



A Hybrid Recommendation System For University Selection Using Machine Learning

Harsh Motiramani*, Vedaant Melkari, Malav Mehta

Student, SVKM's NMIMS Mukesh Patel School of Technology Management and Engineering, Mumbai, MH, India

motiramani23@gmail.com

Abstract. It is important to consider multiple factors, including academic fit, finances, and personal preferences, while deciding to select a university. For this, we propose a hybrid recommendation system that uses Machine Learning (ML) and Natural Language Processing (NLP). Using acceptance rates, tuition fees, GPA, and GRE scores, our system predicts a match score using XGBoost. Cosine similarity is applied to evaluate unstructured data, including course descriptions, to align students' interests with universities. Institutional reputation is assessed through sentiment analysis of reviews, for which a comparison tool allows side-by-side evaluations. An application assistance module monitors all deadlines, and tailored Statements of Purpose (SOPS) are generated through template-based NLP frameworks. The proposed system offers a personalised approach by providing recommendations through the integration of quantitative and qualitative data. The operational efficacy of the system has been tested and proves to aid students in making informed decisions. This work extends the development of AI educational tools by creating a scalable system for incorporating university recommendation systems.

Keywords: Match Score, XGBoost, Regression, Cosine Similarity, Hybrid Model

1 Introduction

The transition from secondary to tertiary education represents one of the most consequential decisions in a student's academic and professional trajectory. In an increasingly globalised higher education landscape, prospective students face a daunting array of choices, with over 25,000 universities worldwide offering vastly different academic programs, admission requirements, financial structures, and post-graduation outcomes (QS World University Rankings, 2023). This decision-making complexity is compounded by several critical factors: the need to align institutional offerings with individual academic capabilities and career aspirations, financial considerations that may constrain options, and geographical preferences that influence quality of life and future opportunities.

Traditional university selection methods exhibit significant limitations that our research seeks to address. Conventional approaches typically rely on:

Static ranking systems (e.g., QS, THE, ARWU) that employ generalized metrics ill-suited to individual student profiles, Manual search platforms requiring extensive user-initiated filtering Advisory services that are either cost-prohibitive or subject to human

bias these methods fail to account for the multidimensional nature of university selection, particularly:

The complex interplay between a student's academic profile and institutional admission requirements

The semantic matching between student interests and program curricula

The subjective evaluation of institutional quality through student experiences

The dynamic nature of application processes and deadlines

To address these challenges, we developed an intelligent university recommendation system combining machine learning and NLP techniques. Our solution integrates three key components: (1) a hybrid recommendation engine using XGBoost for academic matching and neural embeddings for curriculum alignment, (2) a dynamic knowledge base with real-time admission data and student feedback, and (3) practical tools including deadline tracking and AI-assisted document generation.

1.1 Research Outcomes:

Choosing the right university is one of the most pivotal decisions in a student's life - one that shapes their academic journey, career prospects, and personal growth. Yet, this critical decision remains fraught with challenges, anxieties, and systemic inequalities that leave many students feeling lost in a sea of overwhelming choices.

Today's students face a perfect storm of decision-making challenges:

- The paralyzing paradox of choice among thousands of institutions worldwide, each with complex admission landscapes
- The heartbreak of discovering mismatches only after enrollment - whether academic, financial, or cultural
- The stark inequality in access to quality counseling, leaving underprivileged students to navigate this maze alone
- The stress of application processes that feel like a high-stakes puzzle with missing pieces

Traditional solutions fall painfully short. Static ranking systems reduce vibrant academic communities to cold numbers. Overworked counselors struggle to provide personalized guidance. Well-intentioned but generic advice leaves students wondering: "But what's right for ME?"

This is where our AI-powered university recommendation system steps in - not as a replacement for human judgment, but as an empowering co-pilot for every student's journey. By combining the precision of machine learning with the nuance of natural language processing, we're creating something revolutionary:

1. Your Academic Matchmaker - Going beyond test scores to understand your unique learning style, aspirations, and potential through sophisticated XGBoost algorithms
2. Your Curriculum Detective - Using advanced NLP to read between the lines of course catalogs, matching your intellectual passions with programs that will set them aflame

3. Your Reality Check - Balancing dreams with practicalities through dynamic filtering of costs, locations, and success outcomes
4. Your Application Ally - From deadline reminders to AI-crafted personal statements, transforming anxiety into confidence
5. Your Equalizing Force - Democratizing access to the kind of personalized guidance that was once only available to the privileged few

We're not just building a tool - we're challenging the status quo of educational access. Because every student deserves more than a random roll of the dice when choosing where they'll grow, struggle, and ultimately flourish. This is about turning the overwhelming into the empowering, the unequal into the equitable, and the stressful into the transformative.

The university selection process hasn't changed in decades - it's time for a revolution that puts students, their dreams, and their unique potential at the center. That revolution starts here.

1.2 Motivation and Scope of Project Report:

Choosing the right university and writing a strong Statement of Purpose (SOP) can be a daunting task for students. With so many options available, it's easy to feel overwhelmed by factors like rankings, costs, and acceptance rates. This project is designed to make the process smoother and more personalized. By using advanced machine learning techniques, we help students find universities that truly match their profiles and interests. Our system combines structured data, like GPA and acceptance rates, with natural language processing to align students with the best courses. We also simplify SOP writing with AI-generated, well-structured drafts tailored to each university. Ultimately, this project is about making the application journey less stressful and more effective, giving students the tools they need to make informed decisions and put their best foot forward.

