PROJECT REPORT

on

"Comparison of Big Data Tools"

(CSE CC V Semester Mini Project)

December 2021

Submitted to:

Mrs. Garima Sharma

Assistant Professor

Submitted by:

Girija Rathore

Roll. No.: 2015016

CSE-CC-CCIS-V-Sem

Guided by:

Mrs. Garima Sharma

(Resource Person)

Graphic Era

Deemed to be University

1. ABOUT PROJECT

1.1. Introduction

1.1.1. What Big Data Really Is?

Big Data is a term that is frequently encountered in business, the workplace, and everywhere in between. It's been used, overused, and misused so many times that it's hard to know what it really means.

Data is any raw letter or symbol that a computer may store, transmit as signals, or record on media. Raw data, on the other hand, has no value unless it is processed.

Big Data is a vast collection of data that is growing at an exponential rate. It is a data source that is so large and complex that traditional data management solutions are incapable of storing or analyzing it efficiently. Social media, online literature, music, movies, and a rise in the number of sensors have all contributed to the incredible growth in the amount of data that is now available for analysis.

1.1.2. Categories & Characteristics: Big Data Essentials

It's crucial to remember that Big Data isn't just about the amount of data we generate; it's about all the different forms of data as well. Big Data is classified into three broad categories: Structured, Unstructured and Semi-Structured. Each category has different sources and different modes of representation.

(i) <u>Structured Data:</u> It refers to data that has previously been stored in databases in an orderly fashion. It accounts for around 20% of all current data and is commonly employed in programming and computer-related activities.

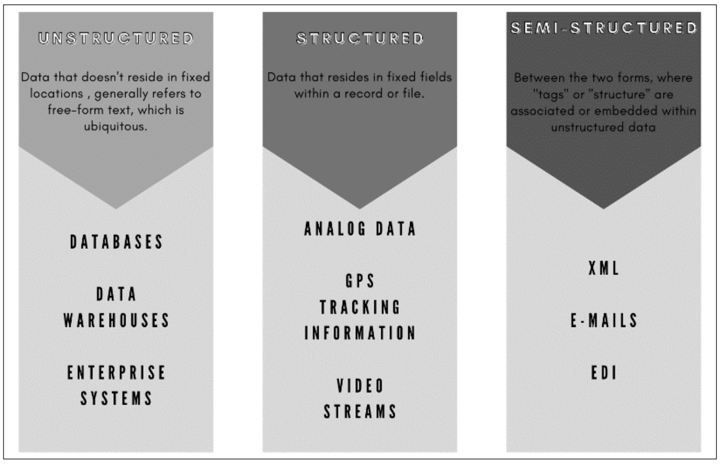

UNSTRUCTURED
Data that doesn't reside in fixed locations , generally refers to free-form text, which is ubiquitous.
DATABASES
DATA WAREHOUSES
ENTERPRISE SYSTEMS
STRUCTURED
Data that resides in fixed fields within a record or file.
ANALOG DATA
GPS TRACKING INFORMATION
VIDEO STREAMS
SEMI-STRUCTURED
Between the two forms, where "tags" or "structure" are associated or embedded within unstructured data
XML
E-MAILS
EDI

Fig. 1 Categories of Data

(ii) <u>Unstructured Data:</u> This type of data, lacks a standardized format and accounts for around 80% of the total data. Until recently, there wasn't much that could be done with it, other than manually storing or analyzing it. The Unstructured Data is further divided into Captured & User-Generated Data.

(iii) <u>Semi-Structured Data:</u> The distinction between Unstructured and Semi-Structured Data has always been hazy, because most semi-structured data seem to be unstructured at first look. Semi-structured data includes information that has certain organizational qualities that make it easier to process.

Big data has at least one, but generally all, of the following characteristics: large volume, rapid velocity (pace of change), a broad variety of types, unpredictable veracity, and potential value.

(i) Volume: Big data, as the term indicates, is huge in size—terabytes, petabytes, or even zettabytes—and expanding at such a rapid rate that measuring its exact size is unfeasible. Volume determines whether data is big or not.

(ii) Velocity: The great bulk of modern data is continually changing, and the development of streaming data from IoT and other sources only accelerates the rate of change and expansion. It shows how fast data is generated and processed for analysis.

(iii) Variety: Big data encompasses any and all types of data, regardless of how or where it was created. It refers to the heterogeneous nature of the data in question.

(iv) Veracity: Big data technology tackles the requirement to validate the quality and dependability of massive volumes of data coming into systems at fast speeds from a variety of sources and formats. It relates to the reliability and unreliability of data.

(v) Value: Big data has enormous potential value if it is handled and shared effectively so that employees can read, evaluate, and apply the ensuing insights to make accurate, confident choices.

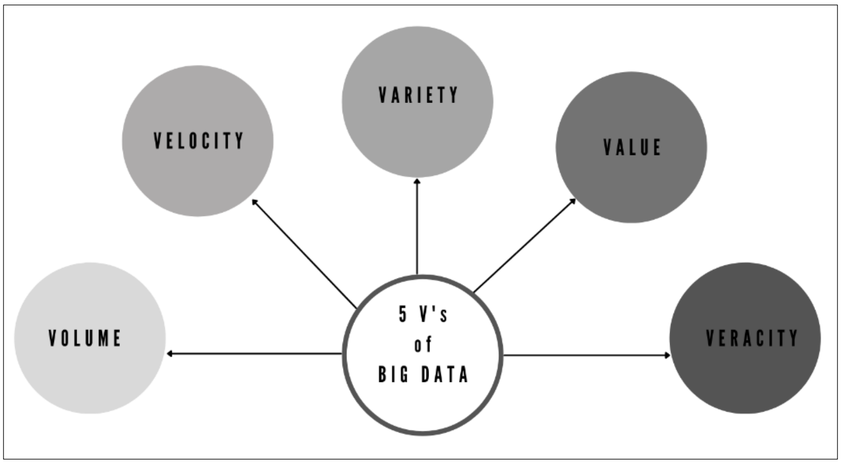


Fig. 2 Characteristics of Big Data

1.1.3. How Does Big Data Work?

The core concept of Big Data is that the more you know about something, the more insights you may get and use to make a choice or discover a solution. Most of the time, this procedure is totally automated - we have such sophisticated technologies that perform millions of simulations to provide us with the best potential result.

The requirement to manage so much data necessitates a very reliable and well-structured infrastructure. It will need to analyze massive amounts of data in a short period of time, which might overburden a single server or cluster. This is why a well-thought-out infrastructure is required to support Big Data.

The three main actions behind Big Data are Integration, Management and Analysis.

(i) Integrate: Big Data is constantly acquired from many sources, and as we speak for massive amounts of information, new tactics and technology to handle it must be created. In certain circumstances, we're talking about petabytes of data pouring into your system, thus integrating such a large volume of data into your system will be difficult. You must receive the data, process it, and format it in the manner required by your business and understandable to your clients.

INTEGRATION
MANAGEMENT
ANALYSIS

Fig. 3 Action Plan for Big Data

(ii) Management: What more could you possibly want for such a massive amount of information? You'll need somewhere to keep it. Your storage solution might be in the cloud, on-premises, or both. You may also choose how your data will be kept so that it is accessible in real time and on demand. This is why an increasing number of individuals are opting for a cloud storage solution since it meets their present computational needs.

(iii) Analysis: Okay, you've received and saved the data, but it has to be analyzed before it can be used. Investigate your data and utilize it to make key decisions, such as determining which features are most frequently requested by your consumers or sharing research. Do anything you want and need with it - put it to use, since you spent a lot of money to set up this infrastructure, therefore you should put it to use.

1.1.4. Getting Started with Big Data Tools

If you're going to work with various types of Big Data, you'll need to think about how you're going to store it. One of the reasons Big Data was labelled as "Big" is that it was too large for existing systems to handle. Gigabytes of data may now be scaled to terabytes and beyond.

Big data analytic tools are software programs that help with data collection and extraction from massive volumes of data. A good data storage provider should provide you with both an infrastructure to run all of your big data tools and a place to store, query, and analyze your data.

