Dear Intern

Project report is an inherent component of your internship. We are enclosing a reference table of content for the project report. Depending on the internship project (IT/Non-IT, Technical/Business Domain), you may choose to include or exclude or rename sections from the table of content mentioned below. You can also add additional sections. The key objective of this report is for you to systemically document the project work done.

|  |  |
| --- | --- |
| Internship Project Title | TCS iON 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 Ashish |
| Name of the Institute | ICT Academy |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Start Date | End Date | Total Effort (hrs.) | Project Environment | Tools used |
| 18/02/2021 |  |  |  | Python |

**TABLE OF CONTENT**

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

**ACKNOWLEDGEMENT**

Successful completion of an internship requires help from many persons. First of all, I would like to express my gratitude to my Industry Mentor **Ms. Himalaya Ashish** for support and guidance throughout this project.

I am great full to **ICT Academy of Kerala** who is the reason for my internship project. Their guidance and support were my motivation to this internship.

I am thankful to the Almighty God and also to my friends whoever helped me to complete the project.

SHARON SHAJI

**INTRODUCTION**

A side effect is usually regarded as an undesirable secondary effect which occurs in addition to the desired therapeutic effect of a drug or medication. Side effects may vary for each individual depending on the person's disease state, age, weight, gender, ethnicity and general health. Drug side-effects, or adverse drug reactions, have become a major healthcare concern.

Side effects can occur when commencing, decreasing/increasing dosages, or ending a drug or medication regimen. Side effects may also lead to non-compliance with prescribed treatment. When side effects of a drug or medication are severe, the dosage may be adjusted or a second medication may be prescribed. Lifestyle or dietary changes may also help to minimize side effects.

 Drug discovery is a time-consuming and laborious process. [Developing drugs](https://en.wikipedia.org/wiki/Drug_development) is a complicated process, because no two people are exactly the same, so even drugs that have virtually no side effects, might be difficult for some people. Also, it is difficult to make a drug that targets one part of the body but that does not affect other parts, the fact that increases the risk of side effects in the untargeted parts. The identification of potential severe adverse side-effects is a challenging issue at many stages of the drug development process. Some side effects are discovered after a drug has been approved. However, most drugs become safer over time, as more people use them, and more information is collected.

**PROJECT OVERVIEW**

Classification model – Build a model that classifies the side effects of a drug.

Create a model that classifies the trial data of a drug based on their age, gender and race by creating a dataset of 4,00,000 patients containing the details of their Name, Age, Gender, Race and Side effects.

**OBJECTIVE**

The objective of this project is to build a classification model that classifies the side effects of a particular drug by age, gender and race.

**DATA DESCRIPTION**

Dataset is created with the help of ‘faker’ in python. Dataset has 4,00,000 rows and 5 columns and the features are;

* **Name:** Name of the patients.
* **Age:** Age of each patient.
* **Gender:** Gender of the patients (M-Male and F-Female).
* **Race:** Race of each patient (Asian, Black, Hispanic and White).
* **Side effect:** This feature shows whether the particular drug has side effect for each patient (yes or no).

**METHODOLOGY**

Methodologies used in this project are

* Importing the necessary libraries
* Creating dataset by using faker
* Preprocessing
* Splitting the dataset
* Classification model creation

**ACCURACY MEASURES**

1. **ACCURACY SCORE**

It is the most intuitive performance measure and it is simply the ratio of correctly predicted observations to the total observations.

1. **PRECISION SCORE**

It is the ratio of correctly predicted positive observation to the total predicted positive observations.

1. **RECALL SCORE**

It is the ratio of correctly predicted positive observations to the all observations in actual class.

1. **F1 SCORE**

It is the weighted average of precision and recall scores. Therefor this score takes both false positive and false negative into account.

**PYTHON; THE TOOL USED**

python is an interpreted, [high-level](https://en.wikipedia.org/wiki/High-level_programming_language) and general-purpose programming language. Python's design philosophy emphasizes [code readability](https://en.wikipedia.org/wiki/Code_readability) with its notable use of [significant indentation](https://en.wikipedia.org/wiki/Off-side_rule). Its [language constructs](https://en.wikipedia.org/wiki/Language_construct) and [object-oriented](https://en.wikipedia.org/wiki/Object-oriented_programming) approach aim to help [programmers](https://en.wikipedia.org/wiki/Programmers) write clear, logical code for small and large-scale projects. python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured, object oriented and functional programming. Python is often described as a "batteries included" language due to its comprehensive [standard library](https://en.wikipedia.org/wiki/Standard_library).

