NULL
AND q.WP_auditory_resources IS NOT NULL
OR q.WP_manual_response IS NOT NULL;
```