Choosing the appropriate tool, the first time will help you save time and avoid difficulties, but you don't have to make this decision on faith. Keep in mind that there is no such thing as the "greatest" big data platform. Each of these applications caters to a distinct set of requirements, so it's critical that you pick the big data solution that best matches your demands.

1.2. Problem Statement

The goal of this study is to compare different big data analytics tools for beginners in this field. For the purposes of this study, several factors were compared, such as data handle size, data types, data loading, predictive capabilities, user interface, add-ons, results display, data review, and so on. These parameters were chosen because all of the aforementioned parameters may be done by the user in each software. Different software may perform better in some areas than others, and some software may have an edge over others in certain circumstances. As a result, this study evaluates the selected software and recommends the best efficient software for big data analysis based on these factors.

1.3. Who will be Benefitted?

Big Data has become a necessary component of any organization in order to improve decision-making and obtain a competitive advantage over competitors. As a result, Big Data tools like Tableau and Orange are in great demand. Companies are searching for experts who know how to use them to get the most out of the data generated within their walls.

These data tools aid in the management of large data sets as well as the identification of patterns and trends within them. Anyone interested in working in the Big Data sector should familiarize oneself with these technologies, their applications, and their comparative advantages.

2. NEED FOR COMPARISON

2.1. Factors of Comparison

While looking out for a pin, in a stack of hay, you might not actually start from one end and go to the other without wasting your efforts and times, that too unnecessarily. While choosing a Big Data Tool, you might want to have a basis of comparison so you get what you want to, without much delay.

Some factors for comparing the existing Big Data Tools that we've considered are: mode of software, technical requirements, data size & type, ease of access from developer's and customer's aspect, prediction capabilities, result presentation, exporting output, pricing, training programs for learners and some other additional remarks.

2.2. Description of Datasets

2.2.1. US Traffic Accidents Dataset

This is a countrywide car accident dataset, which covers 49 states of the USA. The accident data are collected from February 2016 to Dec 2020, using multiple APIs that provide streaming traffic incident (or event) data. Currently, there are about 1.5 million accident records in this dataset.

This dataset can be used for numerous applications such as real-time accident prediction, studying accident hotspot locations, casualty analysis and extracting cause and effect rules to predict accidents, or studying the impact of precipitation or other environmental stimuli on accident occurrence. The dataset covers 49 states of the US. The data, provided in terms of a CSV file contains several attributes, some of which are: ID, Severity, Description, Street, City, Temperature, Visibility, etc.

2.2.2. Titanic Disaster Dataset

The sinking of the Titanic, remains one of the most famous shipwrecks in history. On the morning of April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew. While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.

This classic dataset, perfect for getting started with exploratory analysis and building binary classification models can be used to predict survival preferably using decision trees. Dataset covers passengers only, not crew. Some of the features of the dataset include, survival, class, name, gender, survival status, etc.

2.2.3. Wine Quality

This dataset is related to red variants of the Portuguese "Vinho Verde" wine. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

The datasets can be viewed as classification or regression tasks. The attributes are ordered and not balanced (e.g. there are much more normal wines than excellent or poor ones).

The dataset includes various input variable, such as fixed acidity, citric acid, volatile acidity, pH, sulphates, etc. and the output variable is the quality on an index of 0 to 10.

3. TOOLS AT A GLANCE

3.1. Big Data Tools: Beginner’s Guide

Companies all across the world have come to recognize the importance of their data. Many businesses are launching data science efforts in order to find new methods to leverage value. That's why big data technologies are now a must-have, and data engineering is one of the most in-demand IT skills today.

There are a number of big data technologies available to assist businesses with analytics. Some are all-in-one solutions, while others specialize on a particular area, such as data visualization or data integration. Data storage and administration, data cleansing, data mining, data analysis, data visualization, and data integration are all covered by Big Data Tools, which might overlap with data software.

3.2. Tool 01: Tableau

3.2.1. Retrospective Perspective

Tableau Software is a business intelligence-focused interactive data visualization software firm based in the United States. It began in Mountain View, California, in 2003 and is now based in Seattle, Washington. Salesforce purchased the firm in 2019.

Christian Chabot, Pat Hanrahan, and Chris Stolte, the company's founders, were all researchers at Stanford University's Department of Computer Science. From 1999 to 2002, they focused on visualization tools for examining and analyzing relational databases and data cubes, and the firm began as a commercial outlet for Stanford research.

To build graph-type data visualizations, Tableau products query relational databases, online analytical processing cubes, cloud databases, and spreadsheets. An in-memory data engine may also be used to extract, store, and retrieve data.

3.2.2. Technical & Functional Pre-requisites

Tableau supports various software versions, namely, Tableau Desktop, Tableau Online, Tableau Prep, Tableau Server and Tableau Public. Each of these versions have their very own technical specifications. We'll be dealing with Tableau Public.

Tableau Public works well with both, Windows and MacOS. We can also access the tool via web browsers.

Technical Requirements for Windows:

(i) Microsoft Windows 8/8.1, Windows 10 (x64)
(ii) 2 GB memory
(iii) 1.5 GB minimum free disk space
(iv) CPUs must support SSE4.2 and POPCNT instruction sets

Technical Requirements for Mac:

(i) macOS Mojave 10.14, macOS Catalina 10.15, and Big Sur 11.4+
(ii) Intel processors
(iii) M1 processors under Rosetta 2 emulation mode
(iv) 1.5 GB minimum free disk space
(v) CPUs must support SSE4.2 and POPCNT instruction sets

Supported Formats:

Google Sheets, JSON files, Microsoft Excel 2007 or later, OData, PDF, comma separated value (.csv) files

3.2.3. Working Components

Here's what the startup page of Tableau Public looks like. On the left pane of the interface, we can either import or connect our use case dataset. The next pane from left contains three subdivisions, Pages, Filters and Marks which control the visualization or the Viz. The central sheet pane is our worksheet space where the visuals are created. Once done, the user can Publish their workspace publicly with the option mentioned on Top-right corner.

