# Artificial Intelligence-Powered Web Analytics Tool For University: Visitor Categorization & Behavioral Data Analysis For Strategic Outreach.

Mr. Agni Tanmaya Behera
*Dept. Of Computer Science And Engineering*
*GIET University,Gunupur,Odisha,India*
tanmayabehera@giet.edu

Dr. Raghvendra Agarwal
*Dept. Of Computer Science And Engineering*
*GIET University,Gunupur,Odisha,India*
raghvendra@giet.edu

***Abstract*—The rising concern on educational websites for student enrollment has boosted the understanding and importance of visitor behavioral patterns. Although Google Analytics (GA) tool is widely used to gather visitor information and interaction, its use in institutional and other organization contexts remains confined to explanatory data metrics offering little visitor's practical insights. The current study recommends one AI-Powered system which transforms unprocessed data towards forecasting, category oriented as well as advisory oriented for tactical policy-making. Collecting six-month data belonging to GA activity along with interaction records derived from different institutional websites, this system incorporates unsupervised clustering to detect unique visitor segments, supervised modeling to predict conversion probability plus chronological projection to anticipate traffic patterns surrounding related to essential institutional activities. A novel Interaction Quality Measure (IQM) is proposed in order to evaluate visitor interaction via merging session duration, browsing level, and activity completions into a unified metric. Results illustrate that AI-Powered GA study exposes deeper interactional trends compared with traditional GA reporting, facilitating targeted information refinement, location-based engagement prioritization, and enhanced conversion techniques. This proposed approach provides a scalable, data-secure system intended for educational organizations seeking to utilize web-based data insights in role of a strong tactical resource rather than inactive reporting tool.**

***Keywords*—Google Analytics, Visitor Categorization, Visitor Behavioral Insights, Machine Learning, Forecast Modeling, Interaction Quality Measure(IQM), Higher Education, Web Analytics.***

## I. INTRODUCTION

The organizational websites have developed starting from existing as one inactive information repository into a main connection point linking different organizations and their stakeholders. In higher education within a specific organization along with institutional websites remain commonly the primary stage of interaction intended for prospective learners, guardians as well as contributors. This ability to grasp how these visitors communicate with web content serves as critical for refining end-user experience, fine-tuning policy-making as well as accomplishing institutional targets such as enrollment expansion or curriculum promotion [1].

Google Analytics (GA) remains among the highly used tools intended for monitoring and reporting website activity

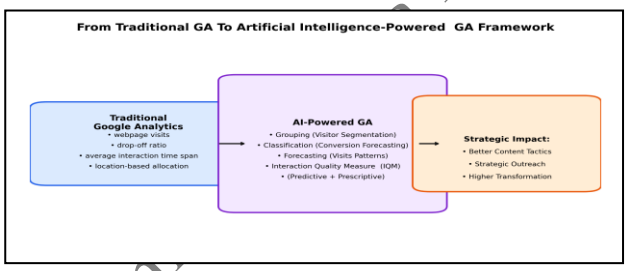


Fig 1. Proposed AI-Powered Web Analytics Tool.

metrics. This provides a variety consisting of parameters and attributes such as webpage visits, drop-off ratio, average interaction time span and location-based allocation—which authorize administrators in order to track website effectiveness. Within educational contexts, earlier analyses have applied GA to recognize frequently visited pages, monitor changes in visitor activity across period, and evaluate the outcome from happenings such as enrollment notices. However, the majority of such implementations, persist explanatory, delivering an outline about the thing that already taken place instead of actionable forecasting or customized tactical recommendations [2].

This restricted range causes significant unexplored opportunities within GA platform datasets. Current computational data-learning techniques are able to enhance Google Analytics' capabilities via exposing unseen interactional patterns forecasting prospective visitors' activities as well as segmenting viewers for specific interaction. For example, clustering-based techniques have the capacity to cluster visitors based on mutual activity-based traits, allowing material customization categorization, models can predict the probability of important activities like form submission and time-series projection has the ability to visit fluctuations near significant occasions. Such anticipated Artificial Intelligence-Powered improvements transform data insights out of one inactive reporting tool towards a dynamic decision-assisting system [3].

