Prediction of Absenteeism in the workplace(Group 4A)

# Technical Design Document

Version 1.0

RESTRICTED DISTRIBUTION

The information is standard Company Confidential, but due to its sensitivity, it has restricted distribution and viewing within iNeuron.

Document Version Control

|  |  |  |  |
| --- | --- | --- | --- |
| Date Issued | Version | Description | Author |
| 17thth July 2020 | 1.0 | Initial Draft | Subrata |
| 27th Aug 2020 | 2.0 | Updated | Rukaiya |
| 27th Aug 2020 | 3.0 | Reviewed and Updated | Subrata |

Contributors

The content of this document has been authored with the combined input of the following group of key individuals.

|  |  |
| --- | --- |
| Name | Section Worked Upon |
| Rukaiya | Initial Draft |
| Subrata | Final Draft |

Document Classification

|  |  |
| --- | --- |
| Classification | Company Confidential |
| Definition | Information is Group confidential and needs to be protected |
| Context | Where the loss of information confidentiality would result in significant harm to the interests of the organisation, financial loss, embarrassment or loss of information |

Contents

[1 Technical Design Document 2](#_Toc44177265)

[1. Introduction 5](#_Toc44177266)

[High level objectives 5](#_Toc44177267)

[2 Workflow Overall 6](#_Toc44177268)

[Application Flow 6](#_Toc44177269)

[Exception Scenarios Overall 6](#_Toc44177270)

[3 Workflow Data Ingestion and File Conversion 8](#_Toc44177271)

[3.1 Technical solution design 10](#_Toc44177272)

[3.2 Exceptions Scenarios 10](#_Toc44177273)

[4 Stats Based EDA 11](#_Toc44177274)

[4.1 Steps 11](#_Toc44177275)

[4.2 Technical solution design 11](#_Toc44177276)

[4.3 Exceptions Scenarios Module Wise 11](#_Toc44177277)

[5 Graph-Based EDA 12](#_Toc44177278)

[5.1 Technical solution design 12](#_Toc44177279)

[5.2 Exceptions Scenarios Module Wise 13](#_Toc44177280)

[6 Library Based Utils 14](#_Toc44177281)

[6.1 Technical solution design 14](#_Toc44177282)

[6.2 Exceptions Scenarios Module Wise 14](#_Toc44177283)

[7 Data Transformers( Pre-processing steps) 15](#_Toc44177284)

[7.1 Technical solution design 15](#_Toc44177285)

[7.2 Exceptions Scenarios Module Wise 15](#_Toc44177286)

[8 ML Model Selection 16](#_Toc44177287)

[8.1 Technical solution design 16](#_Toc44177288)

[8.2 Exceptions Scenarios Module Wise 16](#_Toc44177289)

[9 Model Tuning and Optimization 17](#_Toc44177290)

[9.1 Technical solution design 18](#_Toc44177291)

[9.2 Exceptions Scenarios Module Wise 18](#_Toc44177292)

[10 Testing Modules 19](#_Toc44177293)

[10.1 Technical solution design 19](#_Toc44177294)

[10.2 Exceptions Scenarios Module Wise 20](#_Toc44177295)

[11 Prediction Pipeline 21](#_Toc44177296)

[11.1 Technical solution design 22](#_Toc44177297)

[11.2 Exceptions Scenarios Module Wise 23](#_Toc44177298)

[12 Deployment Strategy 24](#_Toc44177299)

[12.1 Technical solution design 25](#_Toc44177300)

[12.2 Exceptions Scenarios Module Wise 26](#_Toc44177301)

[13 Monitoring 27](#_Toc44177302)

[13.1 Technical solution design 27](#_Toc44177303)

[13.2 Exceptions Scenarios Module Wise 27](#_Toc44177304)

[14 Logging 28](#_Toc44177305)

[14.1 Technical solution design 28](#_Toc44177306)

[14.2 Exceptions Scenarios Module Wise 28](#_Toc44177307)

[15 Hardware Requirements 29](#_Toc44177308)

[15.1 Requirements for model training 29](#_Toc44177309)

[15.2 Requirements for model testing 29](#_Toc44177310)

# Introduction

The goal here is to predict the hours a employee will be absent from office. The user will pass on some of the required details about an employee and the model will predict how many hours they will be absent from office.

This project shall be delivered in two phases:

Phase 1: Prepare a machine learning model to predict absent hours.

Phase2: Integration of UI to all the functionalities.

The technical design document gives a design blueprint of the Absent hours project. This document communicates the technical details of the solution proposed.

In addition, this document also captures the different workflows involved to build the solution, exceptions in the workflows and any assumptions that have been considered.

Once agreed as the basis for the building of the project, the flowchart and assumptions will be used as a platform from which the solution will be designed.

Changes to this business process may constitute a request for change and will be subject to the agreed agility program change procedures.

**Note: All the code will be written in python version 3.8**

## High level objectives

The high-level objectives are:

1. Enable reading/loading of data from csv and convert them into pandas dataframe(details mentioned in the Data Ingestion Section).
2. Perform statistical analytics of the data .
3. Perform graphical analysis for the data .
4. Perform data cleaning operation with all the steps required .
5. After data cleaning showcase the graphical analysis once again for comparison.
6. Choose the appropriate ML model for training.
7. Perform model Tuning.
8. Give option for prediction.
9. Give option for bulk prediction.
10. Give option to retrain the model by uploading new data.
11. Cloud deployment.

**Phase 1:** Create the absent hour prediction model.

**Phase 2:** Create UI & Cloud Deployment

**Data Description:**

It is a dataset of absenteeism of ‘A Courier Company’ in Brazil which was recorded from July 2007 to July 2010. It has 21 attributes and 740 instances .

Data link: https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work

Citation: Martiniano, A., Ferreira, R. P., Sassi, R. J., & Affonso, C. (2012). Application of a neuro fuzzy network in prediction of absenteeism at work. In Information Systems and Technologies (CISTI), 7th Iberian Conference on (pp. 1-4). IEEE.

Below are some codes and their descriptions in the data

**Reasons for absence:**

|  |  |
| --- | --- |
| **Code** | **ICD Decsription** |
| 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 | patient follow up |
| 23 | medical consultation |
| 24 | blood donation |
| 25 | laboratory examination |
| 26 | unjustified absence |
| 27 | physiotherapy |
| 28 | dental consultation |

**Code of Day of the week :**

|  |  |
| --- | --- |
| **Code** | **Description** |
| 2 | monday |
| 3 | tuesday |
| 4 | wednesday |
| 5 | thursday |
| 6 | saturday |

**Code of Education :**

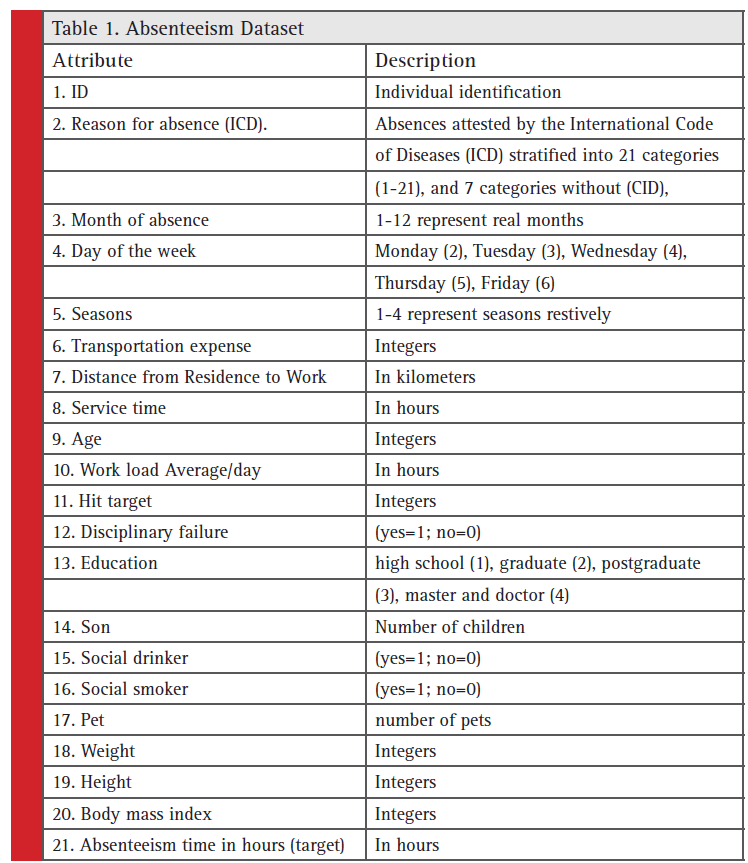
|  |  |
| --- | --- |
| **Code** | **Description** |
| 1 | High School |
| 2 | Graduate |
| 3 | Post Graduate |
| 4 | Doctorate |

**Data Preprocessing**

We transformed the actual outcome attribute titled “absenteeism time in hours” into a categorical column that includes four classes such that “0-2 hrs,”,”2-4 hrs”, “4-8 hrs”, “8-24 hrs”, and “ > 24 hrs” which represents a corresponding amount of time for each class.