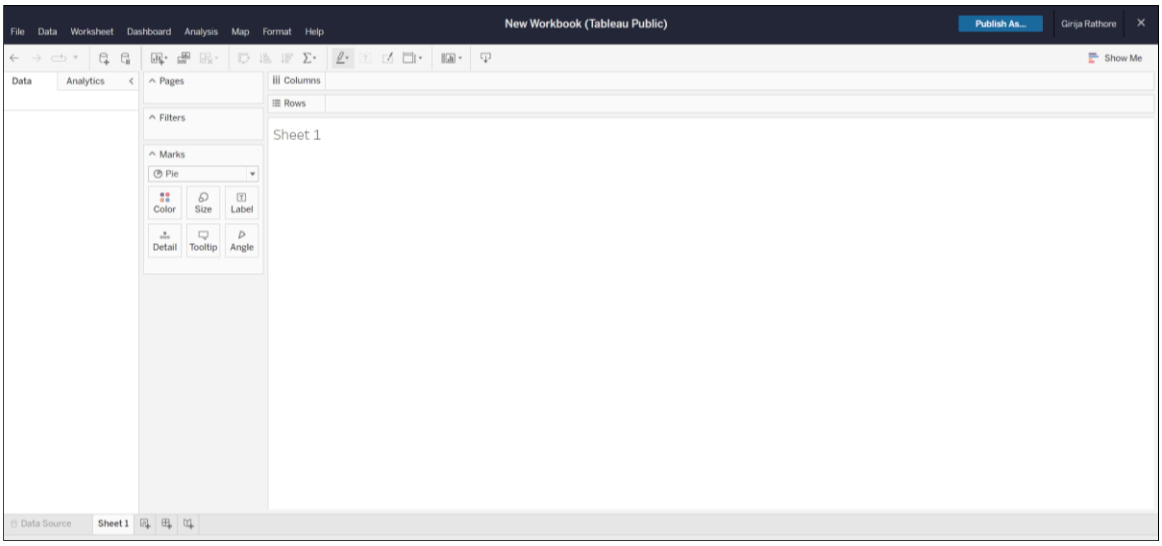
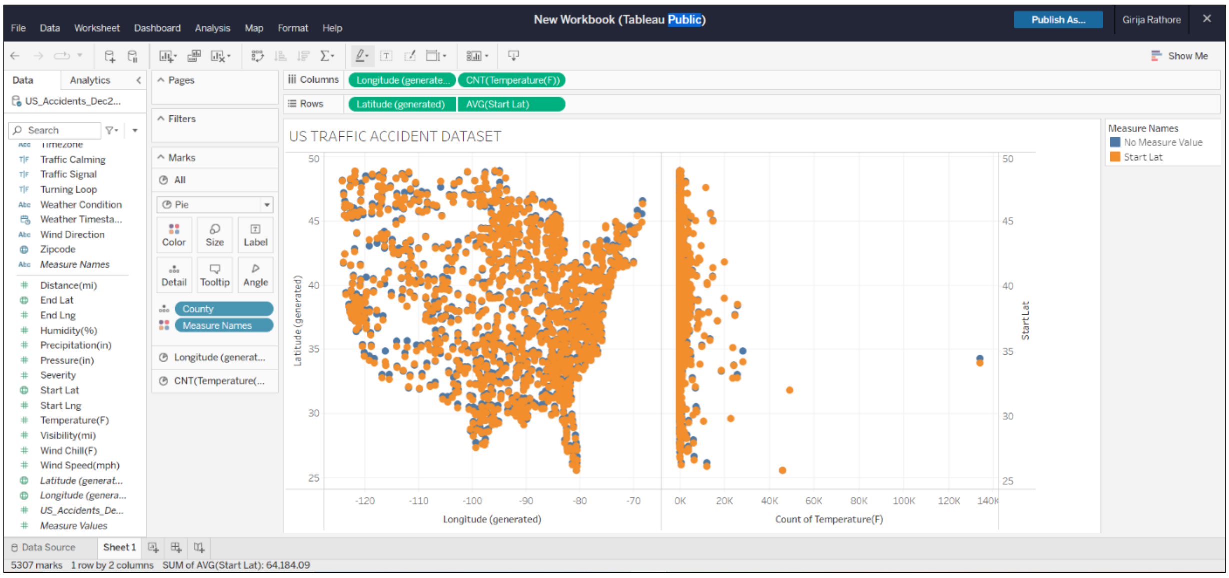


Fig. 4 Start Up User Interface Page for Tableau Public

3.2.4. Implementation & Graphical Visualization

3.2.4.1. Implementation of US Traffic Accident Dataset



5307 marks 1 row by 2 columns SUM of AVG(Start Lat): 64,184.09

Fig. 5 Implementation of US Traffic Accident Dataset in Tableau Public

3.2.4.2. Implementation of Titanic Disaster Dataset

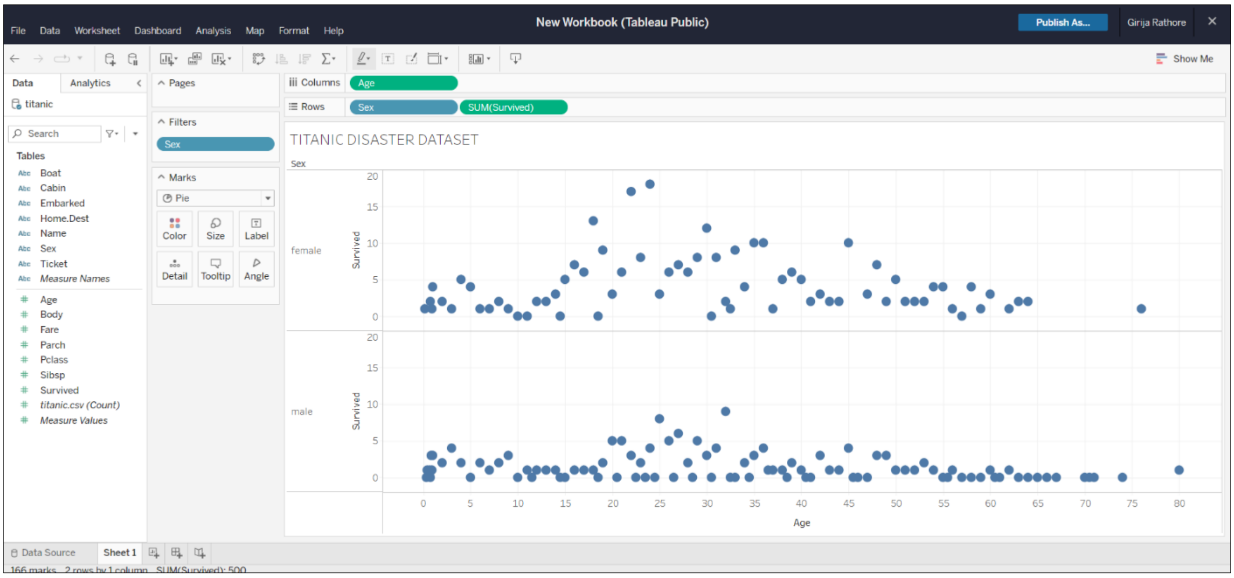


Fig. 6 Implementation of Titanic Disaster Dataset on Tableau Public

3.2.4.3. Implementation of Wine Quality Dataset

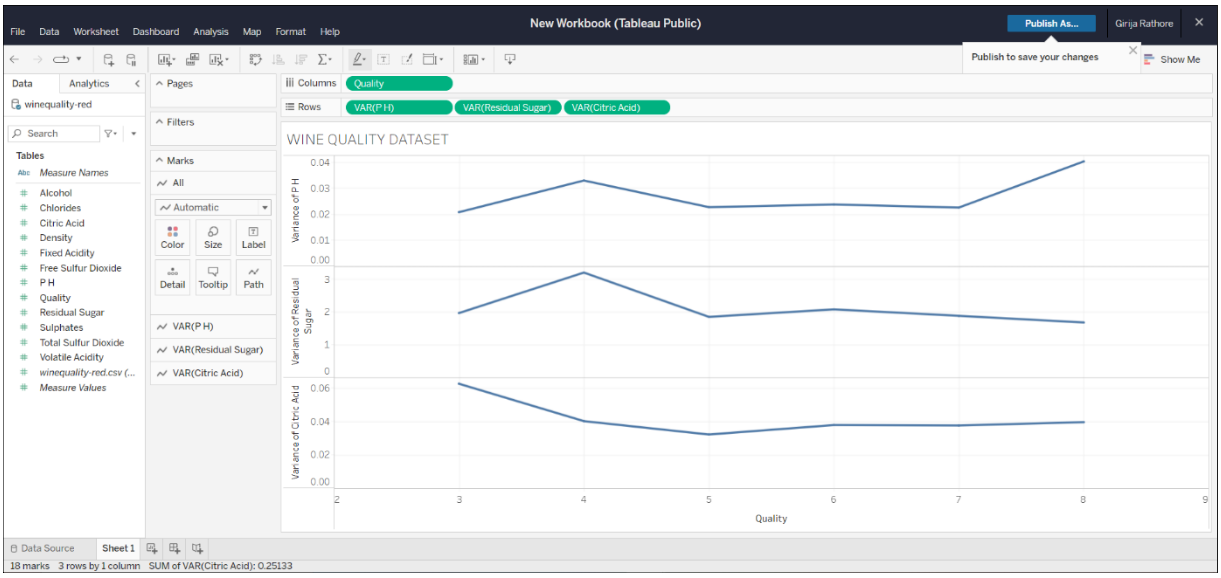


Fig. 7 Implementation of Wine Quality Dataset on Tableau Public

3.3. Tool 02: RapidMiner

3.3.1. Retrospective Perspective

RapidMiner, originally known as YALE (Yet Another Learning Environment), was created by Ralf Klinkenberg, Ingo Mierswa, and Simon Fischer at the Technical University of Dortmund's Artificial Intelligence Unit in 2001. Rapid-I, a firm created by Ingo Mierswa and Ralf Klinkenberg in the same year, has been driving its growth since 2006. The software's name was changed from YALE to RapidMiner in 2007. The startup changed its name from Rapid-I to RapidMiner in 2013.

It covers all elements of the machine learning process, including data preparation, results visualization, model validation, and optimization, and is used for corporate and commercial applications, as well as research, education, training, rapid prototyping, and application development. RapidMiner is based on an open core architecture.

3.3.2. Technical & Functional Pre-requisites

Java-based RapidMiner Studio is platform-independent and runs on every platform for which an appropriate Java Runtime Environment (JRE) is available. For reading and writing data, RapidMiner can connect to all relational database systems offering a fully compliant JDBC driver. Through its supported operators, RapidMiner Studio can connect to a variety of NoSQL databases, cloud connectors, and file types.

<u>Recommended System Specifications:</u>

(i) Quad core
(ii) 3GHz or faster processor
(iii) 16GB RAM
(iv) 100GB free disk space
(v) Operating System

<u>Recommended Operating Systems:</u>

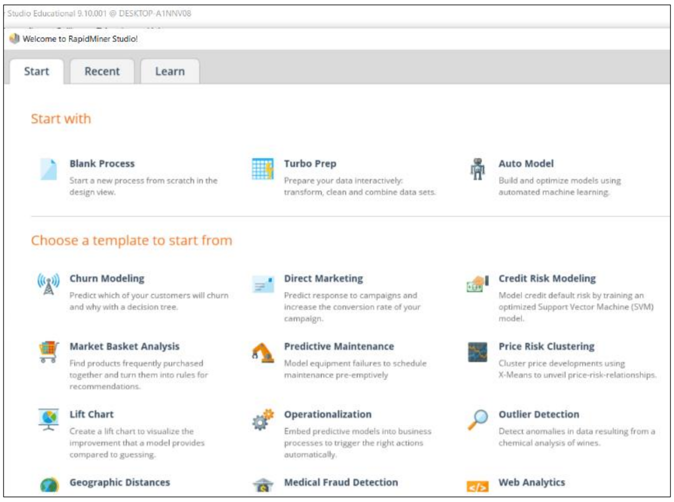
(i) Windows 7, Windows 8, Windows 8.1, Windows 10 (64-bit highly recommended)
(ii) Linux (64-bit only)
(iii) MacOS X 10.10 - 10.15
(iv) Java platform

<u>Compatible Databases</u>

(i) Oracle
(ii) Microsoft SQL Server
(iii) MySQL
(iv) Teradata
(v) MongoDB
(vi) Cassandra

3.3.3. Working Components

Just after the installation of the software, you'll be prompted with this dialog box. You can either start a Blank Process, or choose the other two options. Turbo Prep lets you prepare your data before applying any algorithmic measure to it, that too interactively. The Auto Model option lets you build & deploy automated ML algorithms on datasets.


Studio Educational 9.10.001 @ DESKTOP-A1NNV08
Welcome to RapidMiner Studio!
Start
Recent
Learn

Start with

Blank Process Start a new process from scratch in the design view.

Turbo Prep Prepare your data interactively: transform, clean and combine data sets.

Auto Model Build and optimize models using automated machine learning.

Choose a template to start from

Churn Modeling Predict which of your customers will churn and why with a decision tree.

Direct Marketing Predict response to campaigns and increase the conversion rate of your campaign.

Credit Risk Modeling Model credit default risk by training an optimized Support Vector Machine (SVM) model.

Market Basket Analysis Find products frequently purchased together and turn them into rules for recommendations.

Predictive Maintenance Model equipment failures to schedule maintenance pre-emptively

Price Risk Clustering Cluster price developments using X-Means to unveil price-risk-relationships.

Lift Chart Create a lift chart to visualize the improvement that a model provides compared to guessing.

Operationalization Embed predictive models into business processes to trigger the right actions automatically.

Outlier Detection Detect anomalies in data resulting from a chemical analysis of wines.

