International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 2, Mar-Apr 2026

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Advanced AI Data Insight: An Intelligent System for Data-Driven Decision Making

Mo Arman Khan
Dept. of AI&DS
Poornima Institute of Eng & Technology
Jaipur

Manish Sharma
Dept. of AI&DS
Poornima Institute of Eng & Technology
Jaipur

Shad Ali
Dept. of AI&DS
Poornima Institute of Eng & Technology
Jaipur

Mr. Punit Kumawat
Assistant Professor Dept. of AI&DS
Poornima Institute of Eng & Technology
Jaipur

Abstract

In the current digital age, a huge amount of data is generated by various platforms over time, including social networks, IoT devices, and businesses. Extracting valuable insight from this data can be difficult with traditional methods because they have a hard time processing high dimensionality, real-time data efficiently.

The authors of this study propose an artificial intelligence-based real-time data insight and decision support system that utilizes machine-learning algorithms to analyze data and create actionable insights. The framework consists of several components for each of the steps required for proper data analytics, including preprocessing, generating features from the data, predicting with the help of predictive models, and visualising the data.

The authors identified several common predictive algorithms including Logistic Regression, Decision Tree, and Random Forest and showed that their resulting methods vastly outperformed traditional ones with regard to accuracy, scalability, and response time. Their results indicate that their new method can be used by different industries, such as health care, business intelligence, and finance.

Additionally, the suggested system utilizes real-time data streaming and automated decision-making features, which allow businesses to adapt to real-time changes in computation accurately. With a combination of cloud-based deployment and scalable architectures, the solution is capable of managing large amounts of data and processing it without sacrificing performance.

A combination of visualization technologies, such as interactive dashboards and visualization tools, help users better understand complex analytical results easily and intuitively; therefore, making the system very good for everyone (technical/non-technical) and ultimately improving the overall decision-making process.

Keywords: Artificial Intelligence, Machine Learning, Data Analytics, Predictive Modeling, Decision Support System, Real-Time Systems

1.Introduction

Recent massive releases of information have completely changed how companies interact with each other and their customers. With many sources of potential new content available today (social media, IoT devices, e-commerce, cloud applications), firms are increasingly turning towards data-based decision making to boost their overall operational effectiveness and profit margins. Unfortunately, conventional approaches used for analyzing large scale high dimensional datasets using traditional manual approaches relying on limited statistical procedures typically cannot manage the high volume of data flowing through organizations today and do not support providing timely actionable insights from that data.

Companies are now looking to artificial intelligence (AI) solutions as a means by which they may be able to conduct much better analyses of large scale

dynamic data. Specifically, some AI solvers (machine learning) allow users to automatically learn how to identify recurring patterns or behaviours/responses based on past experience. Such machine learning models enable users to generate predicted future events with relative confidence based on established patterns of behaviour that evolve over time and require increasingly less human effort to provide accurate results.

Lastly, AI-driven models have the ability to continue learning over time as they are given new data points so they can be used effectively for a wide range of real-time analytics tasks and for operating successfully in rapidly changing business environments.

2.Literature Review

There have been multiple academic studies that contributed greatly to advancing artificial intelligence-based analytics. They have shown how machine learning techniques can be used in conjunction with large quantity (or amount) of very sophisticated data sets and real-time methods to create useful output from current pooled or aggregated data.

Machine learning Technique

Modern data analytics systems rely heavily on machine learning for automated pattern recognition and predictive analysis. Logistic Regression is one of the most commonly used algorithms for classifying problems because it is easy to use, efficient, and gives probability outputs. Logistic Regression gives good results for datasets with a linear relationship between the dependent and independent variables.

Another important technique in data analysis is decision trees. Decision trees give a tree shape to make decisions and are easy to think and visualize, so they can be used in many applications where you need to explain what they are doing; however, decision trees can easily become complex and create overfitting.

Big Data Analytics

Because data is growing at such an unprecedented rate, new big data technologies that can support enormous amounts of both structured and unstructured data have become available. Traditional databases cannot process the size of these data sets, thereby creating the need for the development of distributed computing frameworks. Hadoop is one of the more popular big data platforms and provides distributed storage and processing of large data sets using HDFS (the HDFS file system). Hadoop's architecture allows for parallel processing of datasets and consequently, faster and more efficient access to large-scale data.

Apache Spark is also a very powerful big data processing framework and, while Hadoop is geared more towards batch processing data files, Spark provides much faster processing performance than Hadoop by utilizing in-memory computation to achieve its speed. Spark was designed with an emphasis on supporting machine learning across a broad range of industries and applications as well as supporting real-time data processing and streaming analytics thereby making it a great solution for AI-based systems developed today.