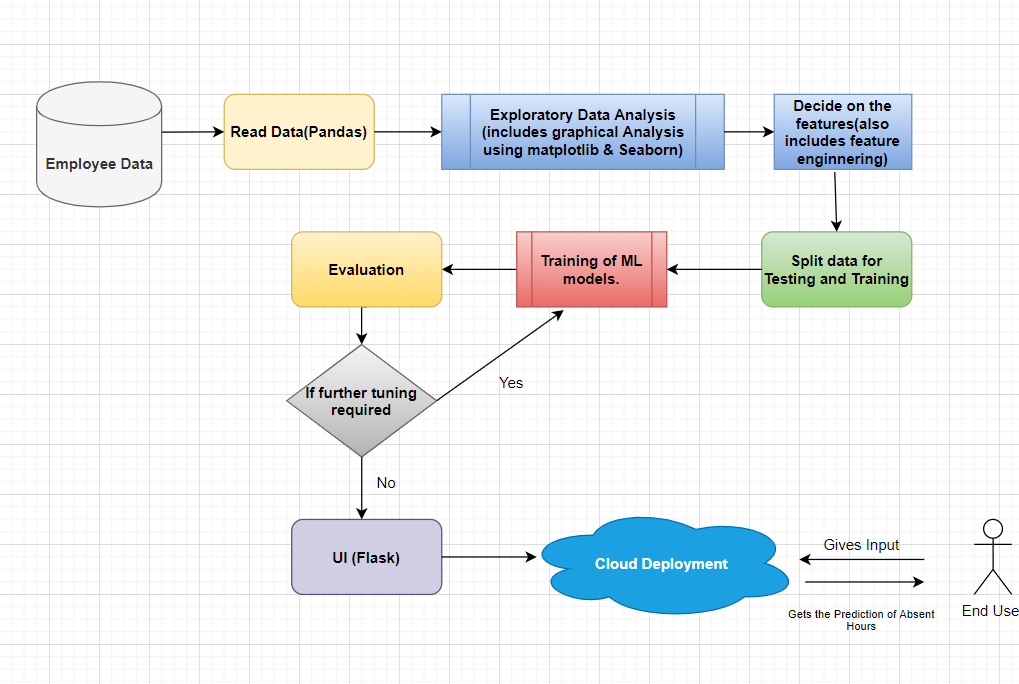
|  |  |
| --- | --- |
| **Absenteeism Time (x)** | **Absenteeism Class** |
| **0<x<=2** | **0-2 hrs** |
| **2<x<=4** | **2-4 hrs** |
| **4<x<8** | **4-8 hrs** |
| **8<x<=24** | **8-24 hrs** |
| **24<x** | **> 24 hrs** |

**Attribute Information:**

****

# Workflow Overall :

## Application Flow



**Above fig shows the flow of model**

## Exception Scenarios Overall

|  |  |  |
| --- | --- | --- |
| **Step** | **Exception** | **Mitigation** |
| User gives Wrong File Format | Give proper error message | Its gives follow instruction message |
| User gives incorrect feature names | Give proper error message | Its gives follow instruction message |

# Workflow Data Ingestion and File Conversion

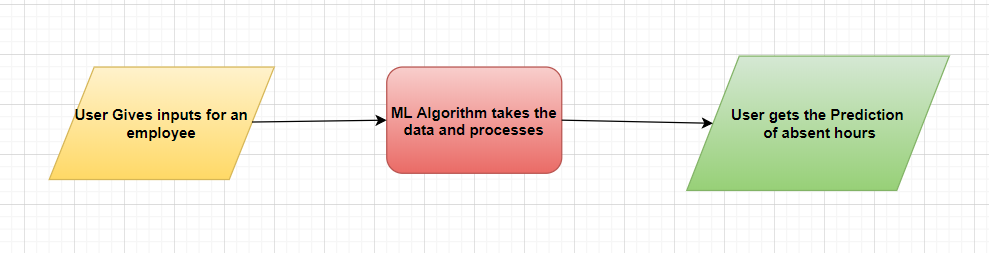
**Data Sources:**

**Phase 1:**

|  |  |
| --- | --- |
| Data Connector Utils | File Conversion Utils |
| Pandas.read\_csv() |  |
|  |  |
|  |  |
|  |  |

**As of now we are fetching information directily from CSV, but in future will try to upload files to DB and then connect it via pandas for further analysis.**

## Technical solution design



**Above Design shows how actual output is got from model and how data is flow**

**Methods/Modules:**

|  |  |
| --- | --- |
| **Custom Module/methods** | **Description** |
| app\_latest.py | Runs the streamlit web page i.e. it runs the user interface to interact with the machine learning model |
| model\_latest.py | Contains the model creation steps. Also retrains the model with new data. |
| predict.py | Runs the predictions on input data. |
| get\_value.py | Converts all the input descriptions to their respective codes |

**Data Profiling :**

After reading the data, following will be done:

1. The number of rows
2. The number of columns
3. Number of missing values per column and their percentage
4. Total missing values and it’s percentage
5. Number of categorical columns and their list
6. Number of numerical columns and their list
7. Number of duplicate rows

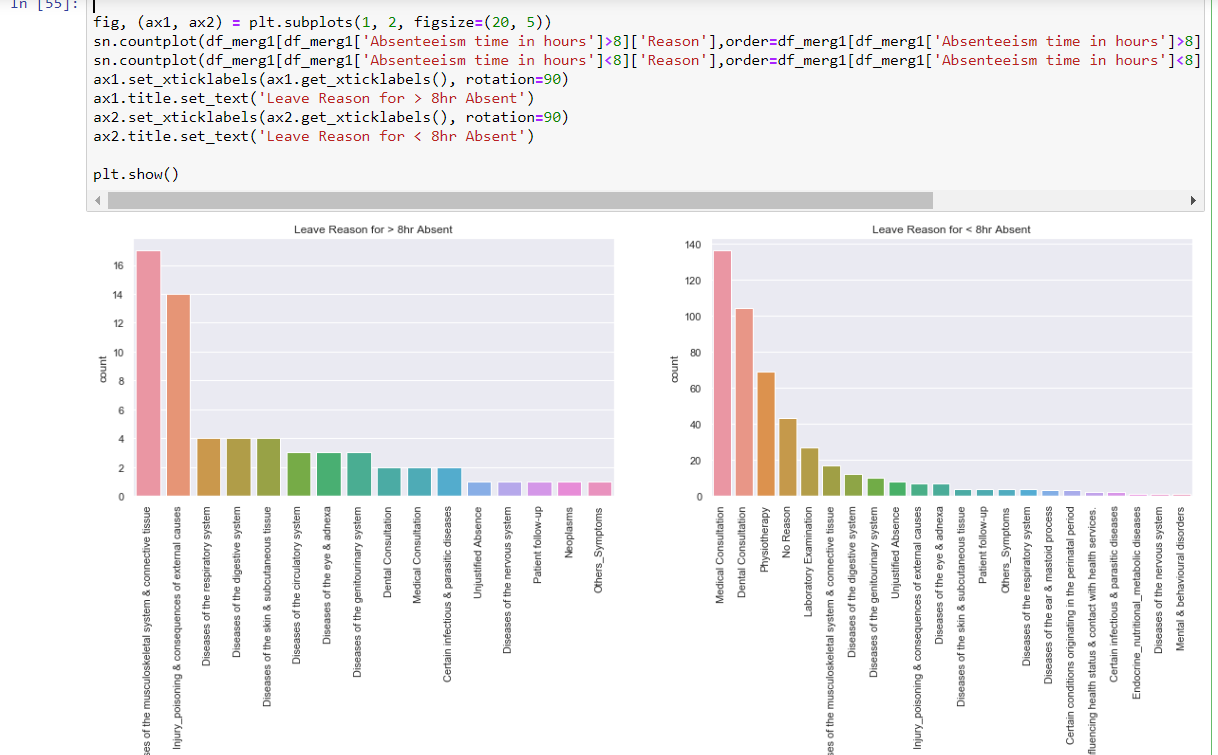
**Graph Based EDA SNAPSHOTS /InSights :**

Count plots |Piecharts for categories |Barplots |Scatterplot

**Insight\_1/Comment**:

We separated the absent hours with less than and greater than 8 hours and it shows clear differences in absent reasons.

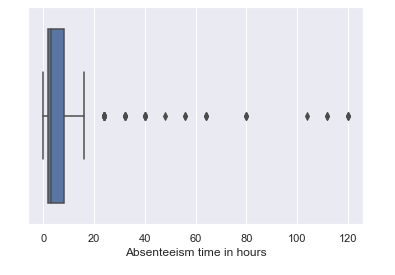
For <8 hrs it is medical & Dental consultation, for >8 hrs it is related to injury and some other issues.



**Insight\_2/Comment:**

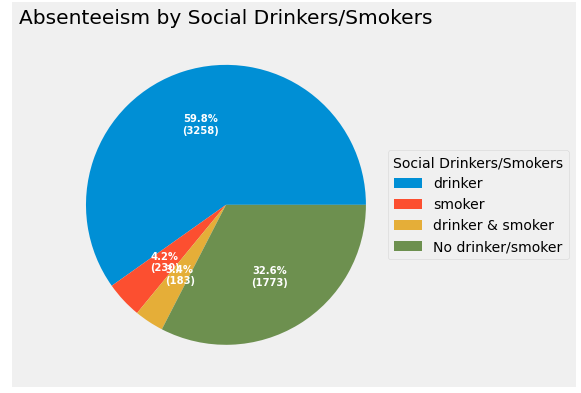
Absent time in hour count is plot using count plot ,majority shows 8 hours i.e. 1 day leave, which makes sense as most of the emloyees takes leave for one day frequently . Reason can be like have to give eolanation to the manager.





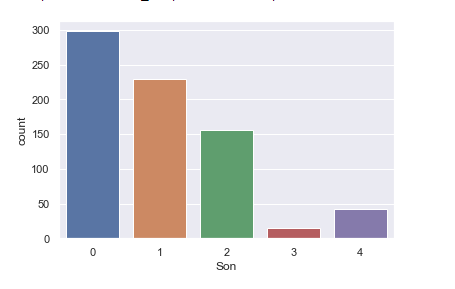
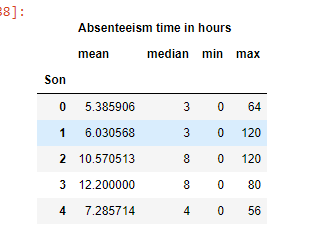
**Insight\_3/Comment:**

Below shows majority of the data belong to drinkers and smokers. People who drink might fall sick or have hangovers the next day which can lead to absence.



