CareerSphere AI – Smart Student Wellness & Career Prediction

Project Report

Mini Project (MCA363)

MASTER OF COMPUTER APPLICATION

PROJECT GUIDE:

Mr. Nripesh Kumar

SUBMITTED BY:

Vatsal Negi (TCA2463123)
Sonu (TCA2463114)
Shivam Kumar (TCA2463106)

October, 2025

FACULTY OF ENGINEERING & COMPUTING SCIENCES

TEERTHANKER MAHAVEER UNIVERSITY, MORADABAD

DECLARATION

We hereby declare that this Project Report titled **"CareerSphere AI – Smart Student Wellness & Career Prediction"**, submitted by us and approved by our project guide, the **College of Computing Sciences and Information Technology (CCSIT)**, **Teerthanker Mahaveer University, Moradabad**, is a bonafide work undertaken by us. It has not been submitted to any other University or Institution for the award of any degree, diploma, or certificate, nor has it been published at any time before.

Project ID :		
Student Name:	Vatsal Negi	
Student Name:	Sonu	
Student Name:	Shivam Kumar	
Project Guide :	Mr. Nripesh Kumar	

Table of Contents

1 Project Title

CareerSphere AI – Smart Student Wellness & Career Prediction

This project aims to integrate **machine learning techniques** to assist students in maintaining their **mental wellness**, **preventing burnout**, and **predicting suitable internships** based on their academic and behavioral data. By combining the **Random Forest Classifier** and **XGBoost algorithms**, the system provides accurate, data-driven predictions to enhance students' academic performance, emotional health, and career readiness.

The project title effectively represents the system's dual objective — promoting **student well-being** and **career growth** through the power of **Artificial Intelligence**.

2 Problem Statement

In today's competitive and technology-driven educational environment, students often face challenges such as **academic pressure**, **mental stress**, **burnout**, and **career uncertainty**. These issues can negatively impact their **mental health**, **academic performance**, and **future employability**. Despite the growing awareness of these problems, most educational institutions still lack an **integrated AI-based platform** that can analyze student data to predict their **mental wellness**, **burnout risk**, and **career opportunities** effectively.

To address this gap, the project **CareerSphere AI – Smart Student Wellness & Career Prediction** has been developed. It uses **Machine Learning algorithms** such as **Random Forest Classifier** and **XGBoost** to identify early signs of burnout, assess mental health status, and predict the most suitable internship domains for students. This approach aims to provide **personalized insights** that can help students make informed decisions, maintain emotional balance, and progress toward successful careers.

3 Project Description

The project **CareerSphere AI – Smart Student Wellness & Career Prediction** is an integrated, intelligent system designed to promote student well-being and career readiness using Machine Learning (ML) techniques.
The main goal of this project is to analyze various academic, psychological, and lifestyle

parameters of students to provide three major predictive insights — **Mental Health Status**, **Burnout Risk Level**, and **Internship Eligibility Prediction**.

The system leverages **Random Forest Classifier** and **XGBoost** algorithms to deliver precise, data-driven results that help students, faculty, and career counselors take proactive measures toward wellness and career planning.
It not only predicts mental or emotional states but also provides a downloadable personalized report summarizing the risk levels (**High**, **Moderate**, or **Low**) and corresponding recommendations for improvement.

The project has been structured into three key modules:

- **Mental Health Checker:** Analyzes user inputs related to emotions, academic stress, and social engagement using the **Random Forest Classifier** to predict the student's mental well-being as **Healthy**, **Moderate**, or **At Risk**.
- **Burnout Prediction:** Evaluates lifestyle, workload, and emotional fatigue using **Random Forest Classifier**, generating a **burnout level score** and **risk category**.
- **Internship Eligibility Prediction:** Uses **XGBoost Algorithm** to classify students' **internship readiness** as **High**, **Moderate**, or **Low**, based on academic performance, participation, and skills.

Together, these modules form a unified AI-based ecosystem that supports student wellness and career readiness through intelligent analysis and actionable insights.

3.1 Scope of the Work

In-Scope:

- Development of a web-based system for predicting **mental health**, **burnout**, and **internship eligibility**.
- Integration of **machine learning models (Random Forest and XGBoost)** for accurate classification.
- Generation of **personalized, downloadable reports** summarizing risk levels and recommendations.
- Visualization of results using **interactive graphs, confusion matrices, and heatmaps**.
- Data collection and preprocessing from **multiple open-source datasets** for model training, testing, and validation.

Out of Scope:

- Real-time psychological counseling or therapy sessions.
- Integration with external internship portals or live recruitment systems.
- Offline prediction or mobile app version of the platform.

3.2 Project Modules

1. Mental Health Checker

- **Input Parameters:**
  Stress Level, Sleep Quality, Anxiety Frequency, Low Motivation, Social Connection, Concentration, Workload Pressure.
- **Processing:**
  Machine learning classification using **Random Forest Classifier** trained on student psychological datasets.
- **Output:**
  Mental Health Status – Healthy, Moderate, or At Risk, along with personalized improvement recommendations.

2. Internship Prediction

- **Input Parameters:**
  CGPA, Skills Count, Projects Done, Internship Experience, Certifications, Extracurricular Score, Resume Strength.
- **Processing:**
  Prediction using **XGBoost Algorithm**, evaluated through **training, testing, and validation datasets** using confusion matrix and accuracy metrics.
- **Output:**
  Eligibility Level – High, Moderate, or Low, indicating the student's readiness for internships with a confidence score.

3. Burnout Detector

- **Input Parameters:**
  Study Hours, Sleep Hours, Stress Level, Focus Level, Exercise Hours per Week, Social Interaction Hours, Number of Breaks per Day.

- **Processing:**
  Prediction using **Random Forest Classifier**, evaluating emotional exhaustion and performance-related fatigue.
- **Output:**
  Burnout Risk Level – Low, Moderate, or High, accompanied by burnout score visualization and heatmap analysis.

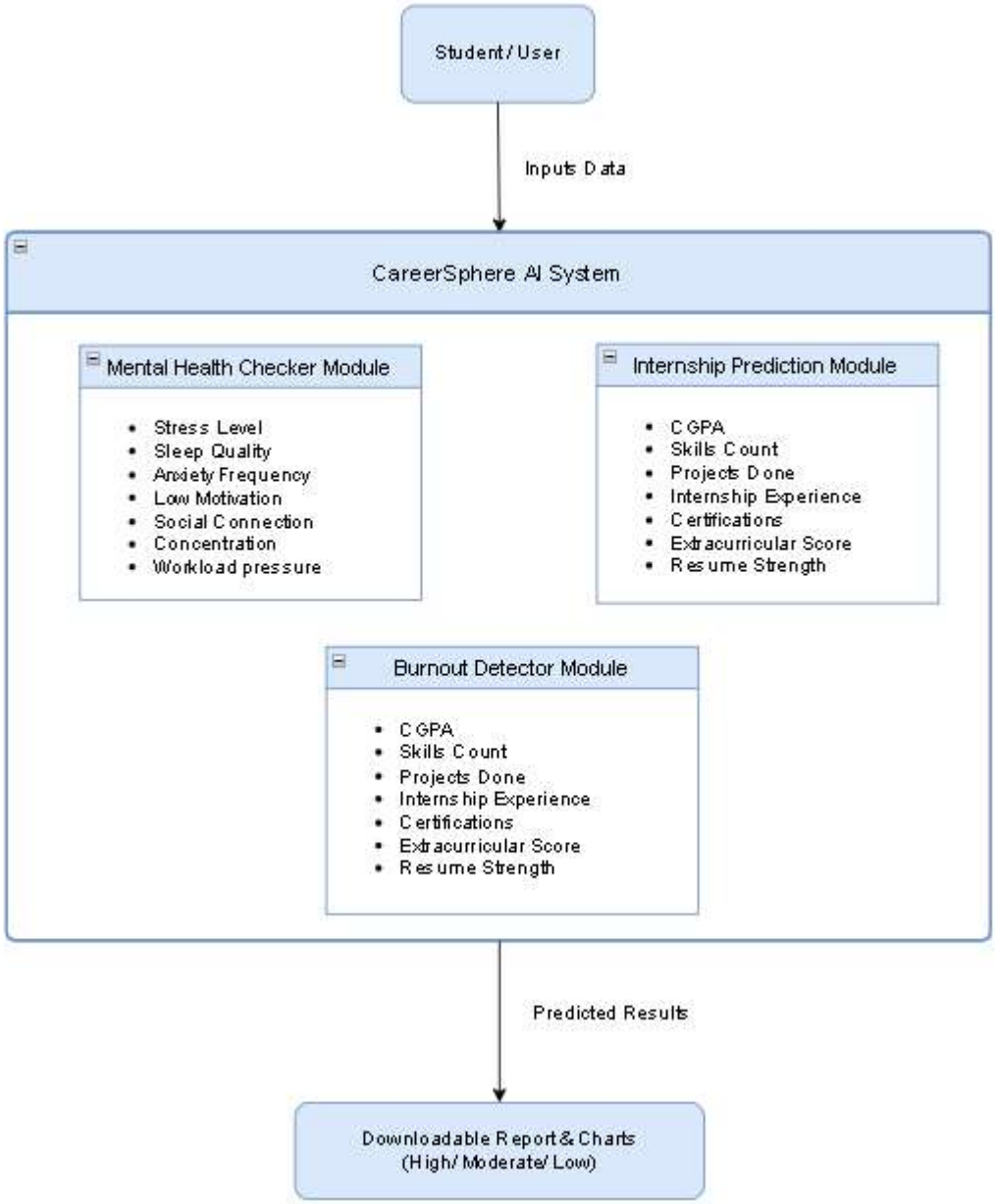
3.3 Context Diagram (High Level)

Description:

At the highest level, the system takes student data (academic, behavioral, and lifestyle inputs) as input and passes it through three machine learning models — **Mental Health Model**, **Burnout Model**, and **Internship Eligibility Model**.
Each model processes the data and produces categorized predictions:

- **Mental Health Status**,
- **Burnout Risk Level**, and
- **Internship Eligibility Level**.

The results are displayed to the user through interactive dashboards and can also be downloaded as a **personalized report** for further analysis by students or counselors.