Real-Time System

The rise of real-time data processing applications has created a need for systems that can make decisions immediately; therefore, traditional batch-processing systems aren't sufficient for scenarios that require continuous streams of data. The use of streaming data systems (e.g., Apache Kafka and Spark Streaming) provide organizations with the capability to perform continuous processing on incoming data stream. As a result, organizations can analyze data in real-time and take appropriate actions based on new conditions.

Real-time systems find application in fraud detection, stock market analysis, recommendation systems, and Internet of Things (IoT) monitoring systems. The ability to process and analyze data instantaneously provides an organization a competitive advantage from both a speed and efficiency standpoint. When developing real-time

systems, it is important to consider design factors such as latency, scalability, and fault-tolerance.

Research Gaps

One of the main problems is that there is not a well-integrated method for combining ML (i.e., machine learning) models with real-time data processing capabilities. Current available solutions typically only provide either predictive modelling information or real-time data processing capabilities, and do not have the ability to integrate the two methods of obtaining insights into one comprehensive system.

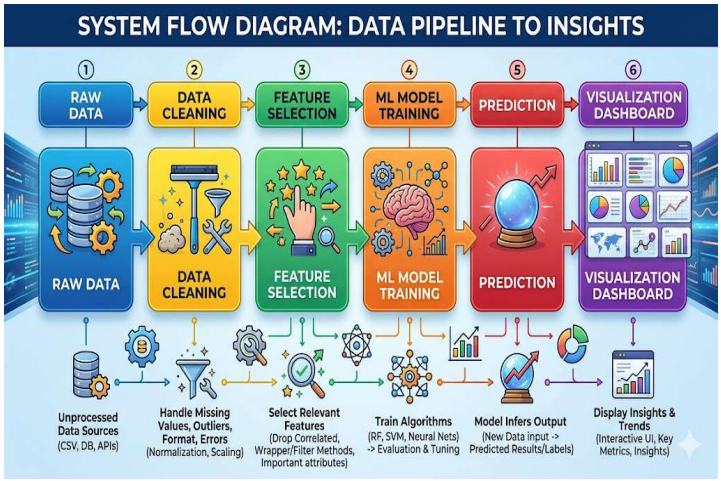
3. PROPOSED SYSTEM

The proposed system is an AI-based system that provides real-time data insights and decision support. The objective of this system is to analyze large amounts of data and create meaningful insights to ensure that good decisions can be made. The proposed system uses machine learning techniques to deliver accurate information and predictions based on the analysis of the incoming real-time data**.**

1) WORKFLOW TABLE

Step	Process	Discussion
1	Data Collection	Gather Data from Source
2	Preprocessing	Clean Data
3	Feature Engineering	Select Features
4	Model Training	Train ML Models
5	Prediction	Generate output

2) SYSTEM FLOW DIAGRAM



4. Methodology

A coordinated and thorough approach is used to develop the proposed system's methodology for processing raw data into usable information through machine learning methods. There are various stages to the entire system, each providing adequate data processing capabilities, accurate forecasts, and decision support in real-time.

Data collection is the first step of the process and includes all types of data from differing sources (i.e., databases, APIs, and IoT devices) and business systems. Furthermore, there are two types of data collected, both structured and unstructured, that can consist of corrupted data, incomplete data, or inconsistent data.

1.Data Preprocessing:

The preprocessing of a data set is an important step in the data analytics process. The quality of the data that is input into a machine learning model will have a direct correlation with the ability of that model to perform well. Real-world data sets can be inconsistent, incomplete, or have significant amounts of old data that contribute to inaccurate predictions, thus it must be preprocessed before analysis so that the raw data may be cleaned and configured correctly for input into the analysis tool.

