Final Year Project

*SKEW*PLAY

FYP Team

Muhammad Ahsan	22F-3255
Syed Noor Ul Ghani	22F-3363
Abdul Qudoos	22F-3851

Supervised by

Sir Rizwan Ul Haq

Co-Supervised by

Dr. Muhammad Usama

Department of Computer Science
National University of Computer and Emerging Science
Chiniot, Pakistan

2025

Abstract

SkewPlay is a web-based, no-code educational platform designed to empower students, teachers, and machine learning (ML) experts to explore and mitigate data imbalance issues in tabular datasets for binary, multiclass, and multilabel classification tasks. Inspired by the book Machine Learning for Imbalanced Data by Kumar Abhishek and Dr. Mounir Abdelaziz (Packt Publishing, 2023), the platform adapts and integrates relevant techniques from the book, including oversampling, undersampling, ensembles, and model calibration, filtered by dataset type.

Key features encompass automatic imbalance and anomaly detection, AI-driven pipeline recommendations, real-time visualizations of data distributions and metrics, statistical explanations with mathematical formulas, inline tutorials featuring book excerpts and video links, guided tours, role-based access, workflow saving and resumption, and PDF report exports. Built with a Vue.js frontend for interactive user interfaces and a Python backend for processing, SkewPlay bridges educational gaps by automating complex ML experiments, fostering hands-on learning without coding barriers.

Table of Contents

LIST OF TABLES

TABLE OF FIGURES

1. Project Vision

1.1. Background and Motivation

Data imbalance is a pervasive challenge in machine learning, where minority classes or labels are underrepresented, leading to biased models with poor performance on critical tasks such as fraud detection or rare disease diagnosis. The book Machine Learning for Imbalanced Data provides a comprehensive guide to techniques spanning oversampling, undersampling, ensembles, deep learning methods, and model calibration. However, applying these requires coding proficiency, limiting accessibility for educational users. SkewPlay bridges this gap by creating a no-code teaching tool that automates experimentation, offers AI-guided recommendations, and embeds book-specific explanations to enhance understanding.

1.2. Problem Statement

Table 1: Problem Statement

Category	Description
Problem	• A common issue in machine learning is where minority classes or labels are underrepresented, leading to biased models and poor performance on critical tasks like fraud detection or rare disease diagnosis. • Traditional ML tools require coding skills and expertise, making it hard for everyone to explore and apply techniques from Machine Learning for Imbalanced Data. • No easy, no-code options for educational experiments, forcing reliance on complex setups that limit hands-on learning and real-time feedback.
Impact	• Results in biased models • Reduced learning outcomes for students • Poor performance in real-world applications like fraud detection.
Affects	• They need simple no-code interfaces for exploring techniques but face complex tools, restricting foundational learning on tabular datasets.
Solution	The system will: • Captures techniques from Machine Learning for Imbalanced Data to promote hands-on ML learning and accountability in experiments. • Builds step-by-step processes for data ingestion, preprocessing, imbalance detection, and applications

	across binary/multiclass/multilabel on tabular data. • Enables AI recommendations, removing coding barriers for secure exploration. • Delivers formula-based explanations to address gaps and foster ethical practices. • Encourages integrity via workflow saving, 'PDF sharing', and visuals; empowers users with metrics to combat data bias in education and apps.

1.3.Objective

Following are the objectives of SkewPlay

- Develop a no-code educational platform integrating book techniques for binary, multiclass, and multilabel on tabular datasets.
- Incorporate AI recommendations, visualizations, and explanations.
- Enhance education through tutorials and explanations.
- Ensure scalability, accessibility, and user features like workflow saving and sharing.

1.4. Business Opportunities

The growing demand for ML education and tools presents a strong opportunity. With the ML market expanding (projected to reach $200 billion by 2025), there is a need for accessible platforms that bridge theory and practice. SkewPlay can capture this by offering a free Basic tier to attract beginners (e.g., via educational institutions) and monetizing through an Advanced paid tier for professionals and advanced users. Integration with Stripe for payments ensures secure revenue. As a companion to the referenced book, it could partner with publishers for cross-promotion. The no-code trend in EdTech allows SkewPlay to stand out, potentially generating income from subscriptions while filling a niche in imbalance-focused learning. Ethical features like bias mitigation add value, appealing to inclusive education markets.

1.5. Inspiration from the book

The platform is a direct companion to Machine Learning for Imbalanced Data, incorporating techniques from the book (e.g., SMOTE variants, ensembles like RUSBoost, DL losses like Focal Loss). Book excerpts, formulas, and case studies are embedded in tutorials and explanations for contextual learning.

1.6. Scope

SkewPlay's scope is limited to a web-based no-code educational platform for handling data imbalance in tabular datasets (CSV uploads). It supports binary, multiclass, and multilabel classification tasks.

1.7. Stakeholder and User Description

1.7.1. Market Demographics

SkewPlay targets the education and ML learning market, which includes:

Age: Primarily 18-35 (college students, early-career professionals).

Education Level: Undergraduate/graduate students in computer science, data science, or related fields; educators with ML knowledge.

Geographic Distribution: Global, but focused on regions with strong tech education (e.g., Pakistan, USA, India).

Occupation: Students (60%), teachers/educators (20%), ML enthusiasts/professionals (20%).

Tech Proficiency: Basic users (beginners with little coding); Advanced users (intermediate ML knowledge).

Market Size: EdTech ML tools are growing; similar platforms (e.g., n8n) have millions of users. SkewPlay niches in imbalance education, appealing to book readers and online courses (e.g., Coursera ML specialization enrollees).

1.7.2. User Environment

Users can interact with SkewPlay through a web-based interface accessible on any device with a modern browser like Chrome or Firefox, though desktops or laptops are preferred for tasks involving data uploads due to larger screens and easier file handling; mobile devices can be used for lighter activities like viewing dashboards or tutorials.

1.7.3. Stakeholder Profiles

1.7.3.1. Supervisor Team

Table 2: Supervisor Team Profile

Representatives	Academic advisors overseeing the project's alignment with educational goals. • Supervisor: Sir Rizwan Ul Haq • Co-Supervisor: Dr. Muhammad Usama
Description	They are involved in supervising activities for the development process.
Type	Primary (internal oversight)
Responsibility	Provide guidance on requirements, review designs and progress, ensure compliance with academic standards, and approve milestones.
Success Criteria	The completion of features which are being committed by the development team at the start of the project.
Involvement	Medium – regular meetings, document reviews, and feedback sessions.

1.7.3.2. Development Team

Table 3: Development Team Profile

Representatives	The student team building the platform • Muhammad Ahsan • Syed Noor Ul Ghani • Abdul Qudoos
Description	They are involved in the development of the project.
Type	Primary (internal creators)
Responsibility	1. Fully involved in the development of the project. 2. Specify requirements like a requirement engineer. 3. Conduct research system features to develop the model. 4. Successful development of proposed
Success Criteria	Successful completion and development of committed features
Involvement	High – Research, development, Documentation of the project.

1.7.3.3. End Users

Table 4: End Users Profile

Representatives	Individuals who directly use the SkewPlay platform for educational or prototyping purposes, including students, educators, and ML enthusiasts in Basic (free) or Advanced (paid) tiers.
Description	End users engage with the platform to learn about imbalanced data handling, experiment with techniques, and apply them in educational or prototyping scenarios.
Type	Primary (external users)
Responsibility	Interact with the platform to explore features, provide feedback on usability, and subscribe to paid tiers for advanced access.
Success Criteria	High user satisfaction
Involvement	High

2. System Requirements Specification

2.1. System Features

SkewPlay offers following features:

2.1.1. User Authentication & Access Control: Secure registration and login via email/password or Google OAuth; session management with auto-logout; tiered access (free Basic tier with limits, paid Advanced tier with full features).

2.1.2. Data Ingestion & Analysis: Upload CSV tabular datasets or select from pre-loaded examples; auto-detects type (binary/multiclass/multilabel), imbalance ratios, and anomalies.

