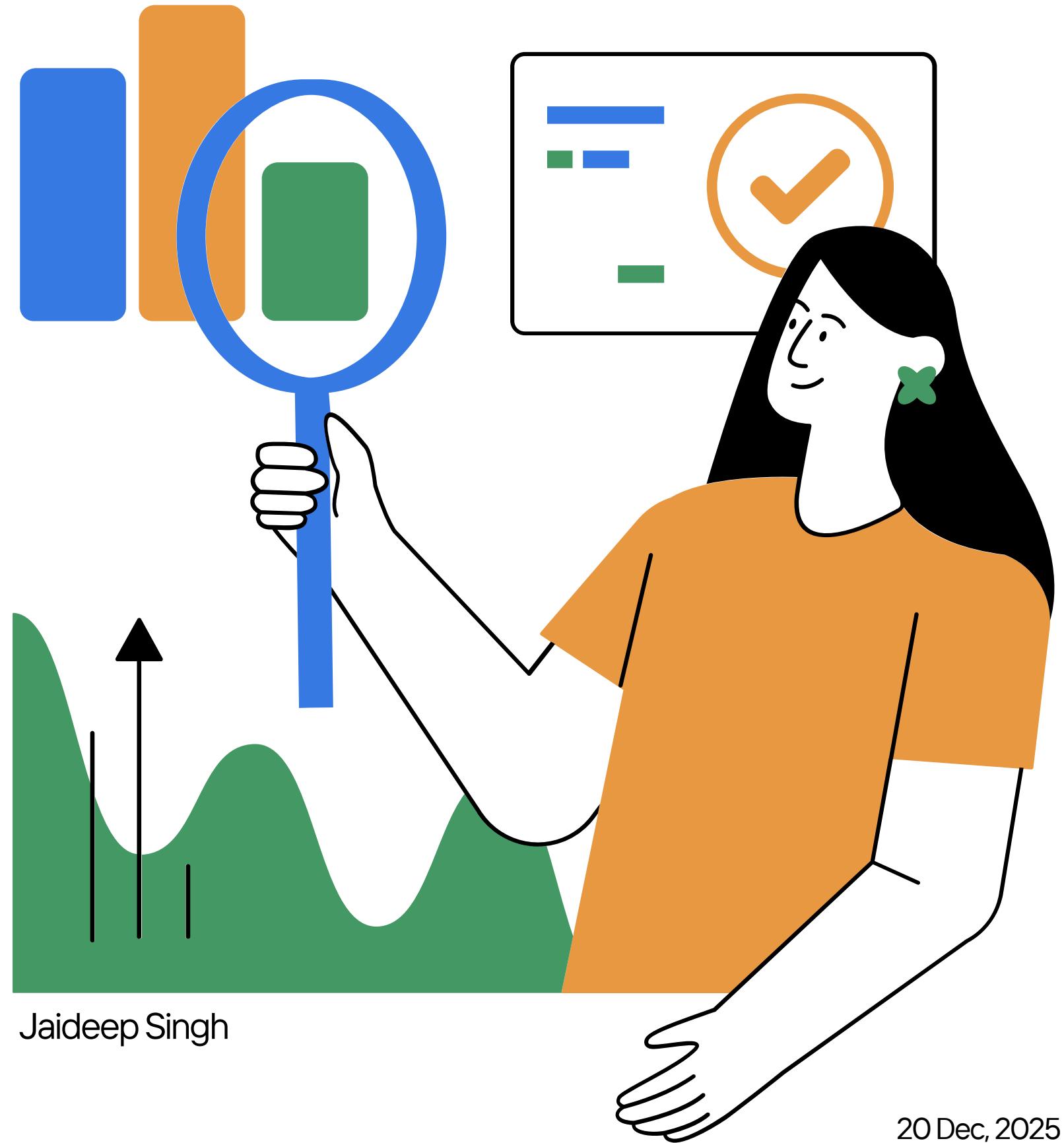
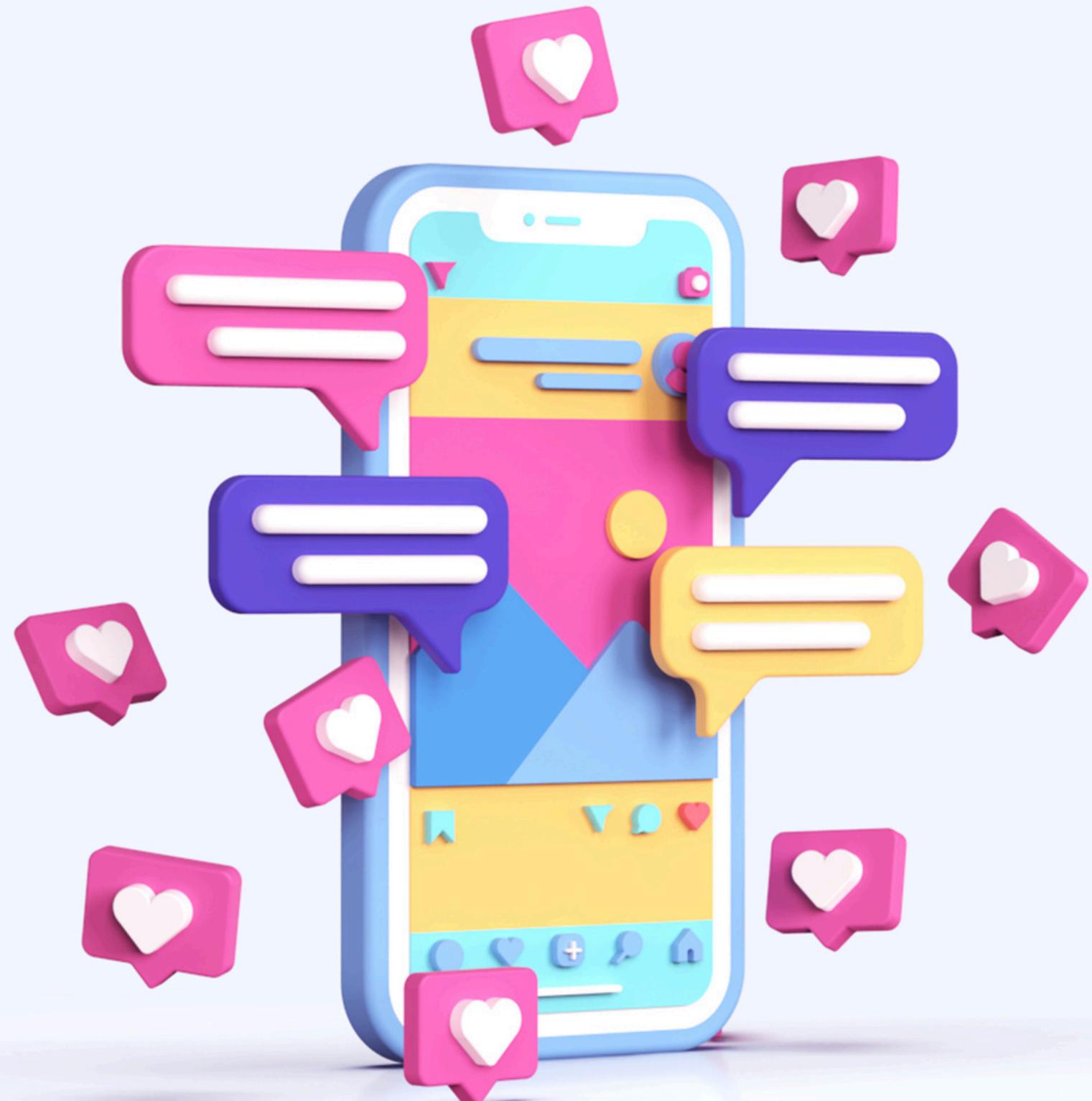
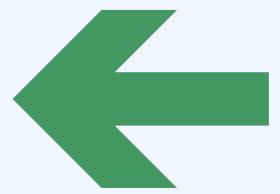


# Predictive Analytics





# What is a InstaPer?

InstaPer is a machine learning based system designed to predict an Instagram user's personality by analysing the hashtags used in their posts. The project leverages the Big Five (OCEAN) personality model that is Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism to map social media behaviour to psychological traits. By extracting hashtags from Instagram content, categorising them into meaningful domains, and applying trained classification models, InstaPer provides an automated and interpretable approach to personality prediction using publicly available data.

# DataSet

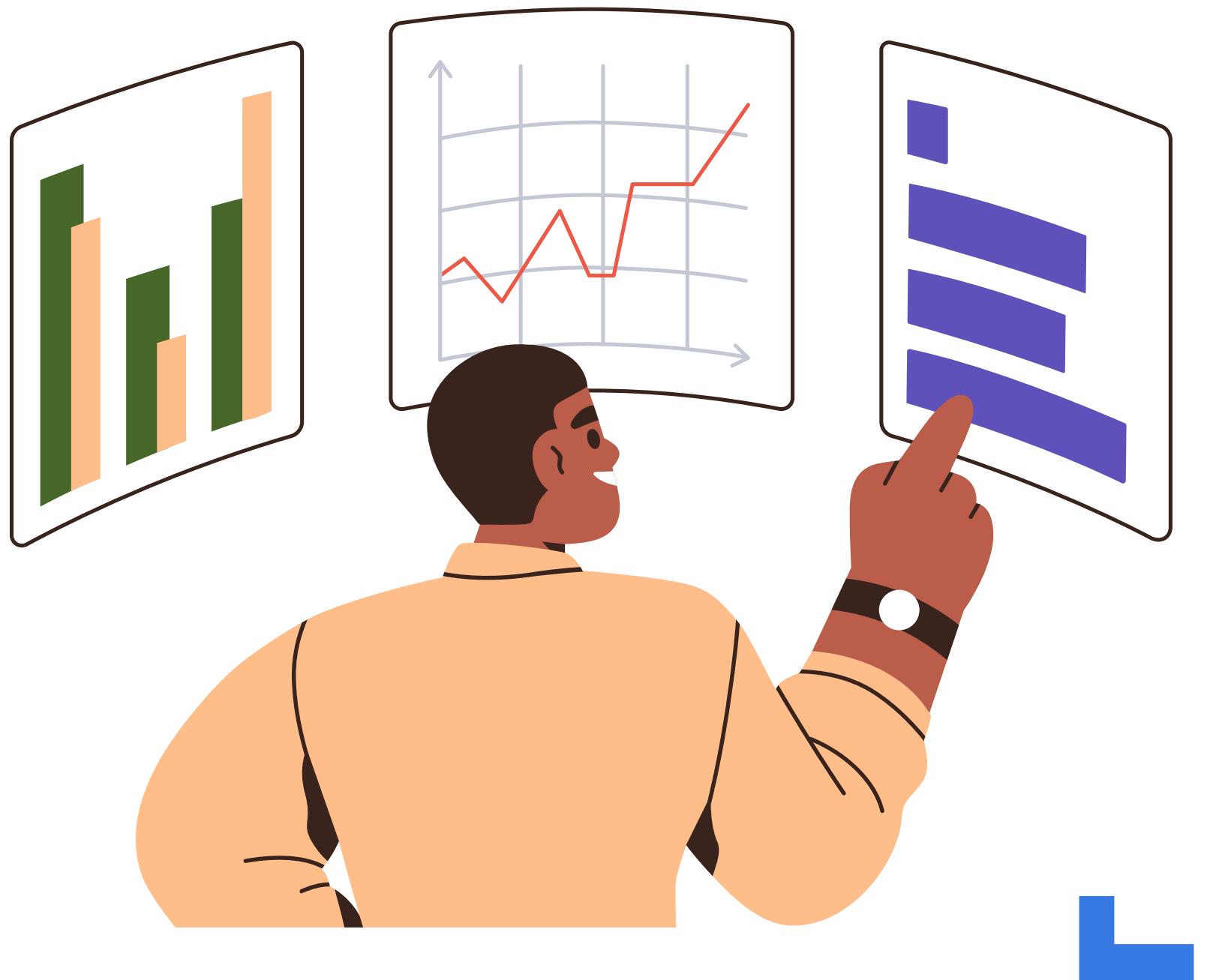
Table\_1(**Post\_table\_with\_user\_id**) dataset was collected using the **Apify** web scraping platform, which enables automated extraction of publicly available Instagram data. Instagram post and reel URLs were used as inputs to scrape metadata such as captions, hashtags, timestamps, likes, and comments. The scraping process focused only on public content, ensuring ethical data collection. Extracted hashtags were cleaned, and structured into a tabular dataset, forming the foundation for feature engineering and personality prediction.

**IEEE DataSet**

**Kaggel DataSet**

<b>Post_table_with_user_id</b>		<b>Hashtags_Table</b>		<b>Personality_Table</b>	
post_id(int)		hashtag(str)		user_id(int)	
user_id(int)		frequency(int)		openness(int)	
caption(str)		category(str)		extraversion(int)	
OwnerFullscreen(str)		sentiment(str)		agreeableness(int)	
OwnerUsername(str)				conscientiousness(int)	
url(str)				neuroticism(int)	
commentsCount(int)				personality_label(str)	
likesCount(int)					
hashtags/0...hashtags/30					
	4353*40	25099*4			3610*7

# DataSet



Table\_2(**Hashtag\_Table**) contains a structured analysis of extracted hashtags. Each hashtag is stored along with its frequency of occurrence, an assigned content category (such as Art, Technology, Fitness, etc.), and its **sentiment polarity**. Sentiment analysis was performed using **TextBlob**, a lightweight NLP library that evaluates the emotional tone of text and classifies hashtags as positive, negative, or neutral. This table serves as a key feature source for mapping Instagram behavior to personality traits.

Table\_3(**Personality\_Table**) represents the final personality feature construction stage. In this table, hashtags from each post are combined and analyzed collectively, and category-wise occurrence scores are calculated for domains such as *Travel & Nature*, *Art & Creativity*, *Technology*, *Fitness*, and others. These category scores are then mapped to the **Big Five (OCEAN)** personality traits, generating quantitative values for **Openness**, **Conscientiousness**, **Extraversion**, **Agreeableness**, and **Neuroticism**. Based on the relative distribution of these five traits, a custom personality label (such as *Adventurous*, *Creative*, or *Balanced*) is assigned to each user.

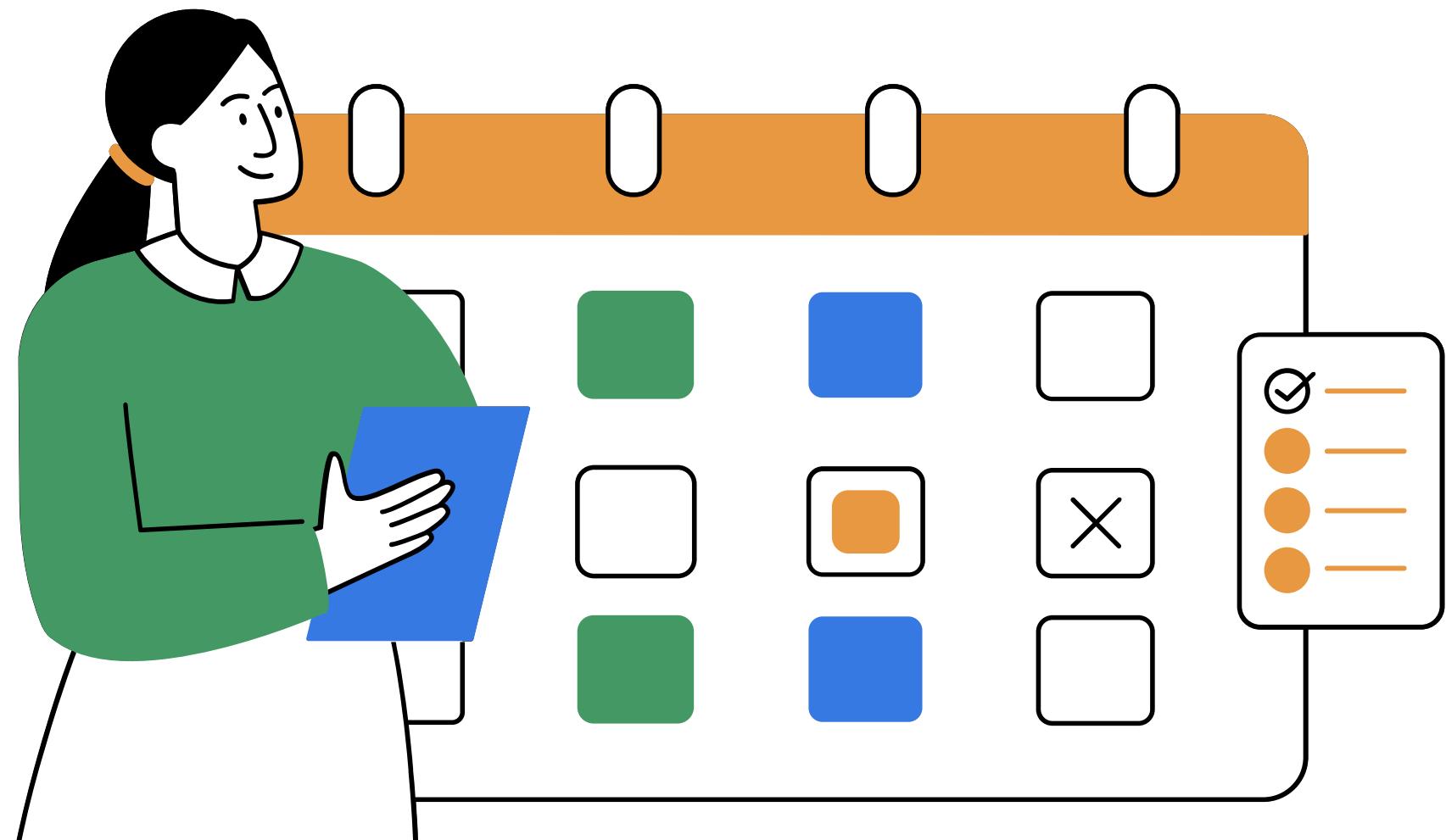
# Machine Learning Modelling

## Features (X) and Target (y)

To build the personality prediction system, the processed dataset was divided into features (X) and target (y).

The input features (X) consist of the computed **Big Five (OCEAN)** personality trait scores derived from hashtag category occurrences.

The target variable (y) is the final personality label (such as *Adventurous, Creative, Balanced*, etc.), which represents the overall personality classification.



## Model Selection

To identify the most effective model, five different machine learning classifiers were trained and evaluated:

- **Logistic Regression** – A linear classification model used as a baseline to understand how well personality labels can be separated using linear decision boundaries.
- **Random Forest Classifier** – An ensemble-based model that combines multiple decision trees to improve accuracy and reduce overfitting.
- **Gradient Boosting Classifier** – A sequential ensemble model that builds trees iteratively, where each new model corrects the errors of the previous one.
- **Extra Trees Classifier** – A highly randomized tree-based ensemble that increases diversity among trees for faster training and reduced variance.
- **Support Vector Machine (SVM)** – A margin-based classifier that attempts to find the optimal hyperplane separating different personality classes.





