# Database and Analytics Programming

1st Suraj Mondal
22169172

2nd Ujjwal Kumar
22164332

3rd Aaditya Gajendragadkar
22158758

***Abstract*—The analysis of huge datasets gathered from multipleinternet sources is covered in this article utilizing the Python programming language and the Jupyter Notebook environment, with unstructured data saved in Mongo DB in JSON format. The purpose of this work is to analyse the features of programming languages used for data analysis, to identify tools and strategies for data cleaning, wrangling, and validation, as well as to pinpoint difficulties in processing large data analytics. The paperalso suggests suitable solutions to these problems, including the use of distributed storage systems, real-time processing and analysis, protection of data privacy, and hiring qualified data scientists and analysts. For organizations looking to use big data analytics for insights and decision-making, the paper offers helpful insights overall. Four machine learning models—logistic regression model, decision tree classifier model, random forest classifier model, and k-neighbour classifier model—are examined on three different datasets.**

***Index Terms*—programming, big data, machine learning, big data analysis, techniques**

## I. Introduction

The motivation behind the paper is to analyse the large datasets that have been collected through different online sources like data.world bank and data.gov. The purpose of performing analysis is to evaluate the different characteristics of the programming languages that are being used to analyse the data. In the context of the current investigation the programming language that have been utilised is python programming language and the platform that will be utilised will be jupyter notebook for performing analysis and Mongo DB for storing the unstructured data in the Json format.

**The objectives of the project are:**

- To analyse the characteristics of the programming languages that are used to perform the analysis.
- To evaluate the different techniques and tools for managing the data for the purpose of data cleaning, wrangling and validation.
- To identify the challenges that are associated with the processing of the big data analytics.
- To purpose the appropriate countermeasures for the challenges that are associated with the processing of the big data analytics.

## II. RELATED WORK

### A. *Characteristics of the programming languages that are used to perform the analysis*

The particular needs of the work, the type of data, and the preferences of the data analysts all influence the programming language that is used for data analysis. However, there are several traits that programming languages have that make thempopular for data analysis:

- Interactive use: A good data analysis language should enable the user to quickly examine and alter data.
- Data structures: Simple to use and modify data structureslike arrays, lists, and dictionaries should be included in the language [1].
- Functions should be supported by the language, enabling users to group frequent data manipulation operations intoreusable code blocks [2].
- Libraries: A comprehensive selection of libraries for dataprocessing, statistical analysis, and visualisation should be made accessible for the language [1].
- Speed: Data analysis operations might be computationally intensive depends upon size of dataset. Consequently,the language must to be able to deal with big datasets effectively [2].
- Open-source: Since many data analysis jobs call for customised code, open-source languages provide programmers the freedom to write and change code as necessary [1].

Python, R, SAS, SQL, and MATLAB are a few examples of programming languages that are frequently used for data analysis. The choice of language will rely on the precise requirements of the assignment because every language has strengths and disadvantages of its own.

### B. *Different techniques and tools for managing the data for the purpose of data cleaning, wrangling and validation*

There are several methods and technologies available to handle the crucial processes of data cleansing, wrangling, and validation in the data analysis process. Here are some typical methods and equipment:

Data cleaning: Finding, fixing flaws, inaccuracies and in- consistencies in the data is called as data cleaning. Typical methods and resources for cleaning data include:

- Regular expressions are patterns that may be used to match and edit text. They may be used to locate andswap out data patterns [3].
- Data profiling: Data profiling entails examining the data to spot trends and irregularities. This can assist in locatingproblems with data quality that need to be fixed [3].

Data wrangling is reorganising and altering data to make it simpler to analyse. Typical methods and resources for data wrangling include:

- Pivot tables: You may aggregate and summarise data using pivot tables.
- Data joining and merging may be used to integrate data from several datasets into a single dataset while dealing with them [4].
- Data transformation capabilities are incorporated into the majority of computer languages used for data

analysis, including grouping, filtering, and sorting [4].