Despite the widespread use of Google Analytics within education, there exists a noticeable gap within investigations which combines cutting-edge Artificial Intelligence techniques into GA-centered visitor analysis aimed at university-level education contexts. Previous research, which is examined as part of the core analysis mainly focused on overall statistics, metrices as well as simple patterns without executing anticipatory or

recommendation-based simulations. This results in admins lacking more in-detailed understandings, which is important to make information-driven measures within actual timeframe [4].

This current study deals with the gap to fulfill through developing a single AI-Powered data insights system which performs based on extracted GA datasets. This suggested technique combines autonomous grouping to identify different visitor categories, trained categorization in order to predict change likelihood and chronological systems in order to forecast visit behavior. In addition, this paper represents a novel Interaction Quality Measure (IQM), integrating with session level, interaction time span as well as critical event activity combines within a single integrated benchmark. This system is evaluated on long-term GA data collections extracted from departmental websites, accompanied by results compared against standard GA monitoring. By integrating different classical analysis with AI techniques, the research wants to supply a single dynamic confidential-aligned model which has the capability to influence meaningful policy-setting decisions within higher-level education as well as beyond [5].

## II. BACKGROUND AND RELATED WORK

*1) Importance of GA in context of Digital World:* Several research concludes google analytics plays an important role in *digital world for* reporting user interaction. The study by Dimitris C. Gkikas and Prokopis K. Theodoridis (2024) explain about the advance analytics and machine learning prediction to engage more customers towards websites and perform*s* informed decisions for making digital market (e.g. e-commerce) more effective and profitable [6]. Roger S. Mission (2023) conducted a comparative study of visitor flows and interaction metrices which says in context of University of Antique (UA), a well known state university in Philippines.The study provides insights for enhancement of web user interface, content relevance, user experience to achieve conversion goals on basics of GA dataset [7]. The next research followed by Bernard J. Jansen, Soon-gyoJung and JoniSalminen (2022) compares google analytics metrices, collected data from 86 websites from 26 countries and 19 industries related to different web services which given results close to same analytics report generated by Google analytics(GA), for some limited attributes like total visits and unique visits . As a result, similar web give *20% lower* (total visits), *40% lower* (unique visits), *25% higher* (bounce rates) and *50% higher* (average session) and these four metrices are commonly co-related to GA [8]. Szu-Chieh Chen, Thomas Chang-Yao Tsao, Ko-Huang Lue and Yafang Tsai (2020) performed a pilot study based enrollment strategies on Medical University, this experiment focuses on students of higher education's visiting departmental website, for analyzing behavioral patterns to choose better institution or departments due to low fertility rate in Taiwan [9].

*2) Drawbacks Of Traditional Google Analytics:* As Google Analytics is widely used worldwide, but it has limitation in context of academics and higher educational organization. Google Analytics is recognized for providing descriptive insights which inherit lacuna in strategic outreach & policy making, but can be achieved by modern AI/ML techniques like: classification, clustering and prediction. For example, Dimitris C. Gkikas and Prokopis K. Theodoridis (2024) has studied on commercial market rather than in context of university, and also have no time-series based visitor forecasting [10]. According to Roger S. Mission (2023), his research is focused more on user content relevance rather than have no advanced predictive modelling [11]. Similarly Bernard J. Jansen, Soon-gyoJung and JoniSalminen (2022) researched on GA and similar web tool but lacking predictive modeling, domain-specific system and major disadvantage is having no actionable insights [12]. Szu-Chieh Chen, Thomas Chang-Yao Tsao, Ko-Huang Lue and Yafang Tsai (2020) researched about enrollment strategies but limted to descriptive dashboards without future planning [13]. This above examples unable to achieve time-series forecasting, visitor categorization, conversion estimation, as making it a weaker decision support system.

*3) Artificial Intelligence in Web Analytics:* Artificial Intelligence is making revolution across multiple industries in the world, mainly to achieve human intelligence mimic system. It plays an transformative role in web analytics by collecting , processing , analzing and interpreting large number of data in real time.

*4) Comparative Studies of Analytics Tools:*

*5) Interaction Quality Measure Approaches:*

*6) Research Gap Identification and Novelty:*

## III. METHODOLOGY

*1) Data Sources*: The dataset we have used in this research was primarily obtained by a custom made web-based visitor analytics application (https://csa-giet.onrender.com/) hosted on the GIET University, Gunupur ( Dept. Of Computer Science And Engineering) web server. This proposed system tracks visitor attributes such as Ip address, browser, os, device_category, source, session-duration, pages per session, click events, geolocation and referral source in real-time and store them in MongoDB database.This timestamp attribute helps in time-series forecasting of traffic volume as well as conversion label supports supervised learning tasks. All above these attributes contributes to behavioral insights [14].

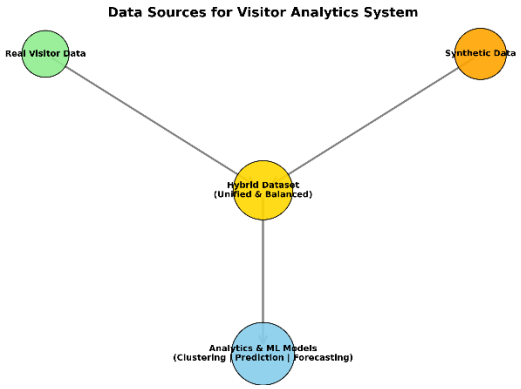


Fig 2: Hybrid data collected from different sources.

Since data volume we generated was intially insufficient, we supplemented it with synthetically created visitors sessions which follow the same schema and data models. The hybrid technique powered us to maintain data authenticity and integrity, ensuring sufficient sample datasets for robust machine learning. The below ( figure 3) shows the data model architecture for visitors flows into system [15].

```
// models/Visitor.js
const mongoose = require('mongoose');

const visitorSchema = new mongoose.Schema({
  visitorId: String,
  ip: String,
  browser: String,
  os: String,
  device_category: String,
  source: { type: String, default: "Direct" },
  latitude: Number,
  longitude: Number,
  sessionDuration: Number,
  pagesPerSession: Number,
  events: Number,
  converted: { type: Boolean, default: false },
  timestamp: { type: Date, default: Date.now }
});

module.exports = mongoose.model('Visitor', visitorSchema);
```

Fig 3 : Data Model Schema used for Visitor-analytics.

*2) Preprocessing:* The raw data we have collected, performed with column mapping techniques on different attributes like session duration, pages per session, click events, conversion label, timestamp, device_category, geolocation to normalize the raw data collected.If any essential columns is not detected, then we have implemented advance logic to attempt fallback i.e it will, search “duration” if “session-duration” is not found as well

as if "event" is not found then logic will search for "clicks" or "key-events" etc…In data cleaning we followed some steps : renaming columns into standardized names of inconsistent data, parsing timestamp with datetime format and converts three key metrices ("session_duration", "pages_per_session","events") to numeric, filling NaN values with 0.0. Categorical column such as device_category and source are normalized with string matching techniques. So, above all steps taken in data preprocessing [16].

*3) Interaction Quality Measure (IQM):* In this research, we have introduced a novel Interaction Quality Measure (IQM) is proposed to combine multiple behavioral insights into a single unified metric.

3.1) Robust Maximal Technique ( 95th percentile) : Let interaction raw merices are d (session_duration),p (pages_per_session) & e (events). The following normalized components are :

$$\tilde{d}=\frac{\min(d, D_{max})}{D_{max}},\ \tilde{p}=\frac{\min(p, P_{max})}{P_{max}},\ \tilde{e}=\frac{\min(e, E_{max})}{E_{max}} \quad (1)$$

▪ To Find the position of 95th percentile (Exclusive):

$$P=\frac{(n+1)\times k}{100} \quad (2)$$

▪ To Find the position of 95th percentile (Inclusive):

$$P=(n-1)\times q+1 \quad (3)$$

▪ General Interpolation Value, for sorted data :

$$P=x_i+f(x_{i+1}-x_i) \quad (4)$$

▪ Choose $D_{max}, P_{max}, E_{max}$ as robust maximal technique (e.g. 95th percentile ), where :

$$\alpha+\beta+\gamma=1 \quad (5)$$

▪ The IQM is Defined as:

$$IQM=\alpha\,\tilde{d}+\beta\,\tilde{p}+\gamma\,\tilde{e} \quad (6)$$

Categorization of IQM threshold for labels :-

- **Low**, if IQM < 0.33
- **Medium**, if 0.33 ≤ IQM < 0.66 (7)
- **High**, if IQM ≥ 0.66

In this study, we approach a novel technique called IQM (Interaction Quality Measure) that combines multiple behavioral patterns into a single metric. The raw metrices session duration ($\tilde{d}$), pages_per_session ($\tilde{p}$) and events ($\tilde{e}$) are normalized between 0 (lowest_interaction_rate) to 1 (highest_interaction_rate) using Robust Maximal Technique (95th percentile) so that extreme outliers are unable to disbalance the scale. The normalization technique is defined in equation (1) : $\tilde{d}=\min(d, D_{max})/D_{max}$ , $\tilde{p}=\min(p, P_{max})/P_{max}$ and $\tilde{e}=\min(e, E_{max})/E_{max}$ ,where $D_{max}, P_{max}$ and $E_{max}$ are chosen as robust maxima(95th percentile) are calculated using equation (2) for exclusive and equation(3) for inclusive, or general formula (equation 4) to find interpolation for sorted data. The IQM (Interaction Quality Measure) in equation (5), is decided on basis on three interactions raw metrices [for example: session_duration (40%), pages_per_session (30%), events (30%)] ,they must add to give result equal to 1 (100%) i.e. in equation (6). In equation (7), this IQM is categorized on basics of threshold i.e. low ( IQM < 0.33) , medium (0.33 ≤ IQM< 0.66) and high (IQM ≥ 0.66).

3.2) Rank Percentile Technique (Alternative) : This is the alternative to robust maximal technique, instead of max cap, we compute rank percentile [0-1]... for individual metric across the dataset as follows :

$$\tilde{d}_{rank}=\frac{rank(d)-1}{N-1},\ \tilde{p}_{rank}=\frac{rank(p)-1}{N-1},\ \tilde{e}_{rank}=\frac{rank(e)-1}{N-1} \quad (8)$$

Hence the IQM is defined as:

$$IQM=\alpha\,\tilde{d}_{rank}+\beta\,\tilde{p}_{rank}+\gamma\,\tilde{e}_{rank} \quad (9)$$

When to use both techniques :

- Robust Maximal Technique : Here values are scaled with respecct to 95th percentile of the distribution which helps to exclude extreme outliers.This technique must be used when data is moderately distributed with some outliers(e.g. session_duration).

- Rank Percentile Technique : Here each metric value is assigned with its percentile rank (ordered manner). This technique must be used when data is highly skewed or when relative order matters more than exact values.

Method 1 : Normalizing Three Metrices using Robust maximal Technique (95%) :

| Sl. | d(s) | p | e | d_rm | p_rm | e_rm | IQM_rm |
|---|---|---|---|---|---|---|---|
| 1 | 45 | 1 | 0 | 0.0143 | 0.0304 | 0 | 0.0148 |
| 2 | 120 | 3 | 1 | 0.0381 | 0.0911 | 0.0498 | 0.0439 |
| 3 | 300 | 5 | 2 | 0.0951 | 0.152 | 0.0995 | 0.1106 |
| 4 | 600 | 8 | 5 | 0.1901 | 0.2432 | 0.2488 | 0.2119 |
| 5 | 30 | 1 | 0 | 0.0095 | 0.0304 | 0 | 0.0103 |
| 6 | 900 | 12 | 8 | 0.2855 | 0.3649 | 0.398 | 0.3238 |
| 7 | 50 | 2 | 0 | 0.0159 | 0.0608 | 0 | 0.0194 |
| 8 | 200 | 4 | 1 | 0.0634 | 0.1217 | 0.0498 | 0.0781 |
| 9 | 400 | 6 | 3 | 0.1268 | 0.1822 | 0.1493 | 0.1502 |
| 10 | 5000 | 50 | 30 | 1 | 1 | 1 | 1 |

Table 1 : Represents the interaction metrices capped at its 95th Percentile.

