AMOD 5610H Project Idea

The project idea will be an approximately 250-750-word description of the broad outline of the project. At this stage students will be expected to have formed groups (if required), identified a potential data set, estimated required hardware or software, and any knowledge they may need to develop over the term.

At this stage you need to find a viable dataset, a group, and an outline of a project.

# Team Member(s)

|  |
| --- |
| **Problem Statement –** The problem is that market conditions today are more difficult than they used to be. This leads to increased workplace stresses. It is therefore fair to assume that unattainable business targets and a high risk of unemployment will increase the level of stress for people. The continued existence of such factors also becomes harmful to the health of an individual. Sometimes this may lead to the mild disease that is not needed, of course. However, the employee may develop a long-term illness, such as being in depression.  **How this project benefit** : From the point of view of the person in charge of the company’s productivity, we will approach the issue so that we will not concentrate on that part of the problem, but we will look at predicting the absenteeism from the job.  More specifically we would like to know when an employee may or may not be supposed to be absent in a given working day for a certain number of hours. Such knowledge will strengthen our decision making by understanding the work process in a way that will help us to avoid a lack of efficiency and enhance the quality of work produced in a particular organization.  **Team Member –** Pramina Patil (0663967) |

# Description of Dataset

In choosing a dataset you need to know what format(s) it is in, whether has statistically significant elements, and whether or not you can transform it to something else as needed. You should also check the dataset for veracity to know if there any errors/outliers that need to be cleaned from your dataset (e.g. if you have a dataset that has a name field, and one entry has “748 1011” as a name, you can’t get the name from that field). I’m not looking for a detailed analysis, just a description of what the data is, and what, if anything, is involved in gathering, processing and cleaning it.

|  |
| --- |
| **Dataset** - The dataset used is the ‘Absenteeism at Work’ dataset. This data is generated from the records of the absenteeism at work from 12th June 2017 to 11th October 2019 across the world.  **Data Dictionary** – The dataset consist of 12 Columns and around 1000 Rows. The columns present in the data - ID, Reason for Absence, Date, Transportation Expense, Distance to Work, Age, Daily Work Load Average, Body Mass Index, Education, Children, Pets, Absenteeism Time in Hours.  **Data Pre-processing** – The data set mostly consist of Integer values. The data needs to categorised. The data needs to be clustered for better performance and results. Also the normalization and standardization is also required during the pre-processing. As of now no cleaning is required but will implement if needed. The data however consist of the outliers so initial level of cleaning is required.  **Purpose of creating the model** - The creation of model will help us to explore whether or not a person displaying certain features is likely to be away from work at certain points in time. |

Hardware EstimateHere you should list how large your dataset it (this could be a size, or a rate), and roughly what hardware you need to do the processing you expect to do. For most projects, the dataset will be on the order of Megabytes -> Terabytes, and can be processed on a single desktop or laptop

|  |
| --- |
| **Machine Configuration –** M4.2x Large (16 GB RAM) – Cloud Machine  **Storage Type –** 256 GB Solid State Drive  **Size of the data** – 37 MB  **Machine OS** – Windows 10 |

# Software Estimate

Identify any likely software tools you’re going to need. What languages and development environments will you need? Do you need a database (e.g. Neo4j, or MS SQL), do you need a processing framework (Spark/Flink), roughly, any major libraries (including visualisation)?

|  |
| --- |
| **Programming Language –** Python 3.6 or Above version  **Cloud Technology –** Amazon Web Services (Sagemaker , Simple storage service (S3))  **Integrated Development Environment –** Jupyter Notebook**.**  **Data Visualization –** Tableau for Data representation**.**  **Version Control System –** GitHub open source  **Libraries** – Pandas, NumPy , Scikitlearn, Seaborn, Matplotlib, Seaborn, Tensorflow and to name a few. |