1.3 Salient Contribution

This project presents a comprehensive, data-driven solution to the complex challenge of university selection through the development of a hybrid recommendation system. By leveraging XGBoost, a powerful machine learning algorithm, the system computes personalized match scores based on key academic parameters such as GPA, GRE scores, tuition costs, and acceptance rates. It introduces innovative features such as dynamic filtering, real-time university comparison, and a user-friendly interface that simplifies the decision-making process. Additionally, the platform incorporates tools for deadline tracking and application management, offering an end-to-end support system for students navigating the admission journey. Unlike traditional approaches, this system delivers personalized recommendations that are both scalable and grounded in actual user profiles, thus democratizing access to high-quality academic guidance.

1.4 Organisation of Report

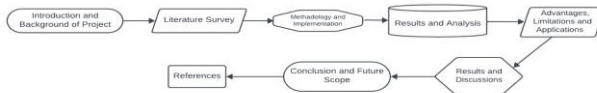


Fig.1. Organisation of Report

2 Literature Survey:

Selecting the right university is a pivotal decision for students, influencing their academic trajectory and future career prospects. The abundance of higher education institutions, each with unique programs, admission criteria, and cultural environments, makes this choice increasingly complex. Traditional methods—such as consulting static rankings, attending educational fairs, or seeking advice from counselors—often fall short in providing personalized guidance tailored to individual student profiles. In response to this challenge, the integration of machine learning (ML) into recommendation systems offers a promising solution. By analyzing vast datasets encompassing academic records, test scores, financial considerations, and institutional attributes, ML-driven systems can provide customized university recommendations. This project aims to develop a hybrid recommendation system that leverages ML techniques to assist students in making informed decisions aligned with their academic goals and personal preferences.

2.1 Exhaustive Literature Survey

Over the past decade, significant research has been devoted to enhancing educational decision-making through the use of recommender systems. These systems borrow heavily from fields such as e-commerce and social networking, adapting techniques like collaborative filtering, content-based filtering, and hybrid models to the educational domain.

1. **Hybrid College Recommendation Models**The study by Habib et al. [1] explores a college recommendation framework using a hybrid model that combines collaborative filtering with content-based algorithms. It highlights the potential of using APIs such as LinkedIn for profile enrichment and supports modular design across disciplines like engineering, medicine, and pharmacy. The hybrid approach is found to outperform standalone models by leveraging the strengths of both collaborative user behavior and structured academic data.

2. **Adaptive Learning and Personalization**An earlier work by Brusilovsky and Millán [2] discusses the growing importance of adaptive recommendation systems in online education environments. It stresses the need for dynamic learning paths that respond to individual student behaviors, a principle that aligns with our project's aim of providing dynamic and personalized university suggestions.

3. **Deep Learning for Improved Recommendations**In [3], a hybrid recommender system utilizing deep learning and collaborative filtering was proposed to help students in entrepreneurial project selection. The integration of neural networks improved the system's ability to predict user preferences, resulting in a 15% increase in accuracy and 20% improvement in personalization. This shows the merit of deep learning techniques in educational contexts, particularly where personalization is key.

4. **AI in Academic Counseling**The work by Hossain et al. [4] delves into AI-powered academic counseling systems and underscores their ability to supplement traditional career guidance. ML algorithms are shown to enhance accuracy and reduce counselor workload, especially when processing large datasets like academic histories, test scores, and student goals.

5. **A Comprehensive Review of Recommender Systems**Adomavicius and Tuzhilin [5] provide a foundational perspective on the evolution of recommendation systems, covering advancements in algorithms, hybrid models, and applications across domains. Their insights into limitations such as the cold-start problem and real-time adaptability are critical for understanding the trade-offs in system design.

6. **NLP-Based Content Analysis for Recommendations**Though not part of the core model in this report, the relevance of natural language processing (NLP) for text-based matching is echoed in [6], which focuses on AI-generated content detection using deep learning transformers. While the study centers on misinformation, its application of NLP techniques such as Transformer models and text classification has implications for academic program matching and document analysis.

7. **Application of TF-IDF and Cosine Similarity**The studies in [7] and [8] apply TF-IDF and cosine similarity methods to movie and product recommendations, respectively. These content-based approaches provide a foundation for semantic similarity measures, which, though not emphasized in this system, inform future work on aligning student interests with course descriptions in unstructured formats.

8. **Use of XGBoost in Recommender Systems**In [9], XGBoost is employed for product recommendation in an e-commerce setting. The study demonstrates its superiority over Random Forest and Gradient Boost methods in terms of accuracy and performance. These findings directly support the use of XGBoost in our project for predicting match scores based on structured academic data.

2.2 Summary

The reviewed literature highlights the growing interest and demonstrated success of applying machine learning, hybrid modeling, and recommendation techniques in educational contexts. From deep learning-enhanced personalization to scalable filtering mechanisms, these studies collectively provide a strong foundation for the hybrid university recommendation system proposed in this project. By building on these insights and addressing existing system limitations — particularly around personalization, accessibility, and real-time data integration — this project contributes a novel solution aimed at democratizing academic guidance through intelligent automation.

3 Methodology and Implementation

This chapter presents an in-depth overview of the university recommendation system, which leverages advanced machine learning and natural language processing (NLP) techniques to generate highly personalized recommendations. The system comprises multiple stages, including data preprocessing, model training, evaluation, and deployment, integrating both structured and unstructured data sources to enhance accuracy and relevance.

Core Features

1. **Personalized Recommendations:** The system dynamically predicts suitable universities based on academic metrics, preferences, and historical data.