2.1.3. Data Preprocessing: No-code tools for feature scaling, categorical encoding, and stratified train/test splitting; includes before/after class distribution visualizations.

2.1.4. Imbalance Handling: Book-inspired methods like over-sampling (SMOTE variants) and under-sampling dataset-type filtering with UI slider tuning.

2.1.5. **Model Configuration & Training:** Select classifiers (e.g., Logistic Regression, RUSBoost ensembles); hyperparameter tuning, baseline comparisons, and multilabel support.

2.1.6. **AI-Driven Recommendations:** The system shall use an AI model to analyze dataset characteristics, with detailed explanations for each recommendation.

2.1.7. **Workflow Management:** Save, load, edit, and resume experiments as JSON; dashboard for workflow organization and access.

2.1.8. **Export & Sharing:** PDF reports with summaries, metrics, plots, and explanations; and CSV exports.

2.1.9. **Educational Resources:** Inline tutorials with book excerpts, video links, and math formula explanations.

2.1.10. **Payment Integration:** Stripe for Advanced tier subscriptions; quota enforcement, billing notifications, and management.

2.2. Functional Requirements

2.2.1. User Authentication and Management

FR-1.1. The system shall allow users to register and log in using email/password via Firebase or Google OAuth.

FR-1.2. The system shall manage user sessions, including automatic logout after inactivity (e.g., 30 minutes).

FR-1.3. The system shall provide a user dashboard displaying personal details, recent experiments, saved workflows, and usage statistics (e.g., experiment run, storage used).

FR-1.4. The system shall assign and enforce user tiers (Basic or Advanced) based on subscription status, dynamically showing or hiding features.

FR-1.5. The system shall allow users to update their profile information (e.g., email, password) and view subscription details.

2.2.2. Data Handling

FR-2.1. The system shall support uploading tabular datasets in CSV format, with validation for file size and structure.

FR-2.2. The system shall provide a library of 5-10 pre-loaded example datasets for users to select and use.

FR-2.3. The system shall automatically detect the dataset type (binary, multiclass, or multilabel classification) based on the target column.

FR-2.4. The system shall calculate and display imbalance ratios (e.g., minority-to-majority class ratios) for the uploaded or selected dataset.

FR-2.5. The system shall perform anomaly detection (e.g., outliers in features) and highlight potential issues in the dataset summary.

2.2.3. Pre-processing

FR-3.1. The system shall provide options for feature scaling (e.g., normalization, standardization) on numerical columns.

FR-3.2. The system shall support encoding categorical features (e.g., one-hot encoding, label encoding).

FR-3.3. The system shall perform stratified train/test splitting to preserve class distributions, with user-configurable split ratios (e.g., 80/20).

FR-3.4. The system shall generate and display visualizations of class distributions before and after preprocessing steps.

2.2.4. Imbalance Analysis and Mitigation

FR-4.1. The system shall filter and display applicable imbalance-handling techniques based on dataset type (e.g., only multiclass-compatible methods for multiclass data).

FR-4.2. The system shall implement data-level techniques such as over-sampling.

FR-4.3. The system shall implement under-sampling techniques.

FR-4.4. The system shall provide tunable parameters for each technique (e.g., sampling strategy ratio via sliders).

FR-4.5. The system shall apply selected techniques to the dataset and generate balanced versions for further modeling.

2.2.5. Model Configuration and Training

FR-5.1. The system shall allow users to select from a list of classifiers (e.g., Logistic Regression, SVM).

FR-5.3. The system shall support ensemble methods (e.g., RandomForest).

FR-5.4. The system shall enable hyperparameter tuning via UI elements (e.g., sliders for learning rate, dropdowns for number of estimators).

FR-5.5. The system shall train a baseline model (e.g., without imbalance

handling) for comparison.

FR-5.6. The system shall train selected models on processed datasets, displaying progress and estimated time

2.2.6. Evaluation and Metrics

FR-6.1. The system shall compute and display standard metrics (e.g., accuracy, precision, recall, F1-score).

FR-6.2. The system shall include imbalance-specific metrics (e.g., G-mean, Brier score).

FR-6.3. The system shall generate confusion matrices for visual comparison of predictions.

FR-6.4. The system shall create interactive plots like ROC curves and Precision-Recall curves.

FR-6.5. The system shall provide side-by-side comparisons of metrics before and after imbalance handling.

2.2.7. Visualization

FR-7.1. The system shall create dynamic charts for dataset summaries (e.g., bar charts for class distributions, pie charts for imbalance ratios).

FR-7.2. The system shall generate plots for evaluation results (e.g., line graphs for metric improvements).

FR-7.3. The system shall ensure visualizations are accessible (e.g., alt text for images, high-contrast colors).

FR-7.4. The system shall allow users to export individual visualizations as images or include them in reports.

2.2.8. AI Recommendation Engine

FR-8.1. The system shall use a pre-trained ML model to analyze dataset characteristics (e.g., imbalance ratio, size, type).

FR-8.2. The system shall recommend optimal pipelines (e.g., "Use SMOTE with RandomForest for high imbalance").

FR-8.3. The system shall provide explanations for recommendations, referencing book examples or dataset traits.

FR-8.4. The system shall allow users to accept, modify, or reject recommendations and apply them to the workflow

2.2.9. Workflow Management

FR-9.1. The system shall save entire workflows (e.g., data selection,

preprocessing, techniques, models) as JSON objects in Firebase.

FR-9.2. The system shall allow users to load saved workflows for replaying or editing.

FR-9.3. The system shall support resuming interrupted sessions with preserved states and results.

FR-9.4. The system shall display a list of saved workflows in the dashboard, with options to delete or rename.

2.2.10. PDF Export

FR-10.1. The system shall generate PDF reports including dataset summaries, applied techniques, metrics, plots, formulas, and explanations.

FR-10.2. The system shall allow exporting raw results (e.g., metrics as CSV) for external use

2.2.11. Educational and Help Features

FR-11.1. The system shall provide inline tutorials with excerpts from the book Machine Learning for Imbalanced Data.

FR-11.2. The system shall include links to external videos or resources for further learning on selected techniques.

FR-11.3. The system shall display "why/how" popups with statistical explanations and mathematical formulas for each method or result

2.2.12.Payment and Tier Management

FR-12.1. The system shall integrate with Stripe for processing Advanced tier subscriptions (e.g., monthly payments).

FR-12.2. The system shall allow users to upgrade from Basic to Advance via the dashboard, with immediate feature unlocking.

FR-12.3. The system shall enforce tier-based restrictions (e.g., quota limits on experiments, access to advanced algorithms).

FR-12.4. The system shall notify users when quotas are reached (e.g., "Upgrade to Advance for unlimited access").

FR-12.5. The system shall handle subscription management (e.g., cancel, renew) and display billing history.

2.3. Non-Functional Requirements

2.3.1. Performance

NFR-1.1. The system shall handle datasets up to 1GB with processing time < 5 minutes on standard hardware.

NFR-1.2. The system shall support up to 50 users at once without crashes

2.3.2. Security

NFR-2.1. The system shall use secure logins (e.g., passwords and Google sign-in) and encrypt all data.

NFR-2.2. The system shall not store personal or sensitive information without permission.

2.3.3. Reliability

NFR-3.3. The system shall achieve 90% uptime.

3. Use Case Models

3.1. Use Case Diagram

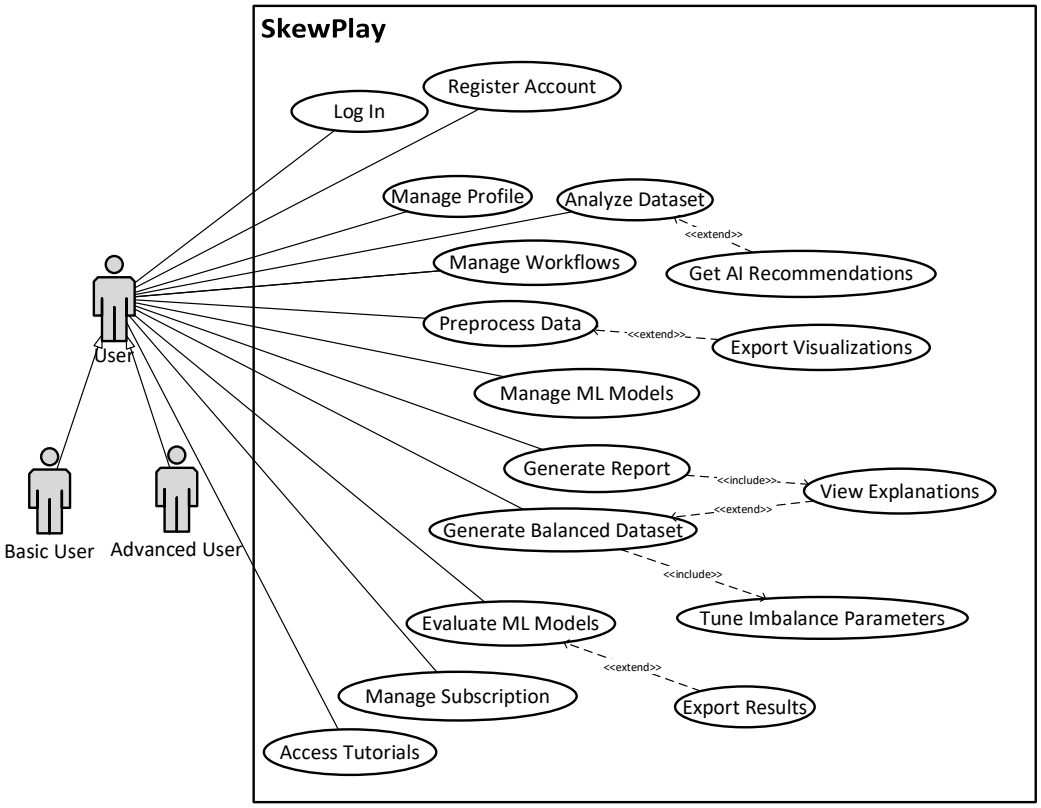

Figure 1: Use Case Diagram

3.2.High-Level Use Cases

Table 5: High-Level Use Cases

Use Case ID	Use Case Name	Actors	Description
UC-03	Manage Profile	User	Permits users to update personal details, view subscription info, and manage sessions.
UC-04	Export Results	User	Exports raw results (e.g., metrics) as CSV for external use.
UC-06	Preprocess Data	User	Applies scaling, encoding, and stratified splitting with visualizations.
UC-07	Tune Imbalance Parameters	User	Adjusts parameters for imbalance techniques via UI sliders.
UC-08	Generate Balanced Dataset	User	Applies filtered imbalance-handling methods to create balanced data versions.
UC-10	Evaluate ML Models	User	Computes metrics, generates confusion matrices, and compares results.
UC-11	Export Visualizations	User	Exports charts and plots as images or includes in reports.
UC-14	Generate Report	User	Creates PDF reports with summaries, metrics, plots, and explanations.
UC-15	Access Tutorials	User	Views inline tutorials with book excerpts and video links.
UC-16	View Explanations	User	Displays statistical and mathematical explanations for methods/results.

3.3. Expanded Use Cases

3.3.1. UC-01: Register Account

Table 6: Expanded UC-01-Register Account

Actors	**User**
Description	Creates a new user account to access SkewPlay's features.
Preconditions	• User is not logged in. • Internet connection available.
Postconditions	• User account is created and assigned to Basic tier. • Confirmation email sent.
Main Flow	1. User navigates to registration page. 2. User enters email, password. 3. System validates input (e.g., unique email, password strength). 4. System creates account in Firebase and assigns Basic tier. 5. System redirects to login page or dashboard.
Alternative Flows	2A. If Google OAuth is selected, system redirects to Google for authentication and auto-fills details.
Exceptions	3E. Invalid input: System displays error (e.g., "Email already in use") and prompts re-entry.

3.3.2. UC-02: Login

Table 7: Expanded UC-02 Login

Actors	**User**
Description	Authenticates user to access personalized features.
Preconditions	• User has a registered account.
Postconditions	• User session is active. • dashboard displayed based on tier.
Main Flow	1. User navigates to login page. 2. User enters email/password. 3. System verifies credentials via Firebase. 4. System starts session (auto-logout after 30 minutes inactivity). 5. System loads dashboard with recent experiments and tier features.
Alternative Flows	2A. Google OAuth: System handles external authentication.
Exceptions	3E. Invalid credentials: System shows error.

3.3.3. UC-05: Analyze Dataset

Table 8: Expanded UC-05 Analyze Dataset

Actors	**User**
Description	Uploads/selects dataset and performs initial analysis.
Preconditions	• User logged in • CSV file ready (if uploading)
Postconditions	• Dataset type, imbalance ratios, and anomalies displayed.
Main Flow	1. User uploads CSV. 2. System validates file (size, structure). 3. System auto-detects type (binary/multiclass/multilabel). 4. System calculates imbalance ratios and anomalies. 5. System displays summary visualizations.
Alternative Flows	1A. Pre-loaded: System loads from library (5-10 examples).
Exceptions	2E. Invalid file: System shows error and rejects upload.

3.3.4. UC-09: Manage ML Models

Table 9: Expanded UC-09 Manage ML Models

Actors	**User**
Description	Configures and trains models.
Preconditions	• Balanced dataset ready.
Postconditions	• Model trained • Baseline comparison available.
Main Flow	1. User selects classifier (e.g., SVM, RUSBoost). 2. User tunes hyperparameters via UI. 3. System trains baseline and selected model. 4. System displays progress.
Alternative Flows	1A. Ensemble: Supports multilabel.
Exceptions	4E. Timeout: System aborts if >2 minutes (for small datasets).

3.3.5. UC-12: Get AI Recommendations

Table 10: Expanded UC-012 Get AI Recommendations

Actors	**User**
Description	Receives AI-suggested pipelines.

Preconditions	• Dataset analyzed
Postconditions	• Recommendations displayed with explanations
Main Flow	1. System analyzes dataset traits. 2. AI engine suggests pipeline (e.g., "SMOTE + RandomForest"). 3. User accepts/modifies.
Alternative Flows	3A. Reject: User proceeds manually.
Exceptions	None

3.3.6. UC-13: Manage Workflows

Table 11: Expanded UC-13 Manage Workflows

Actors	**User**
Description	Saves and manages experiments.
Preconditions	• Workflow in progress
Postconditions	• Workflow saved as JSON in Firebase.
Main Flow	1. User saves current state. 2. System stores in dashboard. 3. User loads/creates/resumes.
Alternative Flows	2A. Delete: Removes from list.
Exceptions	3E. Quota reached: Notifies Basic users to upgrade.

3.3.7. UC-17: Manage Subscription

Table 12: Expanded UC-17 Manage Subscription

Actors	**Advanced User, Basic User**
Description	Handles tier upgrades and payments.
Preconditions	• User logged in.
Postconditions	• Tier updated • Features unlocked/locked.
Main Flow	1. User views subscription in dashboard. 2. User upgrades via Stripe. 3. System enforces quotas/notifies.
Alternative Flows	2A. Cancel: Downgrades to Basic.
Exceptions	3E. Payment fails: System retries or cancels.