Data validation: Data verification entails making sure the data is correct, comprehensive, and consistent. Typical methods andresources for data validation include:

- Data profiling: Data profiling may be used to find flawsand inconsistencies in data [5].
- Data consistency, correctness, and completeness may allbe checked using data quality criteria.
- Finding outliers and other abnormalities in the data maybe done via statistical analysis [5].

Overall, the selection of data management methods and tools will be influenced by the particulars of the work at hand as well as the properties of the data being examined.

### *C. To identify the challenges that are associated with the processing of the big data analytics*

Organisations may find it challenging to get insights from their data as a result of the difficulties associated with processing big data analytics. The following are some typical difficulties encountered while processing large data analytics:

- Volume: Big data analytics involves handling enormous volumes of data, which may be challenging to store and analyse using conventional technology. Large-scale data management need specialised infrastructure and technologies [6].
- Velocity: Due to the extraordinary rate at which data is produced, it might be challenging to analyse and process data in real-time. To swiftly extract insights, organisationsneed to analyse and process data [7].
- Big data analytics necessitates the processing and storageof sensitive data, which is susceptible to online attacks. In order to prevent data breaches and unauthorised access, it is essential to ensure the security and privacy of thedata [6].
- Big data analytics call for specialised knowledge and abilities in order to handle and analyse data. Finding skilled data scientists and analysts who can glean insights from their data may prove difficult for organisations [7].

Overall, specialised infrastructure, tools, and knowledge are needed for handling and analysing big data analytics. To take advantage of the insights that may be derived from big data analytics, organisations must overcome these issues.

### *D. To purpose the appropriate countermeasures for the challenges that are associated with the processing of the big data analytics*

Big data analytics processing presents a number of difficulties that may be overcome with the right defences. The following are some typical workarounds for the difficulties in processing large data analytics:

- Volume: Organisations can utilise distributed storage systems like Hadoop Distributed File System or cloud storage systems like Amazon S3 to handle the volume of data. These systems let businesses to store massive amounts of data across several computers, simplifying the management and processing of massive amounts of data [8].
- Velocity: Organisations may process data in real-timeusing real-time data processing tools like Apache Spark or Apache Flink. These technologies let businesses to process data as it is produced, allowing them to swiftly get insights [9].
- Security and privacy: Organisations can use access re- strictions, encryption, and monitoring systems to guard against cyber risks and secure the security and privacy of data. Additionally, businesses can adopt privacy policies and follow data protection laws like GDPR and CCPA [8].
- Lack of skills: Organisations can engage in training and development programmes to upskill their personnel or recruit data scientists and analysts with the required qualifications to alleviate the skill deficit [9].

## III. METHODOLOGY

### *A. Datasets*

There are total of three datasets that have been used in the study and they have been extracted from data.gov. All threeof the datasets highlights a different issue that can be resolved with the help of the programming languages. Out of thesethree datasets two are semi-structured and one is structured datasets. The structured dataset is named as 'United States COVID-19 Cases and Deaths by State over Time', the nameof the two semi-structured dataset are 'Provisional COVID-19 Deaths by HHS Region, Race, and Age and 'Provisional_COVID-19_Deaths_by_Sex_and_Age.' The semi-structured datasets are in the form of Json and structured data is used in the form of CSV.

The reason behind selection of "Provisional_COVID-19_Deaths_by_Sex_and_Age" dataset for study impact on deaths due to covid-19 along with the pneumonia, and influenza. Based on the information in the dataset it will be easy to identify effect of the corona virus and pneumonia, and influenza disease at the same time [10].

Reason behind selection of 'Provisional COVID-19 Deaths by HHS Region, Race, and Age' for analysis is that it is helpful in analysing the effect of covid-19 pandemic based on the different categories that can be recognized as geographical region, race and age. This particular dataset will allow the understand the covid-19 trends based on the race through which it can be identified that are mentioned within the dataset [11].