Method 2: Normalizing Three Metrices ( $\tilde{d}$, $\tilde{p}$, $\tilde{e}$) using Rank Percentile Technique :

| Sl. | d(s) | p | e | d_rp | p_rp | e_rp | IQM_rp |
|---|---|---|---|---|---|---|---|
| 1 | 45 | 1 | 0 | 0.1111 | 0.0556 | 0.1111 | 0.0944 |
| 2 | 120 | 3 | 1 | 0.2222 | 0.1667 | 0.2222 | 0.2056 |
| 3 | 300 | 5 | 2 | 0.4444 | 0.2778 | 0.3333 | 0.3933 |
| 4 | 600 | 8 | 5 | 0.5556 | 0.4444 | 0.5556 | 0.5306 |
| 5 | 30 | 1 | 0 | 0 | 0.0556 | 0.1111 | 0.0500 |
| 6 | 900 | 12 | 8 | 0.7778 | 0.6667 | 0.7778 | 0.74 |
| 7 | 50 | 2 | 0 | 0.1667 | 0.1111 | 0.1111 | 0.14 |
| 8 | 200 | 4 | 1 | 0.3333 | 0.2222 | 0.2222 | 0.2978 |
| 9 | 400 | 6 | 3 | 0.6667 | 0.6667 | 0.6667 | 0.6667 |
| 10 | 5000 | 50 | 30 | 1 | 1 | 1 | 1 |

Table 2: Represents the interaction metrices scaled on their respective rank.

3.3) Comparative Analysis : From Table (1) and Table (2), we observed that the robust maximal technique scales metrices with respect to 95$^{th}$ percentile, making it more effective in normalizing outliers. However, the rank percentile technique helps in distribution-free normalization in highly skewed datasets. In both methods, IQM scores is generated in between [0-1] range, but rank methods focuses on rank over absolute values. And one more advantage of rank-based method is that it acts as an "analytics tool" for sensitive analysis because highly skewed data unable to affect it .

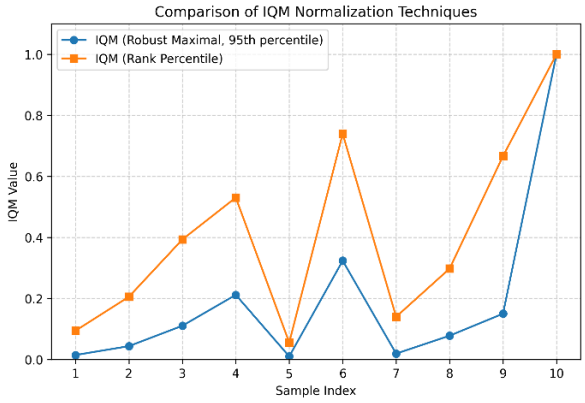


Fig 4: Comparison of IQM Normalization Technique.

3.4) Importance Of Normalization Technique : This above two techniques (Method 1) and (Method 2) validate IQM, which provides a normalized interaction score suitable for machine learning tasks. Beyond descriptive analytics, IQM plays as an engineered feature that enhance clustering, prediction and forecasting which convert analytics into prescriptive intelligence, making a unique contribution to AI-powered web analytics.

*4) System Architecture And Algorithm Overview :* The proposed Artificial Intelligence-Powered Web Analytics Tool For University follows a layered pipeline architecture (Fig. 5), designed to collect, clean, analyse and interpret visitor data,while ensuring data protection enabling advanced decision-support through artificial intelligence.

