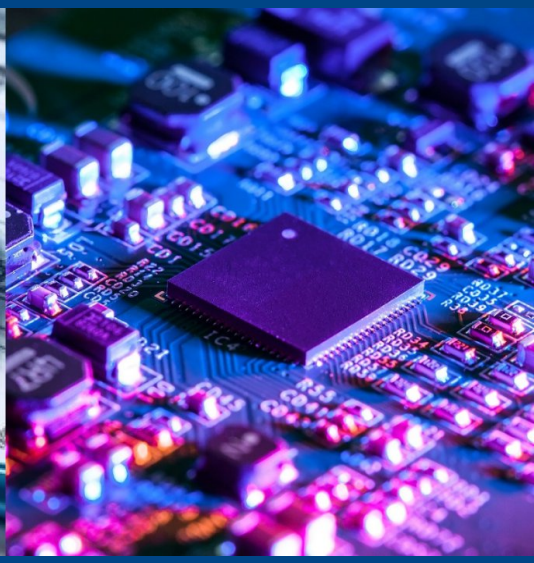
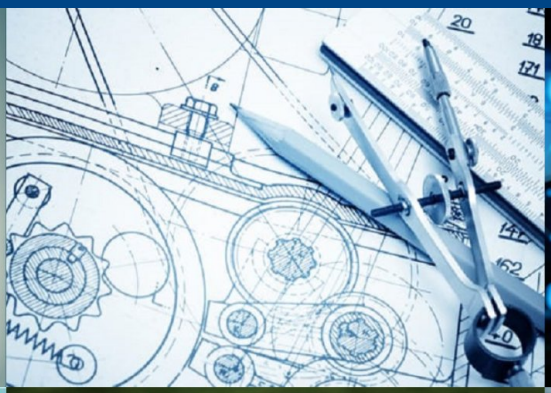
ISSN: 2582-7219

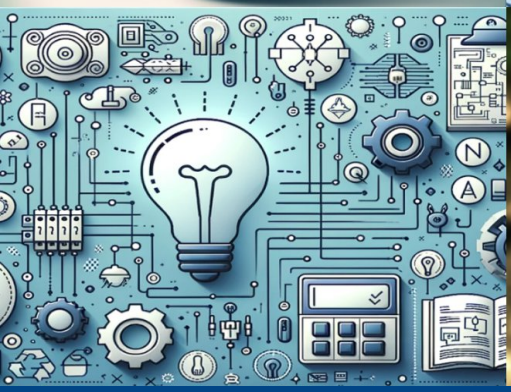




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 9, Issue 4, April 2026**

© 2026 IJMRSET | Volume 9, Issue 4, April 2026| DOI:10.15680/IJMRSET.2026.0904142

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Personality Profiling and Data Leak Prevention Using Python

Harish Gowtham G[1], Dharshan M[2], Manoj Kumar V[3], Sivakumar B[4], Vignesh G[5]

Assistant Professor, Dept. of CSE, PGP CET, Namakkal, Tamil Nadu, India[1]

Student, Dept. of CSE, PGP CET, Namakkal, Tamil Nadu India [2,3,4,5]

ABSTRACT: In the modern digital era, organizations collect large volumes of user data to understand customer behavior and deliver personalized services, but analyzing such data without proper security can lead to privacy issues and data leakage. This paper presents a secure system for personality profiling and data leak prevention using Python, where machine learning techniques like K-Means clustering are used to analyze user data and group individuals into meaningful personas based on behavioral patterns and preferences. At the same time, the system ensures strong data security through encryption techniques and access control mechanisms, including advanced methods such as Attribute-Based Encryption (ABE) and hybrid encryption, which provide fine-grained access control and efficient protection of sensitive information. The proposed system not only enhances the accuracy of user profiling but also maintains data confidentiality and integrity, making it highly useful for applications such as marketing, recommendation systems, and secure data management platforms, thereby demonstrating an effective integration of machine learning and data security to build a reliable and privacy-aware solution.

KEYWORDS Machine Learning, K-Means Clustering, Data Security, Data Leakage Prevention, Attribute-Based Encryption, Hybrid Encryption, User Profiling.

I. INTRODUCTION

In today's data-driven world, organizations rely heavily on user data to understand customer behavior, preferences, and needs. This analysis helps businesses provide personalized services, targeted marketing strategies, and improved user experiences. Personality profiling plays a vital role in identifying different categories of users by analyzing their behavioral and demographic data.

Machine learning techniques enable automated and efficient analysis of large datasets, making it easier to identify patterns and group users into meaningful clusters. Among these techniques, clustering algorithms such as K-Means are widely used for persona prediction due to their simplicity and effectiveness.