The rationale behind selection of 'United States COVID-19 Cases and Deaths by State over Time' is that it will be helpful in assessing how well public health initiatives and programmes are working to stop the COVID-19 virus from spreading. It is feasible to determine if a programme had a good impact by comparing the number of cases and fatalities that happened before and after the execution of the programme or intervention [12].

### B. *Descriptions and justifications of the implemented data processing algorithms*

- The Decision Tree Classifier is a straightforward yet effective machine learning technique that functions by creating a tree-like model of decisions and their potential outcomes. Both classification and regression tasks employit. A decision tree divides the data into more manageable groups in accordance with a list of predetermined criteria[13].

  Justification: Decision Tree Classifier is a well-known method because to its readability, simplicity, and capacityto handle both continuous and categorical data. Additionally, it can tolerate missing values and is resistantto overfitting [13].

- The Random Forest Classifier is an ensemble learningtechnique that builds several decision trees and thencombines their outputs to get a more precise and reliableprediction [14]. All of the decision trees in the forest'spredictions are combined to get the final forecast. Justification: Random Forest Classifier is a well-liked technique since it can handle high-dimensional data withnumerous characteristics and is incredibly accurate andnoise-resistant. Compared to decision trees, it is lessprone to overfitting, and an ensemble of decision treescan produce predictions that are more consistent andtrustworthy [14].
- Logistic Regression: A statistical technique for binary classification tasks is logistic regression. A logistic func-tion is used to model the connection between the predictorfactors and the binary outcome variable [15]. The logisticfunction converts the probability of the binary outcomevariable to the continuous predictor variables. Justification: Because of its ease of use, readability, andeffectiveness, logistic regression is a well-liked method. It may be used for binary classification jobs and can handleboth categorical and continuous data [15]. Additionally, it can tolerate missing values and is resistant to overfitting.
- K-Nearest Neighbors Classifier: A non-parametric lazy learning method used for classification problems is called the K-Nearest Neighbours Classifier. The majority class of the new data point's k-nearest neighbours in the featurespace is used to categorise it [16].

  Justification: The K-Nearest Neighbours Classifier technique is much-liked because of how straightforward it is to use and how well it can handle categorical and continuous data [16]. It is appropriate for complicated and non- linear datasets since it does not make any assumptions about the distribution of the data.

### C. *Justifications for the choice of technologies used in the project*

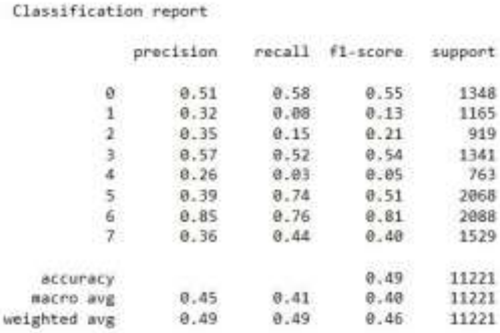
- Python is a well-liked programming language for machine learning and data analysis because of its ease of use, adaptability, and extensive ecosystem of libraries. NumPy, Pandas, Scikit-Learn, TensorFlow, and Keras arejust a few of the many data analysis and machine learningmodules available in Python [17]. These libraries offer high-level abstractions for carrying out challenging ma- chine learning and data analysis tasks, which accelerates and optimises development.
- Jupyter Notebook: Users are able to create and transfer documents using real-time code, calculations, graphics, and text using the free and open-source online application Jupyter Notebook. [18]. Because Jupyter Notebook enables users to write and run code, visualise data, and record their work in a single environment, it is a greattool for data analysis and machine learning.
- MongoDB A common NoSQL document-oriented database called MongoDB offers a scalable and adaptable method for storing and managing data. MongoDB is a great option for data processing because it offers a flexible structure that makes it simple to manipulate and analyse data [19]. MongoDB is also quite scalable and has no trouble handling big datasets and applications with lots of traffic.
- PostgreSQL: The SQL language is utilised and expanded upon by PostgreSQL, an open-source relational database management system (RDBMS). It is renowned for being dependable, strong, and capable of handling complicated data types and queries. The execution plan that the PostgreSQL planner creates for a specific statement is returned by the EXPLAIN command.
- SQL: A domain-specific programming language called Structured Query Language (SQL) is used to handle data stored in relational database management systems (RDBMS) or for stream processing in relational data stream management systems (RDSMS). In order to query the data in a relational database, SQL was created specifically for that purpose. Instead of being imperative like C or BASIC, SQL is a set-based, declarative programming language.

