|  |  |
| --- | --- |
| Name Of The Student | GEETHU ANIL |
| Internship Project Topic | RIO-125: Classification Model- Build a Model that Classifies the Side Effects of a Drug |
| Name of the Organization | TCS ION |
| Name of the Industry Mentor | Himalaya Aashish |
| Name of the Institute | ICT ACADEMY OF KERALA |

|  |  |  |
| --- | --- | --- |
| Date | Day # | Hours Spent |
| 14-3-2021 | 3 | 3.5 |
| Activities done during the day:  I watched and read Project Reference Material,  1**.Data preprocessing steps for machine learning & Data analysis**    **\* FEATURE SELECTION & FEATURE EXTRACTION**  **\* DIMENSIONALITY REDUCTION**  Popular methods for reducing the dimensions of the data  1.principal component analysis(PCA)  2.singular value decomposition(SVD)  **\* DATA VISUALIZATION**  Filling the missing values with mean is important  2.**Taining & Testing data**  Splitting the data in two  Mainly we taken as training data 80% & testing data 20%  3.**Classification VS Regression** | | |

4.**Analytics india mag.com**

In this session explained about 7 types of classification algorithms**,**

1. **LOGISTIC REGRESSION**

**Definition:**Logistic regression is a machine learning algorithm for classification. In this algorithm, the probabilities describing the possible outcomes of a single trial are modelled using a logistic function.

**Advantages:** Logistic regression is designed for this purpose (classification), and is most useful for understanding the influence of several independent variables on a single outcome variable.

**Disadvantages:** Works only when the predicted variable is binary, assumes all predictors are independent of each other and assumes data is free of missing values.

1. **NAIVE BAYES**

**Definition:**Naive Bayes algorithm based on Bayes’ theorem with the assumption of independence between every pair of features. Naive Bayes classifiers work well in many real-world situations such as document classification and spam filtering.

**Advantages:**This algorithm requires a small amount of training data to estimate the necessary parameters. Naive Bayes classifiers are extremely fast compared to more sophisticated methods.

**Disadvantages:**Naive Bayes is is known to be a bad estimator.

1. **Stochastic Gradient Descent**

**Definition:** [Stochastic gradient descent](https://analyticsindiamag.com/a-lowdown-on-alternatives-to-gradient-descent-optimization-algorithms/) is a simple and very efficient approach to fit linear models. It is particularly useful when the number of samples is very large. It supports different loss functions and penalties for classification.

**Advantages:** Efficiency and ease of implementation.

**Disadvantages:** Requires a number of hyper-parameters and it is sensitive to feature scaling

1. **K- Nearest Neighbours**

**Definition:** Neighbours based classification is a type of lazy learning as it does not attempt to construct a general internal model, but simply stores instances of the training data. Classification is computed from a simple majority vote of the k nearest neighbours of each point.

**Advantages:**This algorithm is simple to implement, robust to noisy training data, and effective if training data is large.

**Disadvantages:**Need to determine the value of K and the computation cost is high as it needs to compute the distance of each instance to all the training samples.

### 5. Decision Tree

**Definition:** Given a data of attributes together with its classes, a decision tree produces a sequence of rules that can be used to classify the data.

**Advantages:** [Decision Tree](https://analyticsindiamag.com/hands-on-tutorial-how-to-use-decision-tree-regression-to-solve-machinehacks-new-data-science-hackathon/) is simple to understand and visualise, requires little data preparation, and can handle both numerical and categorical data.

**Disadvantages:**Decision tree can create complex trees that do not generalise well, and decision trees can be unstable because small variations in the data might result in a completely different tree being generated

### Random Forest

**Definition:**[Random forest](https://analyticsindiamag.com/step-by-step-guide-to-reviews-classification-using-svc-naive-bayes-random-forest/) classifier is a meta-estimator that fits a number of decision trees on various sub-samples of datasets and uses average to improve the predictive accuracy of the model and controls over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement.

**Advantages:**Reduction in over-fitting and random forest classifier is more accurate than decision trees in most cases.

**Disadvantages:**Slow real time prediction, difficult to implement, and complex algorithm.

### Support Vector Machine

**Definition:** [Support vector machine](https://analyticsindiamag.com/understanding-the-basics-of-svm-with-example-and-python-implementation/) is a representation of the training data as points in space separated into categories by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

**Advantages:**Effective in high dimensional spaces and uses a subset of training points in the decision function so it is also memory efficient.

**Disadvantages:**The algorithm does not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation.