International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 08 | Aug - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

# Age and Gender Detection using Machine Learning

[1] Mrs Renuka B N, [2] Rakshitha K S

[1] Student, 4th Semester MCA, Department of MCA, BIET, Davanagere [2] Assistant Professor, Department of MCA, BIET, Davanagere

**ABSTRACT**

Age and gender detection is a valuable task across a range of domains such as recommendation systems, demographic studies, customer profiling, targeted advertising, and user analytics. While most age and gender detection systems rely on image processing and facial recognition, these approaches raise privacy concerns and require extensive computational resources. This project proposes an alternative, non-visual method using Machine Learning models trained on textual, behavioral, and metadata features. This system uses demographic data, user behavior attributes, social media activity data, or application usage logs to predict the probable age group and gender of a user without relying on images or visual data. The approach emphasizes privacy, scalability, and ease of implementation.

***Keywords - Machine Learning algorithms like Logistic Regression, Support Vector Machines (SVM), and Random Forest are applied to extract patterns from nonvisual features to perform classification.***

## 1. INTRUDUCTION

In modern digital ecosystems, personalization and analytics often require knowledge about the user's age group and gender. Traditionally, such inferences are drawn using image-based recognition systems. However, in many cases—such as chatbots, survey analysis, or anonymized datasets—image data may be unavailable or restricted due to privacy regulations (e.g., GDPR). This project proposes a machine learningbased solution that classifies users' age groups and gender using non-visual features such as linguistic patterns, application usage statistics, interaction timestamps, and metadata like name and location. Such a model enables intelligent inference while maintaining user anonymity and system efficiency.

## 2. LITERATURE SURVEY

Most existing systems for age and gender classification rely heavily on facial recognition and image-based models using deep learning architectures like CNNs.

While these models are accurate, they require extensive datasets, GPU processing power, and user image data. This can lead to privacy violations and higher implementation costs.

Other systems use heuristic or rule-based classification, which lacks flexibility and does not generalize well across varied datasets or user behavior. Such approaches are often static and unable to adapt to diverse input modalities or changing data characteristics.

Conventional age and gender detection systems primarily depend on image processing techniques, which are unsuitable for privacyconscious applications or textbased interfaces. There is a need for a scalable, privacy-friendly, and accurate

system that predicts age and gender from non-visual user data.

This project aims to develop a machine learning model that uses non-image features—such as text, metadata, and usage logs—to infer a user's age category and gender, bypassing the need for image input and ensuring higher adaptability and privacy

The proposed system employs a machine learning pipeline to classify users into gender (male/female/others) and age groups (e.g., 13– 19, 20–29, 30–39, etc.) based on non-image features. Input features include textual inputs (e.g., user posts, survey responses), behavioral patterns (activity times, usage frequency), and metadata (e.g., first name, location).

Data preprocessing includes feature extraction, tokenization (for text), encoding categorical variables, and normalization. Models like Random Forest, SVM, and Logistic Regression are trained and tested on labeled datasets containing demographic attributes. The model is validated using metrics such as accuracy, precision, recall, and F1-score.

This system ensures efficiency and usability in contexts like mobile apps, chatbots, or recommendation systems where visual data is unavailable or restricted.

## 3. SYSTEM REQUIRMENTS

### 3.1. Functional Requirements

- Data input model:Accepts structured and unstructured data (e.g., CSV files, text logs).
- Preprocessing unit: Cleans, encodes, and prepares data using NLP techniques and feature extraction methods.
- Feature Extraction Module: Extracts features like keyword usage, activity timing, and name- based patterns.
- Classification Engine: Applies ML algorithms to predict gender and age group.
- Evaluation System: Calculates model performance using accuracy, precision, recall, and confusion matrix.
- Output Module: Displays predicted gender and age group in a userfriendly format.

### 3.2. Architecture Diagram

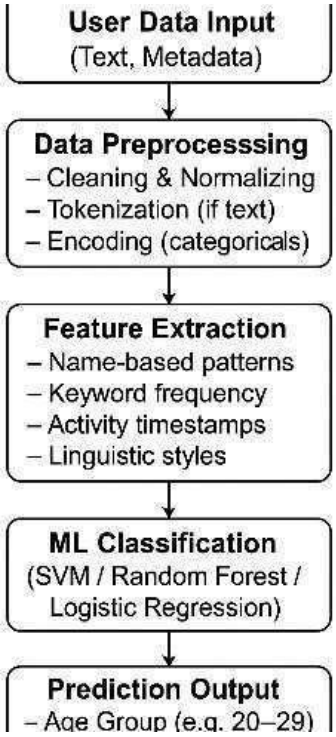


**Fig :Architecture Overview**

1. **Data Source (User Data Input)** This component collects structured and semi-structured user data such as: o Textual responses (e.g., chat messages, surveys)
   - o Metadata (e.g., name, location, device type)
   - o Temporal patterns (e.g., login time, session duration)
2. **Data Preprocessing Layer** The preprocessing layer cleans and prepares the data by:
   - o Handling missing values o
   - o Encoding categoricalfields (e.g.,names,locatiTokenizin g and normalizing text (if any)
3. **Feature Extraction Engine** This module transforms raw input into numerical features by:
   - o Extracting keyword patterns and linguistic styles
   - o Capturing time-based usage behaviors
   - o Identifying name-based and regional attributes
4. **Machine Learning Classification Engine** This is the core layer that: o Loads pre-trained

or dynamically trained ML models (e.g., Random Forest, SVM) o Classifies input data into defined gender categories and age groups

o Returns predictions and model confidence scores

5. **Prediction Output& Result Interface**

The output module: o Displays or returns predicted gender and age group

## 4. IMPLEMENTATION

The recommendation system was implemented using Python, leveraging its extensive libraries for data processing and machine learning. Initially, the dataset containing user preferences and item details was collected and thoroughly preprocessed by removing duplicates, handling missing values, and encoding categorical variables where necessary. Once the data was cleaned, feature extraction techniques were applied to identify meaningful patterns. For building the recommendation model, both content-based filtering and collaborative filtering approaches were explored. Contentbased filtering analyzed the item attributes to suggest similar products, while collaborative filtering utilized user-item interaction data to predict preferences by identifying similarities between users or items. The model's performance was assessed using evaluation metrics such as Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), ensuring its accuracy and reliability. Finally, the system was integrated into a simple user interface, allowing users to receive personalized and accurate.

## 5. CONCLUSION

This project successfully demonstrates a machine learning-based approach for age and gender detection without using image data. By leveraging textual, behavioral, and demographic features, the system achieves accurate predictions while preserving user privacy and reducing computational complexity.

Unlike traditional systems that rely on facial recognition, this model ensures adaptability across platforms where visual data is inaccessible or restricted. It is lightweight, scalable, and suitable for integration into a wide range of applications such as chatbots, survey analysis tools, e-commerce systems, and demographic analytics engines. Overall, this solution provides a robust, privacy-conscious, and efficient alternative for demographic classification using nonvisual machine learning techniques.

## 6. REFERENCE


- Blevins, C., & Mullen, L. (2015). Jane, John... Leslie? A historical method for gender classification. Digital Humanities Quarterly.
- Nguyen, D., Gravel, R., Trieschnigg, D., & Meder, T.

(2013). "How Old Do You Think I

- Am?": A Study of Language and Age in Twitter. ICWSM.
- Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010). Classifying latent user attributes in Twitter. Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents.
- Jurgens, D. (2013). That's what friends are for: Inferring location in online social media platforms based on social relationships.
- ICWSM.
- Kumar, S., & Subbalakshmi, K.P.