4. Activity Diagram

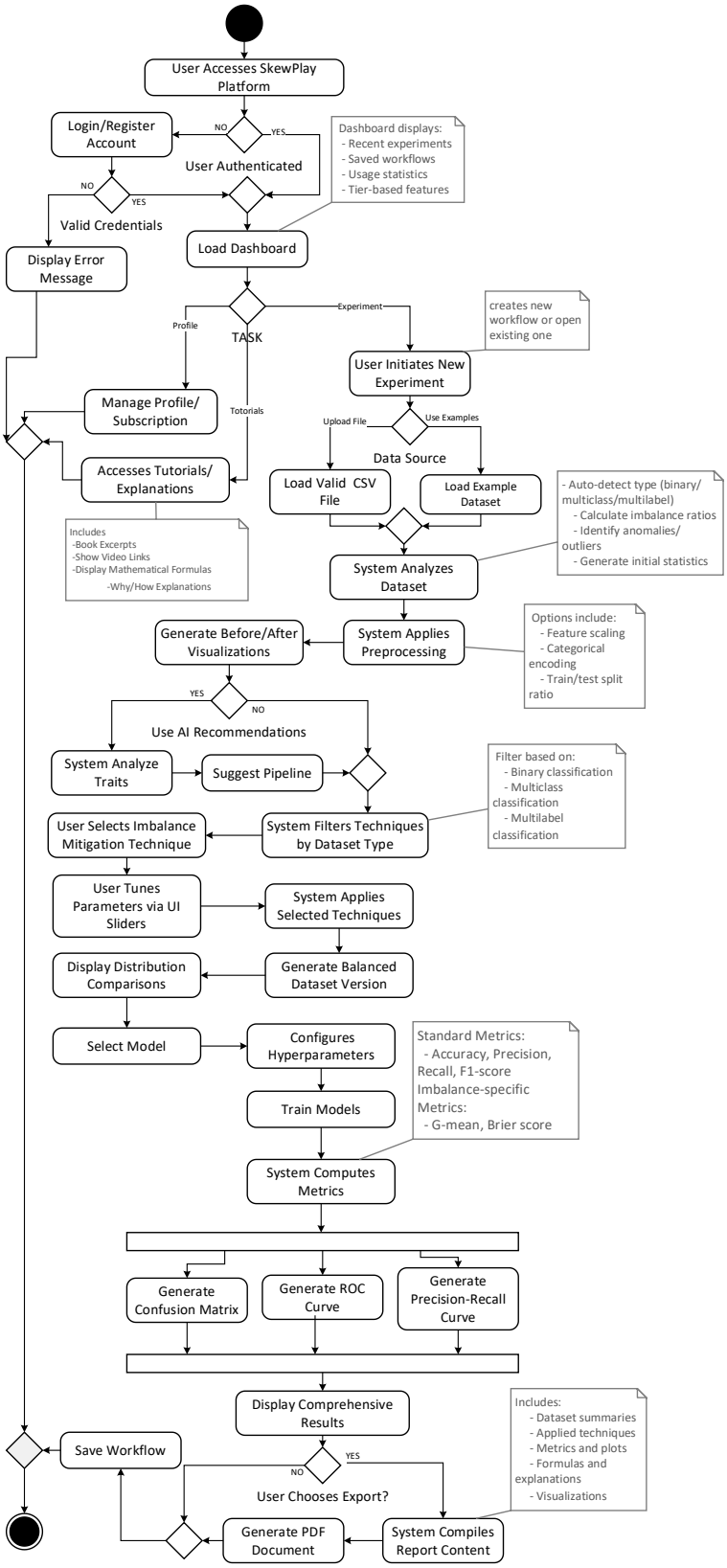

Figure 2: Activity Diagram

5. System Sequence Diagram

5.1. Register Account

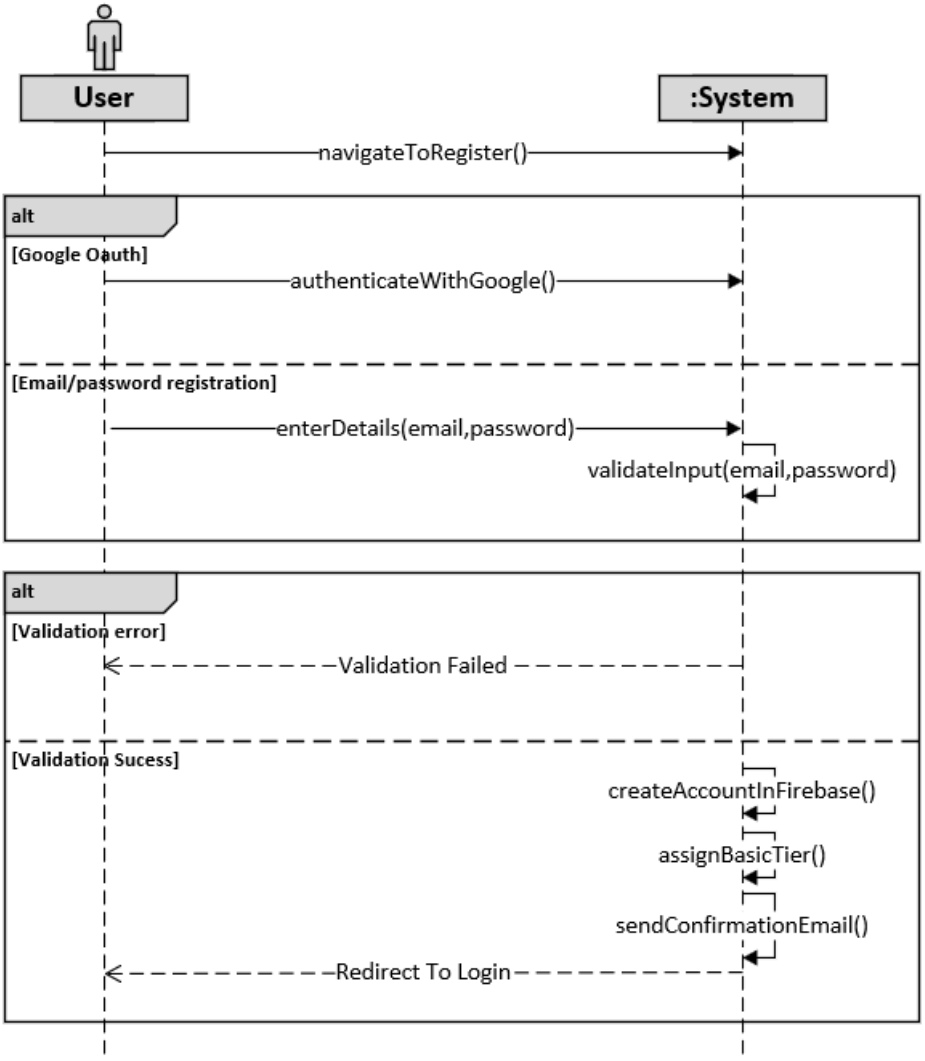

Figure 3: SSD - Register Account

5.2.Login

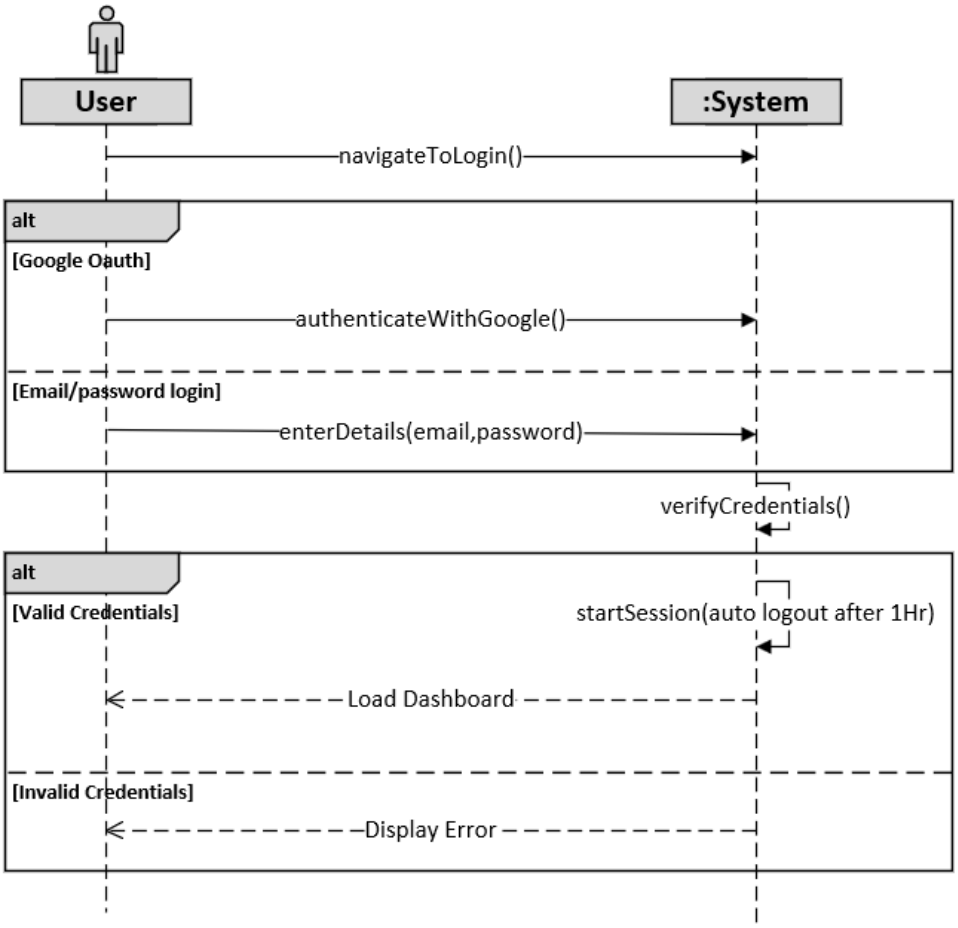

Figure 4: SSD - Login

5.3.Analyze Dataset

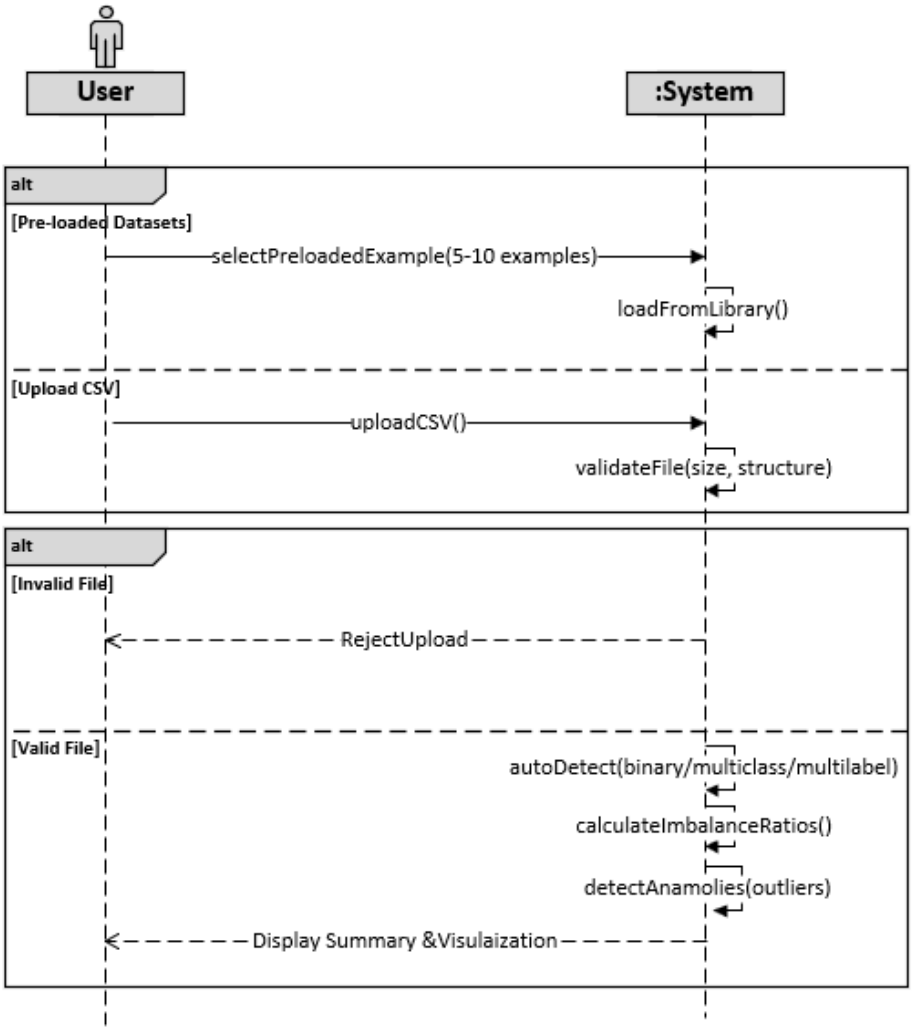

Figure 5: SSD - Analyze Dataset

5.4.Manage ML Models

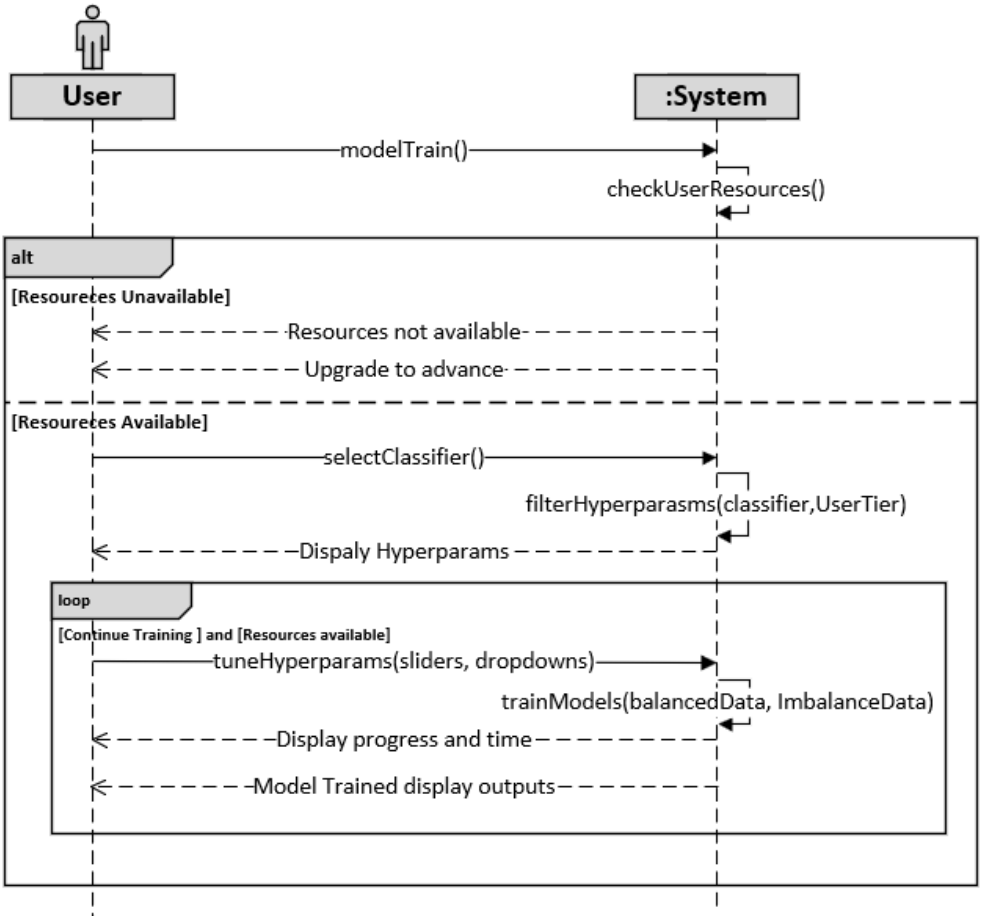