Geographic Distances

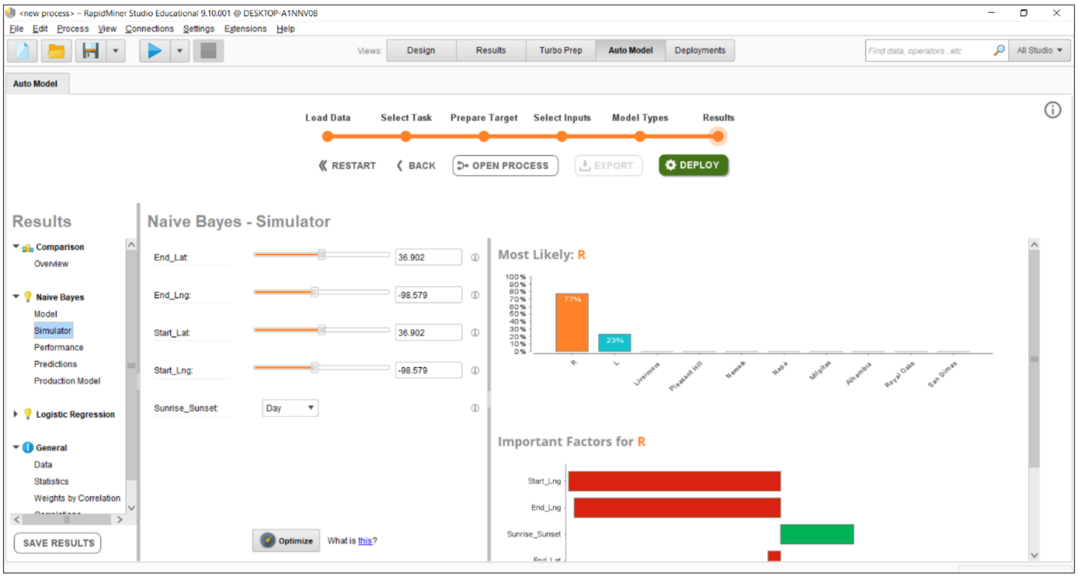
Medical Fraud Detection

Web Analytics

Fig. 8 Startup User Interface of RapidMiner

3.3.4. Implementation & Graphical Visualization

3.3.4.1. Implementation of US Traffic Accident Dataset



End_Lat: 36.902

End_Lng: -98.579

Start_Lat: 36.902

Start_Lng: -98.579

Sunrise_Sunset: Day

Fig. 9 Implementation of US Traffic Dataset on RapidMiner

3.3.4.2. Implementation of Titanic Disaster Dataset

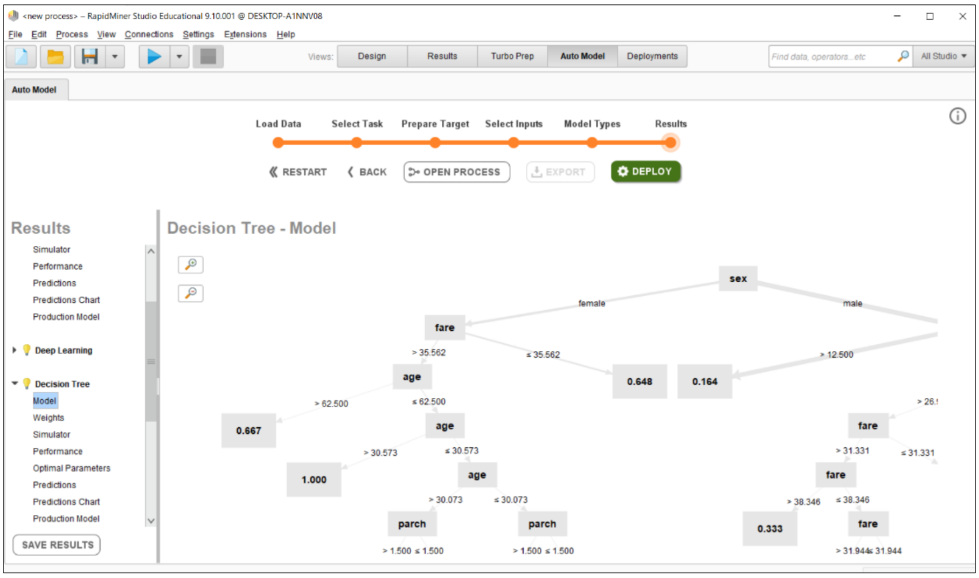
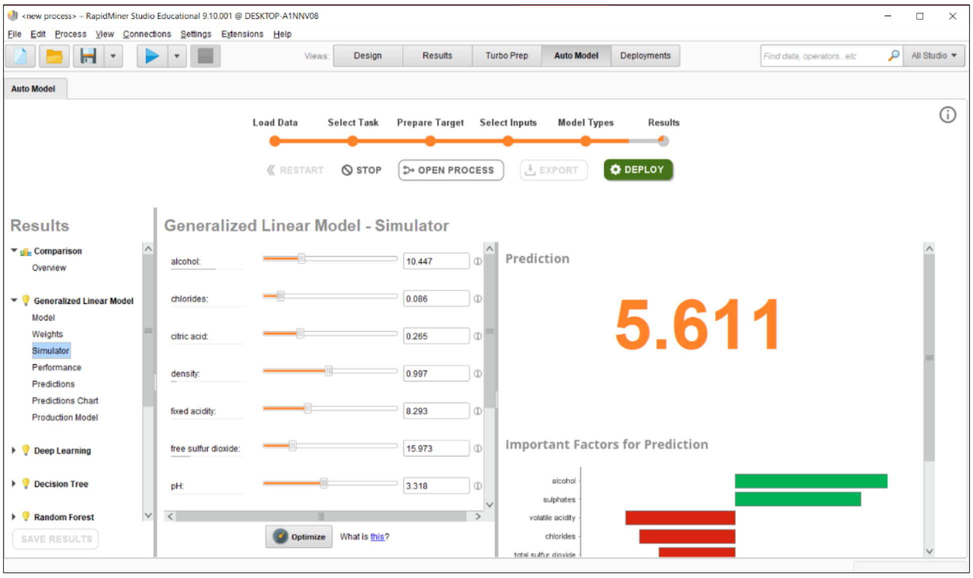


Fig. 10 Implementation of Titanic Disaster Dataset on RapidMiner

3.3.4.3. Implementation of Wine Quality Dataset



alcohol: 10.447

chlorides: 0.086

citric acid: 0.265

density: 0.997

fixed acidity: 8.293

free sulfur dioxide: 15.973

pH: 3.318

Prediction: 5.611

Fig. 11 Implementation of Wine Quality Dataset on RapidMiner

3.4. Tool 03: Orange

3.4.1. Retrospective Perspective

Orange is a C++ core object and routines library that supports a wide range of machine learning and data mining methods, both standard and non-standard. It's a free and open-source application for data visualization, data mining, and machine learning. Orange is a scriptable environment that allows you to quickly prototype new algorithms and test patterns. It's a collection of python-based modules found in the core library.

The origins of Orange may be traced back to 1997, when late Donald Michie developed the concept that machine learning required an open toolbox. We co-organized WebLab97 in lovely Bled, Slovenia, to kickstart the development. The term Workshop came from Michie's concept that the tool should be an online application where individuals could submit data mining code, methods, testing scripts, and data and share them in a collaborative web workspace.

3.4.2. Technical & Functional Pre-requisites

Orange is supported for all three possible Operating Systems, Windows, Mac and Linux. There are no hard & fast system requirements, except for the existence of Python directories and libraries for smooth execution of the tool.

3.4.3. Working Components

After installation of the software, the user is prompted with a dialog box, to either start with a blank new process, open an existing process or view a recent project. The user can also access video tutorials or choose from an existing template, and just change the data source.

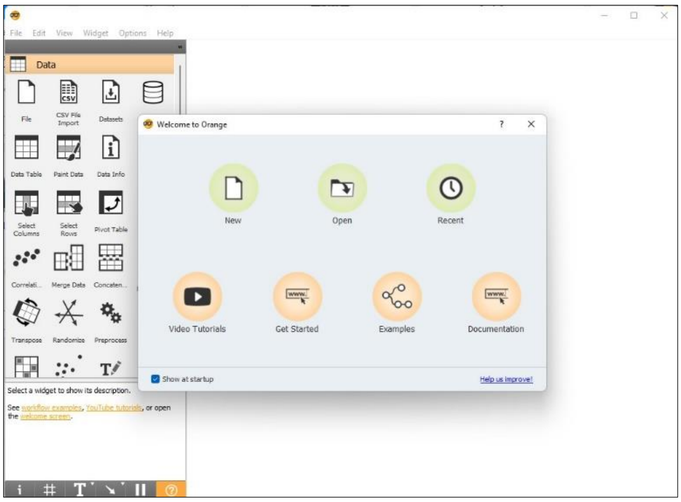
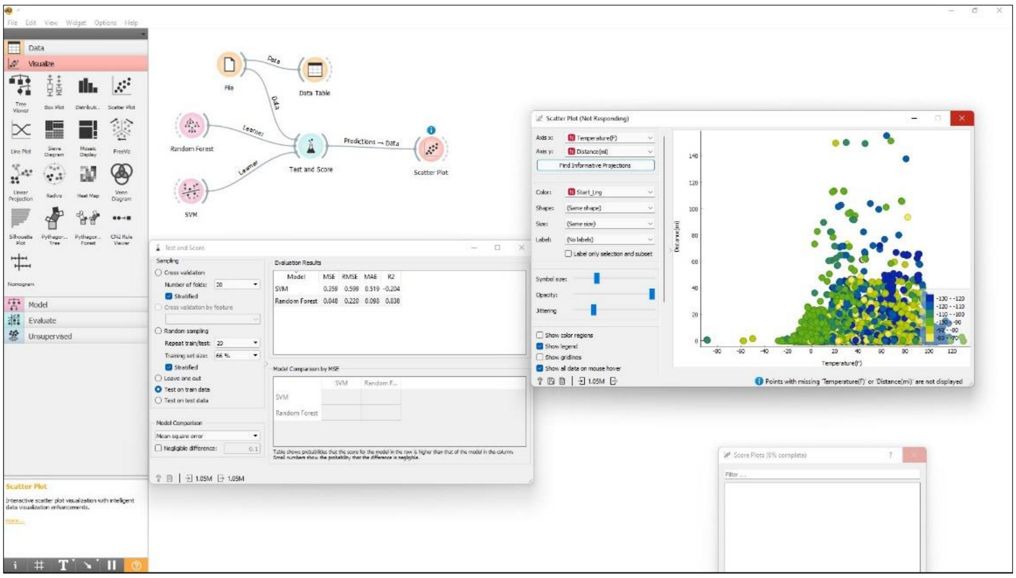


Fig. 12 Startup User Interface of Orange

3.4.4. Implementation & Graphical Visualization

3.4.4.1. Implementation of US Traffic Accident Dataset


Data
Visualize
Model
Evaluate
Unsupervised
File
Data Table
Random Forest
SVM
Test and Score
Scatter Plot
Predictions → Data
Learner
Scatter Plot (Not Responding)
Temperature(F)
Distance(mi)
Find Informative Projections
Start_Lng
Show color regions
Show legend
Show gridlines
Show all data on mouse hover
Points with missing 'Temperature(F)' or 'Distance(mi)' are not displayed
Score Plots (0% complete)
Scatter Plot
Interactive scatter plot visualization with intelligent data visualization enhancements.