**PREPROCESSING OF DATA**

**MISSING VALUES**

Many real-world datasets may contain missing values for various reasons. They are often encoded as NaNs, blanks or any other placeholders. Training a model with a dataset that has a lot of missing values can drastically impact the machine learning model's quality.

Checked for the missing values in the dataset of side effects of drug and found that there were no missing values in it.

**LABEL ENCODING**

Label Encoding refers to converting the labels into numeric form so as to convert it into the machine-readable form. Machine learning algorithms can then decide in a better way on how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.

Since the features ‘Gender’, ‘Race’ and ‘Side effect’ of the dataset are object type, I have done label encoding to those features and as a result the objects are converted to equivalent numeric values.

**FEATURE REDUCTION**

Dropped the feature ‘Name’ as it is unique for each row and has not much importance in model creation.

**FEATURE SCALING**

Feature scaling is a technique to standardize the independent features present in the data in a fixed range. I have done feature scaling using Standard Scaler which makes the values of each feature in the data have zero-mean (when subtracting the mean in the numerator) and unit-variance.

**SPLITTING THE DATA**

The dataset is divided into dependent (target) variable, y and independent variable x.

In my project the feature ‘side effect’ is the target or dependent variable. Then these dependent and independent variables are further split into train and test data such that train set has 70 percent and test set has 30 percent of the data.

**MODEL CREATION**

* **LOGISTIC REGRESSION**

Logistic regression is a classification algorithm, used when the value of the target variable is categoricalin nature. Logistic regression is most commonly used when the data in question has binary output, so when it belongs to one class or another, or is either a 0 or 1.

Values of accuracy measures for logistic regression are;

|  |  |
| --- | --- |
| Accuracy score | 0.5010 |
| Precision score | 0.5011 |
| Recall score | 0.2703 |
| F1 score | 0.3511 |

* **RANDOM FOREST CLASSIFFIER**

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction.

Accuracy measures of random forest classifier are

|  |  |
| --- | --- |
| Accuracy score | 0.5005 |
| Precision score | 0.5001 |
| Recall score | 0.4862 |
| F1 score | 0.4930 |

* **DECISION TREE CLASSIFIER**

Decision Trees are a type of Supervised Machine Learning where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split.

The accuracy measures of decision tree classifier are

|  |  |
| --- | --- |
| Accuracy score | 0.5008 |
| Precision score | 0.5004 |
| Recall score | 0.4723 |
| F1 score | 0.4859 |

* **NAÏVE BAYES CLASSIFIER**

naive Bayes classifiers are a family of simple "[probabilistic classifiers](https://en.wikipedia.org/wiki/Probabilistic_classification)" based on applying [Bayes' theorem](https://en.wikipedia.org/wiki/Bayes%27_theorem) with strong [independence](https://en.wikipedia.org/wiki/Statistical_independence) assumptions between the features and are highly scalable.

Accuracy measures are

|  |  |
| --- | --- |
| Accuracy score | 0.5010 |
| Precision score | 0.5010 |
| Recall score | 0.3064 |
| F1 score | 0.3803 |