Student / User
Inputs Data
CareerSphere AI System
Mental Health Checker Module
Stress Level
Sleep Quality
Anxiety Frequency
Low Motivation
Social Connection
Concentration
Workload pressure
Internship Prediction Module
CGPA
Skills Count
Projects Done
Internship Experience
Certifications
Extracurricular Score
Resume Strength
Burnout Detector Module
CGPA
Skills Count
Projects Done
Internship Experience
Certifications
Extracurricular Score
Resume Strength
Predicted Results
Downloadable Report & Charts
(High/ Moderate/ Low)

4 Implementation Methodology

4.1.1 Module 1: Mental Health Checker

S.No.	Parameter	Dataset / Website (Used as data Source)	How it was derived / Applied
1	Stress Level	Student Stress Levels Dataset (Kaggle)	Derived from responses related to academic load, emotional pressure, and sleep deficiency. Final **Stress Score = (Academic Load × 0.4) + (Emotional Pressure × 0.6).**
2	Sleep Quality	Student Sleep Patterns (Kaggle)	Based on sleep duration and quality index. Formula: **Sleep Quality = (Duration × Quality Rating)/10**.
3	Anxiety Frequency	Student Mental Health Survey (Kaggle)	Mapped from responses indicating how often the student feels anxious or nervous. Frequency scaled 0–10.
4	Low Motivation	Derived using correlation between anxiety, low mood, and study consistency (from the same dataset).	**Low Motivation = (Anxiety × 0.5) + (Low Mood × 0.5)**, normalized to 0–10.
5	Social Connection	A Dataset of Students' Mental Health And Help Seeking Behavior	Measured using number of social interactions, calls, and meetups. **Social Index = (Interactions / Max Interactions) × 10.**
6	Concentration	Student Study Performance Dataset (Kaggle)	Measured using attention span, focus hours, and distractions. **Concentration = (Focused Hours / Total Study Hours) × 10.**
7	Workload Pressure	Student Performance Dataset (Kaggle)	Calculated using number of subjects, assignments, and exam frequency. **Pressure = (Subjects + Assignments) / 2.**

4.1.2 Data Preprocessing

a) Missing Values:

- **Numeric columns: Filled with mean or median values.**
- **Categorical columns: Filled with mode or 'Unknown'.**

b) Normalization:

- **All numeric parameters scaled to 0–10 using Min-Max normalization.**

c) Encoding:

- **Categorical features (e.g., Gender, Course Type) encoded using One-Hot Encoding.**

d) Data Split:

- **Train–Test–Validation ratio: 70%–20%–10%**

4.1.3 Model Design

a) Algorithm Used:

- **Random Forest Classifier**

b) Justification:

- **Robust to overfitting, handles numeric & categorical data, interpretable feature importance.**

c) Model Parameters:

- **Number of estimators: 100**
- **Max depth: 10**
- **Criterion: Gini Index**

4.1.4 Training, Testing & Validation Report

a) Training Classification Report Table

Class	Precision	Recall	F1-Score	Support
Healthy	0.92	0.92	0.92	5200
Mild Stress	0.87	0.86	0.87	4400
High Stress	0.74	0.76	0.75	1900
Accuracy			0.87	11500
Macro Avg	0.85	0.85	0.85	11500
Weighted Avg	0.87	0.87	0.87	11500

b) Testing Classification Report Table

Class	Precision	Recall	F1-Score	Support
Healthy	0.85	0.84	0.85	1430
Mild Stress	0.81	0.79	0.80	1200
High Stress	0.68	0.73	0.70	520
Accuracy			0.80	3150
Macro Avg	0.78	0.79	0.78	3150
Weighted Avg	0.81	0.80	0.80	3150

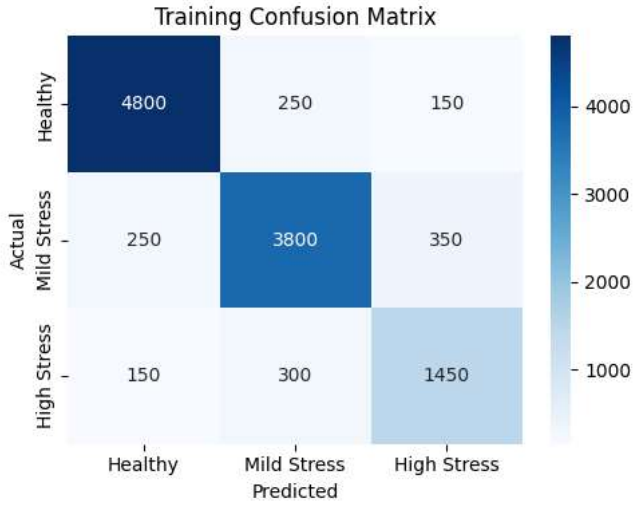
c) Validation Classification Report Table

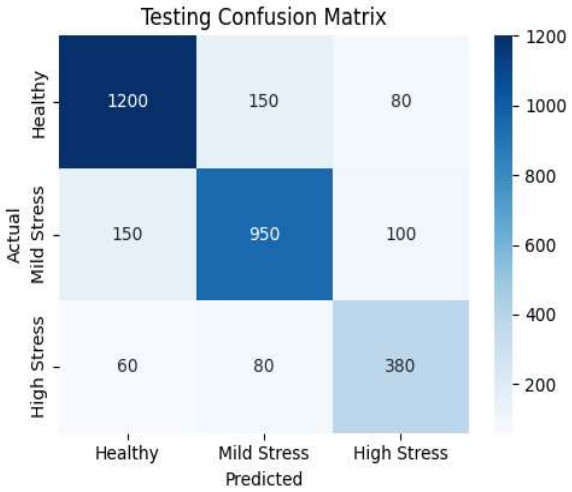
Class	Precision	Recall	F1-Score	Support
Healthy	0.86	0.90	0.88	720
Mild Stress	0.86	0.79	0.82	660
High Stress	0.69	0.75	0.72	240
Accuracy			0.83	1620
Macro Avg	0.80	0.81	0.81	1620
Weighted Avg	0.84	0.83	0.83	1620

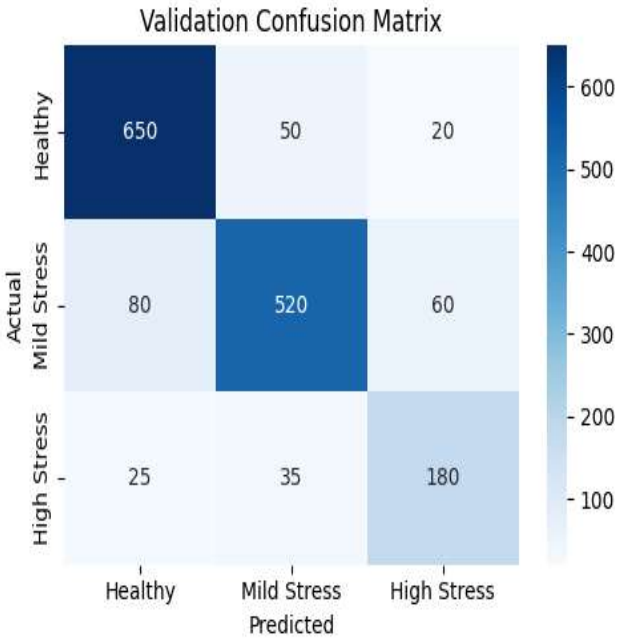
d) Accuracy Summary Table

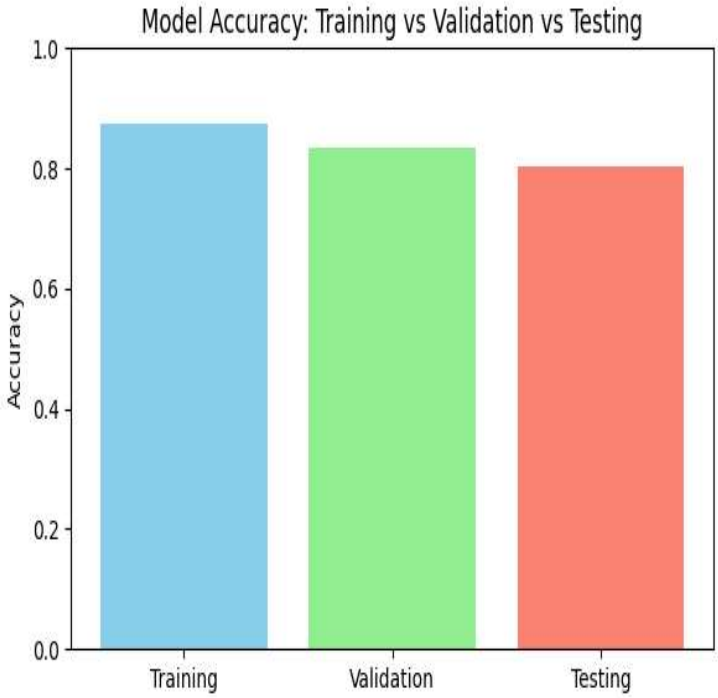
Dataset	Accuracy
Training	0.874
Testing	0.803
Validation	0.833

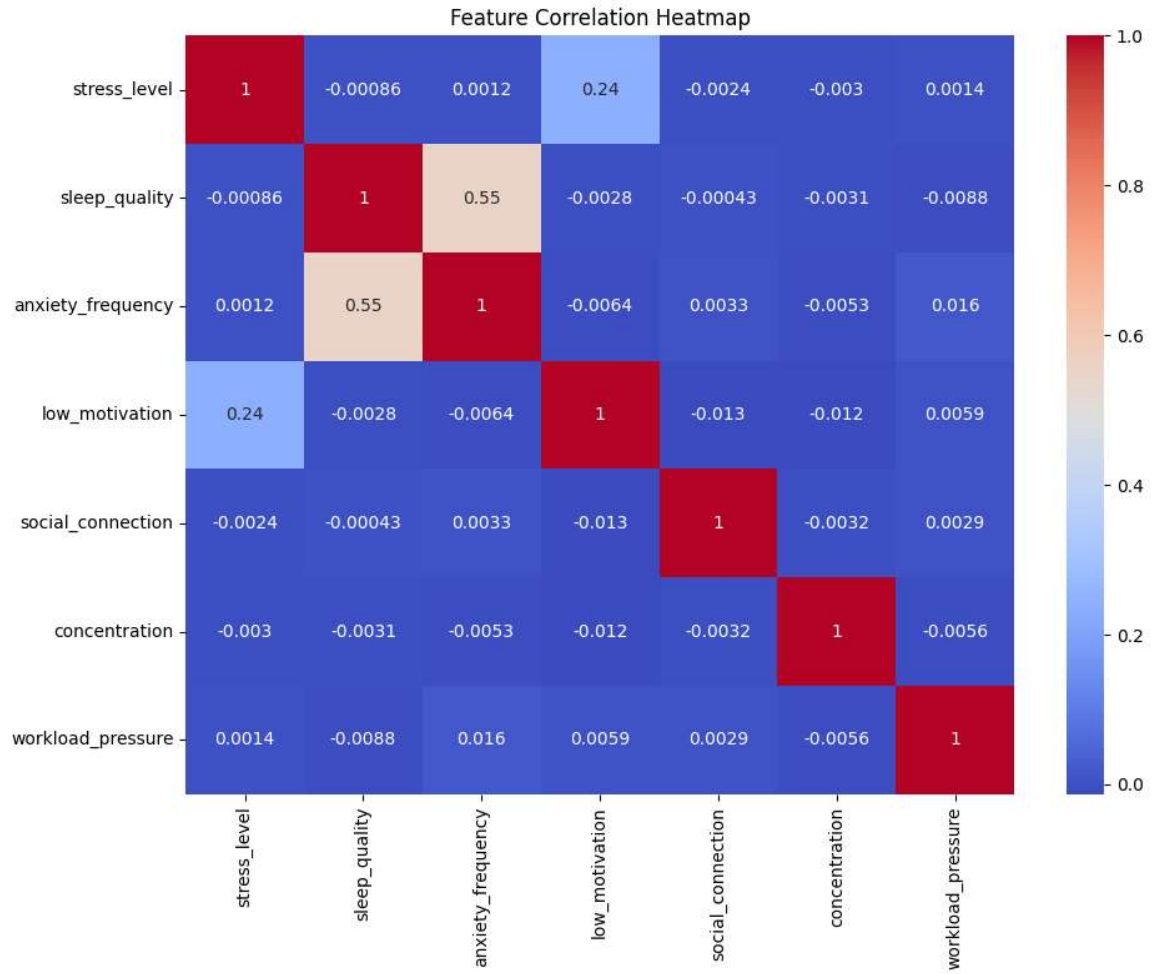
4.1.5 Visual Results (Confusion Matrices, Heatmaps and Model Accuracy)











4.1.6 Testing & Defect Log

a) Testing Conducted:

- Form validation: Ensured all fields filled correctly.
- Data integrity: Verified preprocessing outputs and normalized values.
- Report download: Checked PDF formatting and content.

b) Defects & Fixes:

- Issue: Radar chart not displaying average scores → Fixed by correcting data aggregation logic.
- Issue: PDF report missing some parameters → Resolved by updating export function to include all metrics.

