IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 6, June 2023

DOI: 10.17148/IJARCCE.2023.12683

# A Machine Learning-Based Web Application for Simplifying Data Analysis and Prediction

**Ajay. M[1], Dr. Seedha Devi. V[2], Kumaran. M[3]**

Student M.E (CSE), Jaya Engineering College, Chennai, India[1]

Professor, Department of CSE, Jaya Engineering College, Chennai, India[2]

Professor, Department of CSE, Jaya Engineering College, Chennai, India[3]

**Abstract**: In our rapidly evolving world, technology advancements continue to shape our daily lives, prompting a shift towards modern and simplified techniques to accommodate our busy lifestyles. This project focuses on the utilization of machine learning algorithms, including linear regression, logistic regression, decision trees, SVM, Naive Bayes, KNN, K-means, Random Forest, dimensionality reduction algorithms, gradient boosting algorithms, and AdaBoosting algorithms, to streamline analytical work and prediction tasks. The system offers a user-friendly web interface that facilitates the loading of CSV and Excel data, allowing users to select and apply their preferred algorithm to suit their specific requirements. The system cleans the received data using data cleaning algorithms, and the user is presented with a list of options to assign algorithms to specific columns in the file. Graphs and charts generated by Google Charts based on the output of the predictions can be downloaded by the user. Additionally, the system enables users to visually compare two Excel or CSV files using charts, aiding in data analysis and comprehension. The application is developed using Django, Google Charts, Pandas, NumPy, and a MySQL database. Users can maintain distinct accounts to access their previous analytical work history conveniently. The application supports transforming various types of data into charts, allowing users to select and download the required charts.

**Keywords:** Machine Learning, Big Data, Charts, Data Analysis.

## I. INTRODUCTION

Machine learning (ML) is a branch of research concerned with understanding and developing methods that "develop" - that is, approaches that use data to improve performance on a set of tasks. It is regarded as a component of artificial intelligence. Machine learning algorithms construct a model from sample data, referred to as training data, in order to make predictions or judgements without being explicitly programmed to do so. Machine Learning algorithms are utilised in a wide range of applications, including medical, email filtering, speech recognition, agriculture, and computer vision, where developing traditional algorithms to execute the required tasks would be difficult or impossible. However, not all machine learning is statistical learning. Machine learning is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, and to uncover key insights in data mining projects.

Mathematical optimisation research provides methodology, theory, and application fields to the subject of machine learning. Data mining is a closely connected topic of research that focuses on exploratory data analysis via unsupervised learning. Some machine learning implementations use data and neural networks to replicate the operation of a biological brain. Machine learning is sometimes known as predictive analytics when used to commercial concerns.

The premise behind learning algorithms is that approaches, methods, and inferences that have proven successful in the past are likely to do so again in the future. These conclusions can be straightforward, such "since the sun has been rising every morning for the last 10,000 days, it will probably rise again." likewise for tomorrow morning. They can be subtle, like "Y% chance that undiscovered black swans exist because X% of families have geographically separate species with color variants."

Programs that use machine learning can do tasks without having them explicitly coded. Computers use available data to learn in order to do specific jobs. For straightforward jobs given to computers, it is easy to build algorithms that instruct the device how to carry out all the steps necessary to address the issue at hand; no learning is required on the part of the computer.

  

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 6, June 2023
DOI: 10.17148/IJARCCE.2023.12683

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Additionally, it provides an excellent way for employees or business owners to present data to non-technical audiences without confusion. In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

## II. DATA ANALYTICS

Data analytics (DA) is the process of examining data sets to find trends and draw conclusions about the information they contain. Increasingly, data analytics is done with the aid of specialized systems and software. Data analytics technologies and techniques are widely used in commercial industries to enable organizations to make more-informed business decisions. Scientists and researchers also use analytics tools to verify or disprove scientific models, theories and hypotheses.

As a term, data analytics predominantly refers to an assortment of applications, from basic business intelligence (BI), reporting and online analytical processing (OLAP) to various forms of advanced analytics. In that sense, it's similar in nature to business analytics, another umbrella term for approaches to analyzing data. The difference is that the latter is oriented to business uses, while data analytics has a broader focus. The expansive view of the term isn't universal, though: In some cases, people use data analytics specifically to mean advanced analytics, treating BI as a separate category.