**VISUALIZATIONS OF THE MODELS**

* **DISTRIBUTION PLOT**

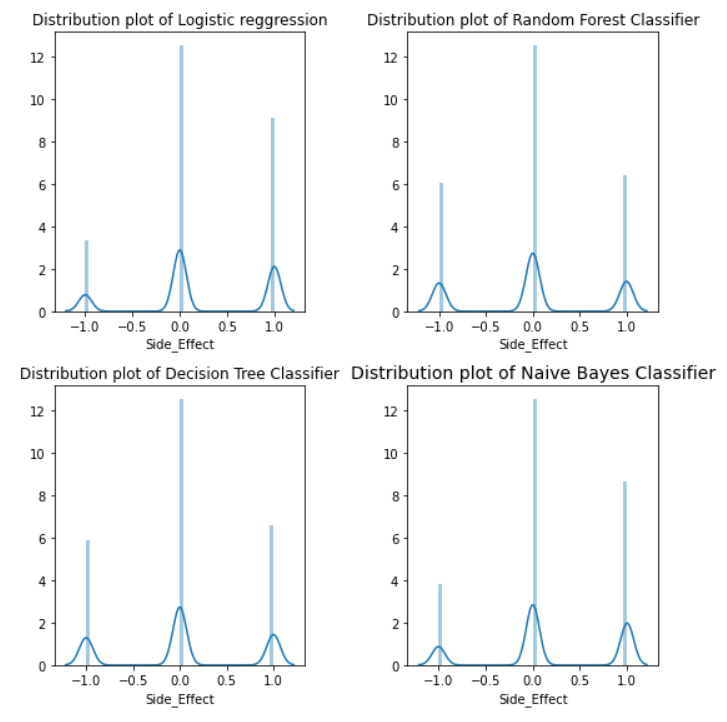
****

Fig 1: *distribution plot of classification models*

* **BAR DIAGRAM**

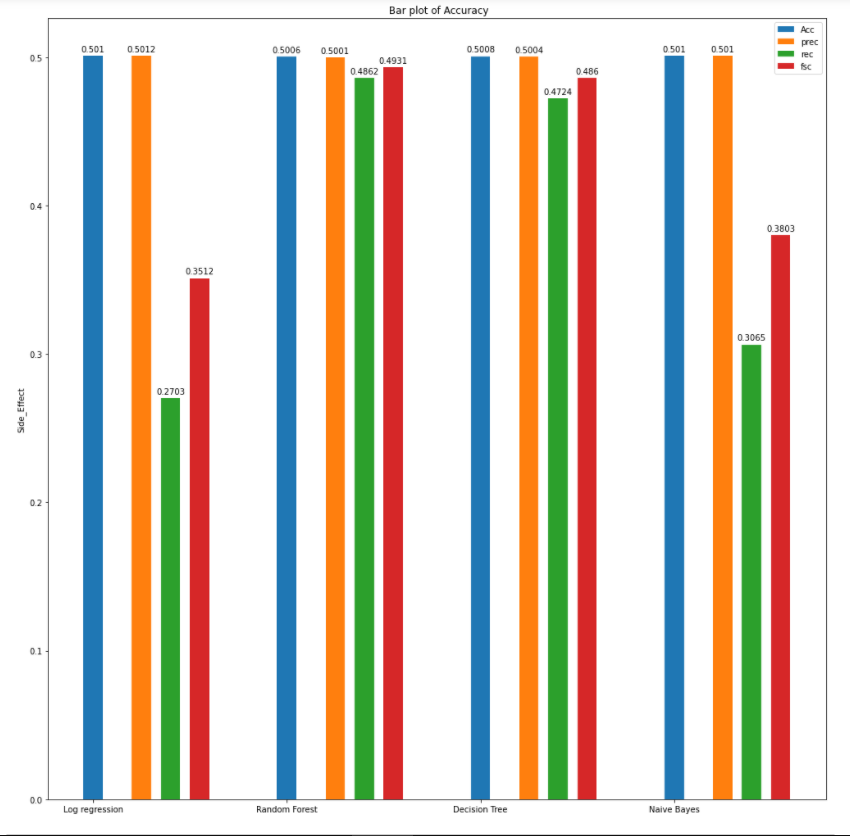


Fig 2: *Bar diagram of accuracy measures*

**CONCLUSION**

Created a dataset of side effects of a particular drug for 4,00,000 patients with their Name, Age, Gender and Race. Done with preprocessing, splitting and model creation for the dataset.

Accuracy score, precision score, recall score and f1 score are used as accuracy measures for evaluating the good model for the data.

Found that accuracy score, precision score, recall score and f1 score are comparatively high for Random Forest classifier, so I conclude Random Forest classifier as the good classification model for my data.

**Link for Project Presentation**

https://www.loom.com/share/e50ae7baf41c464e8adcd81c88aee36d