Pilot Report study

Machine learning is a in demand branch of computer science and engineering which is mainly used for prediction and decision making purposes. In this project, the feasibility to use machine learning algorithms in the medical sector is analyzed. Two tasks are problems which are solved using two sets of machine learning i.e, classification and regression. It is also analyzed how historical data can help in the medical sector to detect the diseases of the patients. The predictions generation, performance of the model depends upon the correctness in the data.

Predictive task means the algorithms or the models which are implemented in order to generate efficient and accurate predictions. Various predictive tasks are performed in this project, which will eventually predict the diseases of the patient. The predictive tasks performed on the problem are classification and regression. In the classification various models such as decision tree classifier, random forest classifier and gradient boost classifier are implemented. The working of each classifier is to detect the target value i.e, whether the person is diabetic or not. The dataset which is used in the implementation of the first task consists of a limited number of rows and a single target value which are given in the attribute Class. Various preprocessing techniques also need to be implemented before building the model. The techniques include data cleaning, splitting the data using divide and conquer techniques etc.

In order to get more efficiency from the machine learning models, NHC could add more data to the dataset. The current dataset contains nearly 1500 rows which is not enough for any machine learning model to train or generate efficient predictions. The data provided does not have greater correlanaltiy with the target which causes problems at the time of feature and target selection. More correlation attributes can increase the performance of the model in various ways.In the second task, a regression based approach was implemented in order to predict the dose of the drug. Due to continuous values present in the target attribute, classification causes many difficulties and very inefficient performance, so regression models such as linear regression, decision tree regressor and Lasso regression models are implemented in the second part. Due to the predefined target value, there is no major requirement for implementing unsupervised machine learning models such as knn, k means etc. these techniques provide detailed knowledge about the feature values and generate labels according to their distance. These unsupervised learning techniques are implemented when there is no target value present in the dataset.

The performance of the machine learning model can be evaluated using various metrics before deployment. Each metric gives the complete details of the correct classification and also provides detailed information about the misclassified values at the time of predictions. The metrics such as accuracy_Score, classification report, confusion matrix are implemented which are used to get the detailed understanding of both correctly classified and misclassified data. However, these metrics do not support the regression models to find the performance of the regression models, r2 score, root mean squared error, mean squared error metrics are implemented. These metrics are efficient and helps better understanding of the machine learning models before deployment