Absenteeism at work

## Assumptions:

* Provided scenario is case of classification problem with label as ‘Absenteeism time in hours', hence solved using classification technique.
* For Evaluation purpose F1\_score and Accuracy matrix is considered.
* All the features resemble distinct numeric re-occurring characteristic, hence considered as categorical.
* As the target variable is numeric time, hence it can also be solved with regression method, But as specified in the dataset It is meant to be solved with classification purpose.

## EDA

EDA is included in Absenteeism\_at\_work.ipnyb along with plots and its conclusions

1. Target

* Target is evaluated along with its distribution for all the provided variables
* Target contains total 19 distinct classes with highly imbalanced data hence need to be balanced.

1. Features

* Checked weather is any missing value: No missing found
* Plots for each feature with respect to label are plotted with total and average value.
* Conclusions are also provided at the same place

## Data Preprocessing

Data preprocessing is included into Absenteeism\_at\_work.ipnyb in separate module.

1. Target Preprocessing:

* Target is highly imbalanced with 19 distinct classes
* Imbalances is removed by merging least occurrence class with mode class value, leading to reduce in total classes of Target along with reduction in imbalances
* Threshold count of 15 is considered to define class as an outlier class.
* Since classes are imbalanced so need to get class weight, to be used in training of model giving importance to dominated class

1. Feature Preprocessing:

* All features are considered as categorical variables due to its numeric distinct reoccurring nature.
* Best features are selected using chi2 feature selection technique, with threshold of 15 for chi value. As this value is greater than 3.7
* Encoding all the features with one-hot encoder
* After encoding feature matrix is converted into sparse matrix.
* Extraction of best data from sparse matrix using PCA. %0 columns are found with highest productivity of target
* Splitting data into training and testing purpose

## Training model

All models ae used are located in into Absenteeism\_at\_work.ipnyb in separate section

* After data preprocessing and feature extraction multiple models are selected with descent result.
* Xgboost , Randomforest and RidgeClassifier are selected in vanilla version
* Hyperparameter are selected using GridSearchCV.
* An voting ensemble model is generated using all the above individuals model, which outperforms over all the individual model.

## Evaluations

A separate section of evaluation is also provided in the into Absenteeism\_at\_work.ipnyb file

* Since model is classification model, hence Precision, recall, F1\_score and Accuracy are considered as evaluation parameters.

## Deep NN model

Along with classical machine learning model an ANN is also created. Since data size is small, hence results with NN are not very descent. Model is included in the same file with separate section

* All the preprocessing tasks are performed as mention above.
* Target label is One hot encoded as per requirement of NN, since classification approach is used.
* An dense NN model is created with normalization and dropout layer.
* Relu activation function is used in each layer to generate the model.
* For weight updating “ADAM” optimizer is used
* Optimum epoch and batch size is used for model training
* Model is compiled and fitted over training data.
* Graph of accuracy changing along with model training is shown
* Predicted values from the model are reconverted into original format.
* Model is evaluated using same metrices as used in classical model.