## Performance Comparison

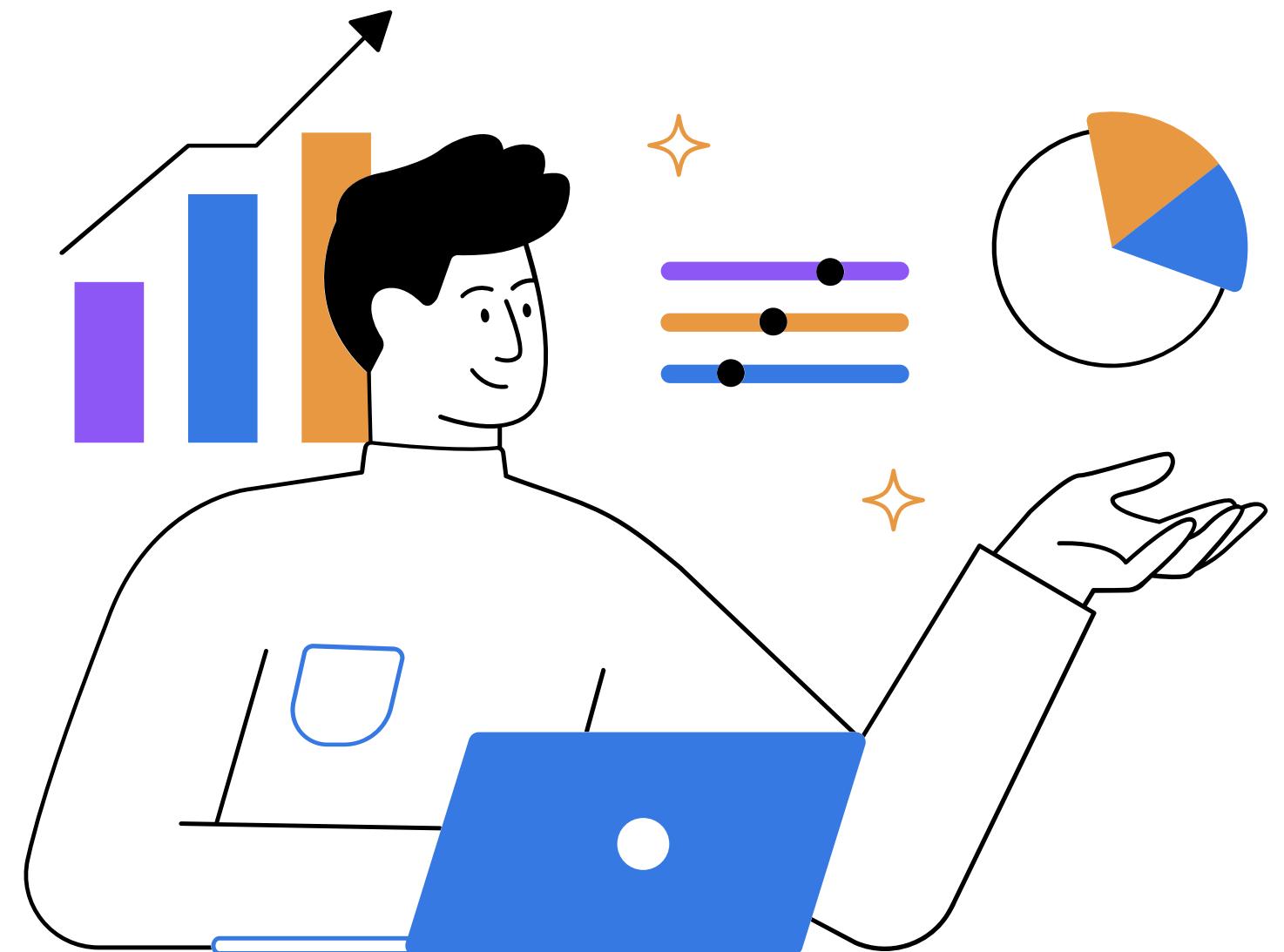
### Model Accuracy Comparison

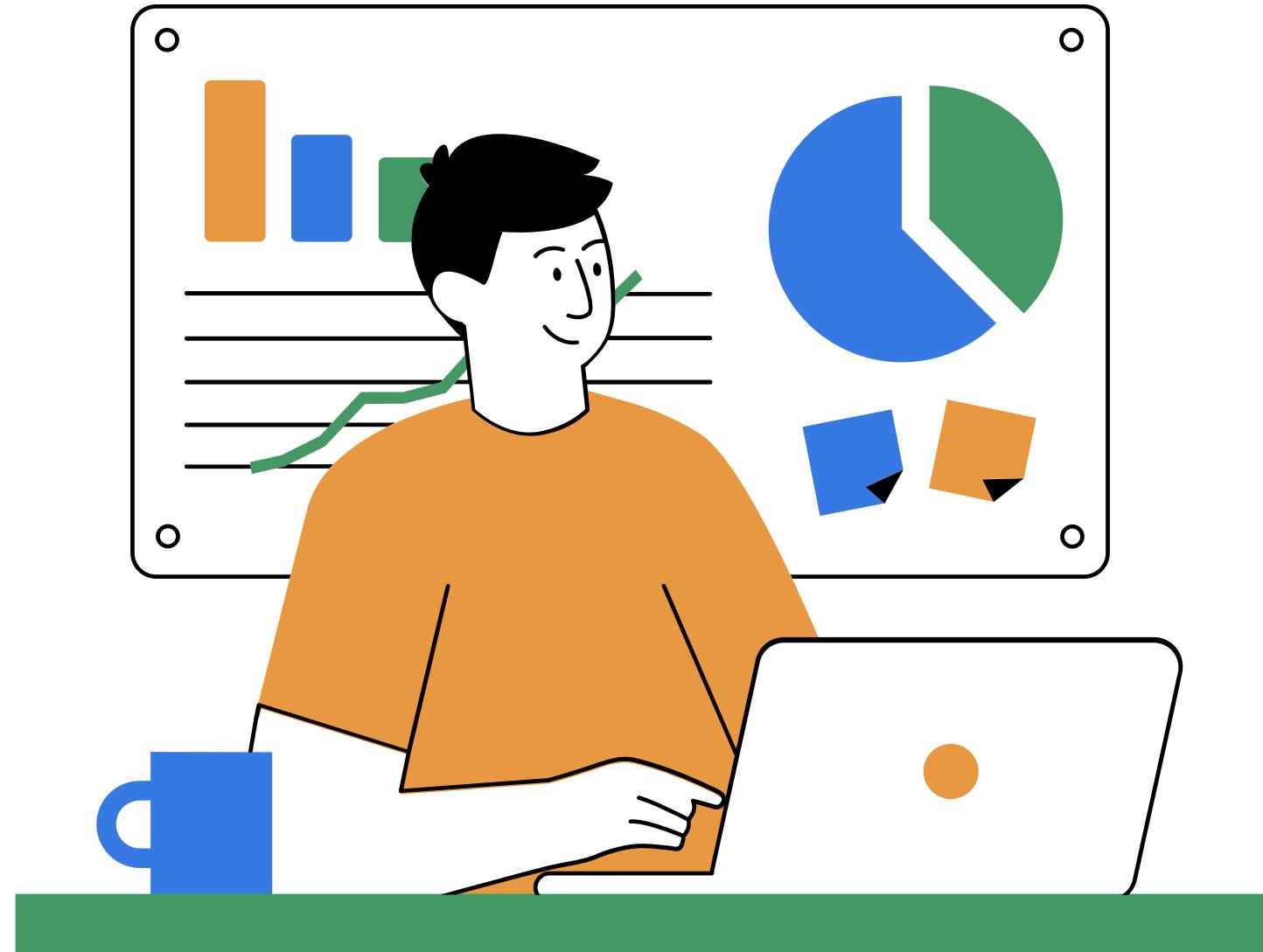
Model	Accuracy
Gradient Boosting	99.58%
Extra Trees	98.06%
Random Forest	97.09%
Logistic Regression	96.54%
SVM	94.18%

The **Gradient Boosting Classifier** emerged as the best-performing model. This model performed better because it effectively captures non-linear relationships between OCEAN personality traits and personality labels, while also focusing on hard-to-classify patterns through iterative error correction. Its ability to balance bias and variance made it particularly suitable for structured, feature-engineered data like InstaPer.

# Applications

- Digital Marketing & Brand Targeting: Brands can use InstaPer to understand audience personality types and deliver more personalised and effective marketing campaigns.
- Content Recommendation Systems: Personality-based insights can help recommend content aligned with user preferences, improving engagement and user experience.
- Social Media Analytics: The system provides deeper behavioural insights beyond likes and comments by analysing personality traits from hashtag usage.
- Psychological & Academic Research: InstaPer can assist researchers in studying online behaviour patterns and their correlation with personality traits using publicly available data.
- User Profiling & Personalisation: Platforms can leverage personality labels to enhance personalisation in feeds, notifications, and interaction strategies.

