**AMOD – 5610H – Big Data Major Research Paper**

**Project Idea and Summarization**

**Absenteeism Prediction**

**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.

**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.

**What exactly is Absenteeism**

Absence from work during normal working hours, which results in temporary incapacity to carry out routine work.

**Dataset and Data Dictionary**

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.

|  |  |
| --- | --- |
| **Column Name** | **Column Data Type** |
| ID (Individual Identification) | Integer |
| Reason of Absence  1 - Certain infectious and parasitic diseases. 2 – Neoplasms  3 - Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism.  4 - Endocrine, nutritional and metabolic diseases.  5 - Mental and behavioral disorders.  6 - Diseases of the nervous system.  7 - Diseases of the eye and adnexa.  8 - Diseases of the ear and mastoid process.  9 - Diseases of the circulatory system.  10 - Diseases of the respiratory system  11 - Diseases of the digestive system.  12 - Diseases of the skin and subcutaneous tissue.  13 - Diseases of the musculoskeletal system and connective tissue.  14 - Diseases of the genitourinary system.  15 - Pregnancy, childbirth and the puerperium.  16 - Certain conditions originating in the perinatal period.  17 - Congenital malformations, deformations and chromosomal abnormalities.  18 - Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified.  19 - Injury, poisoning and certain other consequences of external causes.  20 - External causes of morbidity and mortality.  21 - Factors influencing health status and contact with health services.  22 – Others are ICD | Integer |
| Date | Date Type (Date-Month-Year) |
| Transportation expense | Integer |
| Distance to Work Place (Kilometers) | Integer |
| Age of the Employee | Integer |
| Daily work load average. | Integer |
| Body mass index | Integer |
| Education (Highest) 1- High School 2 – Graduate 3 – Postgraduate 4 – Masters and Doctors | Integer |
| Number of children | Integer |
| Pets owned | Integer |
| Absenteeism Time in Hours | Integer |

**Technology Dexterity**

* **Software Requirement**
* **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
* **Hardware Requirement**
* **Machine Configuration –** M4.2x Large (16 GB RAM) – Cloud Machine
* **Storage Type –** 512 GB Solid State Drive