**Final Year Project: Interim Report**

**Dylan Drein | 13344741**

1. Goals of your Project
   1. Research existing material relating to the project brief. (MouStress paper) and determine the best approach to take for this project based on work already carried out.
   2. Learn Python and the basic principles of the programming language.
   3. Become proficient with Python’s numerous packages and related analytical libraries and software (Scikit-learn, NumPy, SciPy, Jupyter Notebooks etc.)
   4. Research statistical analysis approaches, learn various statistical methodologies and determine how best to apply them in a Machine Learning context to carry out the aim of measuring/detecting stress in the user using mouse movements.
   5. Schedule meeting with a Maths advisor.
   6. Clean up data and remove non-essential fields from the data set to optimise analysis and ideally reduce file size as much as possible. Also split original data file into separate files as required.
   7. Analyse data using previously researched statistical approaches, attempt to find meaningful results from this analysis i.e. Correlation between mouse movement and respondent’s situation and stress levels.
   8. Based on results from the previous step, attempt to apply Machine Learning principles to the data set.
   9. Summarise all of the above, plus findings, in a thesis report.
2. Overview of Background

Previous work in this field has been carried out at the University of Berkeley using an approach which aimed to measure stress levels in calm vs. stressful environments with computer mice, using statistics and a model of a mass-spring-damper system for the arm. There were 49 participants, with results suggesting that within-subject mouse-derived stress measures are quite strong. These results suggested that were was scope for more research to be carried out in this area with potentially interesting results.

1. Progress to Date

Initially I carried out extensive research into previous papers on this topic in an effort to determine the pros and cons of previous approaches and the best methods for myself to use in this project.

After getting a good grasp of Python, I began setting up the necessary packages and software on my laptop to carry out the programming and analysis side of the project. (Scikit-learn, NumPy, SciPy, Jupyter Notebooks etc.)

To prepare for the Machine Learning aspect of the project I chose to do several Machine Learning tutorials. These involved applying machine learning principles to a flower dataset to estimate which type of Iris flower a given set of parameters described. Another tutorial applied machine learning principles to a set of images of numbers, to estimate which number between 0-9 a given image represented.

Once I received the data I also had to search for an editor that is able to handle such a large text file (20gb). I chose EmEditor.

I have met with a mathematics lecturer from UCD to get an important insight and strong knowledge of various statistical approaches and how I can best apply them to this dataset to get meaningful results.

Now that I have the data I have begun to figure out what fields are necessary to keep and what can be discarded. I have split the main file into separate files using Python scripts. I have also begun to form an idea of which statistical approaches that I have encountered in my initial research will be most applicable to the data. Patrick Murphy, the lecturer from UCD, has been sent a sample of the dataset so I am waiting for a reply with his own insight as to what approaches may yield the best results.

1. Problems Encountered

Initially, I found the lack of source material in this area, with the exception of one research paper, a bit of an issue as there was not a lot of background information from different sources to work from.

Having to learn Python was another challenge I encountered as I had only ever used Java as my main programming language before. I used a lot of online resources like CodeCademy to teach myself this, as well as some resources I have picked up from peers who also taught themselves Python.

Similarly, although I have studied Maths for 3 years it had been a while since I had focused largely on statistics, so there was a lot of learning to do to get up to a level required for the analysis in this project. I used online resources to teach myself these skills or to refresh what I had already studied before in my college maths modules.

The size of the original data file has proven to be a large problem to overcome, with most text editors unable to handle such a large file. Although I have downloaded a special text editor, EmEditor, to handle the file it has still proven difficult to manage. I have now split the file into smaller more manageable files (by user) using a Python script.

1. Planned Next Steps

The next major steps involve getting into the ‘nitty gritty’ of the source data set and starting to do some hard analysis on the information that I can glean from it.

There is some further data cleaning to be carried out as there are some duplicate mouse events that appear to have been logged in error, which may affect the results of analysis. I also need to account for some disparities that may have arisen in the time logs of the mouse events and ensure that they match up to the actual times in which the lab exams took place.

I plan to apply the various methodologies I have covered in my earlier research such as analysis of multivariate data, clustering, grouping mouse movement by user, grouping the mouse events themselves, analysis of mouse displacement between start point and click event etc. Afterward, depending on what the data tells me I plan to apply machine learning principles to the data.

If it seems necessary, or if it seems like it might reveal some further useful information, it has been discussed that I may incorporate the user's exam results into the analysis to compare these with my findings and attempt to find any correlation or link between the two. This aspect may not be incorporated due to the privacy issues that may arise for the users.

I also plan to begin writing my report as I have now reached a point at which I can start talking about my progress to date, as I hope to have the report completed with plenty of time before the deadline to allow for changes and editing that might be required.