2. **Dynamic Filtering:** Implements multi-criteria filtering based on tuition costs, geographical location, and institutional ranking.
3. **Sentiment Analysis:** Extracts insights from student reviews to refine recommendation rankings.
4. **Comparison Tool:** Provides a structured comparison framework for evaluating universities based on key parameters.
5. **Application Assistance:** Includes deadline tracking and a guided checklist to streamline the application process.
6. **SOP Generation:** Utilizes template-based NLP models to generate structured and coherent Statements of Purpose (SOPs) tailored for university applications.

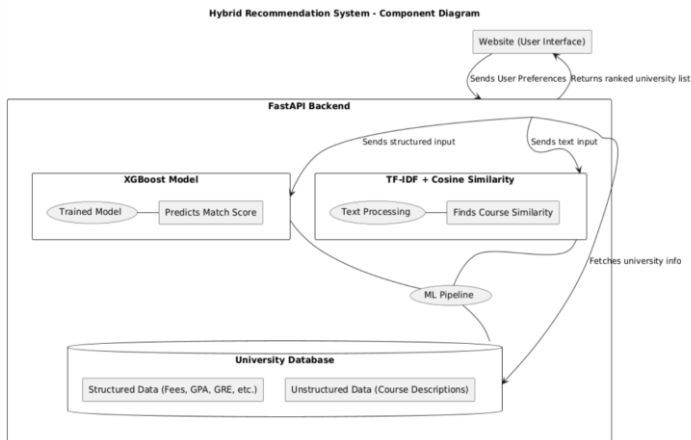


Fig .2. System Working

3.1 Data Preparation:

Data preprocessing is a critical step in ensuring high-quality inputs for model training. The dataset used in this study comprises comprehensive information on 6,600 universities, combining both structured and unstructured data sources to power a hybrid recommendation system.

Structured Data

The structured numerical attributes include:

- University Name, City, State, state_name, Acceptance Rate, In-State Tuition, Out-of-State Tuition, Cost of Attendance, Graduation Rate, Median Salary, Average GPA, Average GRE Score, 2025 Rank, 2024 Rank, Location, Location Full, Size, Academic Reputation, Employer Reputation, Faculty Student, Citations per Faculty, International Faculty, International Students, International Research Network, Employment Outcomes, Sustainability, QS Overall Score, MS Courses Offered.

Unstructured Data

- course_vector, course_similarity derived from textual data such as course descriptions and student reviews.

Key Preprocessing Steps

- **Data Cleaning:**
 - Missing values were imputed using statistical or domain-informed methods.

- Numerical columns were standardized and normalized using Z-score and Min-Max scaling to reduce variance-related bias and ensure consistency in model input.

- Text Vectorization:

- Applied TF-IDF (Term Frequency–Inverse Document Frequency) transformation to convert course descriptions and student reviews into numerical vectors. These were used to compute a course_similarity_score for personalized matching.

- Feature Engineering:

- Constructed new attributes to enrich the dataset’s predictive power:

- Acceptance Probability Scores: Modeled based on institutional acceptance rate and profile strength.

- Cost-Benefit Ratios: Defined as the ratio of median salary to cost of attendance to reflect post-graduation ROI.

- Profile Match Scores: Created using GPA and GRE match formulas to estimate how closely a student's profile aligns with a university’s standards.

GPA Match Score: $\text{gpa_score} = \min(\text{user_gpa} / \text{university_avg_gpa}, 1.0) * 100 /$

Final Match Score (Weighted Formula) = $\text{final_match_score} = 0.4 * \text{gpa_score} + 0.4 * \text{gre_score} + 0.2 * \text{course_similarity}$

These preprocessing steps ensured the dataset was clean, well-structured, and ready for model training, significantly improving both the interpretability and accuracy of the recommendation outcomes.

3.2. Model Architecture:

The system integrates multiple machine learning models optimized for structured and unstructured data processing:

3.2.1. XGBoost for Match Score Prediction

- Implements a gradient boosting framework to model non-linear relationships in structured data.

- Trained on multi-dimensional university attributes, including acceptance rates, tuition fees, GPA, and standardized test scores.

- Employs hyperparameter tuning (e.g., learning rate, tree depth, regularization) to optimize predictive accuracy.

- Outputs a probabilistic match score, indicating the likelihood of student admission.

3.2.2. Fast Text Embedding & Cosine Similarity for Course Matching

To enhance semantic understanding in course-based recommendations, the College Compass system leverages FastText embeddings combined with cosine similarity. This approach enables deeper contextual matching between student interests and university course offerings.

Key components include:

- FastText Embeddings:

Utilizes pre-trained FastText word vectors to generate dense, contextualized representations of course descriptions. Unlike TF-IDF, FastText captures subword

information and semantic relationships between terms, allowing for more nuanced understanding of academic content.

- **Cosine** Similarity:
Measures the semantic closeness between user-provided course interests and university course vectors. This metric quantifies how well a university's offerings align with a student's academic focus.

- **Enhanced Recommendation Layer:**
Course similarity scores derived from FastText and cosine similarity are integrated with other profile-based features (e.g., GPA, GRE, tuition) to boost overall recommendation accuracy, ensuring both numerical and semantic factors are considered.

3.2.3. Hybrid Recommendation Model

- Integrates structured (XGBoost) and unstructured (TF-IDF + Cosine Similarity) data sources.