However, the increasing use of personal data raises serious concerns regarding data privacy and security. Sensitive information stored in databases or cloud systems is vulnerable to unauthorized access, data breaches, and leakage. Therefore, it is essential to implement strong security mechanisms to protect user data.

This paper proposes a secure personality profiling system that integrates machine learning techniques with advanced data security methods. The system predicts user personas while ensuring data confidentiality through encryption and access control mechanisms. By combining data analysis with security, the proposed system provides a reliable solution for modern data-driven applications.

II. LITERATURE REVIEW

In the modern data-driven environment, analyzing user data for personality profiling and ensuring data security are two critical challenges. Traditional systems mainly focus on either user behavior analysis or data protection, but rarely integrate both effectively. Several research works have explored machine learning techniques such as clustering, classification, and regression for user profiling and recommendation systems. For instance, Han et al. introduced data mining techniques for discovering user behavior patterns, while Breiman proposed Random Forest methods for predictive analysis. However, these approaches primarily focus on accuracy and often overlook data privacy concerns.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

On the security side, conventional encryption techniques such as AES and RSA are widely used to protect sensitive information. While these methods provide strong data protection, they lack flexibility in handling dynamic access control in distributed environments. To address this, Sahai and Waters introduced Attribute-Based Encryption (ABE), which enables fine-grained access control based on user attributes rather than identities. Further advancements include hybrid encryption techniques that combine symmetric and asymmetric encryption to improve both efficiency and security.

Despite these developments, existing systems still suffer from limitations such as lack of integration between data analysis and security, scalability issues, and insufficient mechanisms for preventing data leakage during processing. Most systems either provide accurate profiling without security or strong security without intelligent analysis. To overcome these limitations, the proposed system integrates machine learning-based personality profiling with advanced encryption techniques. It combines K-Means clustering for user segmentation with secure data handling methods such as ABE and hybrid encryption, ensuring both accurate prediction and data privacy. This integrated approach provides a reliable and secure solution for modern applications.

Relevance to current Research

This project contributes to the fields of machine learning and data security by integrating personality profiling and data leak prevention into a unified system. Implemented using Python, the system applies K-Means clustering to group users based on behavioral patterns while simultaneously protecting sensitive data through encryption and access control mechanisms. By incorporating Attribute-Based Encryption and hybrid encryption techniques, the system ensures secure data storage and controlled access.

Additionally, the system improves decision-making in applications such as marketing and recommendation systems by providing accurate user segmentation. At the same time, it enhances data privacy by preventing unauthorized access and leakage. This combination of intelligent analysis and strong security makes the system highly relevant for modern data-driven environments, including cloud computing, business analytics, and secure information systems.

Summary Table: Related Work

No	Paper / Technique	Authors	Key Points	Relevance to Current Work
1	Data Mining for User Profiling	Han et al.	Identifies user behavior patterns	Basis for persona analysis
2	Random Forest Algorithm	Breiman	Improves prediction accuracy	Supports data analysis
3	AES Encryption	Stallings	Secures data using symmetric encryption	Data protection
4	Attribute-Based Encryption	Sahai & Waters	Fine-grained access control	Secure data access
5	Hybrid Encryption	Various	Combines symmetric & asymmetric encryption	Efficient and secure system

III. METHODOLOGY OF PROPOSED SURVEY

The proposed system methodology consists of four main stages: data collection, data preprocessing, personality profiling, and secure data protection. Initially, the system collects user-related data such as demographic details, behavioral patterns, and preferences. This data is then processed using preprocessing techniques like data cleaning, normalization, and feature selection to ensure accuracy and consistency.

In the next stage, machine learning algorithms such as K-Means clustering are applied to group users into different personas based on similarities in their data. This helps in identifying meaningful patterns and categorizing users effectively. After profiling, the system focuses on data security by implementing encryption techniques such as AES to protect sensitive information.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Further, Attribute-Based Encryption (ABE) is used to provide fine-grained access control, ensuring that only authorized users can access specific data based on defined attributes. A hybrid encryption approach is adopted to enhance both performance and security by combining symmetric and asymmetric encryption methods. Finally, the system ensures secure data storage and controlled access, preventing data leakage and maintaining data confidentiality, integrity, and privacy.

MODULES

The system is organized into several modules, each handling a specific function of the application. The modules are described as follows:

1. User Data Collection Module

Collects user information such as demographic data, preferences, and behavioral details required for analysis.

2. Data Preprocessing Module

Cleans and processes the collected data by removing noise, handling missing values, and normalizing data for better accuracy.

3. Clustering Module (K-Means)

Applies K-Means clustering algorithm to group users into different personas based on similarity in their data.

4. Persona Prediction Module

Analyzes clustered data to identify user behavior patterns and generate meaningful personality profiles.