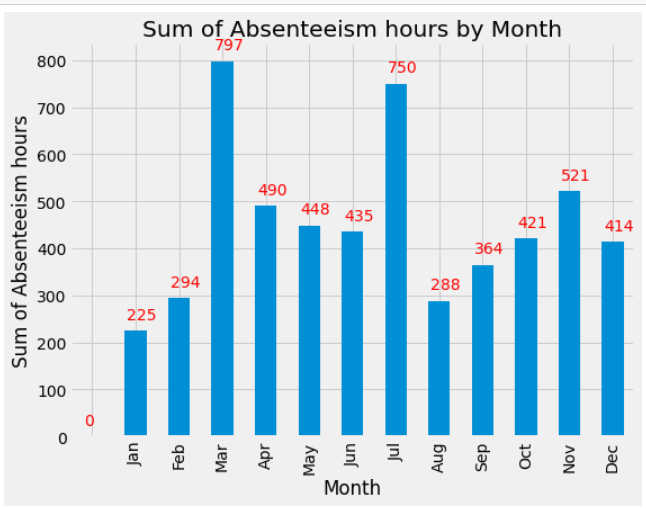
**Insight\_4/Comment:**

Below we tried to check if any impact of number of sons/kids for the employees . Looks like people with 2 or 3 kids have more full day leaves (8 hrs). Probable reason might be more kids means more probability of them falling sick and doctor visits but when kids are around 4 there might be a kid who has grown enough to take care of few things at home. Just a guess !!

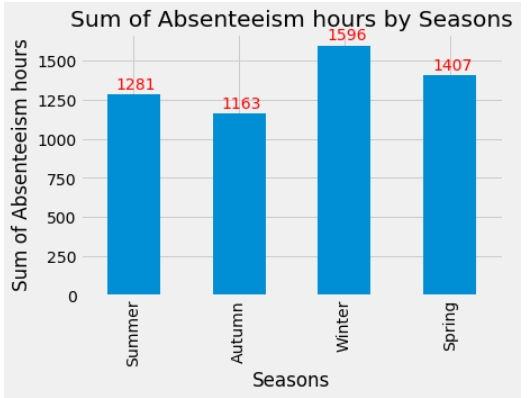
**Insight\_5/Comment:**

From below we came to know about the months which had maximum leaves.March ,July & November are highest. Managers can anticipate and plan accordingly.



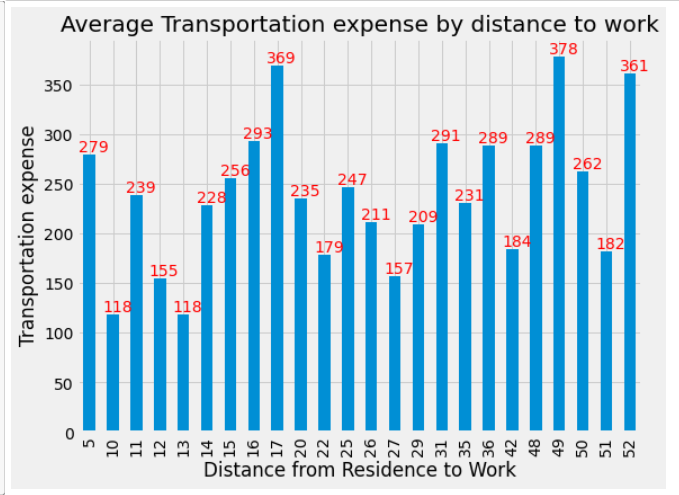
**Insight\_6/Comment:**

From below we came to know about the Seasons which had maximum leaves.Winter & Spring are highest. Managers can anticipate and plan accordingly.



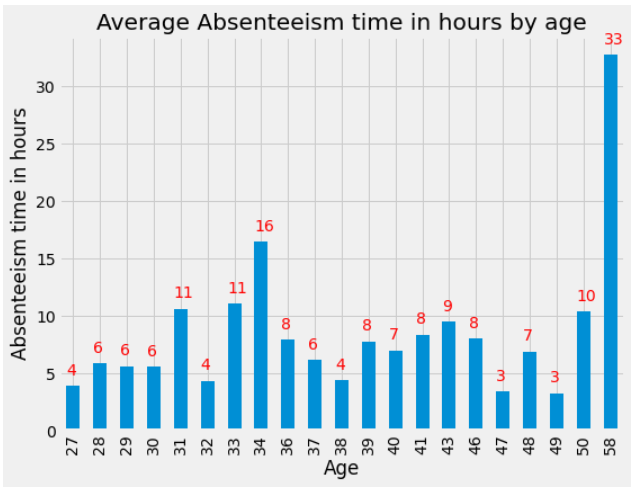
**Insight\_7/Comment:**

Below is the average cost of transport expenses w.r.t. the distance from residence to work.



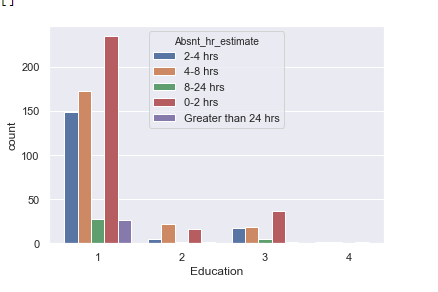
**Insight\_8/Comment:**

Below we tried to check if absent hours increases with Age.



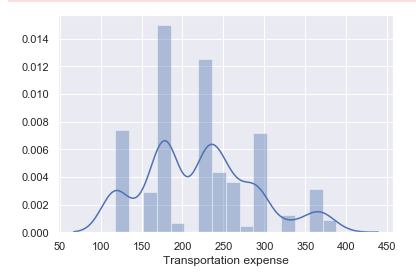
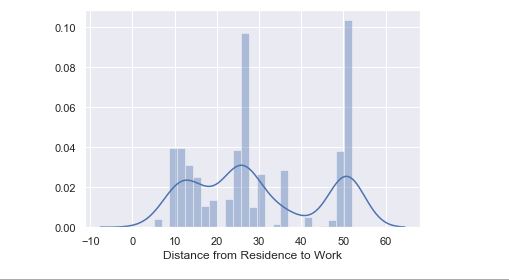
**Insight\_9/Comment:**

Analysing the education level in the data shows that most of the employees with lower level of education has more absent hours. Probably they get frustrated with the level of income as compared to the work load and other family tensions



**Insight\_10/Comment:**

We then analysed the transportation cost & Distance from work of the employees. Assuming that hours of absenteeism will depend on this two parameters as traveling time and cost will make employees hesitant to travel

**Data Transformers( Pre-processing steps)**

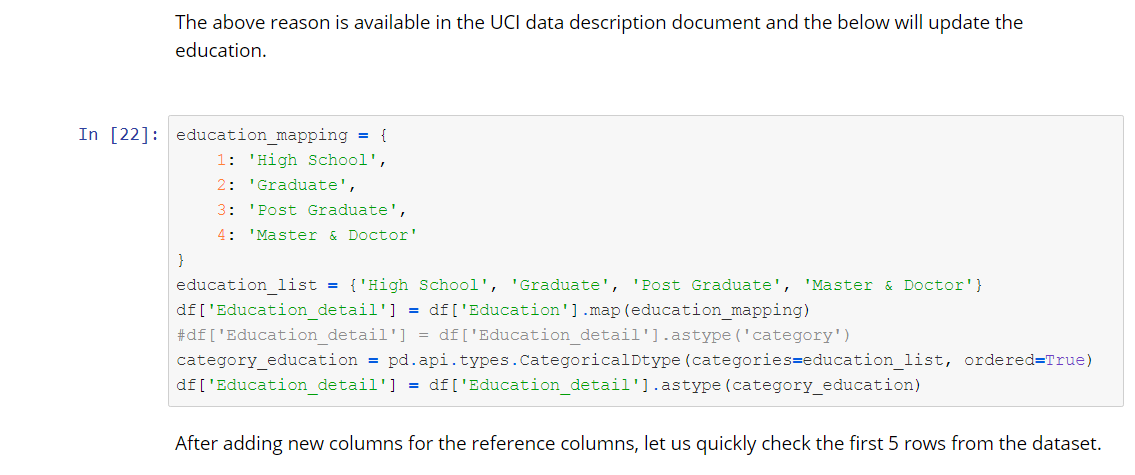
1. Handling Null Values
2. Checking Data Quality
3. Categorical Value Handling (Encoding)
4. Number format handling.

**Checking missing values:**



**Mapping codes to values:**

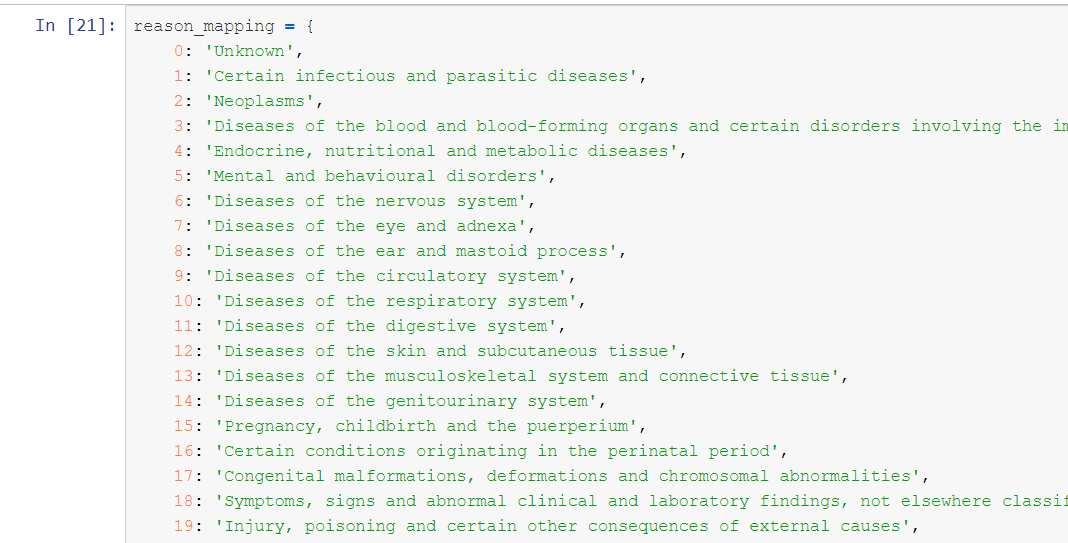
* **Education code Mapping**



* **Season code Mapping**

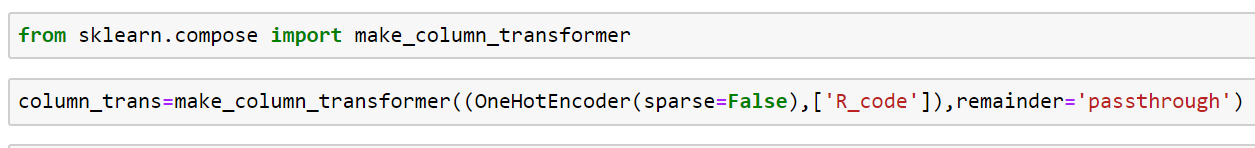


* **Reason code Mapping**



**Encoding of data :**

**As data has categorical variables so we have encoded the values using column transformer and one hot encoding.**



**Number format handling :**

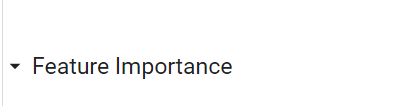
**For this dataset no number format handling was required.**