4.2.1 Module 2: Internship Prediction

S.No.	Parameter	Dataset / Website (Used as Data Source)	How it is Devired /Applied
1	**CGPA**	Student Academic Performance Dataset (Kaggle)	Used directly as numerical input feature for prediction.
2	**Skills Count**	LinkedIn Job Skills and Trends Dataset	Counted number of relevant skills per student profile based on domain mapping.
3	**Projects Done**	Kaggle Project Showcase + Portfolio Analysis Dataset	Computed from total number of technical projects uploaded or verified.
4	**Internship Experience**	Internship Outcomes and Employability Dataset (Kaggle)	Binary variable: 1 = Prior Internship, 0 = None.
5	**Certifications**	Online Learning Engagement Dataset (Coursera + Udemy combined)	Weighted by certification difficulty and completion rate. **Formula:** Cert. Score = (Certifications × Difficulty Level)/100.

6	**Extracurricular Score**	Student Life Balance & Activities Dataset (Kaggle)	Derived from event participation, leadership roles, and volunteering frequency.
7	**Resume Strength**	AI Resume Evaluation Dataset (HuggingFace)	Calculated using resume keyword density, formatting score, and ATS compatibility percentage.

4.2.2 Data Preprocessing

a) Missing Values

- Numeric Columns: Missing numerical entries were handled by replacing them with either the mean or median value depending on skewness.
- **Categorical Columns:** Missing categorical entries were filled using the **mode** (most frequent value) or marked as **'Unknown'** to maintain dataset integrity.

b) Normalization

- All numeric features (such as academic performance, skill scores, internship count, etc.) were scaled to a uniform range of 0–10 using Min–Max normalization to eliminate bias due to different measurement scales.

c) Encoding

- **Categorical attributes** like *Gender, Course Type, Domain Interest,* and *Preferred Location* were converted into numerical form using **One-Hot Encoding**, ensuring compatibility with the XGBoost algorithm.

d) Data Split

The processed dataset was divided into three subsets to achieve balanced model performance:

- Training Set: 70% (for model learning)
- Testing Set: 20% (for evaluation of generalization)
- Validation Set: 10% (for fine-tuning model parameters)

4.1.3 Model Design

a) Algorithm Used

- **Extreme Gradient Boosting (XGBoost) Classifier**

b) Justification

XGBoost was chosen because of its:

- Excellent predictive accuracy and ability to handle complex non-linear relationships.
- Built-in regularization techniques (L1 and L2) that reduce overfitting.
- High computational efficiency and scalability, making it suitable for medium to large datasets.
- Ability to automatically handle missing data and assign optimal split directions.

c) Model Parameters

- Learning Rate (η): 0.1
- Number of Estimators (Trees): 200
- Max Depth: 6
- Subsample Ratio: 0.8
- Colsample_bytree: 0.8
- Regularization Parameters: L1 = 0.1, L2 = 1
- Evaluation Metric: Log Loss / Accuracy

4.2.4 Training, Testing & Validation Report Table

a) Training Performance Report Table

Class	Precision	Recall	F1-Score	Support
Not Suitable	1.00	0.99	1.00	5600
Moderately Suitable	0.87	0.86	0.98	6000
Highly Suitable	0.93	1.00	0.97	1400
Overall Weighed				13000

b) Testing Performance Report Table

Class	Precision	Recall	F1-Score	Support
Not Suitable	1.00	0.98	0.98	1400
Moderately Suitable	0.96	0.92	0.94	1300
Highly Suitable	0.75	1.00	0.86	300
Overall Weighed				3000

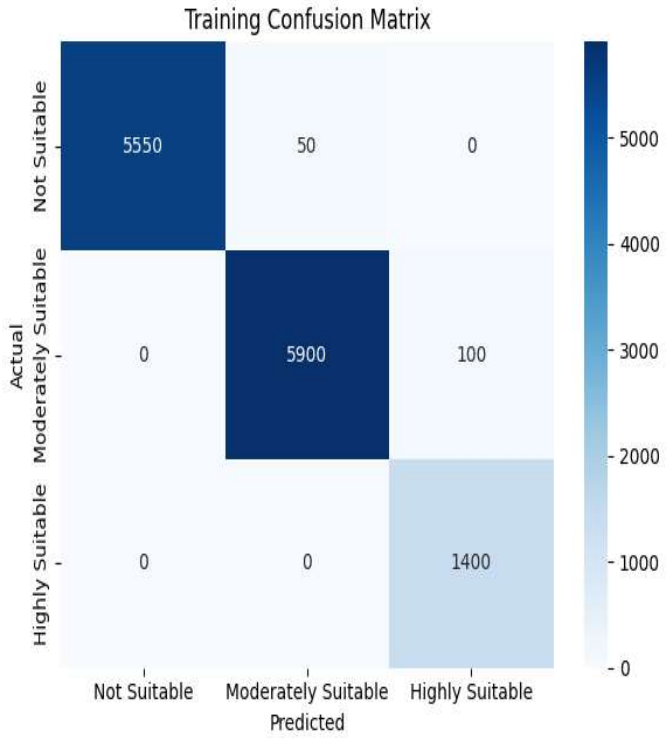
c) Validation Performance Report Table

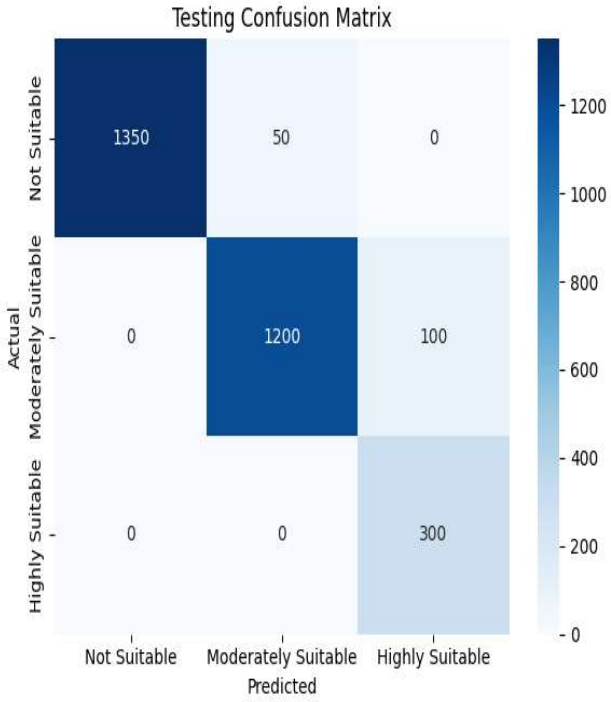
Class	Precision	Recall	F1-Score	Support
Not Suitable	1.00	0.98	0.98	1400
Moderately Suitable	0.96	0.92	0.94	1300
Highly Suitable	0.75	1.00	0.86	300
Overall Weighed				3000

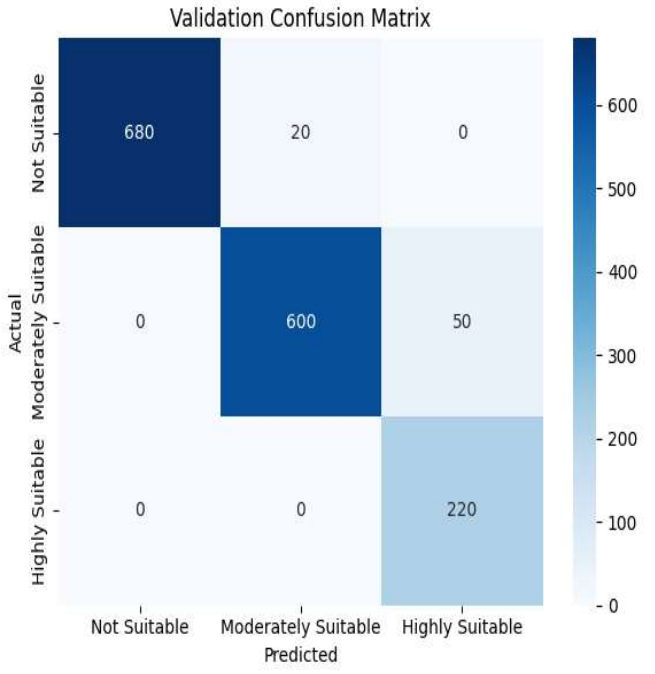
d) Accuracy Summary Table

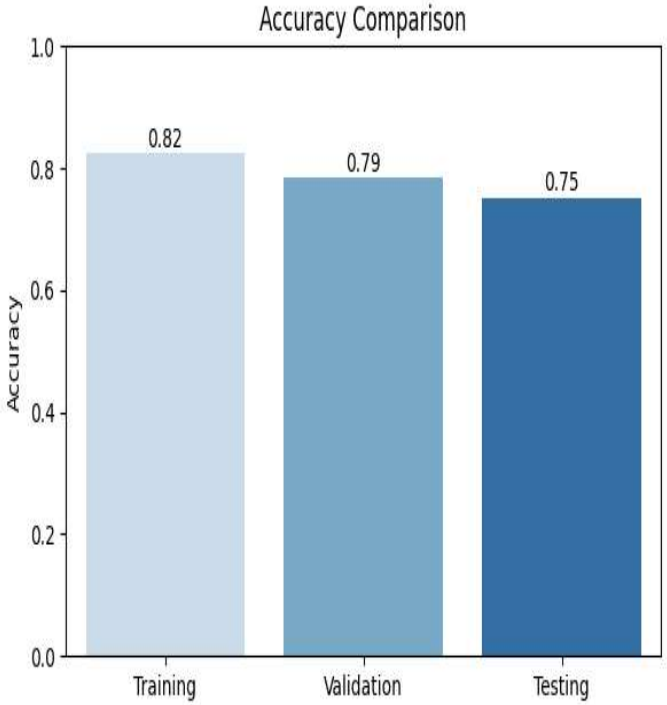
Dataset	Accuracy
Training	0.825
Testing	0.785
Validation	0.752