Test and Score

Sampling
- ○ Cross validation
  - Number of folds: 20
  - ☑ Stratified
- ○ Cross validation by feature
- ○ Random sampling
  - Repeat train/test: 20
  - Training set size: 66 %
  - ☑ Stratified
- ○ Leave one out
- ● Test on train data
- ○ Test on test data

Model Comparison: Mean square error
□ Negligible difference: 0.1

Evaluation Results

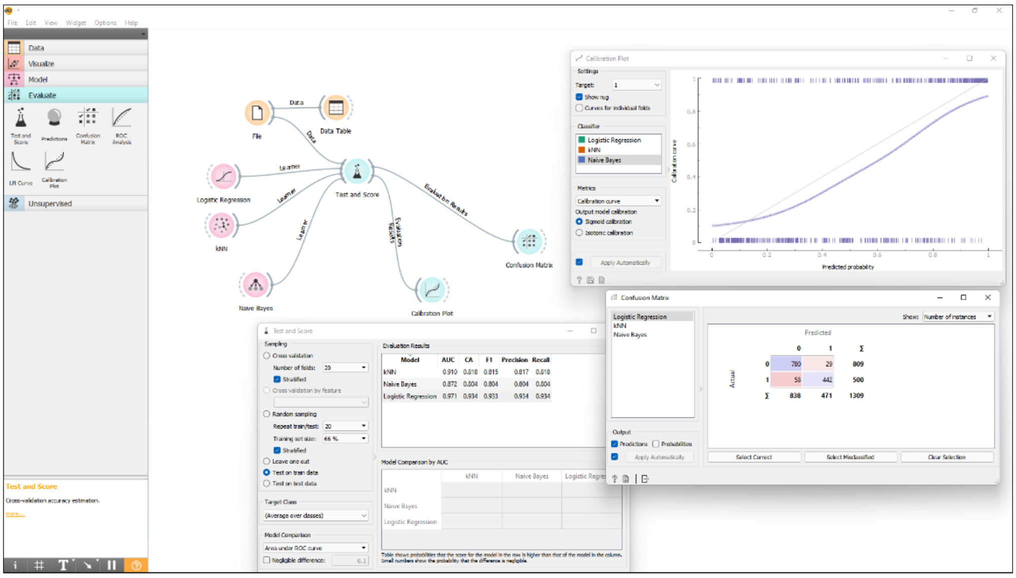
Model	MSE	RMSE	MAE	R2
SVM	0.359	0.599	0.519	-0.204
Random Forest	0.048	0.220	0.098	0.838

Model Comparison by MSE

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

Fig. 13 Implementation of US Traffic Accident on Orange

3.4.4.2. Implementation of Titanic Disaster Dataset



Test and Score

Sampling
- ○ Cross validation
  - Number of folds: 20
  - ☑ Stratified
- ○ Cross validation by feature
- ○ Random sampling
  - Repeat train/test: 20
  - Training set size: 66 %
  - ☑ Stratified
- ○ Leave one out
- ● Test on train data
- ○ Test on test data

Target Class: (Average over classes)

Model Comparison: Area under ROC curve
□ Negligible difference: 0.1

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
kNN	0.910	0.818	0.815	0.817	0.818
Naive Bayes	0.872	0.804	0.804	0.804	0.804
Logistic Regression	0.971	0.934	0.933	0.934	0.934

Model Comparison by AUC

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

Confusion Matrix (Logistic Regression; Show: Number of instances)

Actual \ Predicted	0	1	Σ
0	780	29	809
1	58	442	500
Σ	838	471	1309

Output: ☑ Predictions □ Probabilities

Fig. 14 Implementation of Titanic Dataset on Orange

3.4.4.3. Implementation of Wine Quality Dataset

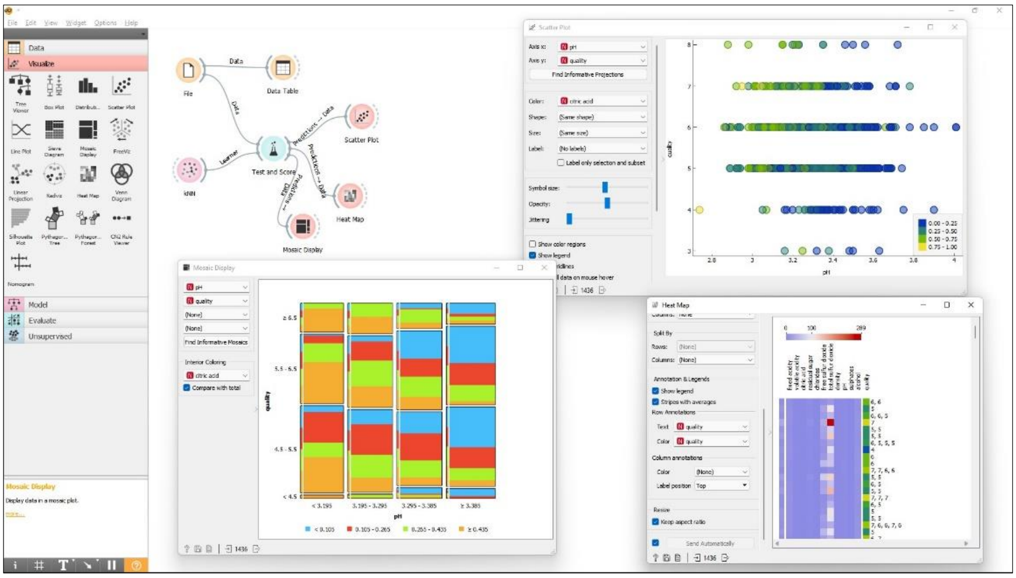


Fig. 13 Implementation of Wine Quality Dataset on Orange

3.5. Tool 04: Python

3.5.1. Retrospective Perspective

Python is a popular high-level programming language in the software industry. To expound on the point, computer language differs significantly from human language.

The advantage of these languages is that programmers can code, interpret, and evaluate to some extent in intelligible syntaxes.

Python is widely used in scientific and academic areas because of its ease of use and straightforward syntax, which makes it easier for those without a technical background to learn. It is also more suited to quick prototyping.

Deep learning frameworks accessible with Python APIs, in addition to scientific packages, have made Python incredibly productive and adaptable, according to engineers from academia and business. Python frameworks for deep learning have gone through a lot of changes, and they're becoming better all the time. In terms of domain fields, ML scientists favor Python.

3.5.2. Working Components

3.5.2.1. Libraries Used

3.5.2.1.1. Numpy

NumPy is a Python package for array processing. It offers high-performance multidimensional array objects as well as array-related tools. NumPy is a useful container for multi-dimensional data in general.

The homogeneous multidimensional array is NumPy's core object. It's a table containing the same datatype elements or numbers, indexed by a tuple of positive integers. NumPy is used to process arrays with the same datatype of values. NumPy makes math operations on arrays and vectorization easier. This considerably improves performance and, as a result, reduces execution time.

3.5.2.1.2. Pandas

Pandas is a Python library that provides high-performance, easy-to-use data structures and data analysis tools for labelled data. Python Data Analysis Library is referred to as Pandas.

Pandas is an excellent tool for wrangling and munging data. It's made to make data processing, reading, aggregation, and visualization as simple as possible. Pandas takes data from a CSV or TSV file or a SQL database and turns it into a data frame, a Python object with rows and columns. In statistical software, such as Excel or SPSS, the data frame is quite similar to a table.

3.5.2.1.3. MatPlotLib

This is without a doubt my favorite and most important Python library. The data shown with Matplotlib can be used to generate stories. Matplotlib is a SciPy Stack library that plots 2D graphs.
Matplotlib is a Python charting toolkit that offers an object-oriented API for integrating charts into programs. It bears a striking resemblance to MATLAB and is written in the Python programming language.

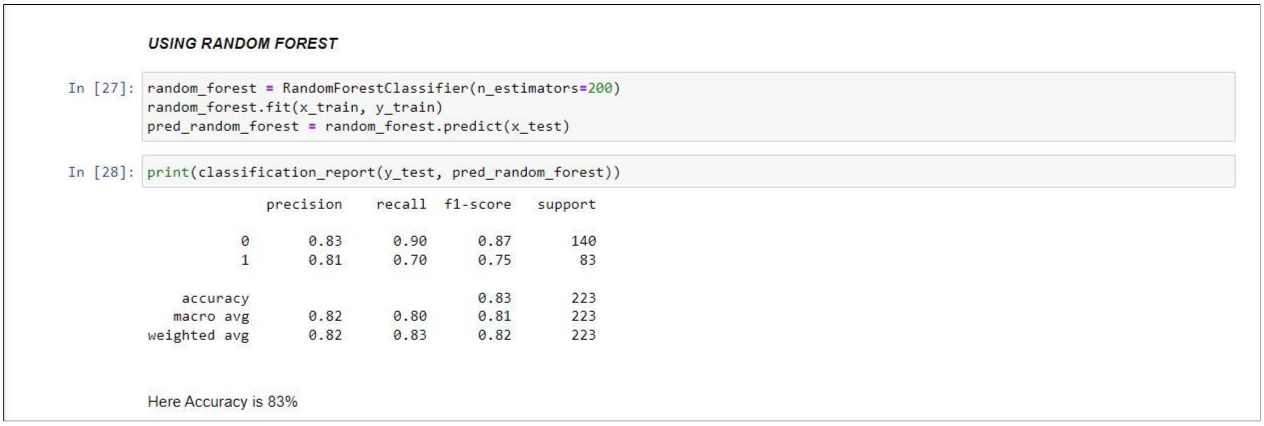
3.5.2.1.4. SKLearn

Scikit Learn is a sophisticated machine learning toolkit for Python that was first introduced to the public as a Google Summer of Code project. SVMs, random forests, k-means clustering, spectral clustering, mean shift, cross-validation, and other machine learning algorithms are included. Scikit Learn also supports NumPy, SciPy, and related scientific processes, as Scikit Learn is part of the SciPy Stack.