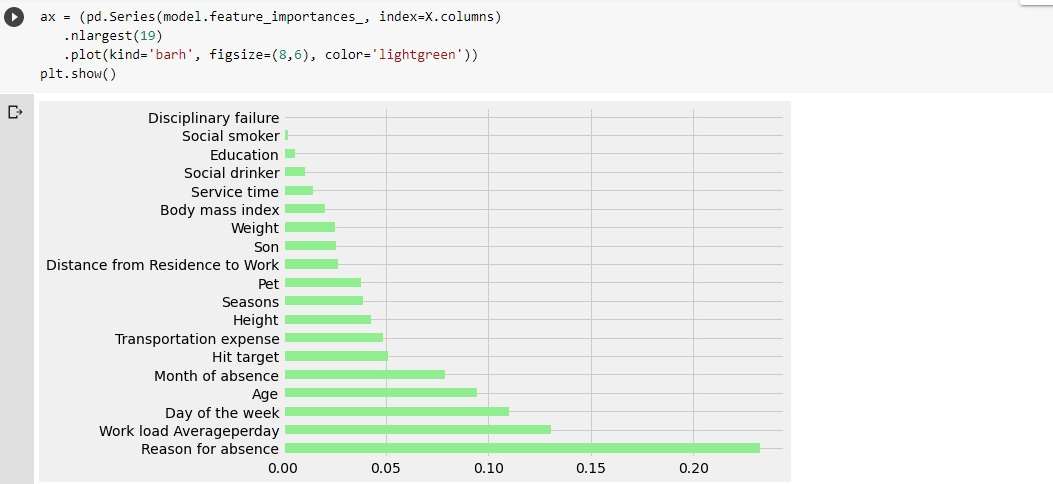
**Feature Selection :**

Identified the below features as important for predicting absent hours :

1. Reason for Absence
2. Distance from Residence to Work (kilometers)
3. Age
4. BMI
5. Transportation expense

The original data set had only values for the categories, we have added category values to the main dataset to help understanding the data and get better insights.





**Above plot shows the which feature is more important and which one is not this plot plays very important role in feature engineering**

**Machine Learning Algorithm Selection:**

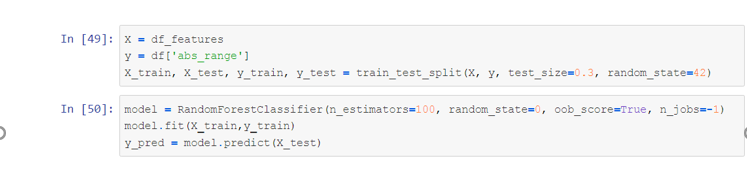
Our goal is to predict a category (Absent hour group) so we will be using a classification algorithm. Also as the data has multiple features and noise so we will be using **Random Forest classifier**.

**Steps of model creation:**

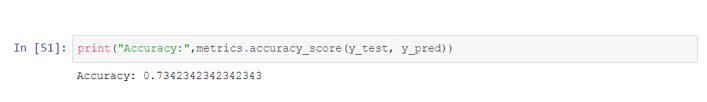
* **Splitting of data for training and testing.**



* **Training the model.**

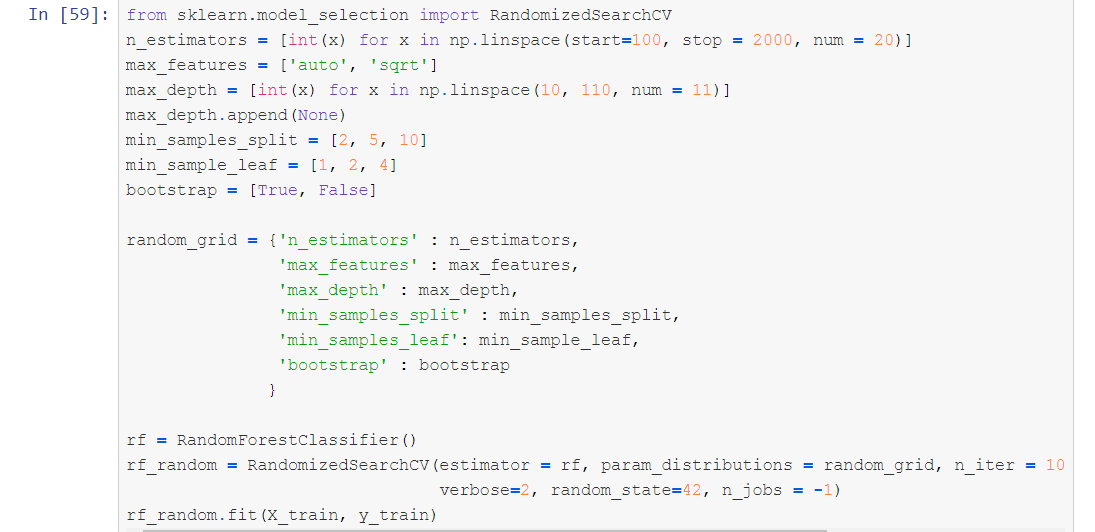


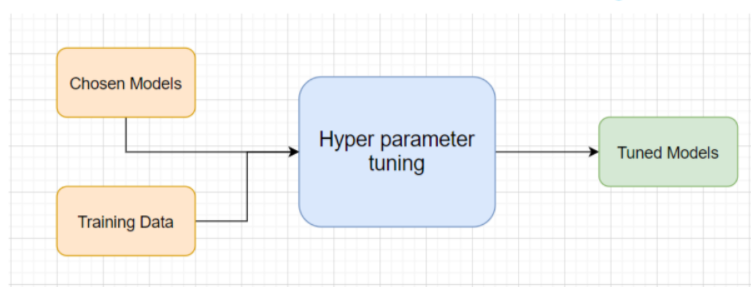
* **Model Evaluation.**



**Model Optimization with HyperParameter Tunning:**

* **RandomizedSearchCV was used to get the optimal parameters.**





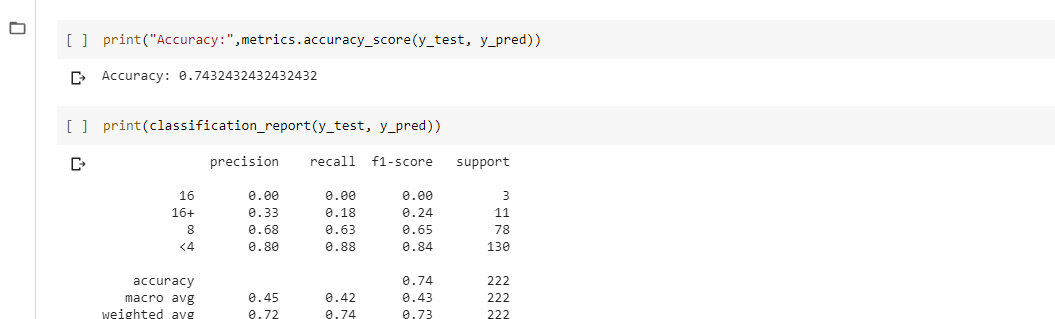
**Testing Modules :**

Divide the training data itself into train and test sets

Use test data to have tests run on the three best models

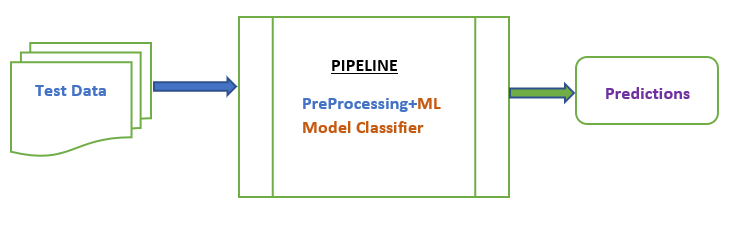
Give the test report

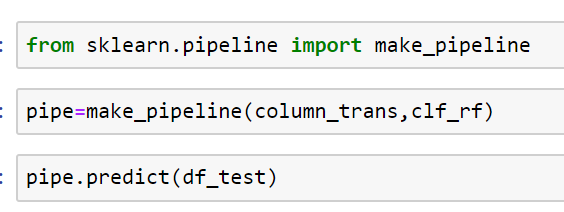
1. Accuracy
2. Precision
3. Recall



**Pipeline:**

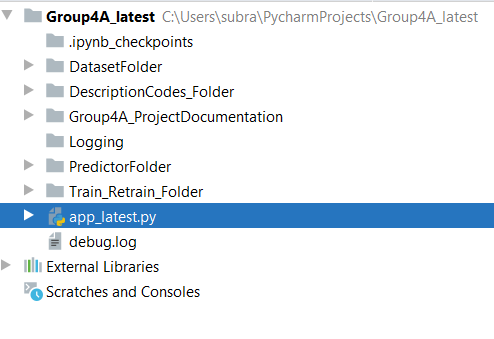
**For transforming (preprocessing) test data and predicting values we have created a pipeline so that we need not pre-process the data again & again.**





**Modular coding(usage of Classes & Instances) :**

**Folder Structure**



* **App\_latest.py :**

**This the main file which runs the application. This script calls all the depending classes for prediction , retraining and values.**