5. Encryption Module (AES)

Encrypts sensitive user data to ensure confidentiality and protect it from unauthorized access.

6. Access Control Module (ABE)

Implements Attribute-Based Encryption to control access based on user roles and attributes.

7. Data Storage Module

Stores encrypted data securely in the database, ensuring protection against data leakage.

8. User Interface Module

Provides an interactive interface for users/admin to input data, view results, and monitor system operations.

IV. SYSTEM ARCHITECTURE

The system architecture of the proposed Personality Profiling and Data Leak Prevention system is designed as a layered model that integrates data processing, machine learning, and security mechanisms. The architecture consists of user input, data preprocessing, clustering, and secure storage components. Initially, user data is collected through the interface and passed to the preprocessing module, where it is cleaned and prepared for analysis. The processed data is then sent to the clustering module, where K-Means algorithm groups users into different personas based on their behavior and preferences. After profiling, the system applies encryption techniques such as AES to secure sensitive data, and Attribute-Based Encryption (ABE) is used to control access based on user roles. The encrypted data is stored in a secure database, ensuring confidentiality and preventing data leakage. This architecture ensures efficient data analysis while maintaining strong security and privacy protection throughout the system.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

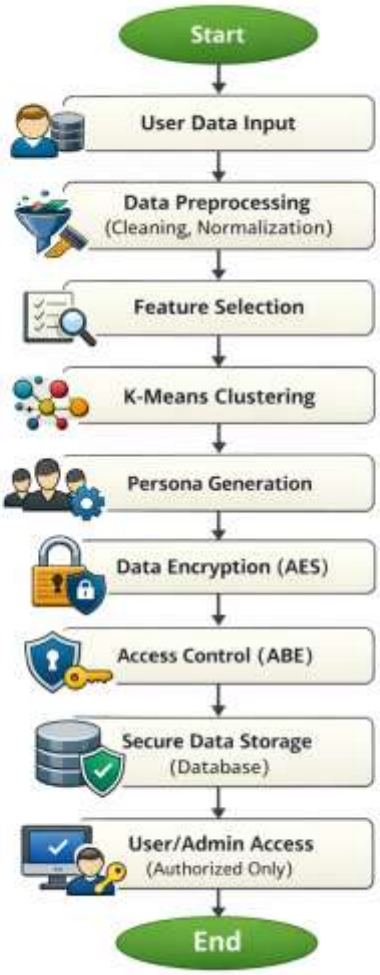


Fig 1 Work Flow

IV. CONCLUSION AND FUTURE WORK

The proposed system successfully demonstrates an effective and secure approach for personality profiling and data leak prevention using machine learning and encryption techniques. By integrating K-Means clustering, the system accurately groups users into meaningful personas based on their behavioral and demographic data, improving decision-making in applications such as marketing and recommendation systems. At the same time, the implementation of strong security mechanisms such as AES encryption and Attribute-Based Encryption (ABE) ensures data confidentiality, integrity, and controlled access. This combination overcomes the limitations of traditional systems that focus only on data analysis or security independently. The system provides a reliable, scalable, and privacy-aware solution for handling sensitive user data in modern digital environments.

Future enhancements can further improve the system's performance, scalability, and intelligence. Advanced machine learning techniques such as deep learning and hybrid models can be integrated to enhance the accuracy of personality prediction. The system can be extended to real-time data processing and deployed in cloud environments for large-scale applications. Incorporating blockchain technology can further strengthen data security and transparency. Additionally, implementing anomaly detection mechanisms can help identify potential data leakage attempts proactively. The development of a web-based or mobile interface for real-time monitoring and user interaction can enhance usability. Including performance metrics such as accuracy, precision, and security efficiency will provide deeper insights and help optimize the system for real-world applications.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

REFERENCES

[1] Breiman, L., "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[2] Han, J., Kamber, M., and Pei, J., "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2011.
[3] Dwork, C., "Differential Privacy," ICALP, 2006.
[4] Sahai, A. and Waters, B., "Fuzzy Identity-Based Encryption," EUROCRYPT, 2005.
[5] Stallings, W., "Cryptography and Network Security: Principles and Practice," 7th Edition, 2017.
[6] Aggarwal, C. C., "Data Mining: The Textbook," Springer, 2015.
[7] Goodfellow, I., Bengio, Y., and Courville, A., "Deep Learning," MIT Press, 2016.
[8] Rivest, R., Shamir, A., and Adleman, L., "A Method for Obtaining Digital Signatures and Public-Key Cryptosystems," Communications of the ACM, 1978.
[9] Goldreich, O., "Foundations of Cryptography," Cambridge University Press, 2001.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com