Figure 6: SSD Manage ML Models

5.5. Get AI Recommendations

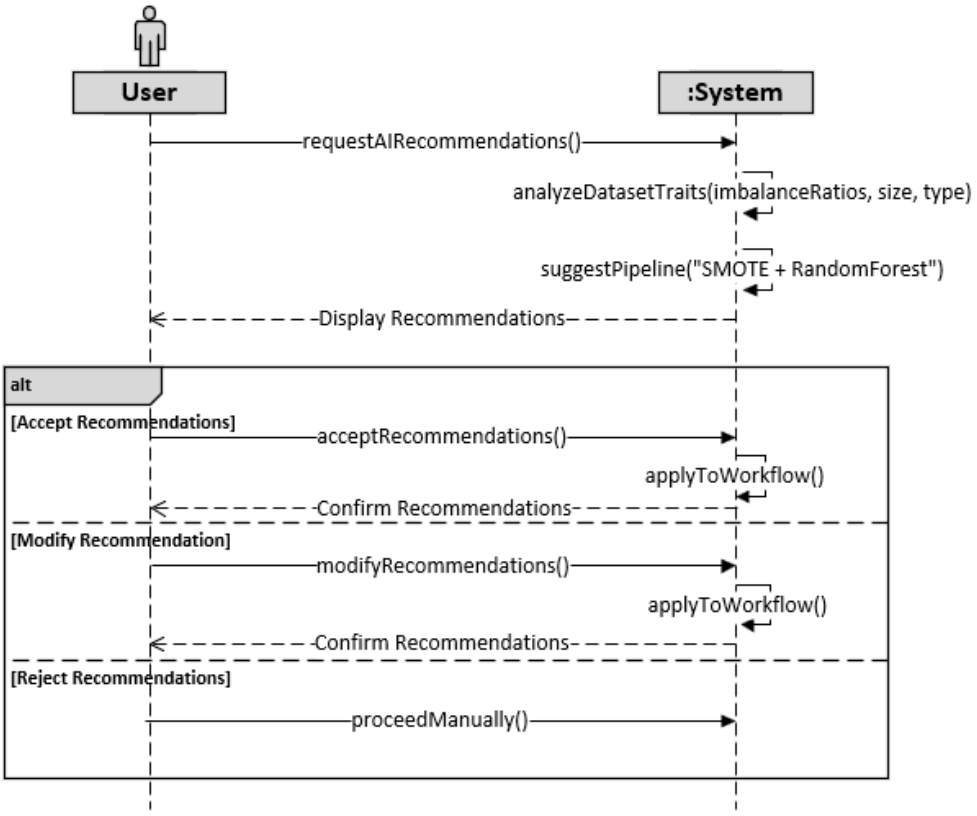

Figure 7: SSD Get AI Recommendations

5.6.Manage Workflows

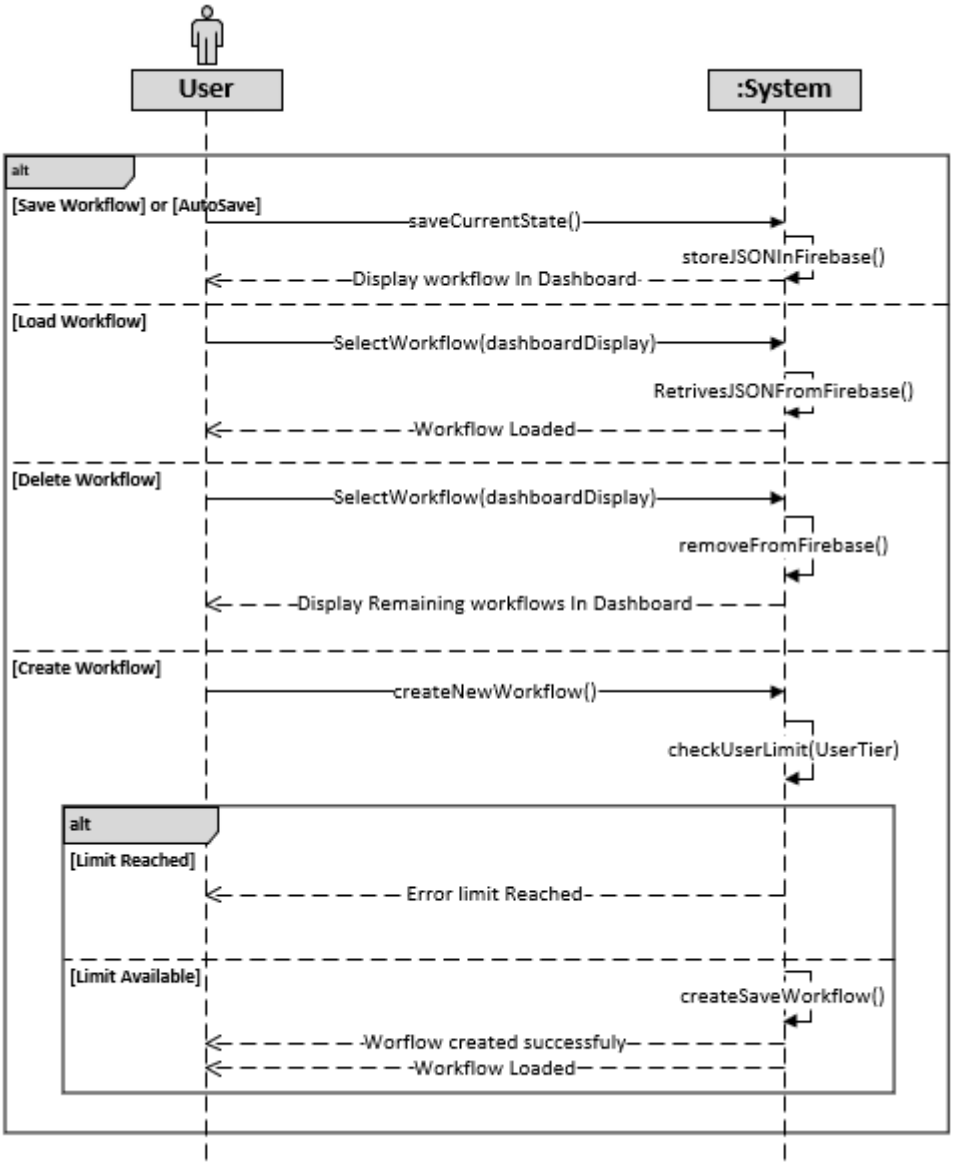

Figure 8: SSD Manage Workflows

5.7.Manage Subscription

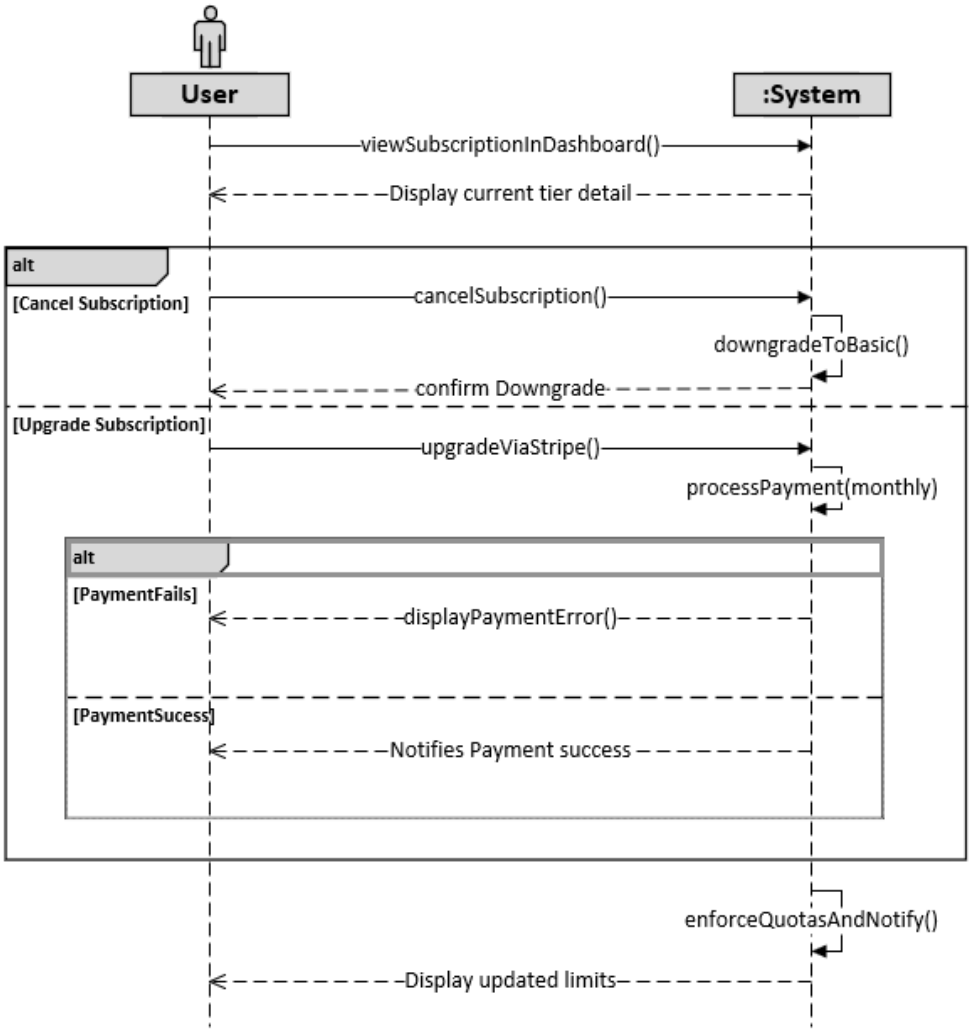

Figure 9: SSD Manage Subscription

6. Data Flow Diagrams

6.1. Level 0 Data Flow Diagram

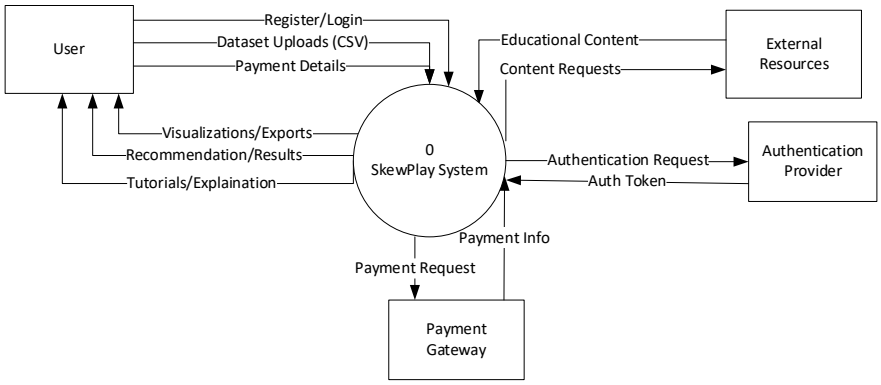