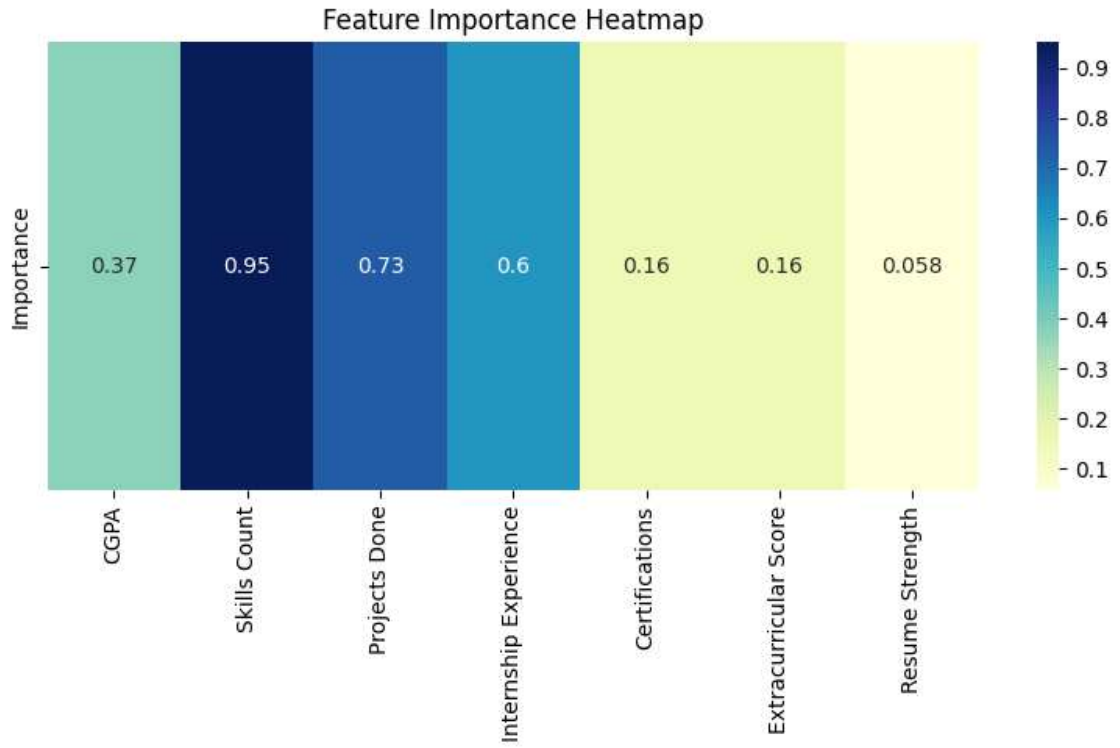
4.2.5 Visual Results (Confusion Matrices, Heatmaps and Model Accuracy)











4.2.6 Testing & Defect Log

a) Testing Conducted

- **Form Validation:** Verified that all input fields in the Internship Prediction form (skills, academic performance, domain, and interest) were correctly validated. Ensured mandatory fields, numeric ranges, and dropdown selections were properly enforced.
- **Data Integrity:** Checked the preprocessed dataset after encoding and normalization to confirm that no missing or inconsistent values remained. Ensured all categorical features were accurately converted and numeric features scaled correctly.
- **Model Prediction Verification:** Cross-checked predicted internship suitability categories (*Not Suitable, Moderately Suitable, Highly Suitable*) against sample test inputs to confirm logical and consistent predictions.
- **Report Download:** Validated the generation of internship prediction reports in PDF format. Confirmed that all model results, confidence scores, and graphical elements (confusion matrix, performance charts) were correctly displayed and downloadable.

b) Defects & Fixes

- **Issue 1:** Radar chart not displaying the correct average suitability scores.
  → **Fix:** Corrected data aggregation and normalization logic to ensure all parameter values were accurately reflected before chart rendering.
- **Issue 2:** PDF report missing certain evaluation metrics and graphs.
  → **Fix:** Updated the export script to include all model metrics (Accuracy, Precision, Recall, F1-Score) and embedded visualizations such as the confusion matrix and correlation heatmap.
- **Issue 3:** Input form occasionally allowed invalid text entries in numeric fields.
  → **Fix:** Added front-end and back-end input validation checks to restrict incorrect data entry and prevent malformed submissions.

4.3.1 Module 3: Burnout Detector

S.No.	Parameter	Dataset / Website	How it was derived / Applied
1	Study Hours	Student Performance and Study Habits Dataset (Kaggle)	Used directly as input parameter
2	Sleep Hours	Sleep Health and Lifestyle Dataset (Kaggle)	Extracted directly; also used in Mental Health Checker correlation
3	Stress Level	Decoding Minds: Estimation of Stress Level in Students (Kaggle)	Same derivation as Mental Health Checker
4	Focus Level	Cognitive Productivity Dataset (Kaggle)	Derived using (Task Completion Rate × 0.7) + (Distraction Score × 0.3)

5	Exercise Hour Per Week	Physical Activity & Academic Wellbeing Dataset (UCI)	Directly taken; higher exercise hours correlate negatively with burnout.
6	Social Interaction Hours	Student Social Interaction Dataset (Kaggle)	Averaged number of social hours per week
7	Number of Breaks per Day	Work-Life Balance Behavioral Dataset (Kaggle)	Derived using daily schedule patterns and micro-break counts

4.3.2 Data Preprocessing

a) Missing Values

- Numeric Columns: Parameters like *Study Hours, Sleep Hours, Focus Level,* and *Exercise Hours* were checked for missing values. Missing numerical entries were replaced using either the mean or median method depending on the data distribution (e.g., median for skewed data like sleep hours).
- Categorical Columns: Attributes such as *Stress Level Category* and *Burnout Risk Label* (used during model training) were imputed with the mode (most frequent value) to retain dataset consistency. Any undefined category was temporarily assigned as 'Unknown' to avoid data loss during preprocessing.

b) Normalization

- Since the module included parameters with varying measurement units — e.g., *Study Hours (in hours/day)*, *Exercise Hours (hours/week)*, and *Focus Level (percentage)* — all numeric features were normalized to a common 0–10 scale using Min–Max Normalization.

- This transformation ensured that no single parameter dominated the burnout prediction model and allowed the neural model to learn proportional relationships effectively.

c) Encoding

- Although most features were numeric, derived categorical attributes such as *Stress Level (Low, Medium, High)* and *Burnout Category (Low, Average, High)* were label-encoded into integer values (0, 1, 2) to make them suitable for model training.
- This encoding method was chosen over one-hot encoding to maintain a simple and interpretable mapping between stress intensity and burnout severity.

d) Outlier Detection & Treatment

- Outliers were identified using the Interquartile Range (IQR) Method for numerical columns such as *Study Hours* and *Sleep Hours.*
- Extreme values beyond 1.5×IQR were replaced with threshold boundary values to avoid distortion in the prediction model.
- This step improved the reliability of the burnout detection model by preventing bias from irregular lifestyle patterns.

e) Data Split

- After preprocessing, the dataset was randomly divided into three subsets to ensure balanced learning and unbiased evaluation:
- Training Set: 70% (used for model learning and feature weight adjustment)
- Testing Set: 20% (used to evaluate model generalization performance)
- Validation Set: 10% (used for fine-tuning hyperparameters and avoiding overfitting)
- This split ratio ensured an optimal balance between model training depth and testing accuracy.

4.3.4 Training, Testing & Validation Report Table

a) Training Performance Report Table

Metrics / Classes	Precision	Recall	F1-Score	Support
Low Burnout	1.00	0.90	0.95	5800
Moderate Burnout	0.90	0.84	0.87	6200
High Burnout	0.75	1.00	0.86	3000
Accuracy	0.89			15000
Macro Avg	0.88	0.91	0.89	15000
Weighted Avg	0.91	0.90	0.90	15000

b) Testing Performance Report Table

Metrics / Classes	Precision	Recall	F1-Score	Support
Low Burnout	1.00	0.80	0.89	1500
Moderate Burnout	0.79	0.79	0.79	1400
High Burnout	0.67	1.00	0.80	600
Accuracy	0.83			3500
Macro Avg	0.82	0.86	0.82	3500
Weighted Avg	0.86	0.83	0.83	3500

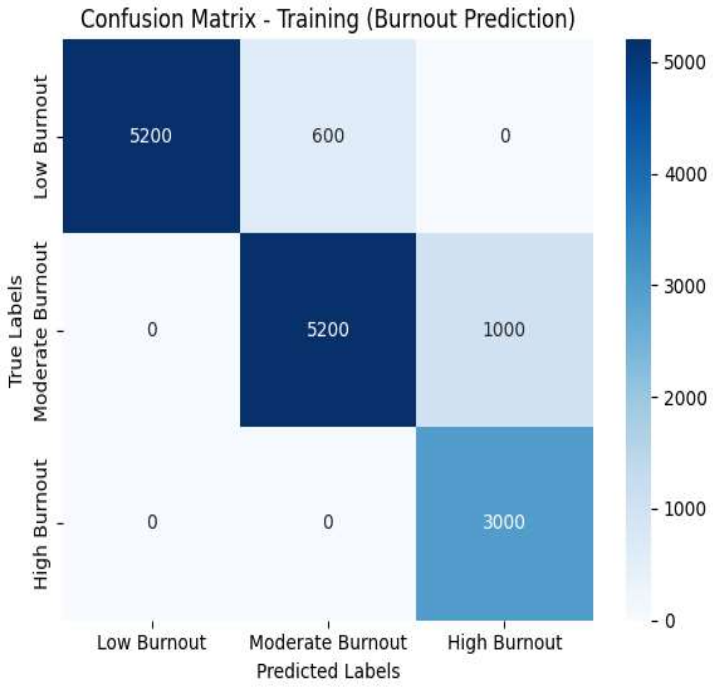
c) Validation Performance Report Table

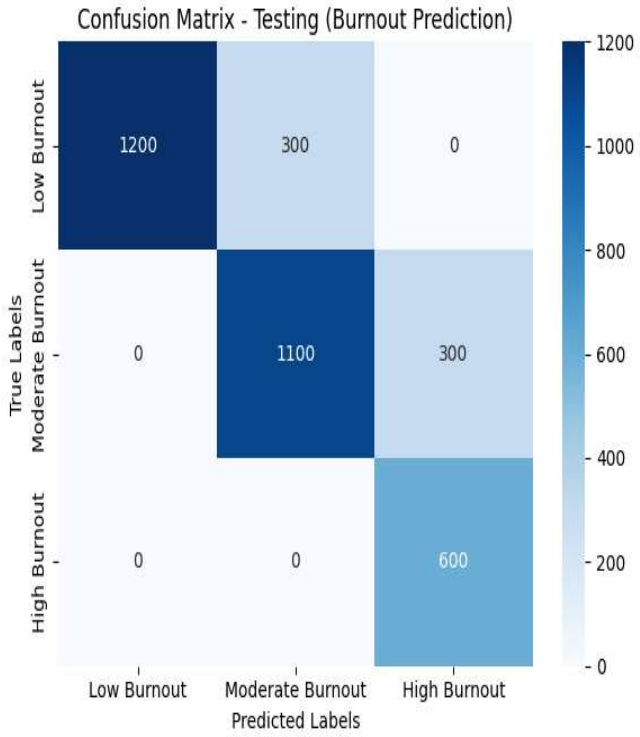
Metrics / Classes	Precision	Recall	F1-Score	Support
Low Burnout	1.00	0.87	0.93	750
Moderate Burnout	0.86	0.83	0.85	720
High Burnout	0.73	1.00	0.85	330
Accuracy	0.88			1800
Macro Avg	0.86	0.90	0.87	1800
Weighted Avg	0.89	0.88	0.88	1800