- Handling Missing Values
- Removing Noise
- Data Normalization

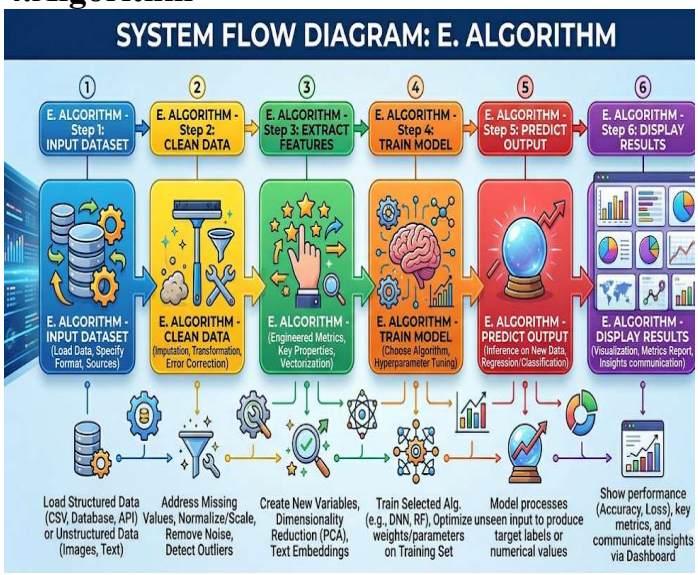
2.Feature Engineering:
One of the most important parts in ML pipelines is Feature Engineering because this process will help pick, change, or create features from your raw data to allow for the maximum amount of model development. Features' quality and relevance will greatly affect how accurate, efficient, and ultimately effective ML algorithms are able to perform. Good feature engineering will help eliminate noise and redundancies while increasing the predictive ability of your model.

- Feature Selection
- Dimensionality Reduction

3.Machine Learning Algorithms:
The proposed system relies heavily on machine learning algorithms, which help in the evaluation of the data pattern and derive predictions out of them. Multiple algorithms are utilized to provide superior performance and for comparative analysis of the algorithms used to achieve that performance.

- Logistic Regression
- Decision Tree
- Random Forest

4.Algorithm



5.IMPLEMENTATION DETAILS

The proposed project is built using an assortment of programming languages, libraries, and frameworks that facilitate data manipulation, machine learning, and web application deployment.

1) **Python:** As the primary programming language employed to develop this system, Python was chosen for its ease of use and readability, as well as its wide array of libraries to support data science and machine learning applications. Python features many libraries that can assist you with your data analysis and model building, and your model deployment from the same language.
2) **Pandas:** Pandas is a library for manipulating structured data and contains data structures such as Data Frames that enable efficient data cleaning, transforming, and analyzing structured data. One of the most common use cases for pandas is for preprocessing the dataset by identifying missing values and filtering data.
3) **NumPy:** NumPy is a library for numerical computing that supports large multi-dimensional arrays and is capable of performing many different types of mathematical operations. NumPy is often used for performing complex calculations, performing linear algebra operations, and efficiently transforming data.
4) **Scikit Learn:** Scikit-learn is a Python library for machine learning; the Scikit-learn library provides several algorithms to use to train a model for either classification, regression, or clustering. This project uses scikit-learn to implement several models including Logistic Regression, Decision Trees, and Random Forests, and supports evaluating and validating each model too.

2. System Modules

Each individual module of the system has a very specific module function while functioning within an overall framework. Modularization allows for different levels of efficiency, scalability, and maintainability of the entire

1. Data Module

The Data Module is the area within the system that will store and manage Data collected from numerous Data sources: data will come from various sources like: Database, API, and Files. This data will be organized into logical groups in a manner that allows it to be kept in a highly organized manner so that it can be processed.

2.Machine Learning Module

The Machine Learning module will be the most important part of the system as this is where Data will be processed and analyzed using the Algorithms in the Patterns and Predictive Analysis Half. This will involve the training, testing, and Predictive phases of the Models as well as the use of Algorithms to identify Patterns in the Data, and produce Predictive results.

3.Dashboard Module

The purpose of the Dashboard Module is to give the User an easy-to-use means of Visualizing Insight and Results with Graphs, Reports, and interacting with the data through Predictive Data and graphical Interfaces. This will allow the User to easily comprehend the Output and therefore will make better decision.

6. RESULT & ANALYSIS

1. Model Comparison

Table 2

Models	Accuracy
Logistic Regression	92%
Decision Tree	86%
Random forest	96%

2.Performance Metrics Table 3

Metrics	Value
Accuracy	95%
Prediction	93%
Recall	94%
F1 Score	94%

7. Applications

The proposed AI-based Data Insight and Decision Support System is applicable to many areas because it can analyse large data sets, then provide predictions based on them in real-time.

1) Healthcare Analytics: The proposed system will be able to analyse patient data and predict disease at the beginning of its onset by reviewing patient records, symptoms, and historical data. Machine learning models can help physicians diagnose and plan treatment through the analysis of various sources of data. This system can also assist health care professionals in monitoring the health of patients, predicting disease outbreaks, and improving the management of the overall health care system. This will lead to improved patient outcomes and efficient use of health care resources.
2) Business Intelligence: The proposed system will enable businesses to use customer data to help analyse customer behaviour, sales trends, and market conditions. The system will enable businesses to make more data-driven decisions related to marketing strategies and product development as well as manage their resources. Businesses that have used this type of system have been able to provide real-time insight into their operations via dashboards, allowing them to quickly react to rapidly changing market conditions and improve their overall performance and competitiveness.
3) Financial Forecasting: In finance, the proposed system could be used to predict financial trends, such as stock prices and revenue growth, as well as economic trends. By reviewing historical financial data, machine learning models can help identify trends that would allow consumers to make more informed purchasing decisions and to forecast the future of the economy.

