Absenteeism at Work Dataset

**Problem Definition:**

The dataset named Absenteeism at work Data Set is available in UCI Machine Learning repository.

*Link:* [*https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work*](https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work)

The data set allows for several new combinations of attributes and attribute exclusions, or the modification of the attribute type (categorical, integer, or real) depending on the purpose of the research. The data set (Absenteeism at work - Part I) was used in academic research at the Universidade Nove de Julho - Postgraduate Program in Informatics and Knowledge Management.

The dataset has 21 attributes which has been discussed as ‘Data Dictionary’ in the my jupyter notebook in detail, later ‘Reason for absence’ attributed has a total of 27 reasons which have been also listed numerically for better understanding.

**Data Analysis:**

Initially I imported the required libraries i.e., pandas, numpy, matplotlib, and seaborn and then imported the data and visualized it.

Data required cleaning, transformation, integration. To find any missing values in dataset, a check was done, and it was seen that dataset does not have any missing values.

The dataset had 740 examples and 21 features which included the target variable (Absenteeism time in hours).

Now it was necessary to select the set of attributes really contributing to the absenteeism at work. To find the correlation between all the other attributes with the target variable a heatmap was created, all attributes having negative correlation with the target variable were dropped. Total 6 attributes were dropped.

Now, a bit of feature engineering was done to extract features from raw data and 8 more features were added to the data which are Age\_Category, smoke\_cat, absenteeism category, Disciplinary cat, drink\_cat, Education\_cat, transportation\_category and distance category.

Now, the dataset has a total of 23 attributes, and we can perform EDA and build machine learning models on it.

### **Exploratory Data Analysis (EDA):**

It is a good practice to understand the data first and try to gather as many insights from it and EDA is done for the same. Exploratory data analysis (EDA) is used by to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

Parameters were analyzed with respect to target variable ‘Absenteeism time in hours’.

* Hours absent vs Social drinker: Both social drinkers and Alcoholics had almost the same number of absent hours.
* Hours absent vs Transportation expense: Expensive and cheap transportation were having almost same, and employees taking very expensive transportation were absent for less hours.
* Hours absent vs Social smoker: Heavy smokers were more absent than social smokers. Employees who smoke tend to have more hours of absence than those who don’t.
* Hours absent vs Age: We can see that young and old employee has the highest numbers of hours absent as compared to the mid age employees. This shows that mid age employees are aware of responsibilities.
* Hours absent vs Education: Employees having education qualification as high school graduates had maximum no. of absent hours following by post graduates then graduates and then doctors. So, as educational qualification increases the hours of absenteeism decreases.
* Hours absent vs Reason of absence: All 27 reasons were plot against the Absenteeism time in hours.
* Hours absent vs Disciplinary action: Disciplinary actions had the highest effect on hours of the absence. When disciplinary measures were taken employees tends to starts to be punctual.
* Hours absent vs Distance from Residence to Work: Surprisingly, employees staying close to the work place had a greater number of absenteeism hours as compared to employees staying far.

**Building Machine Learning Models:**

We split our data into train data and test data in the ratio 70:30 taking our target to be the absenteeism category.

After checking for the balance of target column, we saw that our target column is imbalanced.

Models do not fit well when there is a class imbalance. There are some methods like oversampling, undersampling and mixture. So, to do class oversampling SMOTE (Synthetic Minority Oversampling Technique) is used. It creates new samples along the lines of the existing samples. It had increased the samples of minority classes. The Machine Learning models to find the best solution are:

**1.Naive Bayes:**

Naive Bayes is a simple but surprisingly powerful algorithm for predictive modeling.

Naive Bayes is a classification algorithm for binary (two-class) and multi-class classification problems. The technique is easiest to understand when described using binary or categorical input values.

It is called naive Bayes or idiot Bayes because the calculation of the probabilities for each hypothesis are simplified to make their calculation tractable.

A naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable. Basically, it’s “naive” because it makes assumptions that may or may not turn out to be correct.

After fitting the model, we found that the model accuracy was 0.824, and confusion matrix was also generated.

**2.Logistic Regression:**

Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability.

We can call a Logistic Regression a Linear Regression model, but the Logistic Regression uses a more complex cost function, this cost function can be defined as the ‘Sigmoid function’ or also known as the ‘logistic function’ instead of a linear function.

The hypothesis of logistic regression tends it to limit the cost function between 0 and 1. Therefore linear functions fail to represent it as it can have a value greater than 1 or less than 0 which is not possible as per the hypothesis of logistic regression.

After fitting the model, we found that the model accuracy was 0.833. I generated the confusion matrix and it does somewhat good.

**3.** **Support Vector Machine:**

Support Vector Machine (SVM) is a supervised machine learning algorithm capable of performing classification, regression and even outlier detection. The linear SVM classifier works by drawing a straight line between two classes. All the data points that fall on one side of the line will be labeled as one class and all the points that fall on the other side will be labeled as the second. Sounds simple enough, but there’s an infinite number of lines to choose from. How do we know which line will do the best job of classifying the data? This is where the LSVM algorithm comes in to play. The LSVM algorithm will select a line that not only separates the two classes but stays as far away from the closest samples as possible. In fact, the “support vector” in “support vector machine” refers to two position vectors drawn from the origin to the points which dictate the decision boundary.

After fitting the model, we found that the model accuracy was 0.860.

**4.Random Forest:**

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

After fitting the model, we found that the model accuracy was 0.896, and confusion matrix was also generated.

**Conclusion:**

Random Forest performs the best with accuracy of 0.896 as compared to all other models.

Absenteeism can be avoided before it occurs. Also, the hidden causes of absenteeism could be used to set additional requirement for a new job. Therefore, the percentage of absenteeism could be reduced.

The project is available in my GitHub repository.

*Link:* [*https://github.com/sahil0801/Projects.git*](https://github.com/sahil0801/Projects.git)

Any comment or suggestions for improvement will be helpful.

Thank You!!