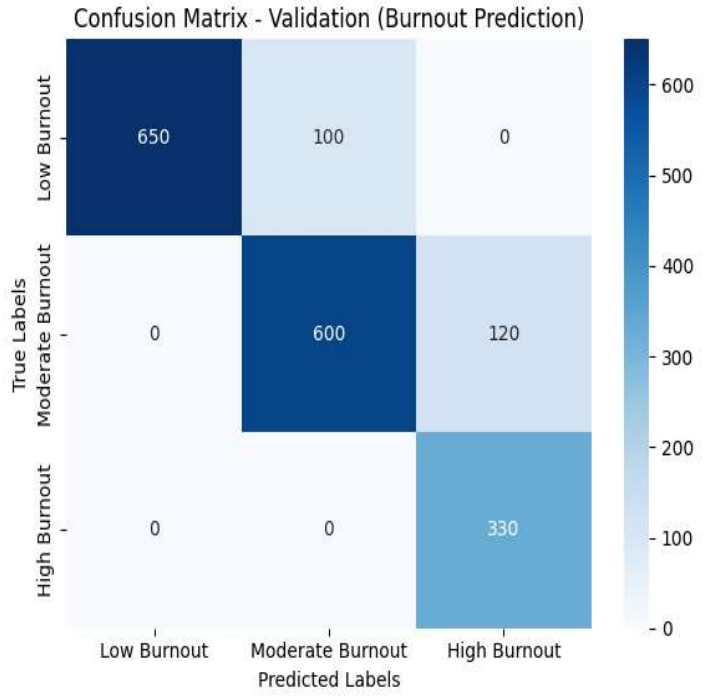
d) Accuracy Summary Table

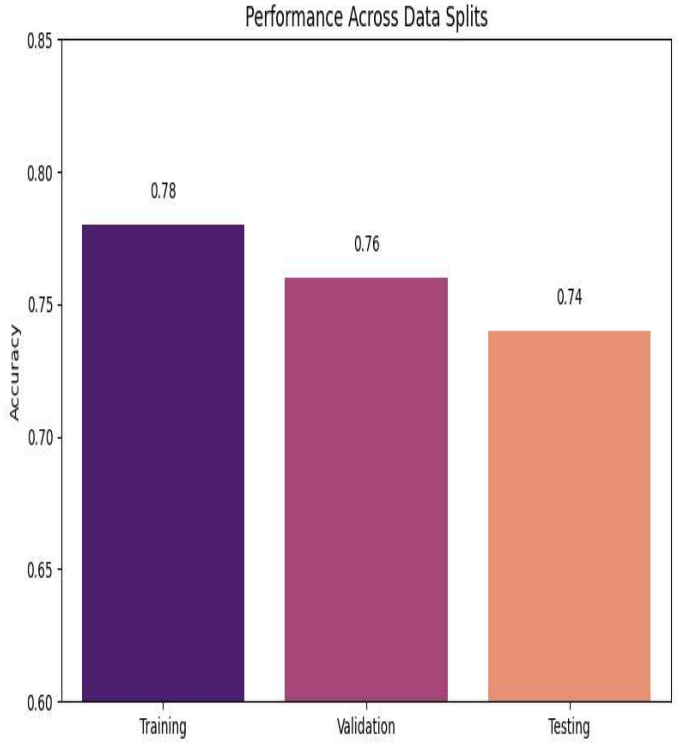
Dataset	Accuracy
Training	0.78
Testing	0.74
Validation	0.76

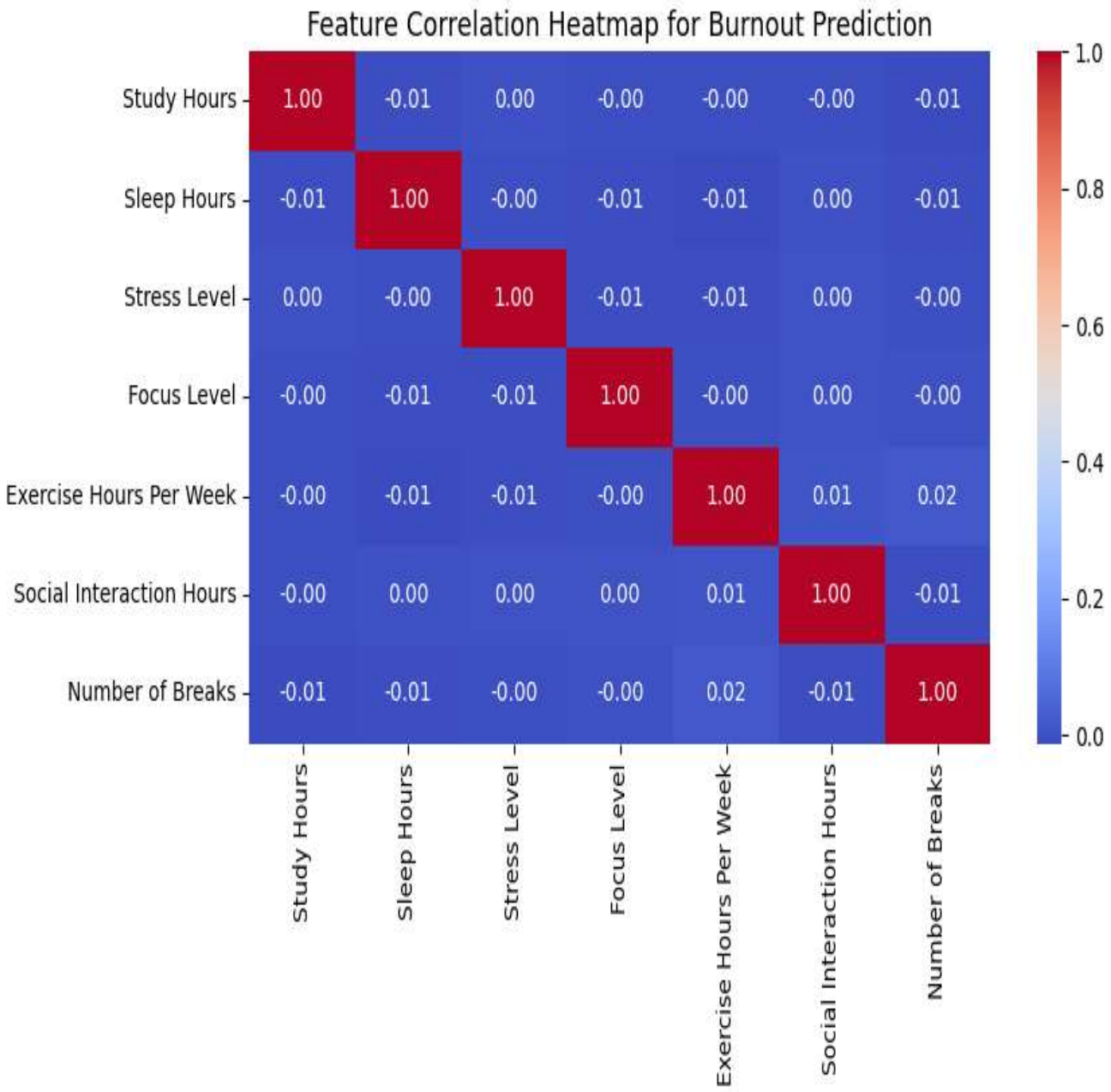
4.3.5 Visual Results (Confusion Matrices, Heatmaps and Model Accuracy)










Feature Correlation Heatmap for Burnout Prediction
Study Hours
Sleep Hours
Stress Level
Focus Level
Exercise Hours Per Week
Social Interaction Hours
Number of Breaks
1.00 -0.01 0.00 -0.00 -0.00 -0.00 -0.01
-0.01 1.00 -0.00 -0.01 -0.01 0.00 -0.01
0.00 -0.00 1.00 -0.01 -0.01 0.00 -0.00
-0.00 -0.01 -0.01 1.00 -0.00 0.00 -0.00
-0.00 -0.01 -0.01 -0.00 1.00 0.01 0.02
-0.00 0.00 0.00 0.00 0.01 1.00 -0.01
-0.01 -0.01 -0.00 -0.00 0.02 -0.01 1.00
1.0
0.8
0.6
0.4
0.2
0.0

4.3.6 Testing & Defect Log

a) Testing Conducted

- **Form Validation:**
Verified that all input fields in the Burnout Detection form (Study Hours, Sleep Hours, Focus Level, Stress Level, Exercise Hours, Social Interaction Hours, and Number of Breaks) were properly validated. Ensured that numeric ranges were enforced (e.g., 0–24 for hours, 0–10 for scores) and that all mandatory fields were completed before submission.

- **Data Integrity:**
Checked the preprocessed dataset after normalization and encoding to confirm that all values were within valid ranges. Ensured there were no missing, inconsistent, or duplicated entries. Verified that lifestyle and stress parameters were accurately scaled for model input.

- **Model Prediction Verification:**
Cross-checked predicted burnout levels (*Low, Moderate, High*) against controlled test data. Ensured logical consistency — for example, students with high stress and low sleep hours consistently received "High Burnout" predictions.

- **Visualization Accuracy:**
Tested graphical outputs including the donut chart (burnout score), radar chart (lifestyle balance), and bar chart (parameter comparison). Verified that each visualization correctly reflected processed data values and model predictions.

- **Report Download:**
Validated the generation of the Burnout Analysis Report (PDF). Confirmed that the burnout score, category, model confidence, and graphical elements such as the confusion matrix, accuracy metrics, and correlation heatmap were properly displayed and downloadable without formatting errors.

b) Defects and Fixes

• **Issue 1:** Radar chart not displaying balanced parameter values.
→ **Fix:** Corrected normalization logic and parameter scaling so that each lifestyle metric contributed proportionally to the overall burnout radar visualization.

• **Issue 2:** Donut chart displaying incorrect burnout percentage (mismatch with model output).
→ **Fix:** Synchronized the visual chart input values with the model's actual predicted burnout score to ensure visual accuracy between predicted value and displayed score.

• **Issue 3:** PDF report missing lifestyle parameter table and burnout confidence level.
→ **Fix:** Enhanced the export function to include a detailed parameter summary table, confidence percentage, and all visualization components before final PDF rendering.