Scikit-learn provides a standard Python interface for a variety of supervised and unsupervised learning techniques. Scikit learn is your go-to for supervised learning models like Naive Bayes and categorising unlabeled data like KMeans.

3.5.3. Implementation & Graphical Visualization

3.5.3.1. Implementation of Titanic Disaster Dataset

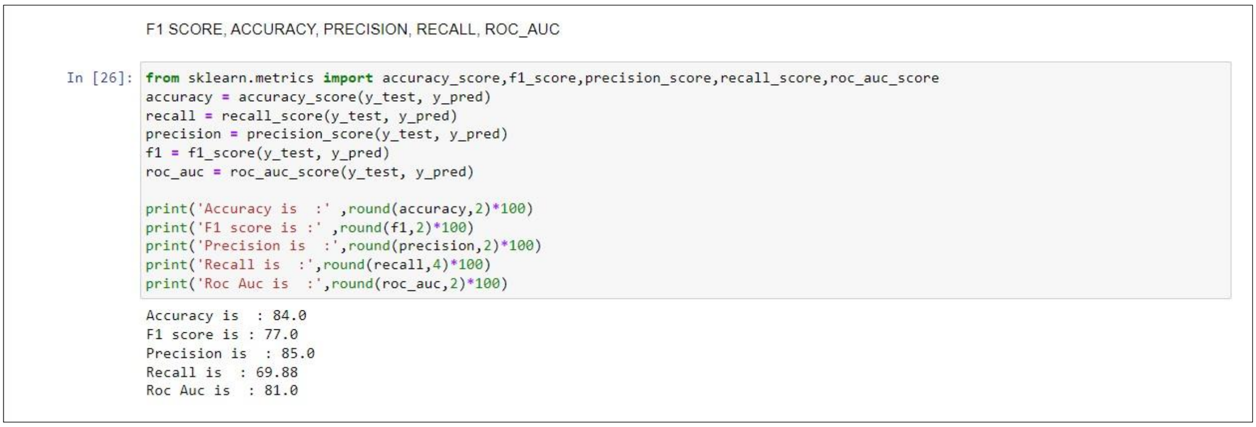


USING RANDOM FOREST

```
In [27]: random_forest = RandomForestClassifier(n_estimators=200)
         random_forest.fit(x_train, y_train)
         pred_random_forest = random_forest.predict(x_test)

In [28]: print(classification_report(y_test, pred_random_forest))

                       precision    recall  f1-score   support

                    0       0.83      0.90      0.87       140
                    1       0.81      0.70      0.75        83

             accuracy                           0.83       223
            macro avg       0.82      0.80      0.81       223
         weighted avg       0.82      0.83      0.82       223
```

Here Accuracy is 83%

Fig. 14 Implementation of Titanic Disaster Dataset using Python

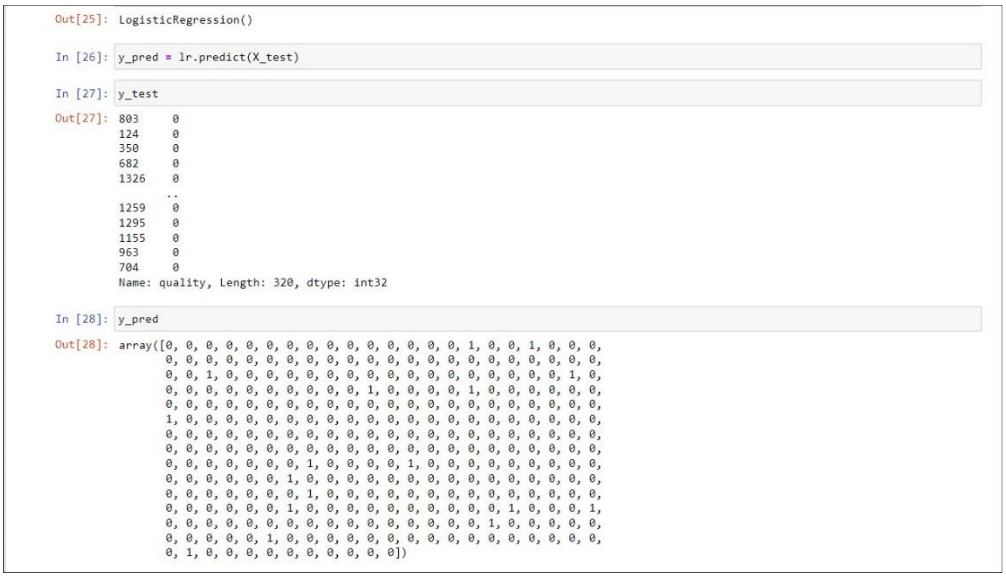


F1 SCORE, ACCURACY, PRECISION, RECALL, ROC_AUC

```
In [26]: from sklearn.metrics import accuracy_score,f1_score,precision_score,recall_score,roc_auc_score
         accuracy = accuracy_score(y_test, y_pred)
         recall = recall_score(y_test, y_pred)
         precision = precision_score(y_test, y_pred)
         f1 = f1_score(y_test, y_pred)
         roc_auc = roc_auc_score(y_test, y_pred)

         print('Accuracy is  :' ,round(accuracy,2)*100)
         print('F1 score is :' ,round(f1,2)*100)
         print('Precision is  :',round(precision,2)*100)
         print('Recall is  :',round(recall,4)*100)
         print('Roc Auc is  :',round(roc_auc,2)*100)

         Accuracy is  : 84.0
         F1 score is : 77.0
         Precision is  : 85.0
         Recall is  : 69.88
         Roc Auc is  : 81.0
```

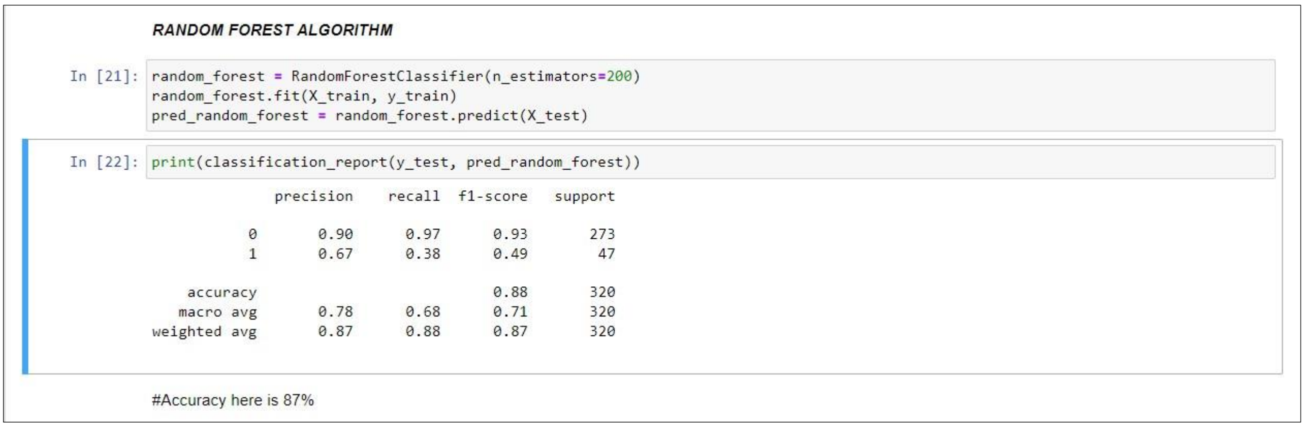
Fig. 15 Implementation of Titanic Disaster Dataset using Python