## IV. Results and Evaluation

### A. *Analysis from 'Provisional COVID-19 Deaths by HHS Region, Race, and Age' dataset*

#### 1) *Logistic regression*

It is identified from the figure 1 classification report that accuracy that has been achieved by the model whitening the data set is 38 percent and during testing the results accuracyof 39 percent was achieved.

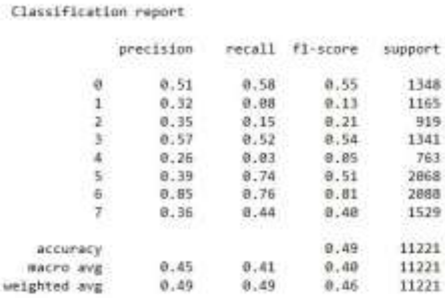


```
Classification report

              precision    recall  f1-score   support

           0       0.51      0.58      0.55      1348
           1       0.32      0.08      0.13      1165
           2       0.35      0.15      0.21       919
           3       0.57      0.52      0.54      1341
           4       0.26      0.03      0.05       763
           5       0.39      0.74      0.51      2068
           6       0.85      0.76      0.81      2088
           7       0.36      0.44      0.40      1529

    accuracy                           0.49     11221
   macro avg       0.45      0.41      0.40     11221
weighted avg       0.49      0.49      0.46     11221
```

Fig. 1. Classification report for Logistic regression

*2) Decision tree classifier*

Based on the results highlighted in the figure 2 it is identifying that accuracy that has been achieved by the model during training the decision tree classifier is 55 percent andit is also identify that accuracy for testing the data is also 49 percent.

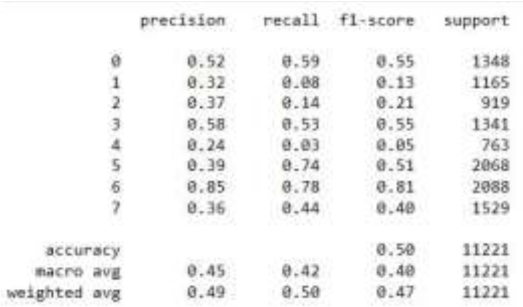


```
Classification report

              precision    recall  f1-score   support

           0       0.51      0.58      0.55      1348
           1       0.32      0.08      0.13      1165
           2       0.35      0.15      0.21       919
           3       0.57      0.52      0.54      1341
           4       0.26      0.03      0.05       763
           5       0.39      0.74      0.51      2068
           6       0.85      0.76      0.81      2088
           7       0.36      0.44      0.40      1529

    accuracy                           0.49     11221
   macro avg       0.45      0.41      0.40     11221
weighted avg       0.49      0.49      0.46     11221
```

Fig. 2. Classification report for Decision tree classifier

*3) Random forest classifier*

The results that are highlighted in the figure 3 classification report shows that an accuracy of 55 percent for the achieved by random forest classified during the time of training and accuracy of 50 percent was achieved while testing the data was percent.