Data analytics initiatives can help businesses increase revenue, improve operational efficiency, optimize marketing campaigns and bolster customer service efforts. Analytics also enable organizations to respond quickly to emerging market trends and gain a competitive edge over business rivals. Depending on the application, the data that's analyzed can consist of either historical records or new information that has been processed for real-time analytics. In addition, it can come from a mix of internal systems and external data sources.

At a high level, data analytics methodologies include exploratory data analysis (EDA) and confirmatory data analysis (CDA). EDA aims to find patterns and relationships in data, while CDA applies statistical techniques to determine whether hypotheses about a data set are true or false. EDA is often compared to detective work, while CDA is akin to the work of a judge or jury during a court trial -- a distinction first drawn by statistician John W. Tukey in his 1977 book Exploratory Data Analysis. Data analytics can also be separated into quantitative data analysis and qualitative data analysis. The former involves the analysis of numerical data with quantifiable variables. These variables can be compared or measured statistically. The qualitative approach is more interpretive -- it focuses on understanding the content of non-numerical data like text, images, audio and video, as well as common phrases, themes and points of view.

Data warehouses typically store structured business data. However, most data in organisations is unstructured data, which occupies about 80 per cent of an enterprise's data [13], [14]. Therefore, it is vital to transform the traditional data warehouse into an efficient unstructured data warehouse. Big data refers to huge data sets characterised by larger volumes and greater variety and complexity, generated at a higher velocity than the normal operational data that an organisation has handled before. As more and more enterprises recognise the values and advantages associated with big data insights, the adoption of big data tools like Hadoop ecosystem is growing. Hence, utilising big data tools as an enhancement to the data warehouse to handle unstructured data besides structured one is a feasible and practical approach to resolve the limitation of the traditional data warehouse and potentially expand its adoption in organisations.

## III. LITERATURE REVIEW

A Big Data Analysis for knowledge based on Machine Learning using Classification Algorithm was created by Assefa Senbato Genale in 2022. This paper approaches the necessity and development made in higher education system using Machine learning algorithm, most commonly the author have involved the concepts and algorithm based on Support Vector Machines and Learning Management systems.[1] The prediction model is the most important thing which is created by the machine, while seeing about the statistical model with music the tempo and medium tempo is received as data from the user. Based on this concept the author has compared it with the educational system with a large amount of data consumption and that is similar to Deep learning.

In 2016 Sergio Ledesma created this work Analysis of data sets with learning conflicts for Machine Learning this algorithm is used to identify the learning conflicts that are intentionally inserted.[3] Next, an artificial neural network is trained and evaluated using the contaminated data set. The algorithm proposed in this work is used in a real-world

DOI: 10.17148/IJARCCE.2023.12683

application to detect problems in a data set for a refrigeration system. It is concluded that the algorithm can be used to improve the performance of machine learning systems.

An Exploratory data analysis of Kyiv city petitions is created by Artur Samvelyan in 2020 and this topic of the analysis of online petitions is not new. The author used Latent Dirichlet Allocation (LDA) with the purpose of automating the extraction of topics in a petitions dataset. In Twitter was analyzed as a platform of e-petition topic discussions and debates as well as the corresponding tweets' sentiments. [2]The conclusions about petition popularity based on the semantics of a textual dataset are made in. Most of the literature on the topic discusses petitions written in the English language with huge textual datasets. We, on the other hand, propose and analyze a small dataset of petitions in the Ukrainian language that is hoped to inspire more research in the study of both low-resource languages and exploratory data analysis with natural language processing given a limited amount of data.

