Dear Intern

Interim project report is an inherent component of your internship. We are enclosing a reference table of content for the interim project report.

The key objective of this report is for you to capture how far you have got in completing the internship work against milestones expected to be achieved within a specific duration and seek the mentor’s feedback. Depending on the internship project and your progress (IT/Non-IT, Technical/Business Domain), you may choose to include or exclude or rename sections or leave some sections blank from the table of content mentioned below. You can also add additional sections. You can refer the project presentation to view the milestones related to your internship project. Please populate milestone# (1 / 2 / 3) and the milestone description in the interim project report based on the milestone for which you are submitting the interim project report.

You can refer the project presentation to view the milestones related to your internship project.

|  |  |
| --- | --- |
| Internship Project Title | RIO-125: Classification Model - Build a Model that Classifies the Side Effects of a Drug |
| Name of the Company | TCS ion |
| Name of the Industry Mentor | Himalaya Aashish |
| Name of the Institute | ICT Academy Kerala |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Start Date | End Date | | Total Effort (hrs.) | | Project Environment | Tools used |
| 23/02/2021 | 15/4/2021 | | 50 | | Jupyter notebook | Excel, Jupyter |
| Milestone # | 2 | Milestone: | | * Encoding is done * Split the dataset into test and train data | | |

**TABLE OF CONTENT**

* Acknowledgements
* Objective
* Introduction / Description of Internship
* Internship Activities
* Approach / Methodology
* Assumptions
* Exceptions / Exclusions
* Charts, Table, Diagrams
* Algorithms
* Challenges & Opportunities
* Outcome / Conclusion
* Enhancement Scope
* Link to code and executable file
* Research questions and responses

**Acknowledgement**

The internship opportunity provided by TCSiON was a great chance for learning and professional development. I express my deepest thanks to the industry mentor for taking part in useful decision & giving necessary advices and guidance, and also express my deepest thanks to all faculty member of ICT Academy of Kerala for their careful and precious guidance which were extremely valuable for my learning.

I consider this as the biggest opportunity that has improve my career development.

**Objective**

Drugs are typically small organic molecules that achieve their desired activity by binding to a target site on a receptor. Drugs can help to treat diseases, but usually come with side effects or adverse reactions. Because of unintended side effects, a great number of approved drugs were even withdrawn from the market. Therefore, recognizing potential side effects helps to reduce costs and avoid risks in the drug discovery.

It could also be helpful for the patients who are buying drugs online to check the side effects of the drugs before buying it.

The main objective of this project to build a classification model that classifies the side effects of a particular drug by age, gender and race. The model need to have good amount of accuracy and have to meet the industry standards.

**Introduction / Description of Internship**

The project guidelines clearly mentioned that we are expected to create a model that classifies the trial data of a drug based on the patients review. At the end of the project we should be able to create a dataset, clean the dataset, sanitize it and preprocess the data to perform data partitioning and handle missing values. Create training and testing sets. Build a classifier and fit the data to the model.

**Internship Activities**

The internship activity is mainly concentrates on how we make up to the objective of the internship. The given resources were very useful to our internship and the day wise plan helps us to calculate the overall time and amount of work to be done each day.

**Approach / Methodology**

The Approach / Methodology used here will be the Linear Strategy which consist in sequential phases with no feedback loops. The project solution is not released until the final phase is reached. This strategy is characterized by clearly defined goal solution and requirements. The pre-defined steps includes data cleaning, data preprocessing , feature processing, splitting to test and train set, applying machine learning algorithms, comparison of machine learning algorithms.

**Assumptions**

By various Exploratory data analysis we can come an assumption that the drug are rated good for the body by chemist, it have a slight side effect of the dataset mainly for depression from the condition attribute mainly concentrate on acne, anxiety, insomnia, birth control, high blood pressure allergies and other mental problem related to brain issues.

**Data Encoding**

Our datset contains categorical values before we split the dataset into two is, train and test data. Here data encoding technique is used to convert categorical data into numerical form.

**Prepare data for training**

Two tasks will be performed which is resultant data is then divided into training and test sets. Training and Testing Data The data is split into training (80%) and testing (20%) data sets through random sampling. The following machine learning techniques were considered in the experiment. As the dataset is labeled properly, it is considered to be used for supervised learning. In order to find out the best machine learning technique, different machine learning techniques were tested

**Algorithms**

1. **Random Forest Classifier**: an ensemble learning based regression model. It uses a model called decision tree, specifically as the name suggests, multiple decision trees to generate the ensemble model which collectively produces a prediction. The benefit of this model is that the trees are produced in parallel and are relatively uncorrelated, thus producing good results as each tree is not prone to individual errors of other trees. The RandomForestClassifier class of the sklearn.ensemble library is used to Classifier problems via random forest. The most important parameter of the RandomForestClassifier class is the n estimators parameter. This parameter defines the number of trees in the random forest.
2. **Bernoulli Naive Baye’s**

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes’ theorem [BernoulliNB](https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.BernoulliNB.html#sklearn.naive_bayes.BernoulliNB) implements the Naive Baye’s training and classification algorithms for data that is distributed according to multivariate Bernoulli distributions; i.e., there may be multiple features but each one is assumed to be a binary-valued variable.

1. **Decision Tree Classifier**

Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves.

**2. The k-nearest neighbors (KNN)** algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It’s easy to implement and understand, but has a major drawback of becoming significantly slows as the size of that data in use grows

**Outcome / Conclusion**

Using above mentioned models we can calculate the different accuracy of models.