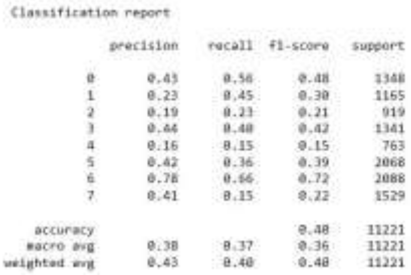


```
              precision    recall  f1-score   support

           0       0.52      0.59      0.55      1348
           1       0.32      0.08      0.13      1165
           2       0.37      0.14      0.21       919
           3       0.58      0.53      0.55      1341
           4       0.24      0.03      0.05       763
           5       0.39      0.74      0.51      2068
           6       0.85      0.78      0.81      2088
           7       0.36      0.44      0.40      1529

    accuracy                           0.50     11221
   macro avg       0.45      0.42      0.40     11221
weighted avg       0.49      0.50      0.47     11221
```

Fig. 3. Classification report for Random Forest classifier

*4) K-neighbours classifier*

It is analysed that accuracy that has been achieved by K-nearest neighbour classifier is 43 percent while training the model and during the process of testing and accuracy of 40 percent was achieved by the model.



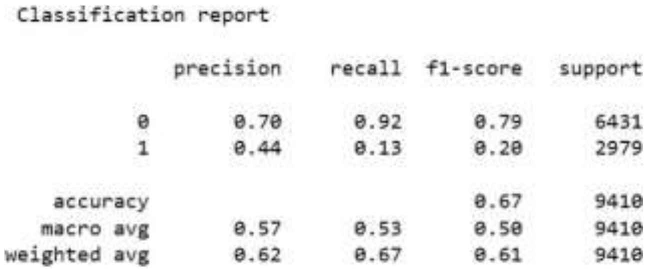
```
Classification report

              precision    recall  f1-score   support

           0       0.43      0.56      0.48      1348
           1       0.23      0.45      0.30      1165
           2       0.19      0.23      0.21       919
           3       0.44      0.40      0.42      1341
           4       0.16      0.15      0.15       763
           5       0.42      0.36      0.39      2068
           6       0.78      0.66      0.72      2088
           7       0.41      0.15      0.22      1529

    accuracy                           0.40     11221
   macro avg       0.38      0.37      0.36     11221
weighted avg       0.43      0.40      0.40     11221
```

Fig. 4. Classification report for K-neighbours classifier

### B. *Analysis from 'United States COVID-19 Cases and Deathsby State over Time' dataset*

*1) Logistic regression*

From the classification report of logistic regression represented in figure 5 it is identify that the model has achieved an accuracy of 67 percent in both cases that is during training and testing the model.

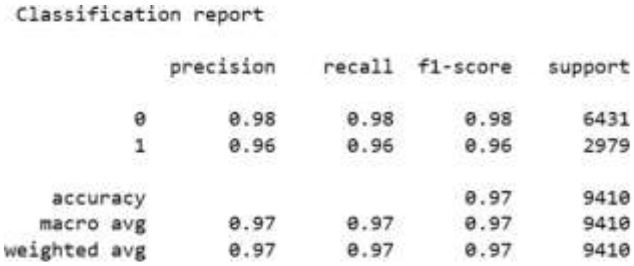


```
Classification report

              precision    recall  f1-score   support

           0       0.70      0.92      0.79      6431
           1       0.44      0.13      0.20      2979

    accuracy                           0.67      9410
   macro avg       0.57      0.53      0.50      9410
weighted avg       0.62      0.67      0.61      9410
```

Fig. 5. Classification report for Logistic regression

*2) Decision tree classifier*

The classification report that is presented in figure 6 highlights that an accuracy of 99 percent was achieved by decision tree classifier while training the model and an accuracy of 97 percent has been achieved by the model in the testing process.

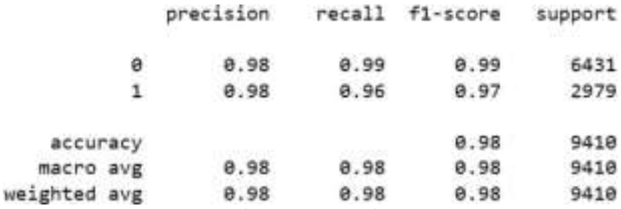


```
Classification report

              precision    recall  f1-score   support

           0       0.98      0.98      0.98      6431
           1       0.96      0.96      0.96      2979

    accuracy                           0.97      9410
   macro avg       0.97      0.97      0.97      9410
weighted avg       0.97      0.97      0.97      9410
```

Fig. 6. Classification report for Decision tree classifier

*3) Random forest classifier*

The classification report that is highlighted in figure 7 represents that an accuracy of 99 percent has been achievedby random forest classifier during the time of training and an accuracy of 98 percent is achieved by the model during testingprocess.