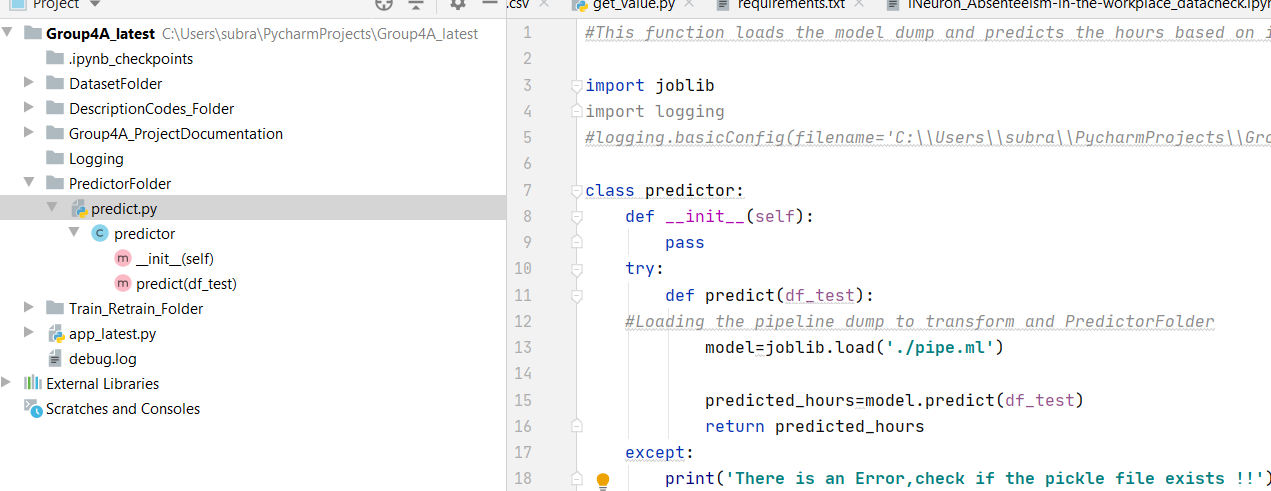
* **Train\_Retrain\_Folder:**

**This folder contains the class which is used to retrain the model**



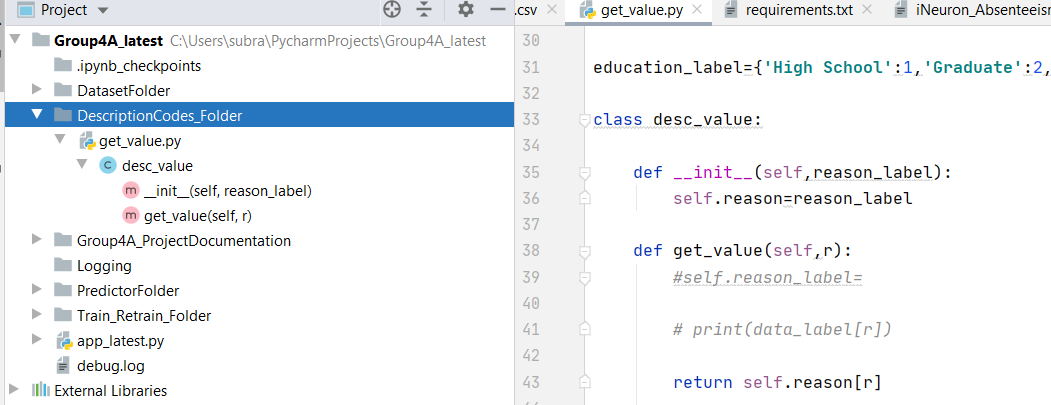
* **PredictionFolder:**

**This folder contains the class which is used to predict the absent hours.**



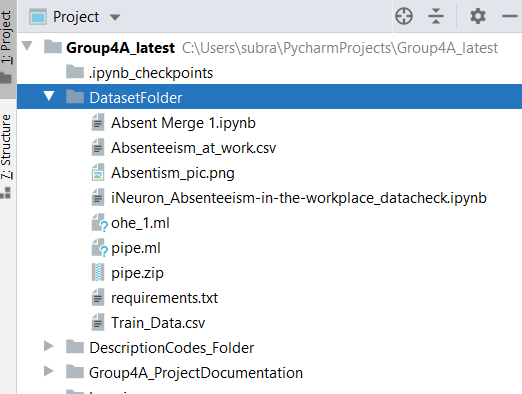
* **DescriptionCodes\_Folder:**

**This folder contains the class which converts the various descriptions to their respective codes.**



* **DatasetFolder:**

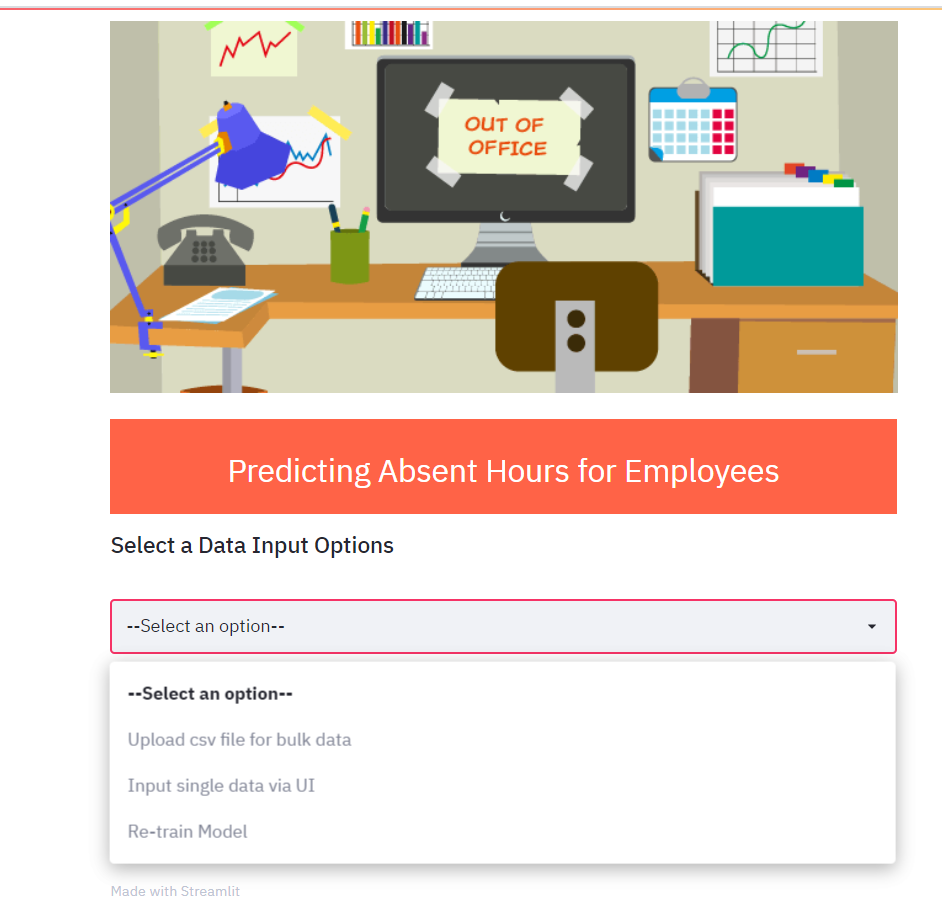
**This folder contains the various files and images used by the application. Even the pickle/dump file of the ML model resides in this folder.**