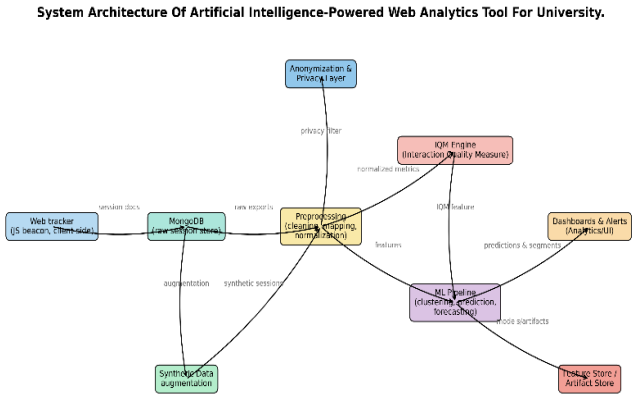


Fig 5: High Level System Architecture of AI-Powered Web Analytics Tool, which creates a pipeline integrating real-time data, preprocessing, feature engineering (IQM), ML analytics, and data visualization dashboard secured by privacy guards.

▪ To activate this above proposed system, enables raw visitor interaction data into actionable insights through a series of systematic steps are as follows :

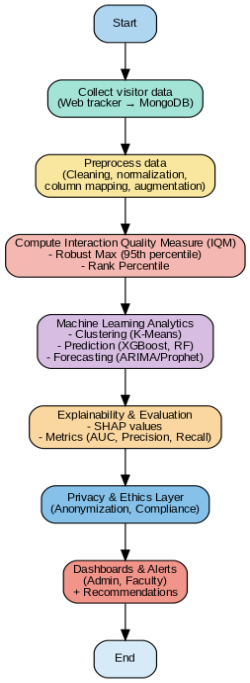


Fig 6: Flowchart of the AI-powered Web Analytics Tool .

Computational Complexity Note :

▪ This above pipeline is efficient, with preprocessing scaling linearly in data size **O(n)**, clustering **O(n*k*i)**, supervised learning **O(t.f.log n)** and forecasting **O(T),** enabling real-time feasibility for large volumed visitor datasets.

4.1) Data Acquisition Layer : The data acquisition Layer acts as an entry point of the proposed AI-Powered web analytics tool. Its primary work is to capture, collect and store raw visitor information across the university's departmental web ecosystem in a structured format, immediately ready for preprocessing and analysis. This layer is integration of both front-end tracking and back-end storage mechanisms to ensure high coverage, real-time responsiveness, and robustness against unstructured data.

▪ There are different components in data acquisition layer are as follows :

- Web Tracker (Client-Side) : It is implemented using Javascript beacon code which is embedded in university's (departmental) websites or web applications. It captures ultrafine grained visitor activities like session [start-end], pageviews and navigations, events (buttons, downloads, links), device/browser fingerprint (OS, device_type, browser_version) and geolocation (through browser APIs, if permitted).Then data packets are sent in JSON format to backend and stored in database.

- MongoDB (Raw session Store) :

4.2) Preprocessing Layer :

- Column Mapping and Data Cleaning :
- Normalization & Encoding :
- Hybrid Data Augmentation :

4.3) Feature Engineering Layer (IQM) :

- Robust Maximal Technique (95$^{th}$ Percentile) :
- Rank_Percentile Technique :

4.4) Machine Learning And Analytics Layer :

- Clustering(K-Means) :
- Supervised Learning(XGBoost, Random Forest) :
- Time-Series Forecasting (ARIMA, Prophet) :
- Explainability (SHAP) :

4.5) Privacy & Ethics Layer :

- Anonymization :
- Consent & Compliance :

4.6) Visualization & Decision Layer :

- Dashboards (Admin & Faculty) :
- Alerts & Recommendations :
- Feature/Artifact Store :

*5) Evaluation Protocols And Metrices:*

*6) Ethical Considerations And Data Privacy:*

## IV. EXPERIMENTS AND RESULTS

*1) EQI Validity Analysis:*

*2) Visitor Segmentation Results:*

*3) Conversion Prediction Results:*

*4) Time-Series Forecasting Results:*

*5) Ablation & Robustness Studies:*

*6) Practical Impact Evaluation (A/B or Simulation):*

## V. DISCUSSION

## VI. CONCLUSION & FUTURE WORK

## VII. REFERENCES