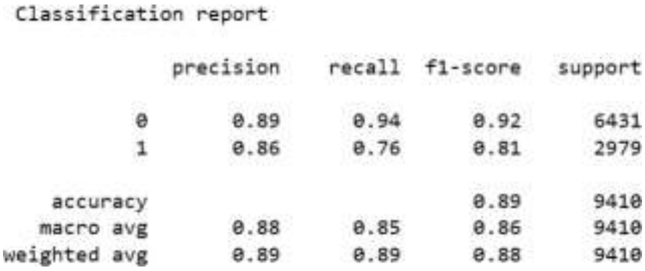


```
              precision    recall  f1-score   support

           0       0.98      0.99      0.99      6431
           1       0.98      0.96      0.97      2979

    accuracy                           0.98      9410
   macro avg       0.98      0.98      0.98      9410
weighted avg       0.98      0.98      0.98      9410
```

Fig. 7. Classification report for Random Forest classifier

### 4) *K-neighbours classifier*

It is interpreted from figure 8 that an accuracy of 92 percenthas been achieved by K-Neighbours classifier during the time of training and all testing the model in accuracy of 88 percent was achieved by the algorithm.



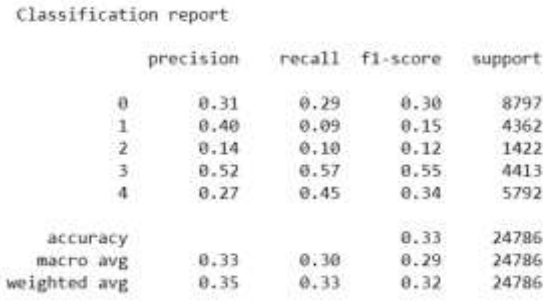
```
Classification report

              precision    recall  f1-score   support

           0       0.89      0.94      0.92      6431
           1       0.86      0.76      0.81      2979

    accuracy                           0.89      9410
   macro avg       0.88      0.85      0.86      9410
weighted avg       0.89      0.89      0.88      9410
```

Fig. 8. Classification report for K-neighbours classifier

## C. *Analysis from ‘Provisional_COVID19_Deaths_by_Sex_and_Age’ dataset*

### 1) *Logistic regression*

It is analysed from the classification report presented in figure 9 that an accuracy of 32 percent has been achieved by the logistic regression during the process of training and an accuracy of 33 percent during the testing process for making predictions.

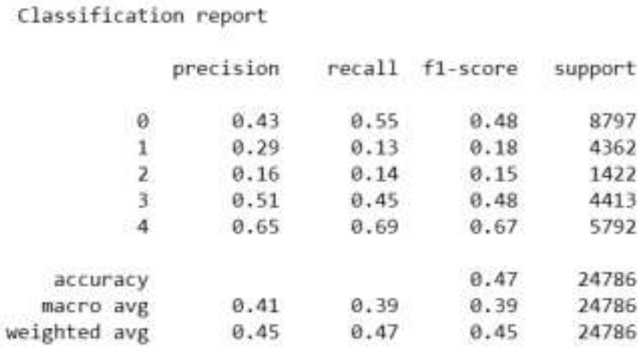


```
Classification report

              precision    recall  f1-score   support

           0       0.31      0.29      0.30      8797
           1       0.40      0.09      0.15      4362
           2       0.14      0.10      0.12      1422
           3       0.52      0.57      0.55      4413
           4       0.27      0.45      0.34      5792

    accuracy                           0.33     24786
   macro avg       0.33      0.30      0.29     24786
weighted avg       0.35      0.33      0.32     24786
```

Fig. 9. Classification report for Logistic regression

### 2) *Decision tree classifier*

The classification report of decision tree classifier that is included in figure 10 represents that an accuracy of 73 percentwas achieved by the model during the process of training and an accuracy of 47 percent has been achieved by the model during the process of testing for the predictions.