**Building Web API/Interface:**

**For this project we have used streamlit library to build Web API.**

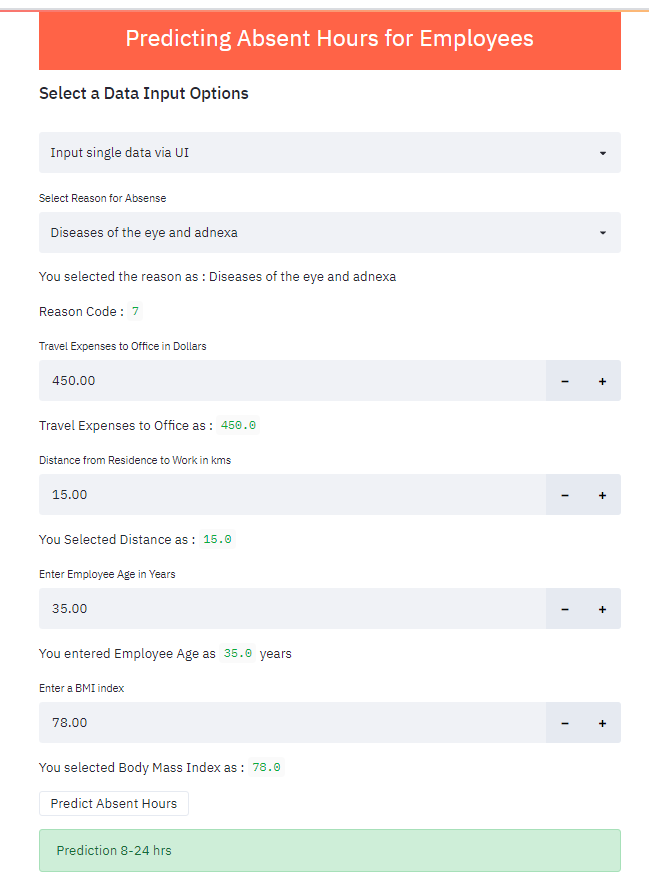




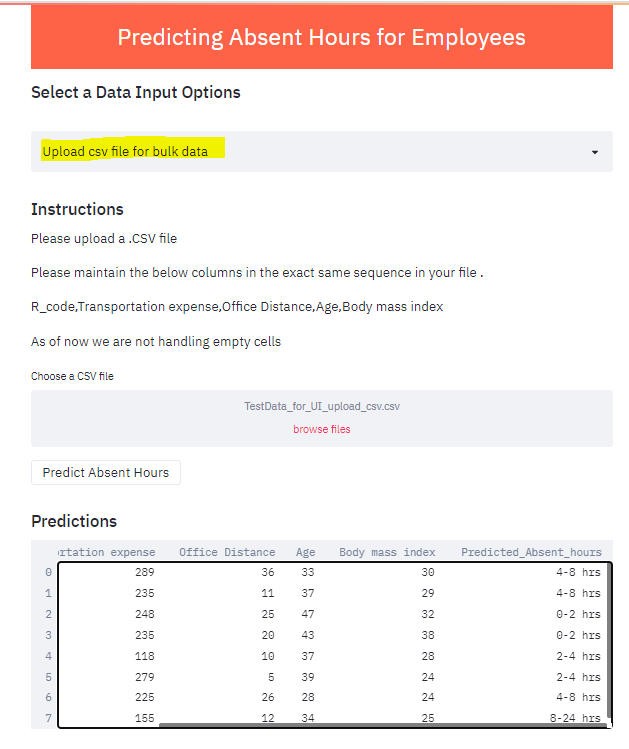
**User options provided via UI :**

**User can do below activity via UI :**

1. **Get a predicted absent hours via single input via UI.**

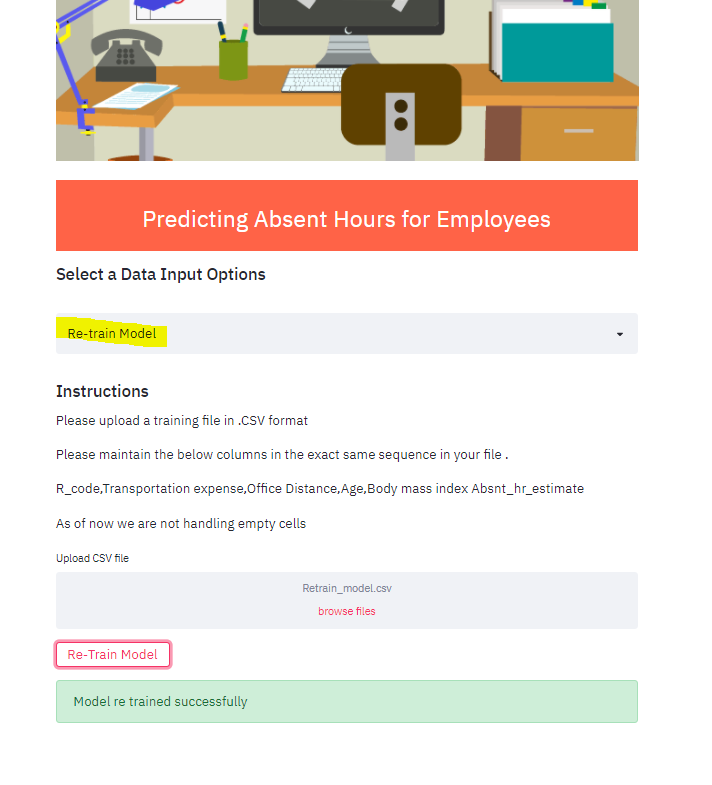


1. **Get Bulk Predictions via CSV file upload.**



**Model Re-Training Strategy/Process:**

**We have provided option via web API to get periodic retraining of model by adding new data. User can upload new data and just click on re-train model.**



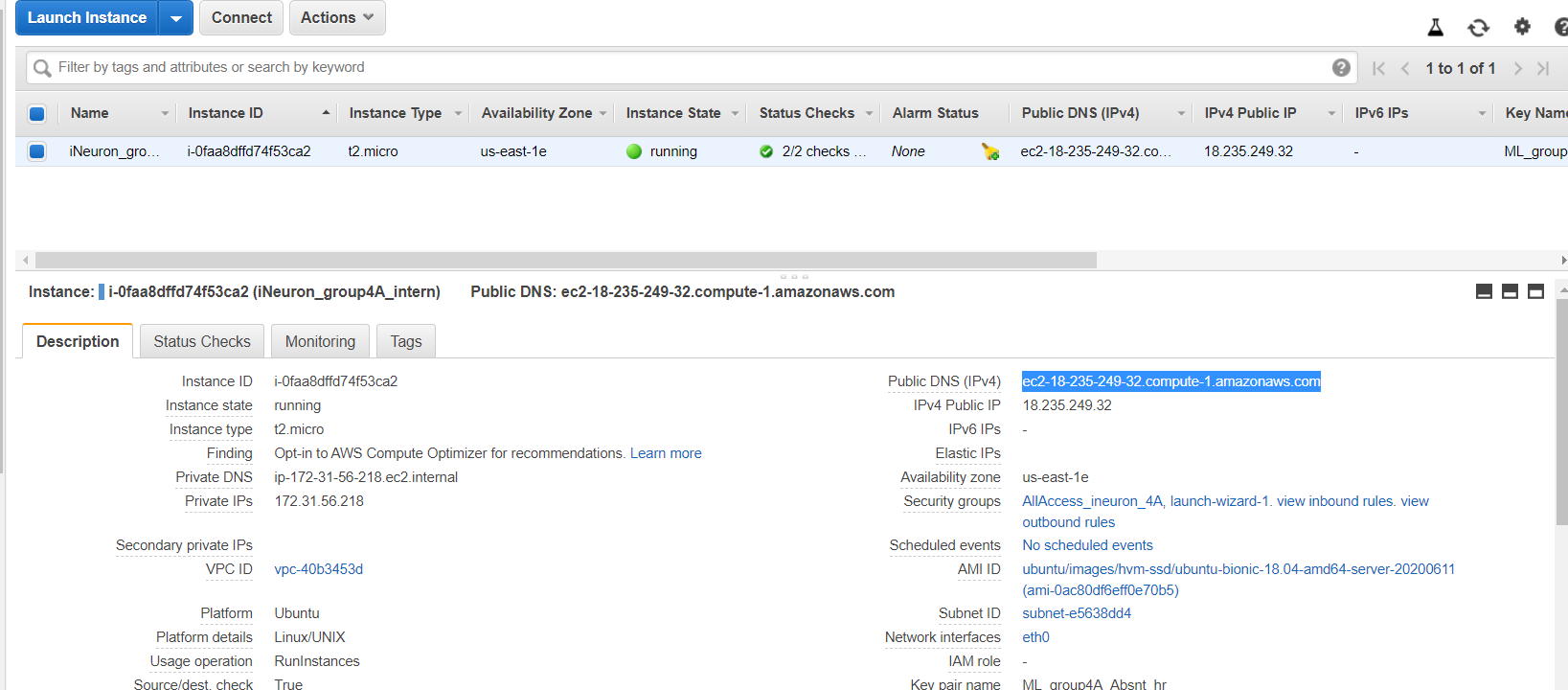
**Cloud Deployment Strategy/Process:**



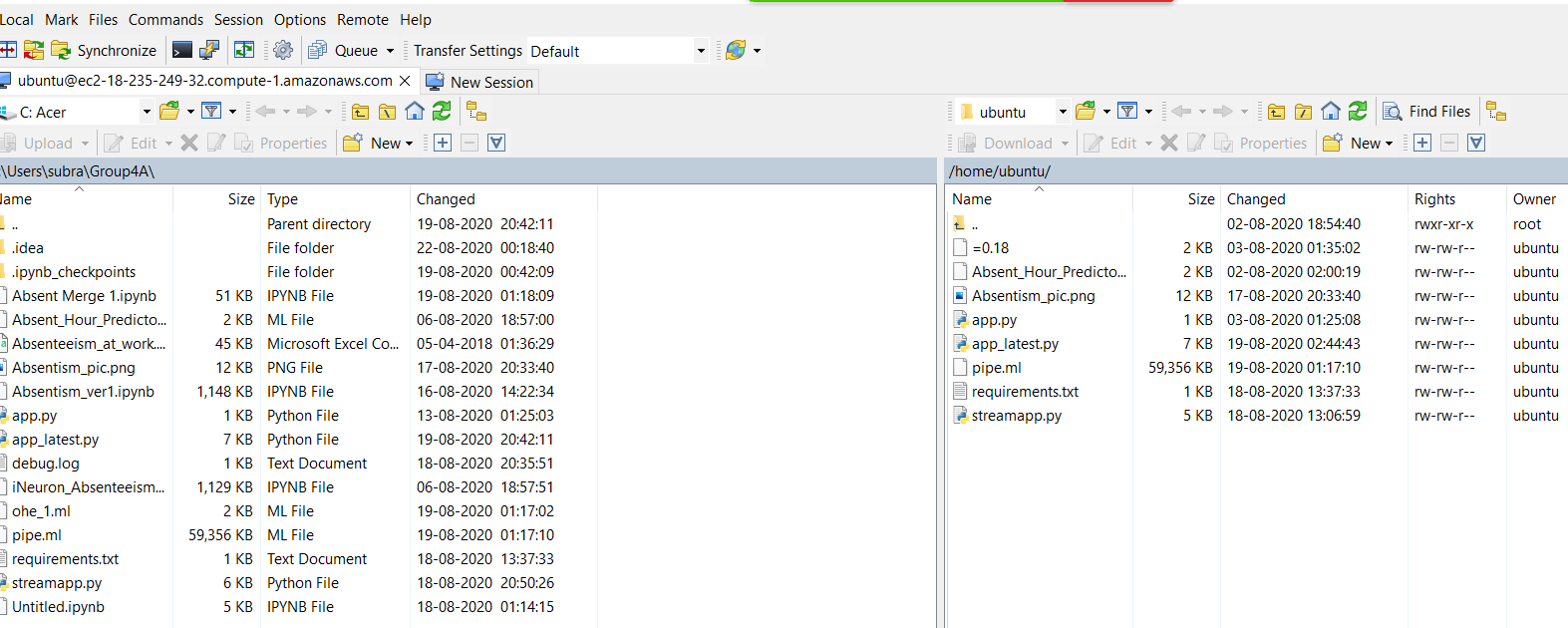
**We used free tier AWS EC2 to deploy the application.**