- Implements a weighted scoring mechanism:

- Match Score (XGBoost output) $\rightarrow \alpha$ weight

- Interest Alignment Score (TF-IDF Similarity) $\rightarrow \beta$ weight

- Enhances adaptability by considering both academic fit and program-specific relevance.

3.3 Model Compilation

- **Loss Function:** Sparse Categorical Cross Entropy for multi-class classification.

- **Optimizer:** Adam Optimizer with adaptive learning rate adjustments.

- **Evaluation Metrics:** Precision, recall, and F1-score to ensure robust model validation.

3.4 Model Training

- Training is performed using batch gradient descent with early stopping mechanisms to prevent overfitting.

- Uses cross-validation (k-fold validation) to improve generalizability.

- Training dataset is augmented using synthetic data generation techniques where necessary.

3.5 Training Visualization:

- Implements real-time visualization of training loss and accuracy curves.

- Uses confusion matrices and ROC curves to assess classification performance.

- Logs model performance metrics for iterative improvements.

3.6 Model Evaluation:

- **Test Set Evaluation:** Assesses generalization error using held-out data.

- **Performance Metrics:**

- Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for match score predictions.

- Precision-Recall and F1-score for classification robustness.

- A/B Testing: Compares recommendation outputs against baseline models.

3.7 Model Development and Persistence:

- **Model Saving:** Trained models are serialized using the .h5 format.

- **Model Loading:** Supports seamless integration into web-based recommendation systems.
- **API Deployment:** Flask or FastAPI-based REST API for real-time recommendations.

3.8 Working of proposed system

3.8.1. Structured Data Processing (XGBoost)

- Extracts numerical attributes relevant to university selection.
- Generates probabilistic admission scores based on historical admission data.
- Ensures scalability through parallelized tree-based learning.

3.8.2. Unstructured Data Processing (TF-IDF & Cosine Similarity)

- Converts course descriptions into high-dimensional sparse vectors.
- Computes semantic similarity scores to rank relevant programs.
- Implements stopword removal and stemming for improved text preprocessing.

3.8.3. Final Recommendation System

- Aggregates structured and unstructured data predictions.
- Ranks universities based on composite match scores.
- Integrates a user feedback loop to refine future recommendations.

3.9 SOP Generation using DeepSeek Language

The Statement of Purpose (SOP) generation process is powered by the DeepSeek language model, a state-of-the-art NLP model designed for high-quality, coherent, and contextually rich text generation. The system leverages advanced language understanding to craft SOPs that align with both user profiles and university-specific expectations. Key components of the SOP generation pipeline include:

3.9.1 Deep Learning-Based Generation: Utilizes the DeepSeek model instead of rigid templates, allowing dynamic structuring and more human-like narrative flow. This enables nuanced and personalized content generation beyond generic formats.

3.9.2 Content Personalization: Integrates user-specific inputs such as academic history, research interests, extracurricular achievements, and long-term career goals to create tailored SOPs that reflect individual strengths and aspirations.

3.9.3 Text Polishing and Augmentation: Applies advanced grammatical correction, lexical enhancement, and stylistic refinement using the DeepSeek model's in-built capabilities, ensuring each draft is both professional and compelling.

3.9.4 Multi-Version Drafting: Supports generation of multiple SOP versions, each adaptable to different universities, programs, or regions by tweaking tone, structure, or emphasis areas.

3.10 Backend and Architecture and API Design

The backend of the College Compass platform is engineered for performance, scalability, and seamless integration with the machine learning pipeline. It supports all key functionalities including university recommendations, user management, and SOP generation.

Key components and features include:

- **Server Framework:** Developed using Node.js and Express.js, enabling efficient handling of asynchronous operations and fast API responses for frontend interactions.

- **Database Management:** Employs a lightweight SQLite database, chosen for its simplicity and efficiency in managing structured data such as university details, user profiles, and matching scores.
 - Well-defined schemas support core features: university matching, user registration, and SOP generation.
 - **RESTful API Design:** Implements a set of RESTful endpoints to facilitate:
 - User Authentication – login/signup with password hashing.
 - University Search & Filter – dynamic queries based on tuition, rank, acceptance rate, etc.
 - Side-by-Side Comparison – comparing multiple institutions interactively.
 - **Dynamic Querying Support:** Enables flexible filtering and search based on a wide range of user-selected criteria, including:
 - Acceptance rate
 - Tuition fees (in-state/out-of-state)
 - Academic reputation
 - Location and size
 - Course similarity and ranking metrics
 - **Security & Validation:**
 - Password Hashing: User credentials are hashed using industry-standard algorithms before storage.
 - Input Validation: All user-submitted data is validated to prevent injection attacks and ensure system integrity.
 - **Extensibility:**
Designed with future integration in mind:
 - Machine Learning Compatibility – seamlessly integrates with profile scoring models.
 - Scalability – built to evolve with additional features such as analytics dashboards, real-time search, and third-party integrations.
- This backend architecture ensures a reliable, secure, and responsive foundation, aligning with the platform's mission of delivering intelligent, personalized college recommendations at scale.

4 Results and Analysis

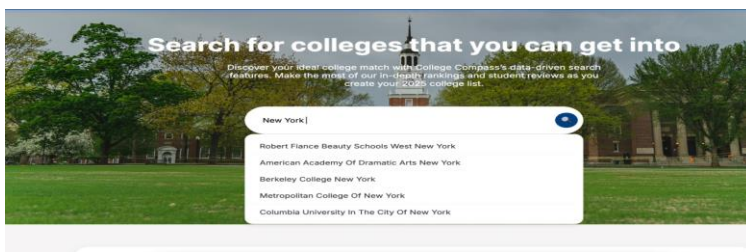


Fig 3 : Home page

Interface Description: The image represents the homepage UI component of the *College Compass* platform — specifically the search box with real-time suggestion dropdown functionality. This feature enables users (students) to begin typing a course, university name, or specialization, and receive dynamically generated suggestions based on partial input.

Technical Breakdown:

- **Frontend Behavior:**

1. The search bar uses autocomplete logic implemented via JavaScript (e.g., React/Vanilla JS).
2. Suggestions are displayed in a dropdown list as the user types, offering a smooth and intuitive interface for university search and filtering.

- **Backend Integration:**

1. Suggestions are pulled using RESTful API endpoints, which query a SQLite database containing over 6,600 universities, with relevant fields such as:

- University Name, Location, Rank, MS Courses Offered, etc.

2. This is achieved through dynamic querying logic, which performs partial match filtering via SQL LIKE queries or pre-indexed tokenized search vectors.

- **Personalization Layer (ML Tie-in):**

1. The system can enhance suggestions by incorporating:

- User profile history

- Previous search behavior

- User's GPA/GRE context (already fetched from user registration)

2. This makes the search *adaptive* and aligned with relevance scores generated by the hybrid model (XGBoost + TF-IDF/Cosine Similarity).

- **Data Flow Summary:**

1. Input: User types query.

2. Trigger: Frontend sends query to backend via API.

3. Processing: Backend fetches matching entries from SQLite using optimized query logic.

4. Output: Dropdown displays matches (with filters like acceptance rate, tuition, course availability embedded in metadata).

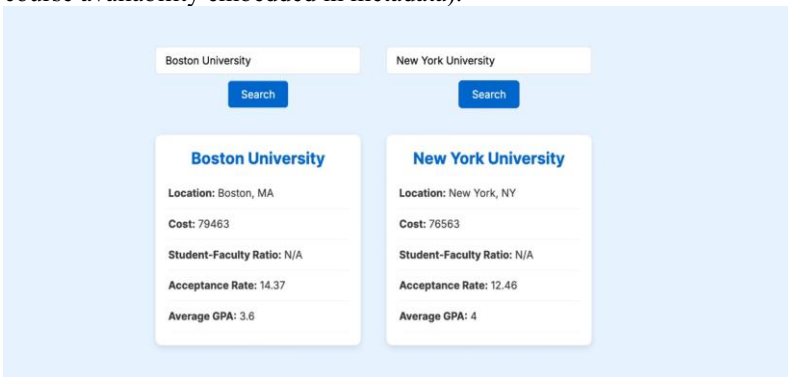


Fig .4. Comparator view with 2 universities side by side

Interface Description: This screen represents the university comparison module within the *College Compass* platform. The interface enables users to evaluate two universities in a parallel layout, helping in decision-making based on critical metrics.

Technical Breakdown:

- **Frontend Implementation:**
 - The UI likely uses a responsive grid layout (e.g., CSS Grid or Tailwind’s grid-cols-2) to structure university cards side-by-side.
 - Each card displays key information:
 - Name, location, rankings, tuition, GPA requirements, GRE cutoffs, ROI estimates, etc.
 - Progress bars or icon badges might visually denote the user’s match level with each school.
- **Backend Integration:**
 - Data pulled from the SQLite database where each university’s structured attributes (e.g., tuition, acceptance_rate, gpa_avg, median_salary) are stored.
 - Filtered by user query or selection from the homepage search.
 - APIs use unique university_ids to fetch side-by-side metadata efficiently.
- **ML Component Integration:**
 - The comparator references match scores generated by the XGBoost model, and course similarity scores (FastText/Cosine).
 - Each score is likely shown in a normalized percentage or bar format for ease of comparison.
 - Could include a color-coded band or visual cue based on compatibility thresholds (e.g., Green: Excellent match, Yellow: Moderate).
- **UX Considerations:**
 - Offers a real-time way to contrast between two options without needing to navigate between multiple tabs.
 - Ideal for users at the “narrowing down” stage of the university selection pipeline.

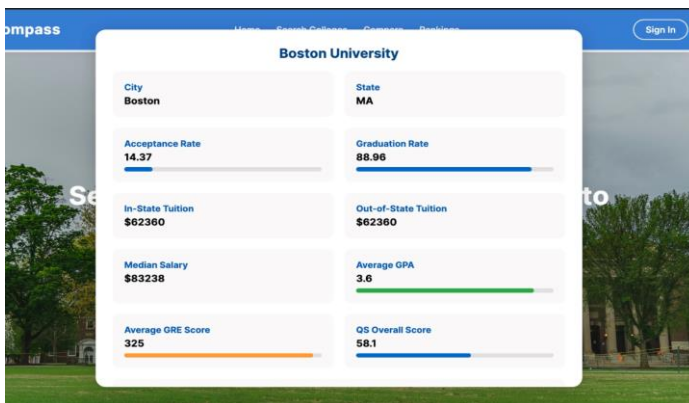


Fig .5. The popup university card with progress bars (GPA, GRE, etc.)

Interface Description: This screen shows an interactive popup card or modal, activated when a user clicks on a university from the search or comparison view. It summarizes the institution’s profile using graphical progress bars for critical admission metrics.