A New Approach to Use Big Data Tools to Substitute Unstructured Data Warehouse by Oras Bake in 2020. This research, we utilised the IBM BigInsights Text Analytics, PostgreSQL, and Pentaho tools, an unstructured data warehouse is implemented and worked excellently with the unstructured text from Amazon review datasets, the new proposed approach creates a practical solution for building an unstructured data warehouse. Data warehouses typically store structured business data. However, most data in organisations is unstructured data, which occupies about 80 per cent of an enterprise's data. Therefore, it is vital to transform the traditional data warehouse into an efficient unstructured data warehouse.[7] Big data refers to huge data sets characterised by larger volumes and greater variety and complexity, generated at a higher velocity than the normal operational data that an organisation has handled before. As more and more enterprises recognise the values and advantages associated with big data insights, the adoption of big data tools like Hadoop ecosystem is growing. Hence, utilising big data tools as an enhancement to the data warehouse to handle unstructured data besides structured one is a feasible and practical approach to resolve the limitation of the traditional data warehouse and potentially expand its adoption in organisations.

In 2020 Jitha Janardhanan created this Data Analytic Tools as an overview The process of extracting interesting patterns from vast database hiding is known as Data mining the choice of best one among abundant data mining tools accessible in the sooqis not easy.[8] A numeral factors demand to be contemplated before formulating a financing in any commercial solution. This article stretches the complete and hypothetical analysis of a few open source data mining tools. The data mining tools and their utilities comparison are conferred here. The paper also discusses about Data Analytics and its importance.

A Web Mining based Patent Analysis and Citation Visualization was proposed by Zhiqiang Liu in the year 2017. [6]This states that Patent data is one of the most valuable reservoir of technical and commercial knowledge. However, patent data of different countries and organizations has been stored separately, which added to the difficulty for patent analysis with these free patent data. By using Web Mining method, we can retrieve the related patent information of a certain field from several patent website as data source. From web pages to structural patent database, we can use these free patent sources to obtain professional patent analysis result and useful knowledge. A citation structure visualization method is introduced, which helps reveal the citation relationships among patents. Public and commercial patent database are two major data sources for patent analysis. Some commercial database, such as Derwent Innovation Index (DII), Dialogue, and Wisdomain, etc, had integrated patent data from different countries and organizations. With the development of the Internet, the free patent websites have become more and more convenient and widely used. In recent years, several patent offices have made some of their databases publicly available on the Internet. While most likely the promoters of free patent information on the Internet may have inventors as potential users in mind, free access to patents may also further the uptake of patent information by probably unintended user groups, such as researchers and policy analysts.

## IV. PROPOSED METHODOLOGY

In this proposed system we are going to implement lot of machine learning algorithm that is majorly used by the people and the Data Analyst. This helps the user to use any algorithm to any of the dataset for what the chart is required. This also do the Data predicted with the help of Data prediction algorithm which is used in Machine Learning. This application use Pandas to clean the data frame and to generate the values and use the Google charts to generate the graphical representation of the data.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 6, June 2023