**Steps(Assuming already aws account is created):**

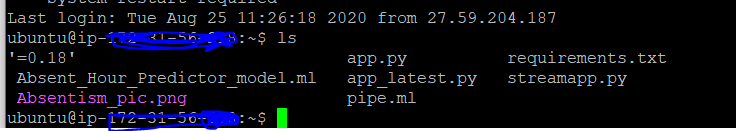
1. **Created AWS EC2 instance**

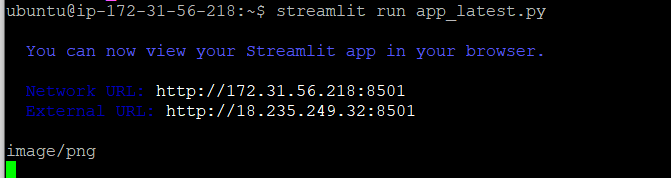


1. **Transferred files from local to AWS using WINSCP.**

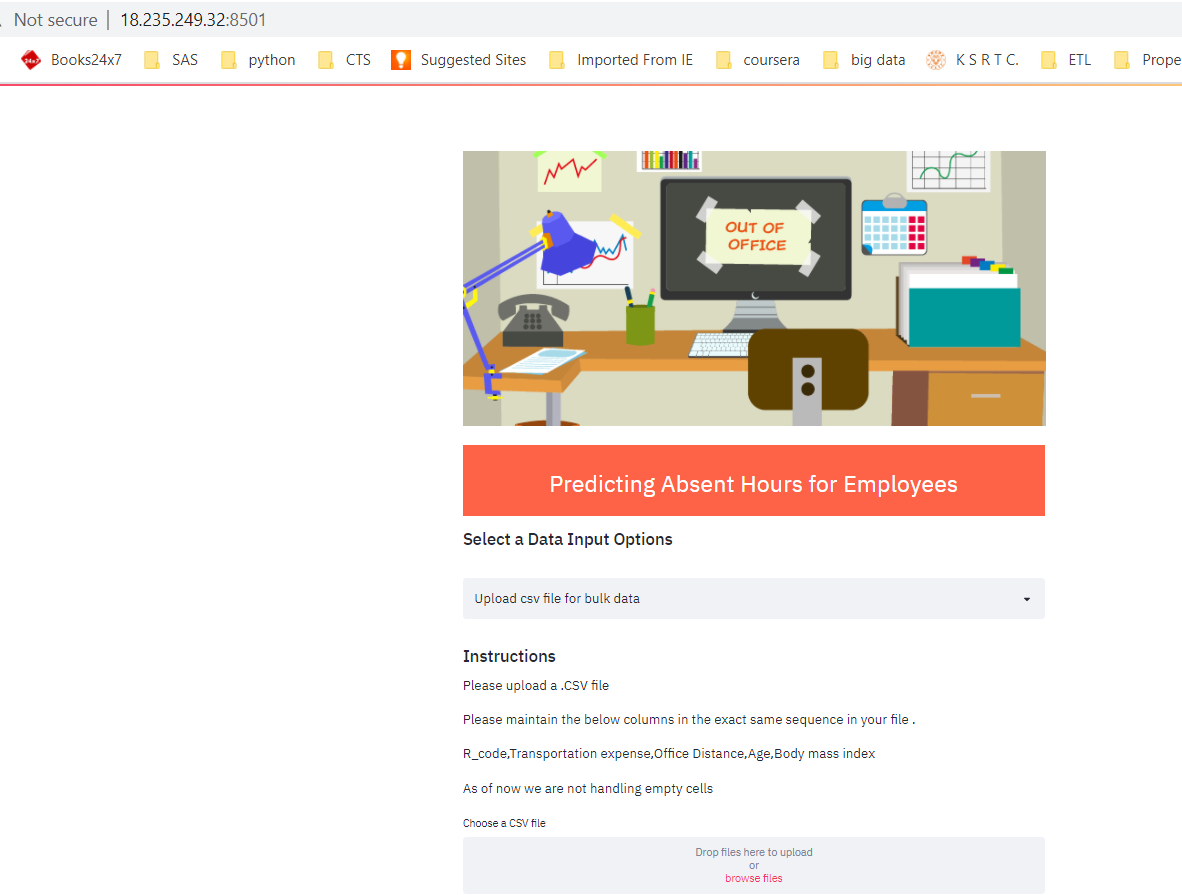


1. **Logged into the instance using PuTtY and run the streamlit job.**

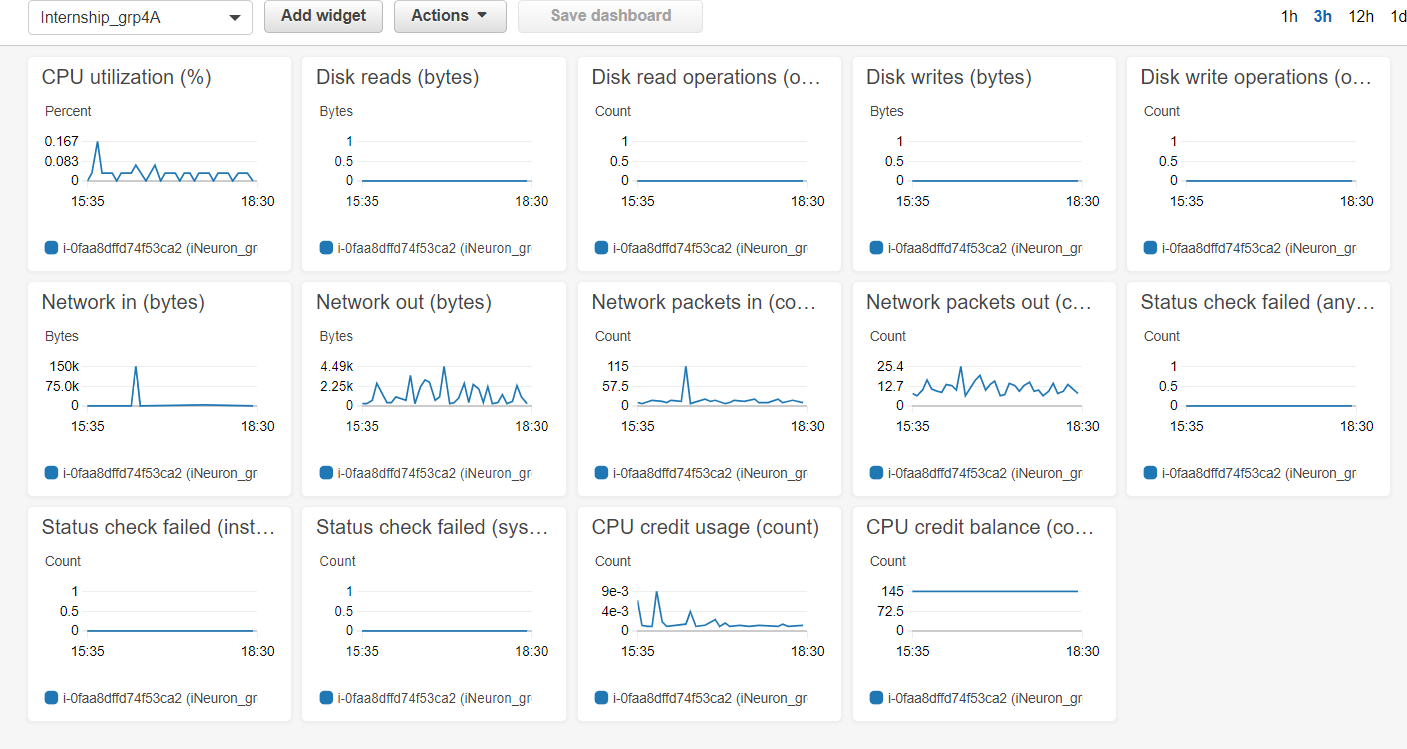




**Above shows the job ran successfully.**

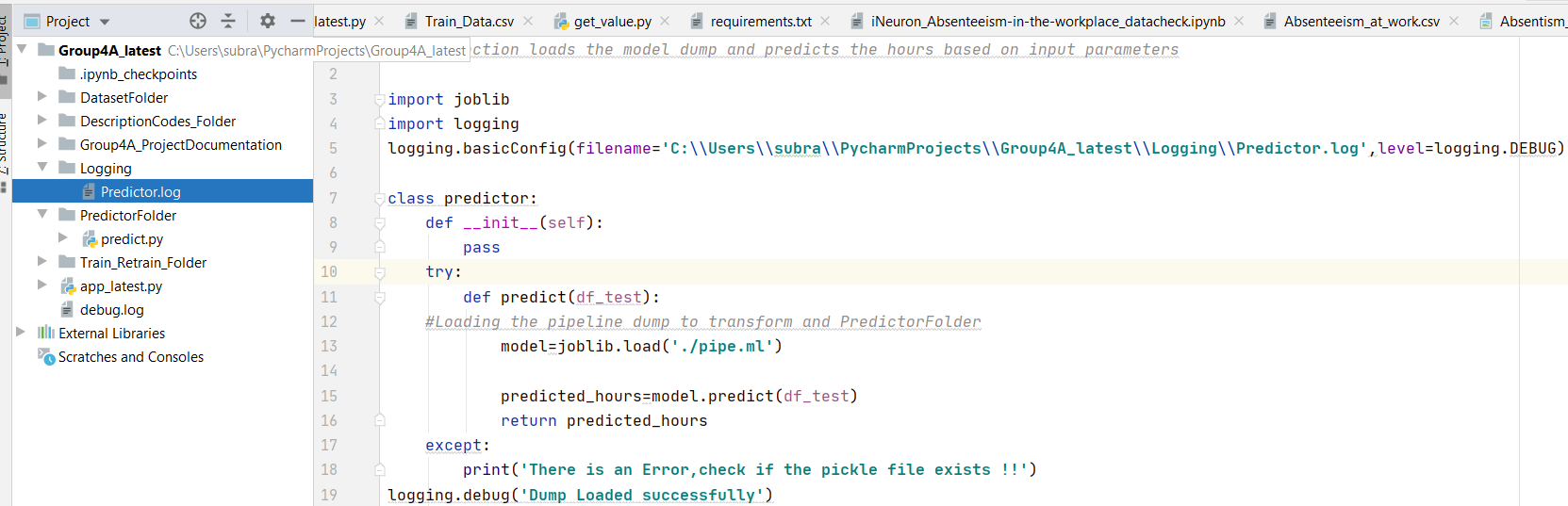


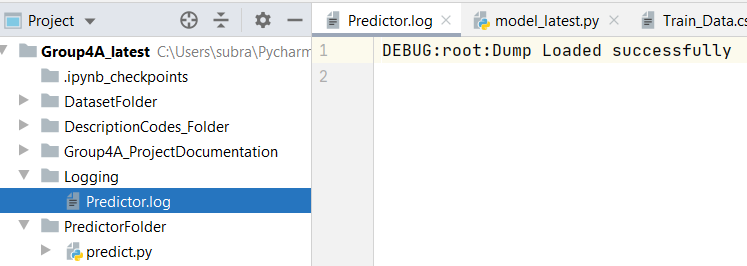
**Dashboarding/Monitoring:**



**Logging:**

**We tried to create logging at the code level by using python logging function. We were able to run that only for the predictor file.**