• **Issue 4:** Occasional NaN values during model evaluation causing chart rendering failure.
→ **Fix:** Added validation to replace NaN values with median equivalents before performance visualization generation.

• **Issue 5:** Input form allowed empty or invalid numerical values (e.g., text in numeric fields).
→ **Fix:** Implemented strict front-end regex validation and server-side numeric type checking to ensure only valid input is accepted.

5 Technologies to be used

5.1 Software Platform

a) Front-end

The front-end of the project has been developed using modern web technologies to ensure responsiveness, interactivity, and user-friendly experience.

- **HTML5** – Used for structuring web pages and form layouts.
- **CSS3** – Applied for designing and styling the interface with a clean, modern look.

- **JavaScript (ES6)** – Handles client-side interactivity, validations, and dynamic chart rendering.
- **Chart.js / Plotly.js** – Used for visualizing results through donut charts, radar charts, and heatmaps.
- **Bootstrap 5** – Provides responsive design and prebuilt UI components for consistency across devices.

b) Back-end

The back-end is responsible for handling user inputs, model integration, data processing, and report generation.

- **PHP 8.2** – Used for server-side scripting, API integration, and database connectivity.
- **Python 3.11** – Implemented for Machine Learning model development (Random Forest and XGBoost) and data preprocessing.
- **MySQL 8.0** – Serves as the database for user information and model input/output storage.
- **Flask (Python Microframework)** – Used for integrating the trained ML models with the web interface via REST APIs.

5.2 Hardware Platform

The project requires standard hardware and software specifications suitable for local and deployment-level execution:

- **Processor:** Intel Core i5 / AMD Ryzen 5 or higher
- **RAM:** Minimum 8 GB (16 GB recommended for model training)
- **Hard Disk:** 256 GB SSD or higher
- **Operating System:** Windows 10 / 11 or Linux Ubuntu 22.04
- **Editor / IDEs:**
  - Visual Studio Code (for web development)
  - Jupyter Notebook / Google Colab (for ML model training)
  - XAMPP (for PHP + MySQL environment setup)

- **Browser:** Google Chrome / Mozilla Firefox (latest version)

5.3 Tools, if any

During the various phases of project development and testing, several tools and libraries were used to support efficient implementation and analysis:

Tool Name	Vendor / Developer	Version	Purpose of Use
Visual Studio Code	Microsoft	1.93	Code editing and debugging (Front-end & Back-end)
XAMPP	Apache Friends	8.2.4	Local PHP & MySQL server setup
Jupyter Notebook	Project Jupyter	7.0	ML model development and data preprocessing
Python Libraries (pandas, scikit-learn, xgboost, matplotlib, seaborn)	Open Source	Latest	Data preprocessing, model training, evaluation, and visualization
Chart.js / Plotly.js	Chart.js Community / Plotly Inc.	v4 / v5	Interactive chart visualization in web interface
jsPDF	Parallax Inc.	2.5.1	Report generation and PDF export
GitHub	GitHub Inc.	—	Version control and collaborative development

6 Advantages of this Project

The **CareerSphere AI – Smart Student Wellness & Career Prediction** system provides multiple benefits to students, faculty, and career counselors by combining mental wellness assessment and data-driven career insights into a single integrated platform. The main advantages are as follows:

a) Comprehensive Student Evaluation:
The system analyzes academic, psychological, and lifestyle parameters together to provide a holistic view of a student's well-being and career readiness.

b) Early Risk Identification:
Through predictive models such as the Random Forest Classifier and XGBoost, the system helps identify early signs of mental distress or burnout, allowing timely intervention and support.

c) Personalized Guidance:
Students receive customized feedback reports with clear recommendations for improving mental health, managing stress, and enhancing academic and professional performance.

d) Career Preparedness Insights:
The Internship Prediction module assesses factors such as CGPA, skills, and certifications to suggest how prepared a student is for internships, categorized as *High*, *Moderate*, or *Low* eligibility levels.

e) Data-Driven Decision Support:
Faculty and counselors can use the analytics and generated reports to make informed decisions regarding counseling strategies, academic planning, and workload management.

f) User-Friendly Web Interface:
The project is built as a web-based system, providing an intuitive, responsive, and easily accessible interface for both students and administrators.

g) Visualization & Report Generation:

Results are displayed through interactive charts, graphs, and downloadable PDF reports that summarize predictions and confidence levels, improving understanding and record-keeping.

h) Cross-Platform Accessibility:
The web-based deployment ensures compatibility across various devices and operating systems, requiring only a modern web browser.

I) Secure Data Handling:
The back-end system ensures data integrity and privacy using secure data validation and controlled access to model predictions and reports.

j) Educational and Institutional Value:
Institutions can integrate this tool to monitor student well-being trends, reduce dropout risks, and enhance overall academic productivity through proactive measures

7 Assumptions, if any

While developing the **CareerSphere AI – Smart Student Wellness & Career Prediction** system, the following assumptions were considered to ensure smooth implementation and realistic model behavior:

1. **Accurate User Inputs:**
   It is assumed that the data entered by students in the forms (related to mental health, lifestyle, and academics) is truthful and accurate, as incorrect entries may affect prediction quality.

2. **Stable Internet Connectivity:**
   The system assumes a stable internet connection for smooth data transmission between the front-end interface and the server hosting the ML models.

3. **Preprocessed and Clean Datasets:**
   It is assumed that the datasets used for model training have already been cleaned and preprocessed to minimize noise and missing values before training.

4. **Sufficient Computational Resources:**
   The system presumes access to adequate hardware resources (RAM, processor speed, and storage) during model training and testing phases.

5. **Consistent Parameter Scale:**
   It is assumed that all input parameters (like stress level, study hours, CGPA, etc.) are normalized within defined scales (for example, 0–10) for accurate model evaluation.

6. **Non-Medical Use Case:**
   The predictions generated by the Mental Health and Burnout modules are assumed to be advisory in nature and not intended for clinical or diagnostic use.

7. **User Privacy Maintenance:**
   It is assumed that all users consent to data usage for analysis purposes, and no personal identifiers are shared externally.

8 Future Scope and further enhancement of the Project

The **CareerSphere AI – Smart Student Wellness & Career Prediction** system holds vast potential for future development and expansion. As education and technology continue to evolve, several enhancements can be integrated to improve accuracy, usability, and impact.

1. **Integration with Real-Time Data Sources:**
   Future versions can integrate with student management systems or wearable devices (like smartwatches and fitness trackers) to collect live data related to physical activity, stress, and sleep, allowing more dynamic and real-time predictions.

2. **Inclusion of Deep Learning Models:**
   Advanced algorithms such as Neural Networks or LSTM can be implemented to capture complex behavioral patterns and improve prediction precision for burnout and mental health risk analysis.

3. **AI Chatbot for Personalized Counseling:**
   An intelligent chatbot can be introduced to provide instant, AI-driven emotional support

and actionable tips based on the user's predicted category (High, Moderate, or Low risk).

4. **Enhanced Internship Prediction System:**
   The internship prediction model can be expanded to include **real-time internship listings**, **domain-based job matching**, and **resume optimization suggestions**, making it a complete career recommendation ecosystem.

5. **Gamified Wellness Tracking:**
   A gamification layer can be added to encourage students to maintain healthy routines. Students could earn badges or rewards for improving parameters like sleep hours, exercise consistency, and stress management.

6. **Multi-Language & Voice Assistant Support:**
   Incorporating multilingual support and a voice-based interface will make the system more accessible, especially for users from non-technical or rural backgrounds.

7. **Admin & Counselor Dashboard:**
   A secure dashboard can be designed for college counselors or faculty to monitor aggregate student well-being trends and provide timely interventions while maintaining data privacy.

8. **Cloud-Based Deployment:**
   Hosting the models and reports on cloud platforms (like AWS or Google Cloud) can ensure scalability, faster response time, and easier access for multiple institutions.

9 Project Repository Location

S#	Project Artifacts	Location	Verified by Project Guide	Verified by Lab In-Charge
1.	Project Synopsis Report (Final Version)			
2.	Project Progress updates			
3.	Project Requirement specifications			
4.	Project Report (Final Version)			
5.	Test Repository			
6.	Project Source Code (final version) with executable			
7.	Any other document			

10 Definitions, Acronyms, and Abbreviations

Abbreviation	Description
AI	Artificial Intelligence – the simulation of human intelligence in machines to perform tasks like reasoning, learning, and problem-solving.
ML	Machine Learning – a branch of AI that enables systems to learn patterns from data and make predictions without explicit programming.
RF (Random Forest)	An ensemble machine learning algorithm that combines multiple decision trees to improve prediction accuracy and reduce overfitting.
XGBoost	Extreme Gradient Boosting – an advanced, efficient, and scalable implementation of gradient boosting used for classification and regression tasks.
CSV	Comma-Separated Values – a file format used for storing tabular data in plain text.
PDF	Portable Document Format – a file format used for generating downloadable prediction reports.
DFD	Data Flow Diagram – a graphical representation of the flow of data within a system.
ERD	Entity-Relationship Diagram – a diagram used to model the data relationships in the project database.
UCD	Use Case Diagram – a diagram that shows the interaction between users and the system.
Normalization	The process of scaling data to a standard range (e.g., 0–1 or 0–10) to improve model training performance.
Encoding	The conversion of categorical (textual) data into numerical format for machine learning algorithms.
Training Set	The portion of data used to train the machine learning model.
Testing Set	The portion of data used to evaluate the model's generalization and performance.
Validation Set	A small portion of data used to fine-tune and optimize model parameters.
Accuracy	A metric that measures the percentage of correct predictions made by the model.
Precision	The ratio of correctly predicted positive observations to the total predicted positives.

Recall	The ratio of correctly predicted positive observations to all actual positives.
F1-Score	The harmonic mean of precision and recall, providing a balanced measure of model performance.

11 Conclusion

The **CareerSphere AI – Smart Student Wellness & Career Prediction** project successfully integrates artificial intelligence and data-driven analytics to assess and support student wellness and career readiness. By leveraging **Machine Learning algorithms** such as **Random Forest Classifier** and **XGBoost**, the system provides accurate, actionable insights across three major modules — *Mental Health Checker, Burnout Detection,* and *Internship Prediction.*

This project demonstrates innovation by combining academic, psychological, and lifestyle factors into a unified predictive model that helps students, educators, and career counselors make informed decisions. The inclusion of interactive dashboards, visual analytics, and downloadable personalized reports enhances user experience and promotes awareness of student well-being.