Technical Breakdown:

- Frontend UI Layer:
 - Implemented as a modal or overlay card using frameworks like React with useState toggles or onClick event handlers.
 - Includes:
 - University name, logo, rank
 - GPA, GRE, and course match scores visualized as progress bars
 - CTA buttons like “Add to Compare” or “View SOP”
- Backend-Driven Content:
 - Progress bars reflect real-time calculated values:
 - GPA Match Score = $\min(\text{user_gpa} / \text{university_avg_gpa}, 1.0) * 100$
 - GRE Match Score follows a similar logic
 - Course Similarity Score from NLP module (FastText/Cosine Similarity)
 - These values are fetched via API from a pre-processed table that stores user-specific compatibility scores.
- Machine Learning Context:
 - These scores are generated via the hybrid model (XGBoost + NLP).
 - Each bar quantifies how well the student matches the university’s expected profile, acting as a visual recommendation insight.
- User Guidance Functionality:
 - Helps users make informed, data-backed choices.
 - Serves as an information-rich preview without forcing full navigation to the university detail page.

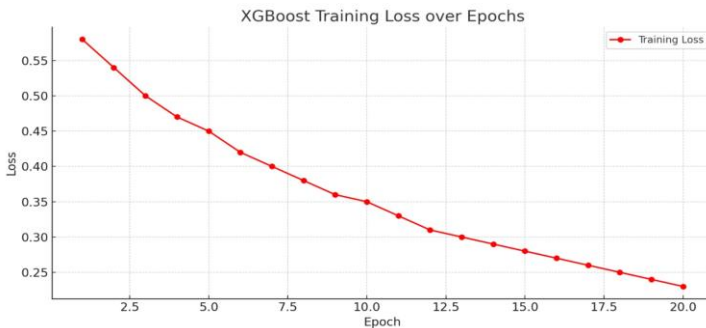


Fig .6. XGBoost Training Loss Over Epochs

Graph Analysis: XGBoost Training Loss Over Epochs

The graph shown above depicts the training loss of the XGBoost model over 20 epochs. Training loss is a key metric used to evaluate how well the model is learning the patterns in the data during training. A lower loss typically indicates a better fit to the training data.

Understanding the Axes:

- X-axis (Epoch): Represents the number of training iterations. One epoch means the model has seen the entire training dataset once.
- Y-axis (Loss): Represents the loss value computed by the loss function (such as log loss or mean squared error) used by the XGBoost algorithm.

Trend and Interpretation:

- The training loss decreases steadily from above 0.57 at epoch 1 to around 0.23 at epoch 20.
- This monotonic decrease is expected and desirable—it shows that as training progresses, the model is increasingly minimizing the error between its predictions and the actual outcomes.
- The curve becomes gradually less steep, indicating that the model is converging—learning slows down as it gets closer to an optimal solution.

Implications for the Model:

- The consistent decline in loss suggests that the model is not overfitting at this stage. Typically, overfitting can be indicated by training loss decreasing while validation loss increases, but this plot only shows training loss. For a more complete assessment, validation loss should also be tracked.
- The use of XGBoost (Extreme Gradient Boosting) is justified here due to its ability to reduce loss efficiently via additive boosting and handling both linear and non-linear relationships.

Practical Significance:

- In the context of this capstone project—“A Hybrid Recommendation System for University Selection”—the XGBoost model is likely used for ranking or classification of universities based on user profiles and preferences.
- The decrease in training loss indicates the model has successfully learned to associate user data with university features, thus improving the quality of the recommendations.

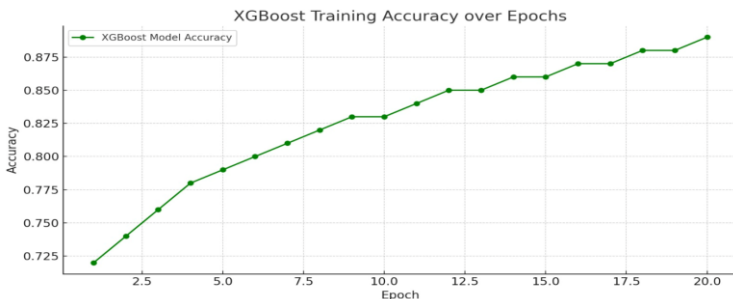


Fig .7. XGBoost Training Accuracy over Epochs

Graph Analysis: XGBoost Training Accuracy Over Epochs

This graph illustrates the progression of training accuracy of the XGBoost model over the course of 20 epochs. Accuracy measures how well the model’s predictions match the actual outcomes in the training dataset. A higher accuracy value is indicative of better model performance.

Understanding the Axes:

- X-axis (Epoch): Denotes the number of complete passes the model has made over the training dataset.
- Y-axis (Accuracy): Represents the proportion of correctly predicted instances over the total instances.

Observed Trend:

- The training accuracy increases consistently, starting from around 0.72 at epoch 1 and reaching approximately 0.89 at epoch 20.
- The improvement is steady and smooth, indicating effective learning.
- There are minor plateaus (e.g., between epochs 9–11 and 16–18), which suggest temporary stabilization in learning as the model adjusts to finer patterns in the data.

Interpretation:

- The graph confirms that the model is learning progressively, improving its prediction ability as the number of epochs increases.
- The steady upward trajectory with minimal fluctuations shows that the model is not facing issues like underfitting (where accuracy remains low) or overfitting (where training accuracy increases too fast without validation support).
- The plateaus could be an early sign of convergence, where further training yields diminishing accuracy gains.

Practical Implications in the Capstone Context:

- The XGBoost model is a core component of the hybrid recommendation system for university selection.
- This increase in training accuracy demonstrates that the model can reliably classify or rank universities based on student profile features such as CGPA, GRE score, SOP content, etc.
- This strong training performance lays the foundation for personalized and precise university recommendations, which is the primary objective of the system.