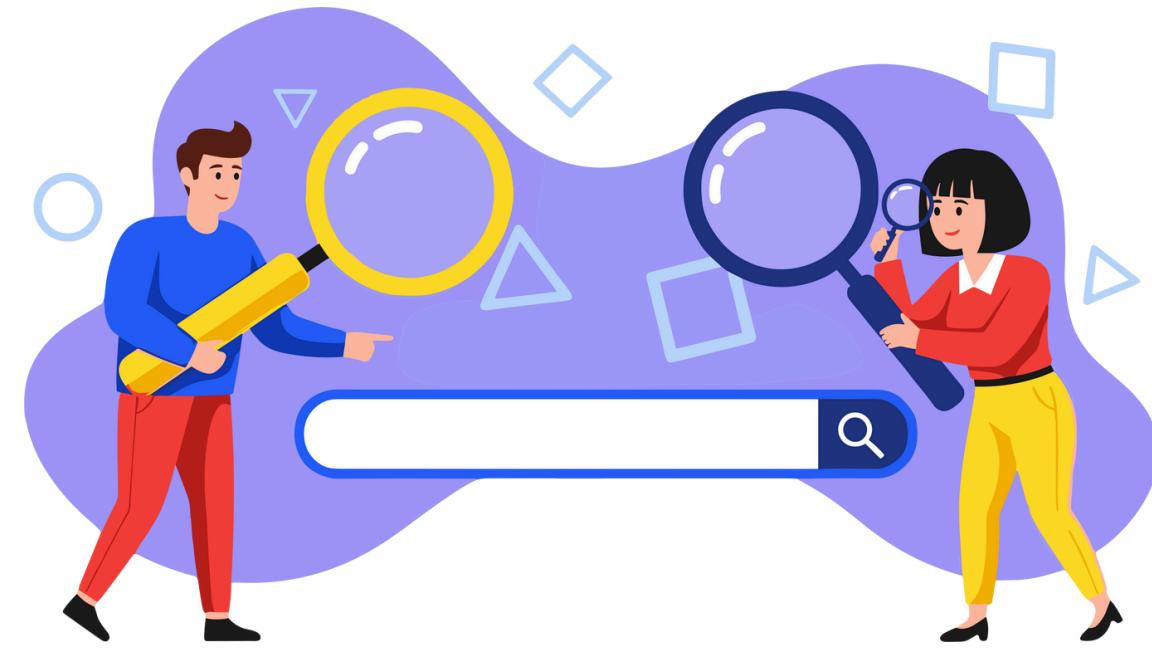




# Learnings

- Gained practical experience in web scraping and ethical data collection using automation tools.
- Learned how to convert unstructured social media text into structured features through NLP and categorization.
- Applied the Big Five (OCEAN) personality model to real-world social media data.
- Compared multiple machine learning models and evaluated them using performance metrics.
- Understood the importance of feature engineering in achieving high model accuracy.
- Developed an end-to-end ML pipeline, from data collection to prediction and evaluation.

# Tools & Technologies



1. **Jupyter Notebook:** Used as the primary development environment for data preprocessing, feature engineering, model training, and evaluation.
2. **Apify:** Utilised for automated web scraping to collect publicly available Instagram post and reel data in a structured format.
3. Machine Learning & NLP Frameworks
  - **Scikit-learn** – Implemented classification models such as Logistic Regression, Random Forest, Gradient Boosting, Extra Trees, and SVM.
  - **TextBlob**: Used for lightweight sentiment analysis of hashtags.
  - **Pandas** : Used for data cleaning, transformation, and feature construction.
4. Mentorship & Academic Guidance: Special guidance and conceptual clarity were provided by Mrs. **Aashima**, supporting the methodological and analytical aspects of the project.
5. Research & Presentation Tools
  - **ChatGPT** – Assisted in research understanding, model explanation, and documentation support.
  - **Canva** – Used for designing and structuring the final project presentation (PPT).

# Thank You

- [www.InstaPer.com](http://www.InstaPer.com)
- IEEE DataSet
- Kaggle DataSet