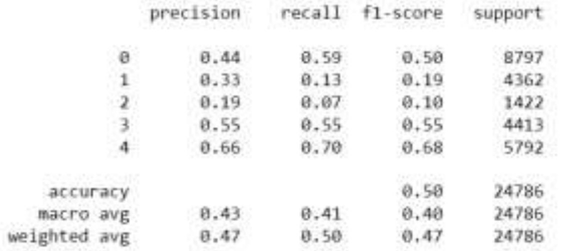


```
Classification report

              precision    recall  f1-score   support

           0       0.43      0.55      0.48      8797
           1       0.29      0.13      0.18      4362
           2       0.16      0.14      0.15      1422
           3       0.51      0.45      0.48      4413
           4       0.65      0.69      0.67      5792

    accuracy                           0.47     24786
   macro avg       0.41      0.39      0.39     24786
weighted avg       0.45      0.47      0.45     24786
```

Fig. 10. Classification report for Decision tree classifier

### 3) *Random forest classifier*

Based on the facts highlight it in figure 11 it is identified that random forest classified achieved an accuracy of 73 percent during training of the model and an accuracy of 49 percentwas achieved by the modern during testing of the modelsprediction.

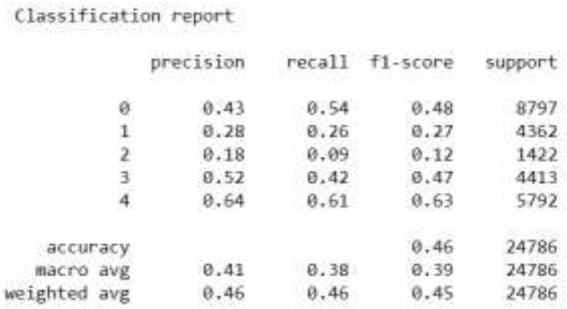


```
              precision    recall  f1-score   support

           0       0.44      0.59      0.50      8797
           1       0.33      0.13      0.19      4362
           2       0.19      0.07      0.10      1422
           3       0.55      0.55      0.55      4413
           4       0.66      0.70      0.68      5792

    accuracy                           0.50     24786
   macro avg       0.43      0.41      0.40     24786
weighted avg       0.47      0.50      0.47     24786
```

Fig. 11. Classification report for Random Forest classifier

### 4) *K-neighbours classifier*

Based on the information presented in figure 12 it is identified that K-nearest neighbour classifier has achieved an accuracy of 55 percent during the process of training and an accuracy of 46 percent was achieved by the model during the process of testing.



```
Classification report

              precision    recall  f1-score   support

           0       0.43      0.54      0.48      8797
           1       0.28      0.26      0.27      4362
           2       0.18      0.09      0.12      1422
           3       0.52      0.42      0.47      4413
           4       0.64      0.61      0.63      5792

    accuracy                           0.46     24786
   macro avg       0.41      0.38      0.39     24786
weighted avg       0.46      0.46      0.45     24786
```