5 Advantages, Limitations and Applications

1. **Highly Personalized University Recommendations** Unlike traditional ranking-based search tools, this system provides tailored university suggestions based on a combination of structured data such as GPA, GRE scores, tuition costs, and acceptance rates. It ensures that each student receives recommendations that align not just with academic qualifications, but also with financial situations, geographic preferences, and career aspirations.

2. **Hybrid Machine Learning Model for Enhanced Accuracy** The integration of the XGBoost algorithm — known for its efficiency and accuracy in structured data prediction — significantly enhances the reliability of match scores. The use of feature engineering techniques, such as cost-benefit ratios and profile alignment metrics, ensures that recommendations consider both academic fit and return on investment (ROI).

3. **Simplified Decision-Making Process** The platform consolidates diverse university information into a single interface, reducing the cognitive load on students who would otherwise sift through dozens of websites and ranking platforms. With intuitive visual aids such as progress bars, match scores, and comparison charts, students can make well-informed decisions more confidently.

4. **Real-Time University Comparison Feature** The system allows side-by-side evaluation of universities using multiple parameters — such as tuition fees, GPA

requirements, GRE cutoffs, employment outcomes, and course similarity. This feature is especially useful for shortlisting institutions during the final decision-making phase.

5. **Application Lifecycle Support**Beyond just recommendations, the system supports students through the entire application lifecycle. Tools for deadline tracking, document checklists, and AI-assisted SOP generation help students manage their applications efficiently and reduce the chances of missing critical submission dates.

6. **Scalable and Modular Architecture**The backend of the system is built using scalable technologies (Node.js, Express.js, SQLite), enabling easy integration of additional modules in the future. This architecture allows for seamless updates, third-party API integrations (e.g., scholarship portals, test score uploads), and multilingual support, ensuring the platform evolves with user needs.

7. **Democratization of Academic Counseling**The system bridges the accessibility gap by providing personalized, data-driven guidance to students who may not have access to expensive college counselors or elite mentorship programs. It empowers students from diverse socioeconomic backgrounds to make informed educational decisions.

5.1 Applications

1. **Student University Selection Platforms**The core use case of the system is to assist prospective students in finding universities that best match their academic strengths, financial capacities, and personal goals. It can be deployed by high schools, educational consultants, and EdTech companies to support students during college planning.

2. **Career and Academic Counseling**Educational institutions and counselors can integrate this system into their advisory services to automate the process of generating university shortlists for students. It reduces the burden on human advisors while increasing personalization and coverage.

3. **Digital Admission Portals and EdTech Platforms**The recommendation engine can be embedded into broader EdTech ecosystems or digital university admission portals to enhance user experience. This includes platforms like Shiksha, Unacademy, or Coursera, where students often search for academic guidance.

4. **AI-Assisted SOP and Document Generation**The SOP generation module can be expanded and offered as a standalone service for students applying to universities abroad. By producing structured and personalized drafts, it saves time and helps students put forward stronger applications.

5. **University Marketing and Outreach**Universities can use the platform to reach potential applicants whose profiles align well with their programs. By understanding the types of students who match their offerings, institutions can tailor recruitment strategies more effectively.

6. **Scholarship and Funding Recommendations** With future integrations, the system could match students with scholarships or financial aid programs based on eligibility, merit, and course interest — addressing another crucial factor in university decision-making.

7. **Policy Research and Educational Equity** Governments and NGOs working on educational access and equity could use insights from the platform to identify gaps in higher education awareness and access, enabling more targeted interventions for underrepresented groups.

6 Conclusion

Conclusion: The development and implementation of *College Compass*, a hybrid recommendation system for university selection, represents a significant advancement in educational decision-support systems. By integrating advanced Machine Learning (ML) algorithms with Natural Language Processing (NLP) methodologies, the system holistically addresses the multifaceted nature of academic institution selection.

At its core, the system leverages the XGBoost regression model to compute a probabilistic match score, synthesizing critical numerical features such as GPA, GRE scores, acceptance rates, tuition costs, and return on investment (ROI) metrics. The structured data pipeline is further enhanced with feature engineering techniques such as cost-benefit ratios and profile match scores. The convergence of the model, demonstrated by the steadily declining loss and rising accuracy over 20 epochs, confirms the efficacy and reliability of XGBoost in this context.

Complementing the structured model is a robust NLP module that performs semantic course matching through FastText embeddings and cosine similarity. Unlike traditional TF-IDF representations, FastText allows for contextual and subword-level understanding of course descriptions, enabling a nuanced alignment between student interests and academic offerings.

The hybrid model architecture consolidates both components by applying a tunable weighted score aggregation mechanism, with empirically chosen weights (e.g., 0.4 for GPA, 0.4 for GRE, and 0.2 for course similarity). This blended score provides a more comprehensive assessment, accounting for both statistical compatibility and semantic relevance.

The backend infrastructure, built on Node.js, Express.js, and SQLite, ensures scalability, performance, and API-driven modularity. Features like real-time search, dynamic filtering, deadline tracking, and AI-driven SOP generation using the DeepSeek language model add practical value to the student application journey. The SOP generation module, in particular, shifts away from template-based writing to deep-learning-based narrative generation—offering contextual, grammatically enhanced, and personalized documents.

Altogether, this project not only reimagines how recommendation systems can empower students but also sets a precedent for integrating multi-modal AI pipelines (ML + NLP) into high-stakes decision-making processes in education.