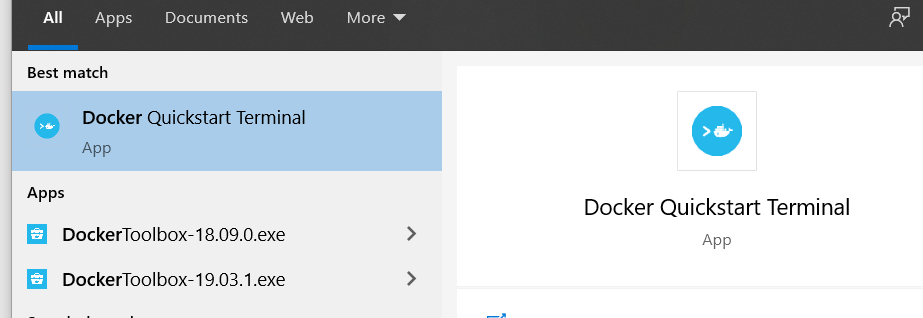


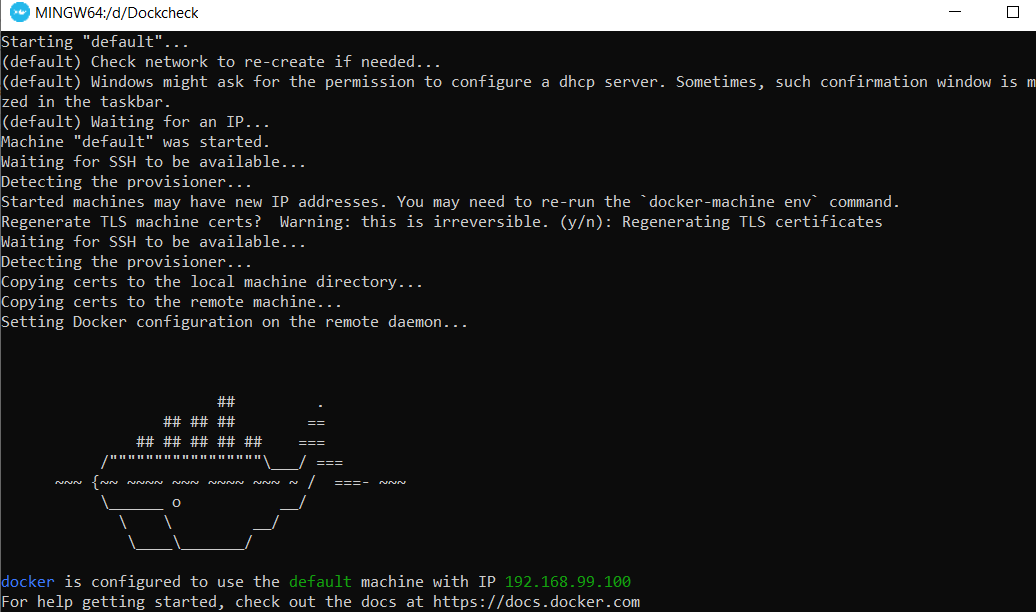
**We were unable to create the log file from the main script i.e. (app\_latest.py).**

**Dockerization:**

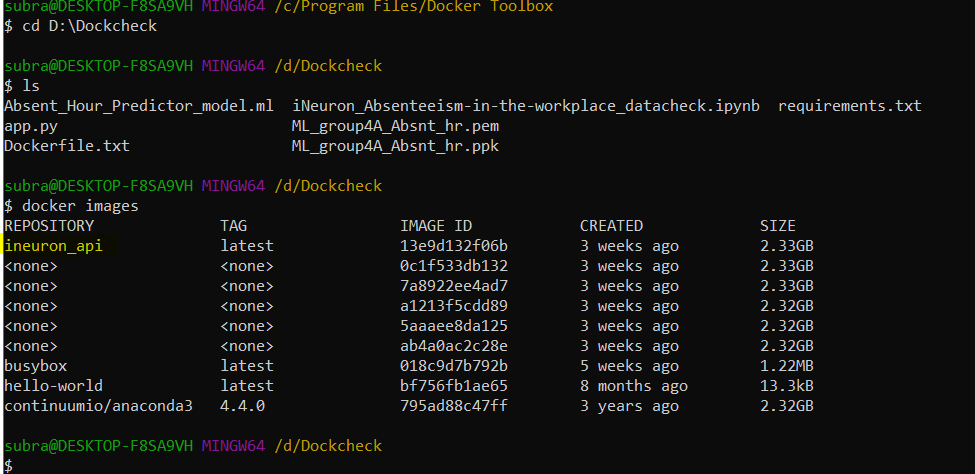
**We tried to create to run a docker image but was getting errors**

**Installed docker:**

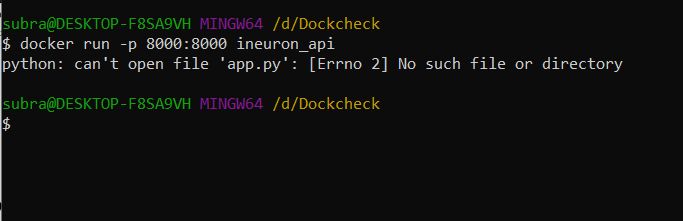




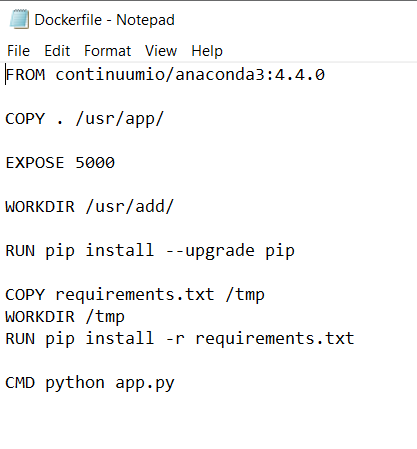
**Created a docker Image:” ineuron\_api”**



**We get error saying app.py file is not present but the Dockerfile has the app.py file which was used to build the image.**



**Docker File used for building the image.**



**Hurdles & Learnings:**

**Hurdles:**

* Deployment part was a big hurdle as previously never done any machine learning stuff beyond Jupyter Notebook.
* Balancing office and this project was also a challenge.
* Collaborating with team members on various modules.

**Learnings :**

* **Practical Machine Learning is different than only running model in Jupyter Notebook using train test split data. Learned about end to end ML.**
* **Learned not only about cloud but also little bit about deployment. Previously had zero knowledge about cloud and deployment.**
* **While working on this project learned many new new things in pandas .**
* **Learned a lot about dumping ML models using pickle & Joblib.**
* **First time built an web application using streamlit in python.**
* **First time implemented modular coding practice (belong to a non coding back ground).**
* **Even though sometimes used to see logs but never really tried building a logging mechanism for my own project. Good learning on this concept also.**
* **Even though was unable to deploy Docker but while trying learned a lot about Docker & containers.**

# Monitoring

Phase 2

No. Of predictions for individual classes

No. of predictions (per day, per hour, per week etc.)

No. of hits

Training data size (number of rows)

Time spent in training

Failures

# Hardware Requirements

## Requirements for model training

The minimum configuration should be:

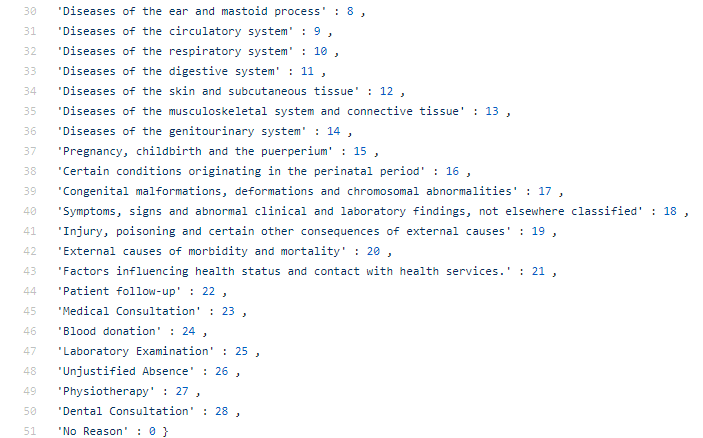
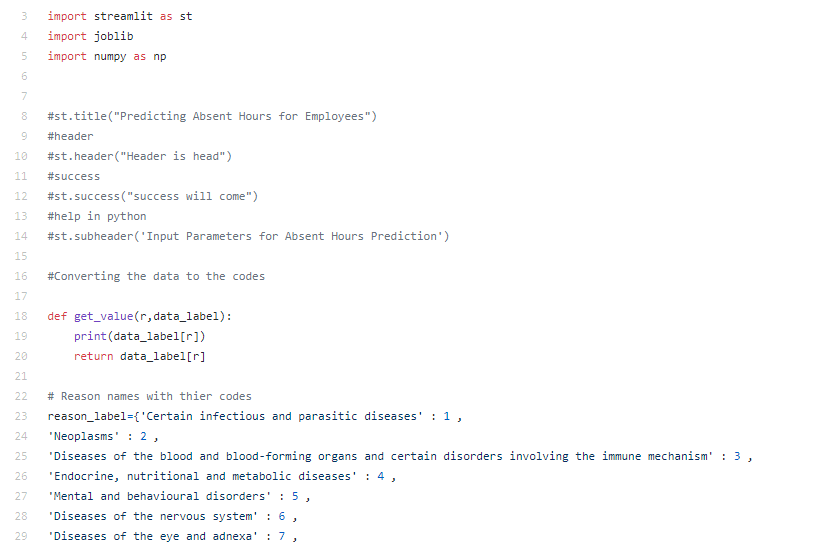
* 8 GB RAM
* 2 GB of Hard Disk Space
* Intel Core i5 Processor

### Requirements for model testing

The minimum configuration should be:

* 4 GB RAM
* 2 GB of Hard Disk Space
* Intel Core i5 Processor

# Sample code and standard to be followed:





**Above code is For UI design .it is created using Streamlit**

In above coding one UI is design in which user having two choices one is uplod file n another one is giving input via ui

In first option user is uplod file on the basis of given instructions

Another one option is giving input via UI n predict Absent Hours