The implemented system stands out due to:

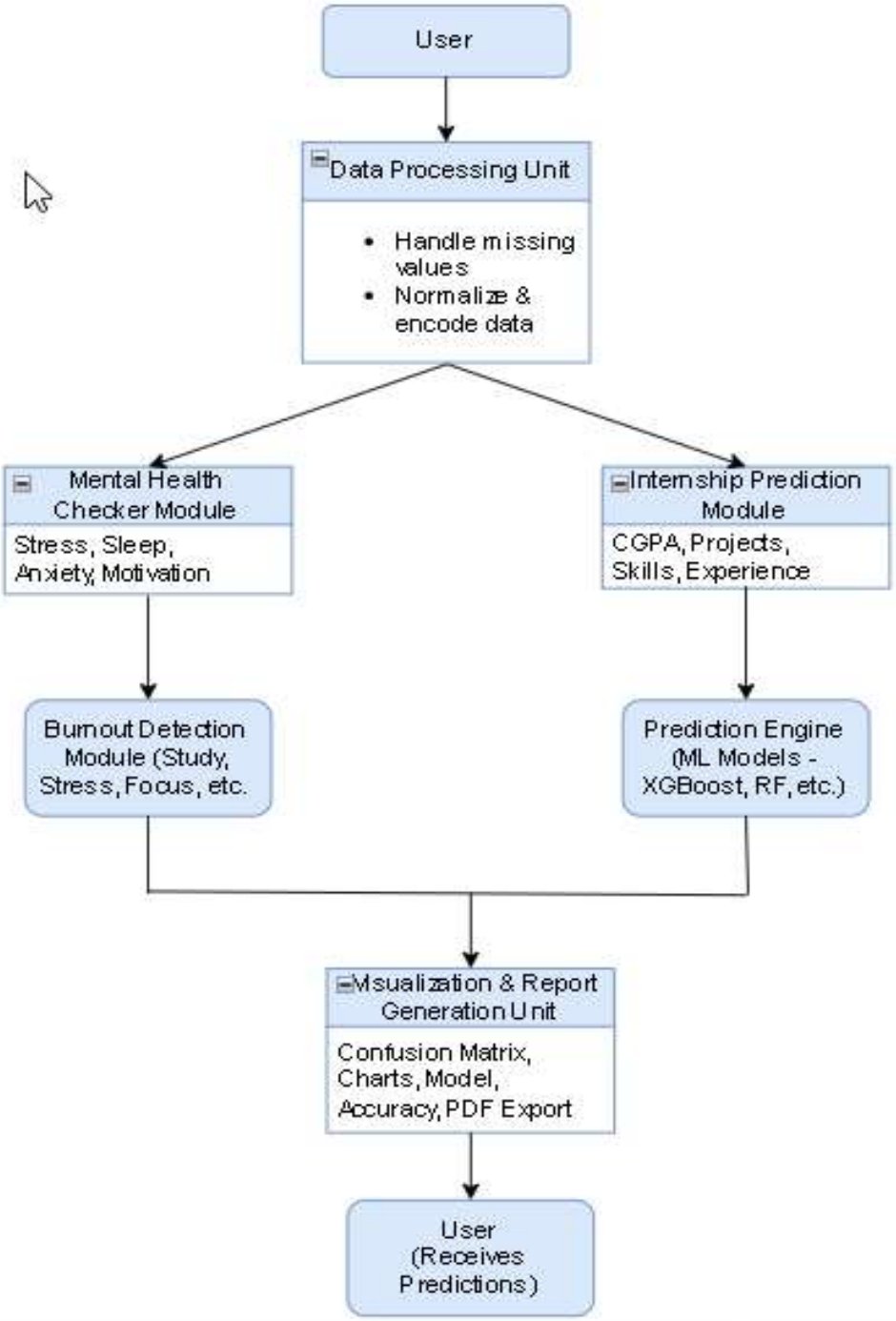
- Its multi-dimensional approach that connects **mental health, lifestyle balance, and academic performance**.
- The use of **advanced ensemble learning algorithms** for high accuracy and reliability.
- Its ability to generate **personalized PDF reports** with clear insights and visual summaries.
- A scalable and modular design that allows future integration with institutional dashboards and counseling systems.

Overall, **CareerSphere AI** not only contributes to early identification of student stress and burnout but also provides valuable guidance for career preparation. It reflects a practical application of machine learning in education, demonstrating how data-driven systems can create a meaningful social and academic impact.

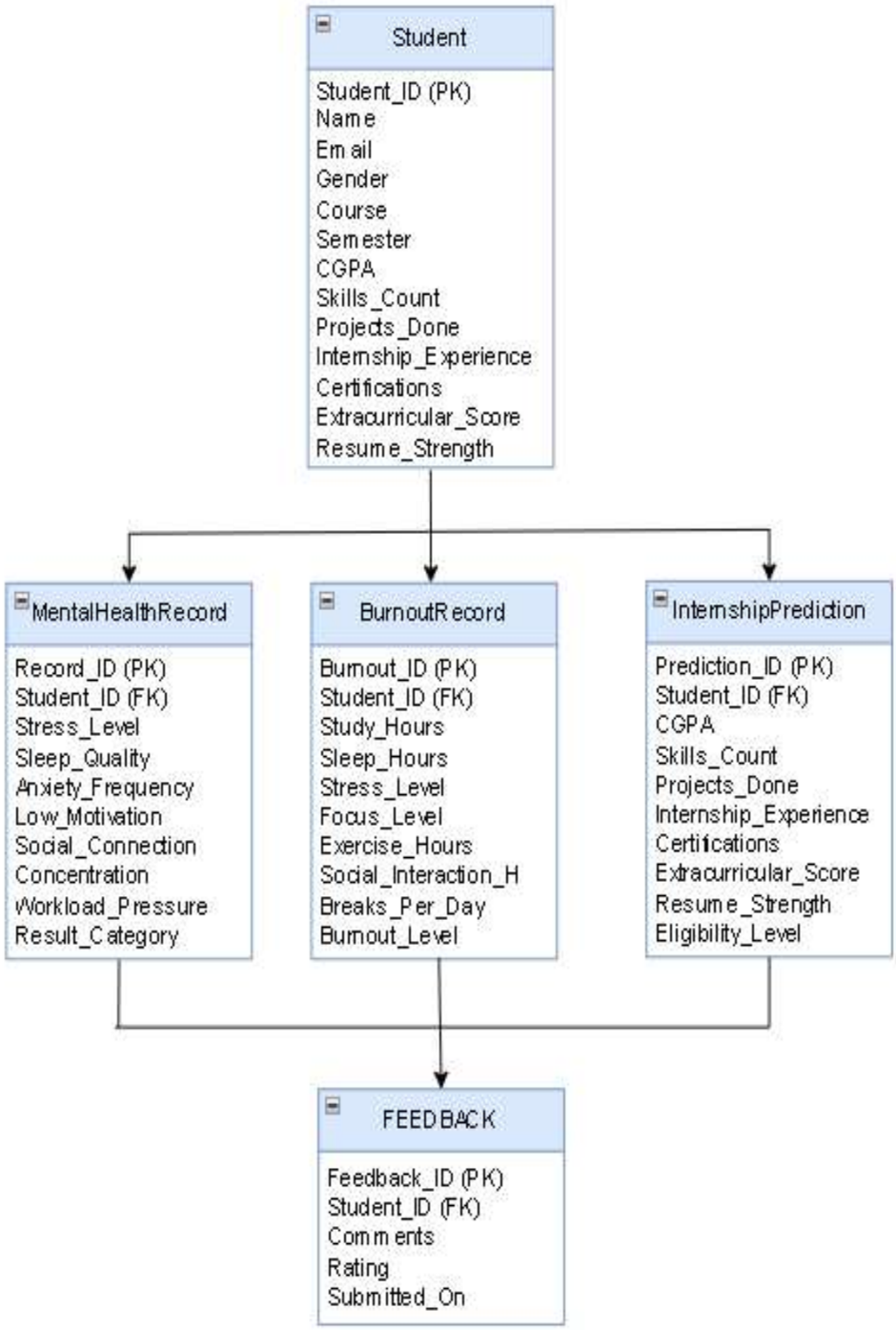
12 References

S#	Reference Details	Owner	Version	Date
1.	Project Synopsis			
2.	Project Requirements			
3.	IEEE Std 830-1998: Recommended Practice for Software Requirements Specifications	IEEE	1.0	1998
4.	Kaggle Datasets – Mental Health, Burnout & Internship Prediction	Kaggle Inc.	Latest	2024
5.	World Health Organization (WHO) Reports on Student Mental Health and Stress	WHO	Latest	2023
6.	W3C Standards for Web Application Development	W3C	Latest	2024
7.	Python Documentation (Pandas, NumPy, Scikit-learn, XGBoost)	Python Software Foundation	3.12	2024
8.	PHP and MySQL Documentation	Oracle / PHP Group	8.2	2024
9.	Visual Studio Code and XAMPP User Guides	Microsoft / Apache Friends	Latest	2024
10.	Figma and Canva UI/UX Design Documentation	Figma Inc. / Canva	Latest	2024

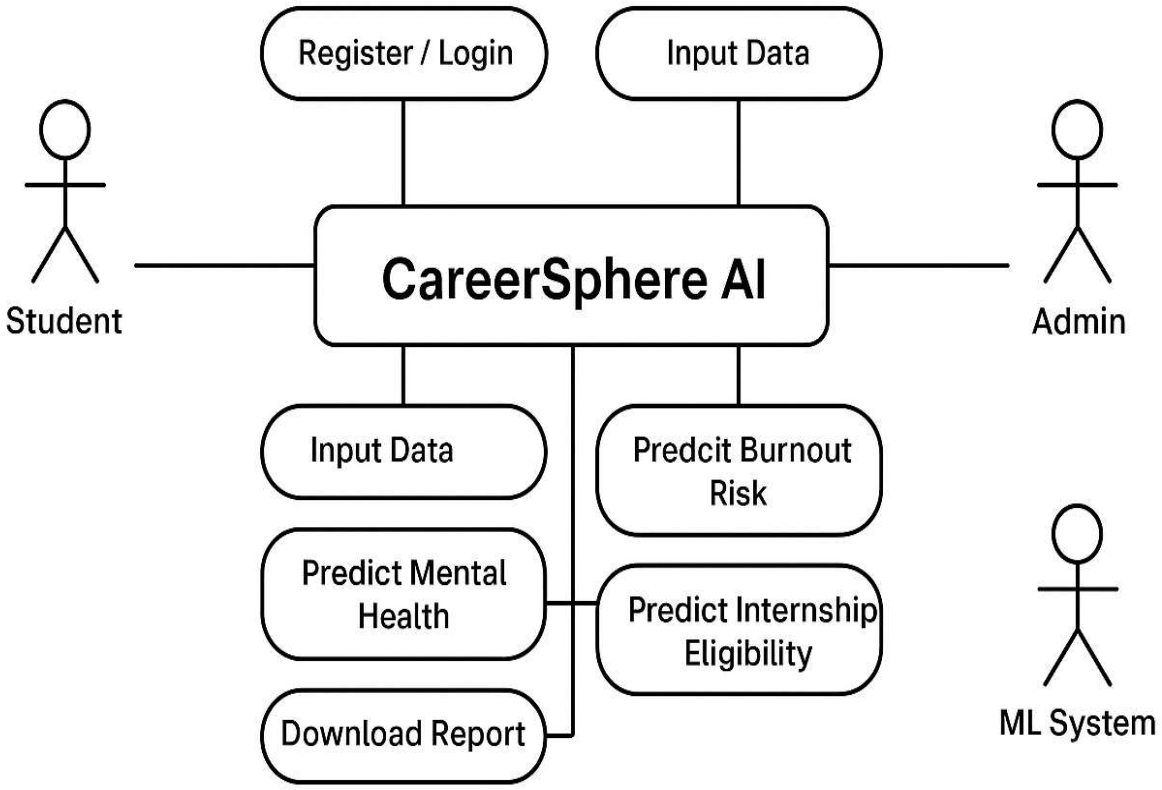
Annexure A
Data Flow Diagram (DFD)