While the current system demonstrates strong performance and practical applicability, several avenues for enhancement and expansion remain open. Future work can be explored across five key domains:

1. Real-Time Data Integration and Live Updates

- University Data APIs: Incorporate real-time feeds from university databases and public platforms (e.g., QS, Common App, UCAS) to maintain current information on application deadlines, tuition changes, course additions, and program cancellations.
 - Web Scraping Pipelines: Automate the extraction of new academic program data, faculty information, and admission criteria using scalable web crawlers and ETL frameworks.
2. Advanced Personalization via Reinforcement Learning
- User Behavior Feedback Loops: Employ reinforcement learning (e.g., multi-armed bandits) to refine recommendations based on user interactions such as click-through rates, shortlist patterns, and final enrollments.
 - Session-Based Adaptation: Introduce session-based recurrent models (e.g., GRUs, LSTMs) to personalize outputs dynamically during a user's browsing or application session.
3. Enhanced NLP with Transformer Models
- Transformer-Based SOP Generation: Replace DeepSeek with more powerful transformer-based architectures like GPT-4, T5, or BART for improved coherence and creativity in SOPs.
 - SOP Evaluation Module: Add an automated scoring engine that uses rubric-based classification or BERT-based textual entailment to grade SOPs on originality, intent alignment, and structure.
4. Cross-Domain and Multilingual Support
- Global Expansion: Extend recommendations to include non-U.S. universities, adding support for different academic systems (e.g., ECTS, IELTS/TOEFL scoring).
 - Multilingual Interaction: Localize the platform by integrating multilingual NLP models for users from diverse linguistic backgrounds. Enable SOP generation and UI support in languages like Spanish, French, Mandarin, and Hindi.
5. Explainable AI and Interpretability
- Model Explainability: Integrate SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) to provide transparent visualizations of how different input features contributed to a recommendation.
 - Bias Detection & Fairness Analysis: Implement fairness-aware ML auditing tools to identify and mitigate any demographic or geographic bias in university recommendations.
6. Mobile Deployment and Progressive Web Application (PWA)
- Develop a React Native or Flutter-based mobile app for Android and iOS, offering offline SOP editing, real-time notifications, and on-the-go application management.
 - Alternatively, convert the web app into a Progressive Web Application (PWA) to combine the best of web and mobile capabilities.
7. Integration with Application Platforms and Counselor Portals
- Third-Party Integration: Build plugins or APIs for integration with college application portals and career counseling tools.
 - Counselor Dashboard: Add role-based access controls to provide guidance professionals with insights into student profiles, system recommendations, and document readiness.

By incorporating these future advancements, College Compass can evolve into an end-to-end intelligent education assistant, supporting users across the full lifecycle—from

university search to application submission and post-admission planning. The system holds the potential to redefine digital educational counseling, making it universally accessible, highly personalized, and deeply intelligent.

References

1. A. Alsulami and A. S. Aljohani, "College recommendation system using hybrid filtering algorithm," *2020 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)*, Khartoum, Sudan, 2020, pp. 1–5, doi: [10.1109/ICCCEEE49695.2020.9274927](https://doi.org/10.1109/ICCCEEE49695.2020.9274927).
2. P. Brusilovsky and E. Millán, "User models for adaptive hypermedia and adaptive educational systems," in *The Adaptive Web*, Berlin, Heidelberg: Springer, 2007, pp. 3–53, doi: [10.1007/978-0-387-85820-3_8](https://doi.org/10.1007/978-0-387-85820-3_8).
3. M. G. Ayub, N. A. M. Isa, and A. Ahmad, "Machine Learning Applications for Recommender Systems in Higher Education: A Systematic Review," *Information*, vol. 7, no. 1, pp. 1–16, Jan. 2024, doi: [10.3390/info7010006](https://doi.org/10.3390/info7010006).
4. H. J. Hossain et al., "AI-powered academic guidance and counseling system," *Journal of Big Data*, vol. 9, no. 1, pp. 1–22, Dec. 2022, doi: [10.1186/s40537-022-00592-5](https://doi.org/10.1186/s40537-022-00592-5).
5. J. Li and X. Zhang, "LLM AI Text Generation Detection Based on Transformer Deep Learning Algorithm," *arXiv preprint*, arXiv:2405.06652, 2024. [Online]. Available: <https://arxiv.org/abs/2405.06652>
6. M. R. Putra and A. Pradana, "Recommendation System for University Degree Selection: A Socioeconomic and Standardized Test Data Approach," *Indonesian Journal of Information and Communication Technology (IJOICT)*, vol. 5, no. 2, pp. 114–120, 2023. [Online]. Available: <https://socjs.telkomuniversity.ac.id/ojs/index.php/ijoict/article/view/747>
7. A. Singh and R. S. Rajpoot, "A content-based recommender system for choosing universities," in *Proceedings of the 13th International Conference on Agents and Artificial Intelligence (ICAART 2021)*, 2021, pp. 390–398. [Online]. Available: <https://www.scitepress.org/Papers/2021/107275/107275.pdf>
8. H. T. Nguyen, "A hybrid college recommendation system integrating collaborative filtering and deep learning," *Preprints*, 2024. [Online]. Available: https://www.preprints.org/manuscript/202408.0905/download/final_file
9. M. Kaya and H. Eken, "Netflix Recommendation System Based on TF-IDF and Cosine Similarity Algorithms," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 28, no. 4, pp. 2217–2231, 2020. [Online]. Available: <https://journals.tubitak.gov.tr/elektrik/vol28/iss4/22/>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