DOI: 10.17148/IJARCCE.2023.12683

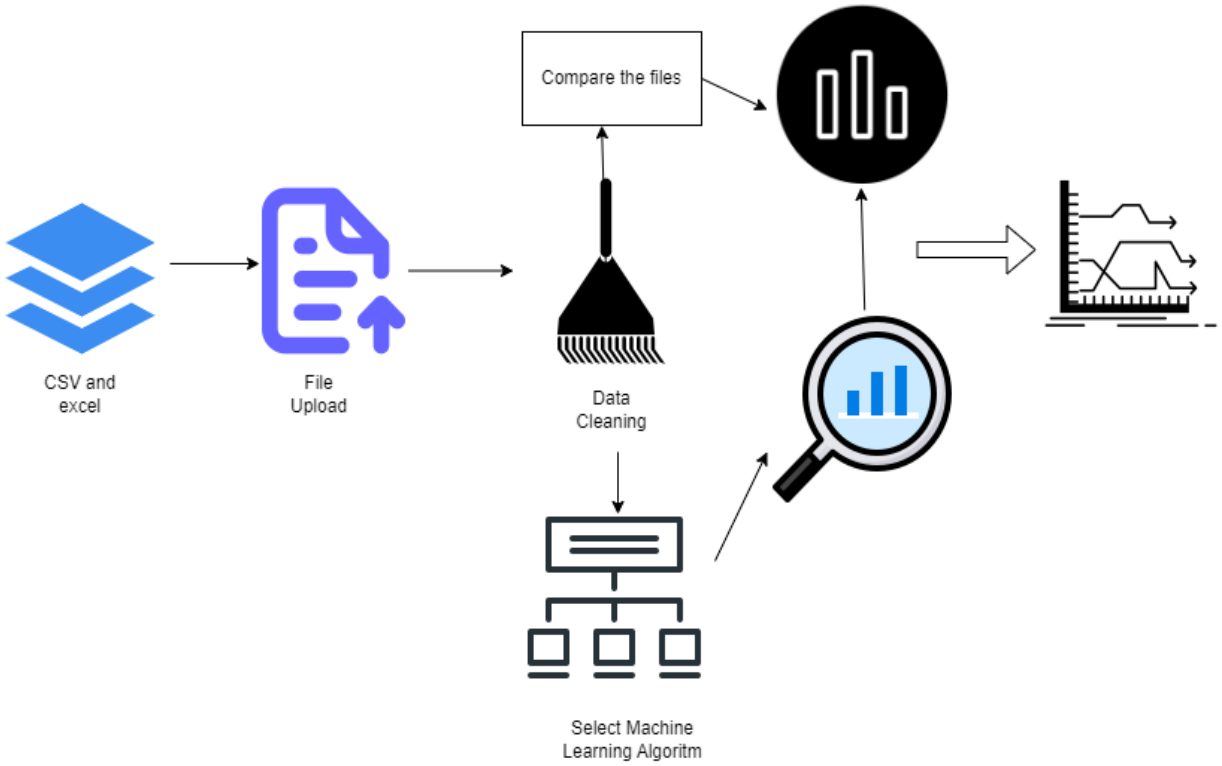


Fig 1: Report Generation and Data Analysis System

All the machine learning algorithm will be implemented in the Python based code and the User interface is created in Python Django which helps the user to upload the CSV and excel files and also to select the columns of the files which need to be compared or graphed. End users are able to download the charts which they generated based on their requirement. Each user will be assigned with the individual login where they can keep track of all their activity and the charts which is generated by the application.

Data cleaning is a crucial process in Data Mining. It carries an important part in the building of a model. Data Cleaning can be regarded as the process needed, but everyone often neglects it. Data quality is the main issue in quality information management. Data quality problems occur anywhere in information systems. These problems are solved by data cleaning. Data cleaning is fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled.

Generally, data cleaning reduces errors and improves data quality. Correcting errors in data and eliminating bad records can be a time-consuming and tedious process, but it cannot be ignored. Data mining is a key technique for data cleaning. Data mining is a technique for discovering interesting information in data. Data quality mining is a recent approach applying data mining techniques to identify and recover data quality problems in large databases. Data mining automatically extracts hidden and intrinsic information from the collections of data. Data mining has various techniques that are suitable for data cleaning.

Data analytics allows organizations to digitally transform their business and culture, becoming more innovative and forward-thinking in their decision-making. Going beyond traditional KPI monitoring and reporting to finding hidden patterns in data, algorithm-driven organizations are the new innovators and business leaders.
By shifting the paradigm beyond data to connect insights with action, companies are able to create personalized customer experiences, build connected digital products, optimize operations, and increase employee productivity.

Executing this tool with the Car purchase Dataset against the SVM algorithm which will analysis and split the users who brought SUV and who not brought SUV.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 6, June 2023