Figure 10: DFD Level 0

6.2.Level 1 Data Flow Diagram

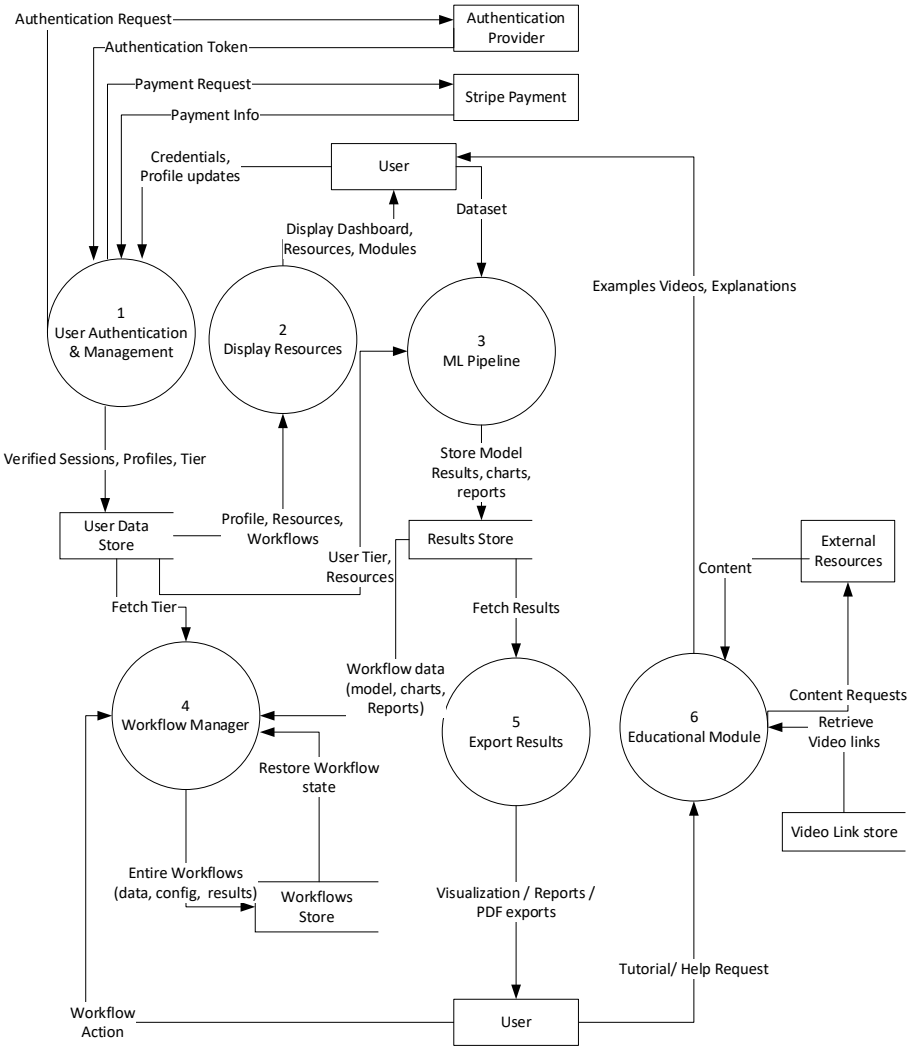

Figure 11: DFD Level 1

6.3.Level 2 Data Flow Diagram

6.3.1. DFD ML Pipeline

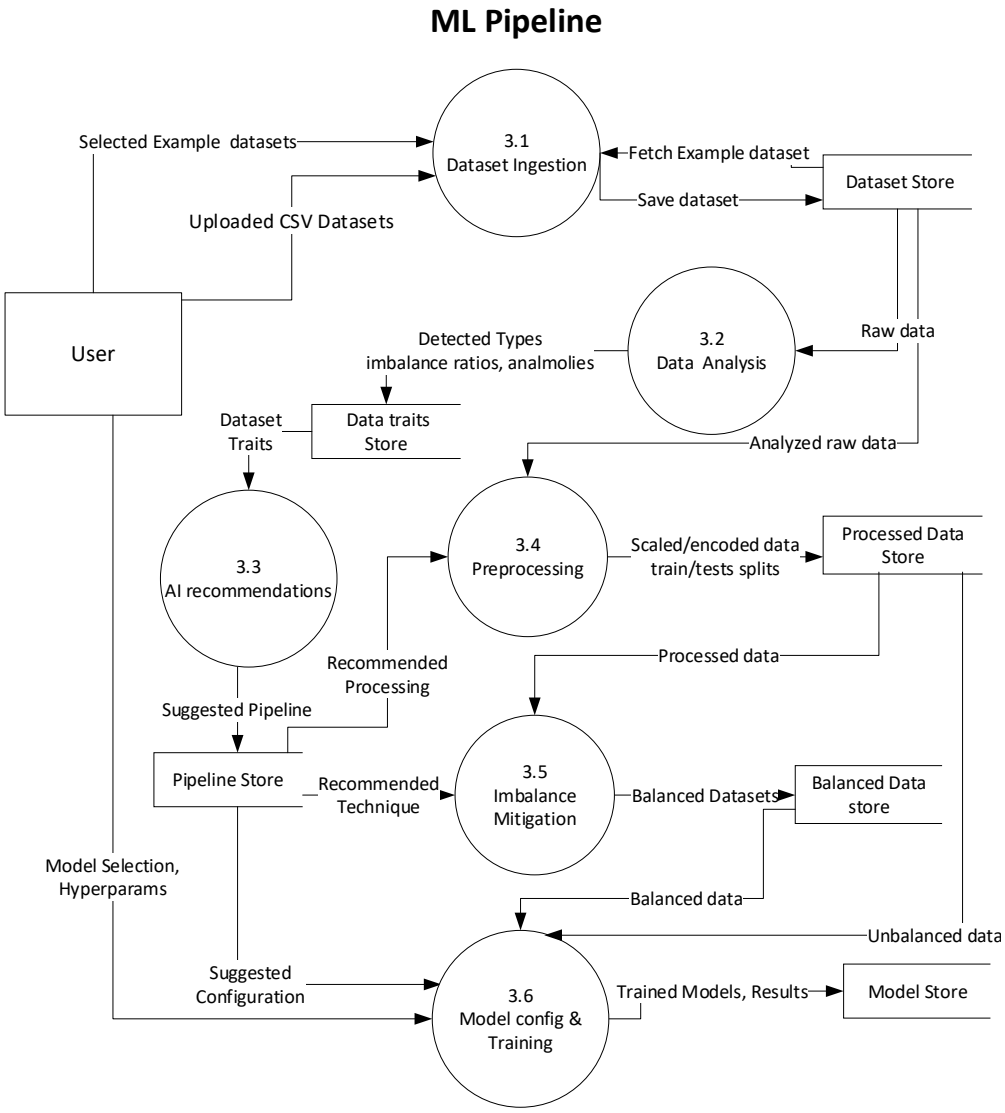

Figure 12: DFD Level 2 -ML Pipeline

7. Component & Package Diagram

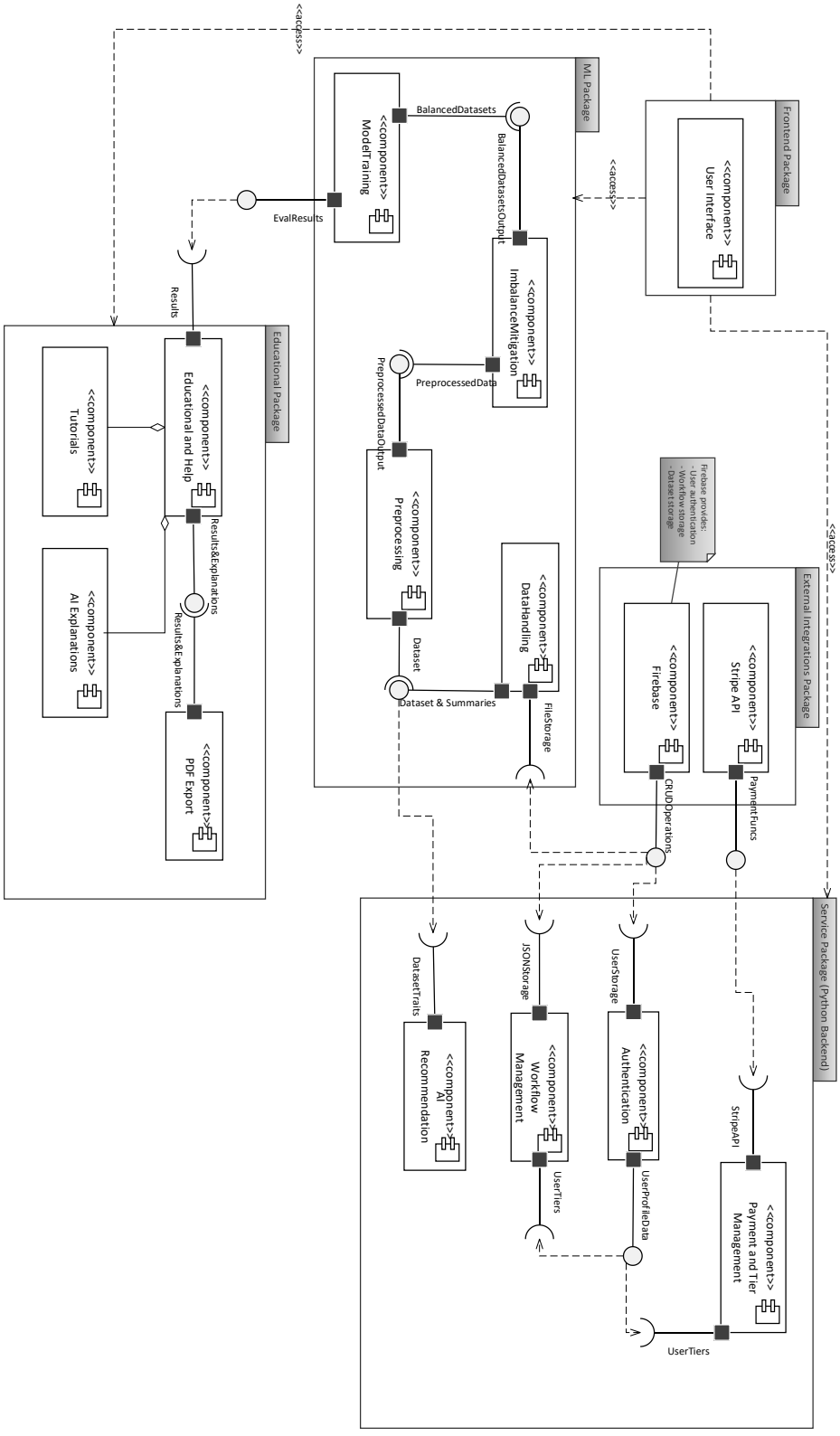

Figure 13 Component & Package Diagram

8. High Level Architecture Diagram

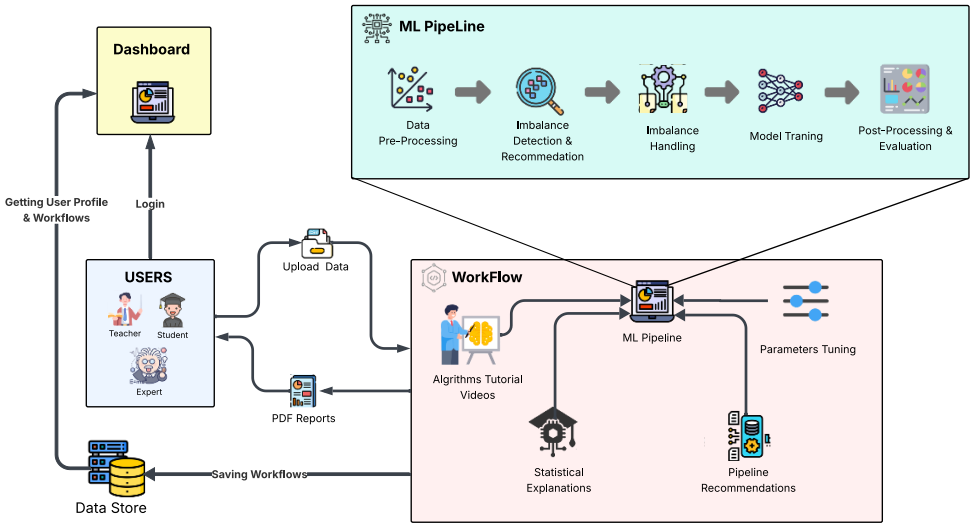

Figure 14: High Level Architecture Diagram

9. References

[1] "TensorFlow Playground," [Online]. Available: https://playground.tensorflow.org/.

[2] "n8n," [Online]. Available: https://n8n.io/.

[3] "KNIME Analytics Platform," [Online]. Available: https://www.knime.com/.

[4] K. Abhishek and M. Abdelaziz , Machine Learning for Imbalanced Data, Packt Publishing, 2023.

[5] J. Demšar, T. Curk, A. Erjavec, Č. Gorup, T. Hočevar, M. Milutinovič, M. Možina, M. Polajnar, M. Toplak, A. Starič, M. Stajdohar, L. Umek, L. Žagar, J. Žbontar, M. Žitnik and B. Zupan, "Orange: Data Mining Toolbox in Python," *Journal of Machine Learning Research*, vol. 14, no. Aug, pp. 2349-2353, 2013.

[6] M. R. Berthold, "KNIME: The Konstanz Information Miner," in *Data Analysis, Machine Learning and Applications*, Springer, 2009, pp. 319-326.

[7] J. Demšar and B. Zupan, "Orange: Data Mining Fruitful and Fun," ResearchGate, 2021.