Fig. 12. Classification report for K-neighbours classifier

## V. CONCLUSIONS AND FUTURE WORK

The analysis of huge datasets using the Python programming language and Jupyter Notebook environment has been covered in this article. We have emphasised the main features of programming languages, such as interactive use, data structures, functions, libraries, speed, and open-source nature, that make them suited for data analysis. The researcher has also looked at a variety of data management methods and technologies, such as data cleansing, wrangling, and validation. However, there are a number of difficulties with processing big data analytics, including volume, velocity, security, and expertise needs. Organisations can use specialist infrastructure, tools, and expertise to overcome these obstacles, including distributed storage systems, real-time data processing, data security standards, and the hiring of qualified data scientists and analysts. Organisations may successfully evaluate enormous datasets and gain insightful data by doing this. Provisional COVID-19 Deaths by HHS Region, Race, and Age, United States COVID-19 Cases and Deaths by State over Time, and Provisional_COVID-19_Deaths_by_Sex_and_Age are the datasets utilised in this study. The classification reports for each model are supplied for each dataset, demonstrating the level of accuracy attained during the training and testing phases. Four machine learning models—logistic regression, decision tree classifier, random forest classifier, and k-neighbour classifier—are examined.The results show that the models, with minor differences basedon the model and dataset employed, obtained good levels of accuracy for the majority of datasets.

## VI. REFERENCES

[1] Hao, Jiangang, and Tin Kam Ho. "Machine learning made easy: a reviewof scikit-learn package in python programming language." Journal of Educational and Behavioral Statistics 44.3 (2019): 348-361.

[2] Chen, Rung-Ching, et al. "Selecting critical features for data classifi- cation based on machine learning methods." Journal of Big Data 7.1 (2020): 52.

[3] Ilyas, Ihab F., and Xu Chu. Data cleaning. Morgan Claypool, 2019.

[4] McGregor, Susan E. Practical Python Data Wrangling and Data Quality." O'Reilly Media, Inc.", 2021

[5] Tawfik, Gehad Mohamed, et al. "A step by step guide for conducting a systematic review and meta-analysis with simulation data." Tropical medicine and health 47.1 (2019): 1-9.

[6] Hariri, Reihaneh H., Erik M. Fredericks, and Kate M. Bowers. "Uncertainty in big data analytics: survey, opportunities, and challenges." Journal of Big Data 6.1 (2019): 1-16.

[7] Bag, Surajit, et al. "Big data analytics as an operational excellence approach to enhance sustainable supply chain performance." Resources, Conservation and Recycling 153 (2020): 104559.

[8] Chehri, Abdellah, Issouf Fofana, and Xiaomin Yang. "Security risk modeling in smart grid critical infrastructures in the era of big data and artificial intelligence. "Sustainability 13.6 (2021): 3196.

[9] Jo, Jeong Hoon, et al. "Emerging technologies for sustainable smart city network security: Issues, challenges, and countermeasures." Journal of Information Processing Systems 15.4 (2019): 765-784.

[10] Data.gov. "Provisional COVID-19 Deaths by Sex and Age" 2023.

[11] Data.gov. "Air Traffic Passenger Statistics" 2023.

[12] Data.gov. "Provisional COVID-19 Deaths by HHS Region, Race, and Age" 2021.

[13] Tangirala, Suryakanthi. "Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm." International Journal of Advanced Computer Science and Applications 11.2 (2020): 612-619.

[14] Parmar, Aakash, Rakesh Katariya, and Vatsal Patel. "A review on random forest: An ensemble classifier. "International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018. Springer International Publishing, 2019.

[15] Connelly, Lynne. "Logistic regression." Medsurg Nursing 29.5 (2020): 353-354.

[16] Kumbure, Mahinda Mailagaha, Pasi Luukka, and Mikael Collan. "A new fuzzy k-nearest neighbor classifier based on the Bonferroni mean." Pattern Recognition Letters 140 (2020): 172-178.

[17] Hao, Jiangang, and Tin Kam Ho. "Machine learning made easy: a review of scikit-learn package in python programming language. "Journal of Educational and Behavioral Statistics 44.3 (2019): 348-361.

[18] Wang, Jiawei, Li Li, and Andreas Zeller. "Better code, better sharing: on the need of analyzing jupyter notebooks." Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering: New Ideas and Emerging Results. 2020.

[19] Chauhan, Anjali. "A review on various aspects of MongoDB databases." International Journal of Engineering Research Technology (IJERT) 8.05 (2019): 90-92.