DOI: 10.17148/IJARCCE.2023.12683

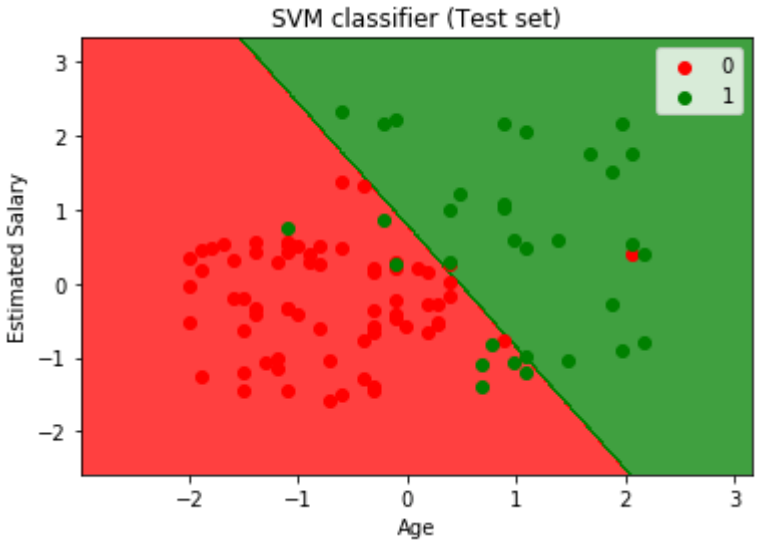


Fig 2: SVM Classifier Algorithm output

As we can see in the above output image, the SVM classifier has divided the users into two regions (Purchased or Not purchased). Users who purchased the SUV are in the red region with the red scatter points. And users who did not purchase the SUV are in the green region with green scatter points. The hyperplane has divided the two classes into Purchased and not purchased variable.

## V. CONCLUSION AND FUTURE ENHANCEMENTS

We have succeeded in our aim to develop a system that can be used to generate the reports and compare the data on a UI based Application. Which have the ability to read all the types of CSV and excel files. They are being generated based the required columns and also this System can compare two files and give results in the visual repetition manner. This system works based on majorly used machine learning algorithm which makes user to easily analysis the data sets which they need to analysis by just uploading the files and choose which algorithm they need and finally they can download the result which is generated by the system, so all the end users can use this in a friendly way.

We can expand this in the future to incorporate other machine learning techniques, as well as provide the ability to submit two files, compare the data, and produce reports. Additionally, we may link the session to the cloud to keep the user's extensive history and data in a highly safe manner using encryption. So that it enables the files to a secured location like cloud storage explorers which helps to process huge sized files. we can use PySpark to process faster results because it uses big data and hadoop architecture.

## REFERENCES

[1] Assefa Senbato Genale “Big Data Analysis for knowledge based on Machine Learning using Classification Algorithm”, 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS).
[2] SERGIO LEDESMA “Analysis of data sets with learning conflicts for Machine Learning”, International Journal of Research in Engineering, Science and Management, 2018.
[3] Gennadiy Kyselov “Exploratory data analysis of Kyiv city petitions”, International Conference of Research in Engineering, Science and Management, 2020.
[4] Oras Baker “A New Approach to Use Big Data Tools to Substitute Unstructured Data Warehouse”, International Conference of Research UTC in Engineering, Science and Management, 2021
[5] Irene Martín-Morató “A case study on feature sensitivity for audio event classification using support vector machines”, 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)
[6] Zhiqiang Liu, Donghua Zhu “Web Mining based Patent Analysis and Citation Visualization”, 2017 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS).
[7] Amarjeet Rawat “Sentiment Analysis of Covid19 Vaccines Tweets Using NLP and Machine Learning Classifiers”, 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)
[8] Kelly, J. E. (2015). Computing, cognition and the future of knowing. IBM.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 6, June 2023
DOI: 10.17148/IJARCCE.2023.12683

[9] Roberts, P. (2010). Corralling Unstructured Data for Data Warehouses. Business Intelligence Journal, 15(4), 50-55.
[10] Gupta, V., & Rathore, N. (2013). Deriving Business Intelligence from Unstructured Data. International Journal of Information and Computation Technology, 3(9), 971-976.
[11] Gonzalez, S. M., & Berbel, T. d. (2014). Considering unstructured data for OLAP: a feasibility study using a systematic review. Salesian Journal on Information Systems, 14, 26-35.
[12] Tekadpande, S., & Deshpande, L. (2015). Analysis and Design of ETL process using Hadoop. International Journal of Engineering and Innovative Technology (IJEIT), 4(12), 171-174.
[13] Kelly, J. E. (2015). Computing, cognition and the future of knowing. IBM.
[14] Roberts, P. (2010). Corralling Unstructured Data for Data Warehouses. Business Intelligence Journal, 15(4), 50-55.