II. THEORETICAL FEASIBILITY OF THE TECHNOLOGY

Machine Learning

Machine learning is a method by which IT systems use pattern recognition to solve problems. Therefore, LUCC can identify trends and create alternatives with the current methodologies and existing knowledge. Machine learning operates from experiences and concludes these experiences [2].

Types of Machine Learning

There are 3 types of Machine Learning techniques,

1. Supervised

Machine learning supervised is the search for algorithms that give rise to generic hypotheses from externally supplied circumstances that predict future circumstances [3].

2. Unsupervised

Unsupervised learning is an advanced model methodology where users don't have to track the model. Rather, it helps the model to work itself to discover previously undetected patterns and information [4].

3. Reinforcement

The computer is setting up the options. The agent learns how to deal with the unexpected. AI could improve learning by implementing game-like approaches [5].

A statistical model is built for a specific operation or predicting the value. Modeling helps to solve business problems. A model is a simple equation. It has input variables and output variables. The independent variables give us the output value. Explanatory variables are independent variables. The dependent variable is output.

As problems arise, the model must evolve. Modeling is a complicated process. Algorithms are used to implement models. Algorithms can be derived from models. Algorithms can be classified into supervised, unsupervised, and reinforcement learning [6].

A. Random Forest

Classification is predicting the pre-defined class label of the given instance. Classification belongs to supervised machine learning. Classification methods can be classified into two types: binary and multi-class. The machine learning classifiers include logistic regression, KNN, SVM, decision tree, Naïve Bayes, and Random forest (2020). One of the supervised classifications and regression algorithm is Random Forest. The algorithm produces a greater number of trees. Random forests are groups of decision trees. It acts as an ensemble. Decision trees classify the class of an instance. The highest vote from the class will be the final class label. Each decision tree generates its class label.

The decision will be made based on the highest number of decision tree classifiers. More trees in the forest improve the accuracy of the model [7].

Figure 25 (Decision Tree)

B. How Random Forest works

Classifiers include both predictors and responses. All predictor variables are used to predict the response variable. Random forests are a collection of decision trees. Random forest techniques select predictors randomly and predict the desired response variable. Random forest considers the average decision tree result. Each decision tree predicts the response variable. The most important variable will be the final decision of the random forest [8].

The decision tree is built with the entire dataset and all predictors whereas a random forest includes multiple decision trees built with a part of the dataset. The decision tree classifier is overfitted. It classifies by building the model with the training instance. Overfitting occurs for predicting the new instance. This algorithm is powerful. The output combines multiple decision trees.

Information gain takes place at each node. It predicts the instance where the information gain is high. The search continues until further information is not gained [9].

Decision trees are weak learners. Random forests also train the same weak learners. It uses more than one classifier and combines the outputs of all classifiers. It works similarly to the decision tree. Each decision tree does not provide a complete predictive model. It pulls data for random instances and learns to classify instances. Multiple decision trees are combined to derive the final decision. It improves the classification accuracy and reduces bias [10].

C. Advantages of Random Forest

The algorithm is a powerful ensemble learning (bagging) algorithm. It uses the ensemble learning method. It has many trees and utilizes a subset of data. It uses the combined output of multiple trees. It reduces overfitting as well as variance. It increases accuracy. It can be used for solving both classifications as well as regression problems. It can also work with continuous and categorical variables. It automatically handles the missing values. It follows a rule-based approach. It does not require feature scaling. It handles the nonlinear parameters effectively than curve-based methods. It outperforms the nonlinear independent variables.

It efficiently handles the missing values as well as outliers. It is a stable method. The method is having less impact on noise [11].

D. Limitation of Random Forest

The random forest model has more decision trees. It creates 100 trees in Python. More trees increase the computational power and resources that are available. Besides this, trees or random forests have several other limitations such as: Using any loss function such as cross-entropy for regression or classification. It can be harder to interpret, interpret RFs can be memory-hungry, and it may not be effective in predicting the other three classifications. Our models need longer training [12].

E. Framework

Two statistical tools will be used to collect data, R and Python. Both are open-source statistical tools. The two tools add new libraries with every release. Python is the tool primarily used for data analysis, and R is a general tool. Python and R are both powerful programming languages used for data science [13].

Factors	**Why it matters**
Cost	Cost association with any work is a burden, fortunately, most of the data analytical tools are open source and free of cost.
Easy to use	Any analytical tool should be user-friendly, and it is very necessary even business point of view. Here in data science most of the tolls are user friendly and easy to use.
Easy to learn	It is a vital point that data analytical tools should be easy to learn. The most popular and famous tools are quite easy to learn.
Community Support	Community support for any IT-based programming language or package is very important, community support is very essential for the development of IT.
Range of libraries	Libraries made it easy for coding work. A data scientist can use and escape a huge coding headache.
Computational speed	Computational speed is a core requirement for data analytics. We need both hardware and tools to be speedy and stable for data statistics because huge hidden calculations are there in data & image processing.
Memory consumption	Saving memories is an important function to create and built an application. Tools that usage fewer memories are always fast, and they become more popular and do well in business too.
Visualization capability	In data science, visualization is an important part. People want to understand from pictorial message most of the time because that is attractive fast and easy to gain knowledge.
System integration	In data science or IT, any system and package that comes with more integration capacity with another system gets more popularity and useful

F. R-code

R is an open-source implementation of the R programming language. The language is used for statistical computing and graphical illustration of results. It is the best tool for performing statistical analysis on the dataset [14].

Advantages

R language has some advantages, such as…

It contains a variety of libraries and tools for specific data operations. The library has various commands for data pre-processing, analysis, visualizations, etc.

R expands packages and libraries, adding more packages and libraries to the R environment. It keeps research and analysis of new methods and techniques.

IDE environment for building applications. Code highlight, compilation, help, visualizations of data, etc.

Interacts with a variety of operating systems, like Unix, Windows, etc.

It offers a web-based dashboard for data analysis and visualization [15].

G. Python

The language was released in the year 1991 and got a good reputation due to its simple nature.

It provides support for various applications such as data science, image processing, websites and back end services, and so on [16].

Advantages

The following are the main advantages of Python such as It has Simple syntax which can be easily understood

Wider usage of the tools that enable data analysis to clarify queries more quickly through collaboration

It supports third-party modules as well as libraries

It provides support for both procedural as well as to object-oriented language [17]

R code and Python

Both languages are more popular with data scientists. The data from the Stack Overflow Python Survey of 2019 shows that Python is used in 4th place. R is in 16th place.