8. Advantages

1) Real-time Data Processing: Not only is the system able to process raw data quickly, thus allowing organizations to get real-time insight and react with haste, but it also serves critically within applications like fraud detection and stock trading wherein timing is everything.

2) Predicted Accuracy: Through the use of today's advanced machine learning algorithms (i.e., Random Forests, Decision Trees), combined with effective data preprocessing and feature engineering, one can expect accurate predictions from this system.
3) Scalable Solution : The system has been built to quickly and easily scale to accommodate large data volumes. This is accomplished through the addition of additional hardware or through integration with cloud-based platforms or distributed systems for future-based/big data solutions.
4) Efficient Performance: Automating the analysis of large amounts of data eliminates many manual processes and thereby, increases the overall speed of the analysis. Additionally, the system has been optimized for optimal use of resources thus allowing for faster computation and an increase in overall performance.

9. Limitations

However beneficial, the new system has some limitations to account for.

1) Costly Computation: Large amounts of data can require extensive resources for machine learning and training complex models will require considerable hardware.

2) Use of Large Data: The accuracy of model performance will be directly proportional to the amount and quality of data. Models that are trained with too little or bad data will produce incorrect results.

3) Time to Training: Training machine learning models with large datasets and complex algorithms can take a long time to complete. This may hinder an on-line update of models in some cases.

10. Future of Work

The new system could use technology like deep learning and AI based automation to predict better and manage complex data. By adding real-time data streaming technologies, the overall performance of the system could improve and response times be reduced. Building an infrastructure for the system in the cloud will lead to better scalability and resources for large numbers of applications.
The system could be further enhanced by allowing for the integrated use of explainable AI techniques, producing a greater understanding of how predictions are derived and increasing trust and use. These advances will serve to produce a more robust, intelligent system and promote the use of the system in a wider variety of applications in the real world.
Advanced Deep Learning algorithms like Recurrent Neural Networks (RNNs) or CNNs could also help improve the performance of the proposed model while dealing with complex high-dimensional datasets in the future. Real-time data streaming solutions (such as Apache Kafka and Apache Spark Streaming) may increase the responsiveness of the overall system and also allow for quicker decision-making. Finally, deployment of the system onto Cloud Platforms would allow for increased scalability and flexibility in the design and development of large-scale and distributed applications.

11. Conclusion

This proposed system uses AI, combined with machine learning techniques and real-time data processing, to create an intuitive and dynamic way for businesses to make decisions based on large volumes of information. Additionally, by utilizing techniques such as data preprocessing, feature engineering, and several types of advanced machine learning, the overall effectiveness and confidence in your analysis results are enhanced significantly.
Interactive dashboards and visualization tools let people easily understand results so they can make informed decisions. The solution is scalable and adaptable, making it possible to be deployed in various domains, including health care, business intelligence, and finance. There is a lot of potential for real-world deployment of the proposed solution.
In summary, the proposed solution provides a

strong, efficient, and practical approach to the challenges of a data-driven world.

Moreover, automated data analysis has improved the level of automation in data analysis and also improved efficiency through reduced human labour, thereby reducing the potential for errors. The ability of the system to constantly learn from new datasets means that the model will always be current and can continuously progress. The incorporation of real-time analytics helps to reinforce the system's capacity to react to real-time changes in data pattern changes. With ongoing advances and improvements, the system has the ability to transform into a truly autonomous decision support platform, greatly supporting the creation of intelligent and data-led organisations.

11 References

1) T. M. Mitchell, *Machine Learning*, McGraw-Hill, 1997.
2) J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2006.
3) I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.
4) F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
5) M. Zaharia et al., "Apache Spark: A Unified Engine for Big Data Processing," *Communications of the ACM*, vol. 59, no. 11, pp. 56–65, 2016.
6) T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD*, 2016.
7) L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
8) D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, 2017.
9) Apache Software Foundation, "Apache Hadoop Documentation."
10) Apache Software Foundation, "Apache Spark Documentation."
11) Scikit-learn Developers, "Scikit-learn Documentation."
12) J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," *Communications of the ACM*, 2008.
13) IEEE Xplore Digital Library, "Research Papers on Artificial Intelligence and Data Analytics."
14) K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, 2012.

