Phase 2:Innovation

In this section we need to put our design into innovation to solve the problem. We have to consider incorporating advanced machine learning algorithms for predictive analysis or anomaly detection in the big data.

Algorithms used in big data analysis:

Random Forest:

Random Forest is perhaps the most popular classification algorithm, capable of both classification and regression. It can accurately classify large volumes of data.

Generalized Linear Model (GLM) for Two Values:

The Generalized Linear Model (GLM) is a more complex variant of the General Linear Model. It takes the latter model's comparison of the effects of multiple variables on continuous variables before drawing from an array of different distributions to find the "best fit" model.

Gradient Boosted Model (GBM):

The Gradient Boosted Model produces a prediction model composed of an ensemble of decision trees (each one of them a "weak learner," as was the case with Random Forest), before generalizing. As its name suggests, it uses the "boosted" machine learning technique, as opposed to the bagging used by Random Forest. It is used for the classification model.

K-Means:

A highly popular, high-speed algorithm, K-means involves placing unlabeled data points in separate groups based on similarities. This algorithm is used for the clustering model

Prophet:

The Prophet algorithm is used in the time series and forecast models. It is an open-source algorithm developed by Facebook, used internally by the company for forecasting.

Steps to be followed in the cloud database :

Step 1: Data Collection and Preparation

Data Sources Identification: Identify the sources of the data you plan to analyze. This could include internal databases, external APIs, or third-party data providers.

Data Extraction and Integration: Extract data from various sources and integrate it into a unified format suitable for analysis.

Data Cleaning: Cleanse the data to handle missing values, outliers, and inconsistencies.

Data Transformation: Prepare the data for analysis by performing tasks like normalization, aggregation, or feature engineering.

Step 2: IBM Cloud Setup

IBM Cloud Account Creation: If not already done, create an IBM Cloud account.

Select Database Service: Choose the appropriate IBM Cloud Database service based on your data requirements (e.g., IBM Db2, IBM Cloudant, or IBM Db2 on Cloud).

Provision Database: Set up the database instance on IBM Cloud, configure security settings, and ensure proper access control.

Step 3: Data Loading and Storage

Data Ingestion: Load the prepared data into the IBM Cloud Database. Depending on the volume, you may need to implement data ingestion pipelines.

Data Security: Implement encryption and access controls to protect sensitive data.

Data Backup and Recovery: Establish backup and recovery mechanisms to safeguard data.

Libraries used in ibm cloud database(python):

```
import pandas as pd

from sklearn.model_selection import train_test_split

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import accuracy_score

import ibm_db

import ibm_db_sa

from sqlalchemy import create_engine
```

Steps in cloud database to work with datasets:

Let us consider the data is using a catalogue of journal articles from 51 different journals published during 2015. Articles are published in different languages, by different publishers and under different licences.

Import:

1. Download the CSV files from any websites.
2. Start a New Database **Database -> New Database**
3. Start the import **Database -> Import**
4. Select the file to import (start with articles.csv)
5. Give the table a name that matches the file name (articles, journals, licences, languages publishers), or use the default
6. Since the first row has column headings, check the “First row contains column names”- box
7. Under “Fields separated by”, check “Comma”. Ensure ‘Ignore trailing Separator/Delimiter’ is left *unchecked*.
8. Also, under “Fields enclosed by”, ensure that “Double quotes if necessary” is left *checked*.
9. Press **OK**
10. When asked if you want to modify the table, click **OK**
11. Set the data types for each field and INTEGER for fields with numbers:
12. Click **OK**