3.5.3.2. Implementation of Wine Quality Dataset



```
Out[25]: LogisticRegression()

In [26]: y_pred = lr.predict(X_test)

In [27]: y_test

Out[27]: 803     0
         124     0
         350     0
         682     0
         1326    0
                ..
         1259    0
         1295    0
         1155    0
         963     0
         704     0
         Name: quality, Length: 320, dtype: int32

In [28]: y_pred

Out[28]: array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0,
                0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,
                0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
                0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1,
                0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
                0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0])
```

Fig. 16 Implementation of Wine Quality Dataset using Python



RANDOM FOREST ALGORITHM

```
In [21]: random_forest = RandomForestClassifier(n_estimators=200)
         random_forest.fit(X_train, y_train)
         pred_random_forest = random_forest.predict(X_test)

In [22]: print(classification_report(y_test, pred_random_forest))

                       precision    recall  f1-score   support

                    0       0.90      0.97      0.93       273
                    1       0.67      0.38      0.49        47

             accuracy                           0.88       320
            macro avg       0.78      0.68      0.71       320
         weighted avg       0.87      0.88      0.87       320
```

#Accuracy here is 87%

Fig. 17 Implementation of Wine Quality Dataset using Python

4. Comparative Analysis

Tools & Criterion	Tableau	RapidMiner	Orange	Python
Usability	**Simple to Use**	**Easy to Use**	**GUI + CLI Compatible**	**Complicated as coding is required**
Speed	**Works fast on any machine**	**Requires more memory to operate**	**Slower while handling large data sets**	**Slower when comes to visualization**
Visualization	**Many visualization options**	**Comparatively less options than Tableau**	**Interactive Visualization**	**Time consuming data visualization using open libraries such as MatPlotLib, etc.**
Supported Algorithms	**Not used to implement algorithms**	**Classification & Clustering**	**Classification, Clustering and Regression**	**Most Machine Learning Algorithms can be implemented**
Data Set Size	**Supports any data set**	**Supports large and small data set**	**Supports small dataset efficiently**	**Supports any dataset**
Memory Usage	**Less Memory**	**Requires more memory**	**Requires more memory**	**Less Memory**
Primary Usage	**Business Intelligence**	**Data Mining, Predictive Analysis**	**Machine Learning, Data Visualization**	**Programming Language**
Interface Type Supported	**GUI**	**GUI**	**GUI/ CLI**	**CLI**
Data Sources	**Often Excel or CSV files**	**Often Excel or CSV files**	**Often Excel or CSV files**	**Streaming Data or Imported data files**
Data Model Adaptability	**Simple Data Models**	**Data Models focusing on aggregated datasets**	**Simple Data Models**	**Handles streaming data efficiently**
Dashboards	**Customized Dashboards**	**Customized Dashboard**	**Customized Dashboard**	**Complex Visualization Landscape**
User Friendliness	**Interactive**	**Simple, Easy to learn**	**Simple, Easy to Learn**	**Requires Programming Knowledge**
Pricing	**Student License Available**	**Student License Available**	**Free, Open Source**	**Free, Open Source**

5. Conclusions

In this project, a few big data tools were elucidated along with their features of several tasks. Big data provides extremely effective supporting mechanisms for the acquisition of extremely complicated and big data sets. This legal obligation paves the path for the development of several big data tools.

The project began by defining Big Data, including its definition, categories, and characteristics such as volume, velocity, variety, veracity, and value. Following this discussion of traits and classifications, a quick introduction to the aim of big data tools was given along with the description of datasets that we've used. We then dug further into each tool, examining its advantages and disadvantages, dos and don'ts, and addressing every imaginable how, why, and what.

The beginner's guide to Big Data Tools ends with a conclusion and the future of Big Data, with numerous challenges preceding its existence.

5.1. Future Scope

Today, Big Data is having an impact on the IT sector as few other technologies have. Sensor-enabled machinery, mobile devices, cloud computing, social media, and satellites create vast amounts of data that assist various firms enhance their decision-making and push their business to the next level.

"Big data has the potential to revolutionize the way governments, corporations, and academic institutions do business and make discoveries, and it's likely to impact how everyone lives their daily lives," says Susan Hauser, Microsoft's corporate vice president.

Data is the most significant development in the business since Steve Jobs developed the personal computer. As previously said, data is created at such a quick rate that traditional databases and other data storage systems will eventually fail to store, retrieve, and identify correlations between data. Through the utilization of commodity hardware and distribution, big data technologies have addressed the difficulties associated with this new big data revolution. Google, Yahoo!, General Electric, Cornerstone, Microsoft, Kaggle, Facebook, and Amazon are among the companies that are heavily investing in Big Data research and initiatives.

Big data isn't new, but it's suddenly gaining traction as more people digitize their lives. "People are walking sensors," said Nicholas Skytland, a NASA project manager in the Space Life Sciences Directorate's Human Adaptation and Countermeasures Division. Taking the average of all the estimates offered by top big data industry analyst and research firms, it can be determined that around 15% of all IT enterprises would migrate to cloud-based service platforms, with this market predicted to increase by 35% between 2015 and 2021.

5.1.1. Challenges

Big data has collected in recent years in a variety of fields, including health care, government administration, retail, biochemistry, and other multidisciplinary scientific projects. Big data is widely seen in web-based applications, such as social computing, online text and documents, and internet search indexing. Internet search indexing includes ISI, IEEE Xplorer, Scopus, and other services, whereas social computing comprises social network analysis, online communities, recommender systems, reputation systems, and prediction markets.

To meet the difficulties, we must be familiar with a variety of computational complexity and methods for analyzing large amounts of data. Similarly, many computational approaches that work well with little data confront substantial difficulties when applied to large data. Data storage and analysis, knowledge discovery and computational difficulties, scalability and data visualization, and information security are the four primary areas of big data analytics issues.

6. References

(i) “Top 10 Python Libraries for Data Science”, https://towardsdatascience.com/top-10-python-libraries-for-data-science-cd82294ec266, October 2021

(ii) “Python vs Tableau- Which One Is Better For Data Science?”, https://www.mindbowser.com/python-vs-tableau/, November 2021

(iii) “Comparative Study of Data Visualization Tools”, https://www.researchgate.net/publication/344425307_Comparative_Study_of_Data_Visualization_Tools, December 2021

(iv) “Comparative Analysis of Tools for Big Data Visualization and Challenges”, https://link.springer.com/chapter/10.1007/978-981-15-2282-6_3, December 2021

(v) “Comparison of Data Analysis Tools: Excel, R, Python and BI Tools”, https://towardsdatascience.com/comparison-of-data-analysis-tools-excel-r-python-and-bi-tools-6c4685a8ea6f, December 2021

(vi) “RapidMiner User Manual”, https://docs.rapidminer.com/downloads/RapidMiner-v6-user-manual.pdf, December 2021

(vii) “Tableau — A Beginners Guide”, https://medium.com/@sj20997/tableau-c9d6962991ca, November 2021

(viii) “Brief History of Orange, Praise to Donald Michie”, https://orangedatamining.com/blog/2013/10/09/brief-history-of-orange-praise-to-donald-michie/, November 2021

(ix) “Big Data: What Is It and How Does It Work?”, https://www.business2community.com/big-data/big-data-what-is-it-and-how-does-it-work-02265540, November 2021

(x) “Titanic Disaster Dataset”, https://data.world/nrippner/titanic-disaster-dataset/activity, November 2021

(xi) “Red Wine Quality Dataset”, https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009, November 2021

(xii) “US Traffic Accident Dataset”, https://www.kaggle.com/sobhanmoosavi/us-accidents/tasks, November 2021