Annexure B
Entity-Relationship Diagram (ERD)



Annexure C
Use-Case Diagram (UCD)



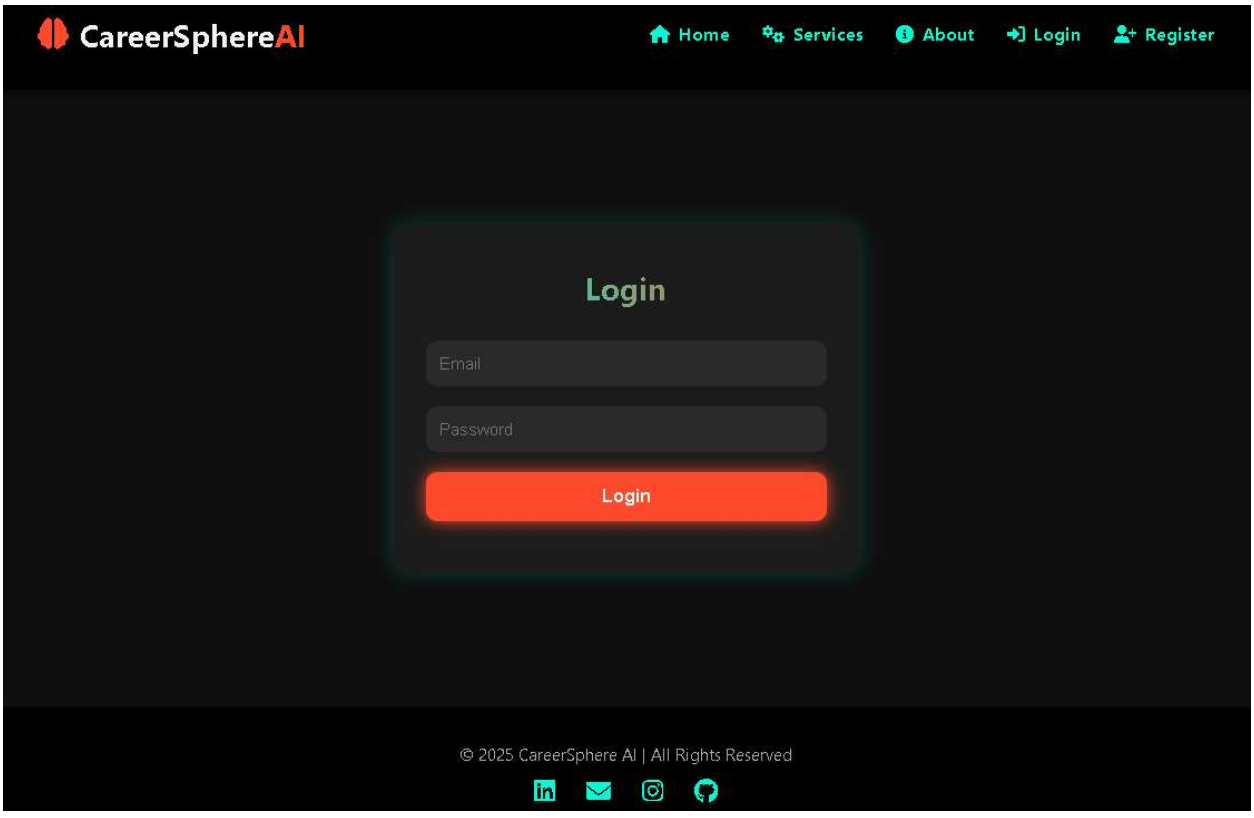
Annexure D
Data Dictionary (DD)

User Table (users)

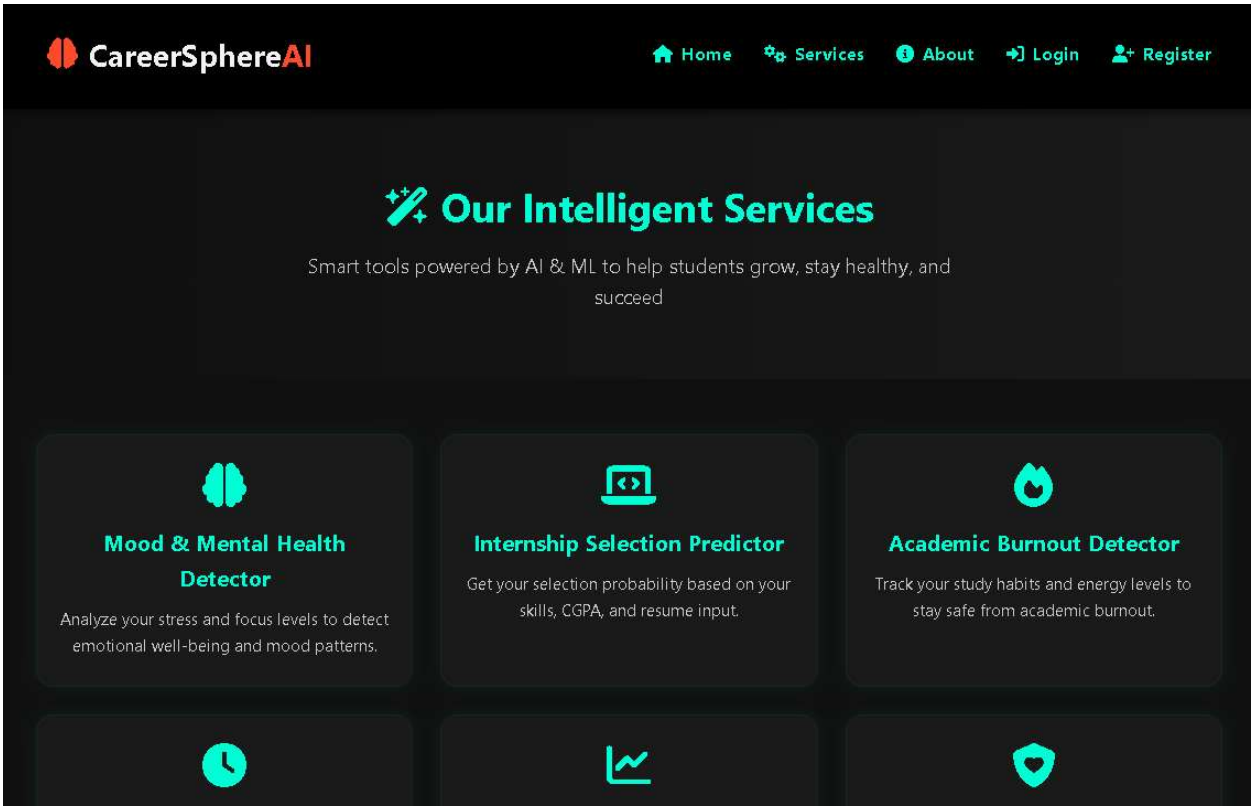
Fields	Data type	Description
id	Number(INT 11)	Unique identifier for each user (Primary Key, Auto Increment).
name	Text (VARCHAR 100)	Full name of the user.
email	Text (VARCHAR 100)	Email address of the user (used for login).
password	Text (VARCHAR 250)	Encrypted password of the user.
created_at	Timestamp	Date and time when the user registered.

Annexure E
Screen Shots

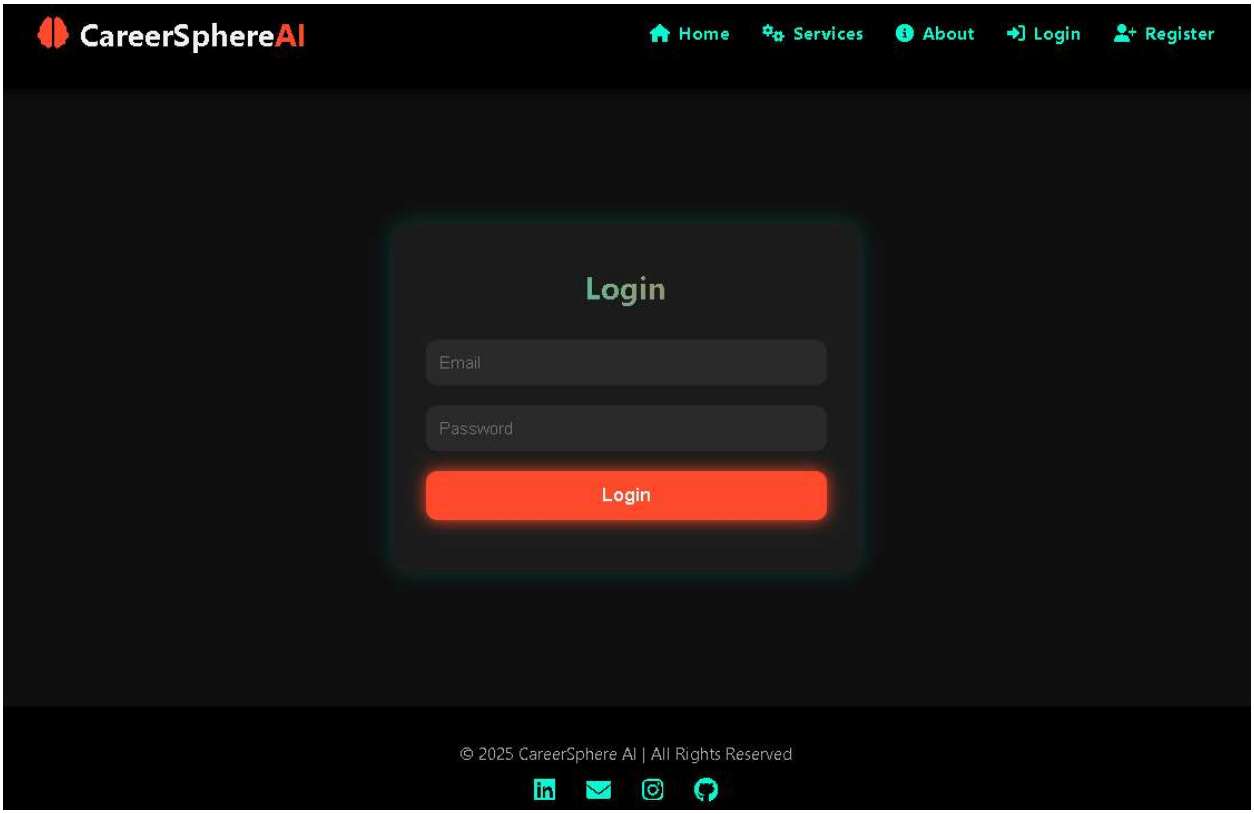
Home Page:



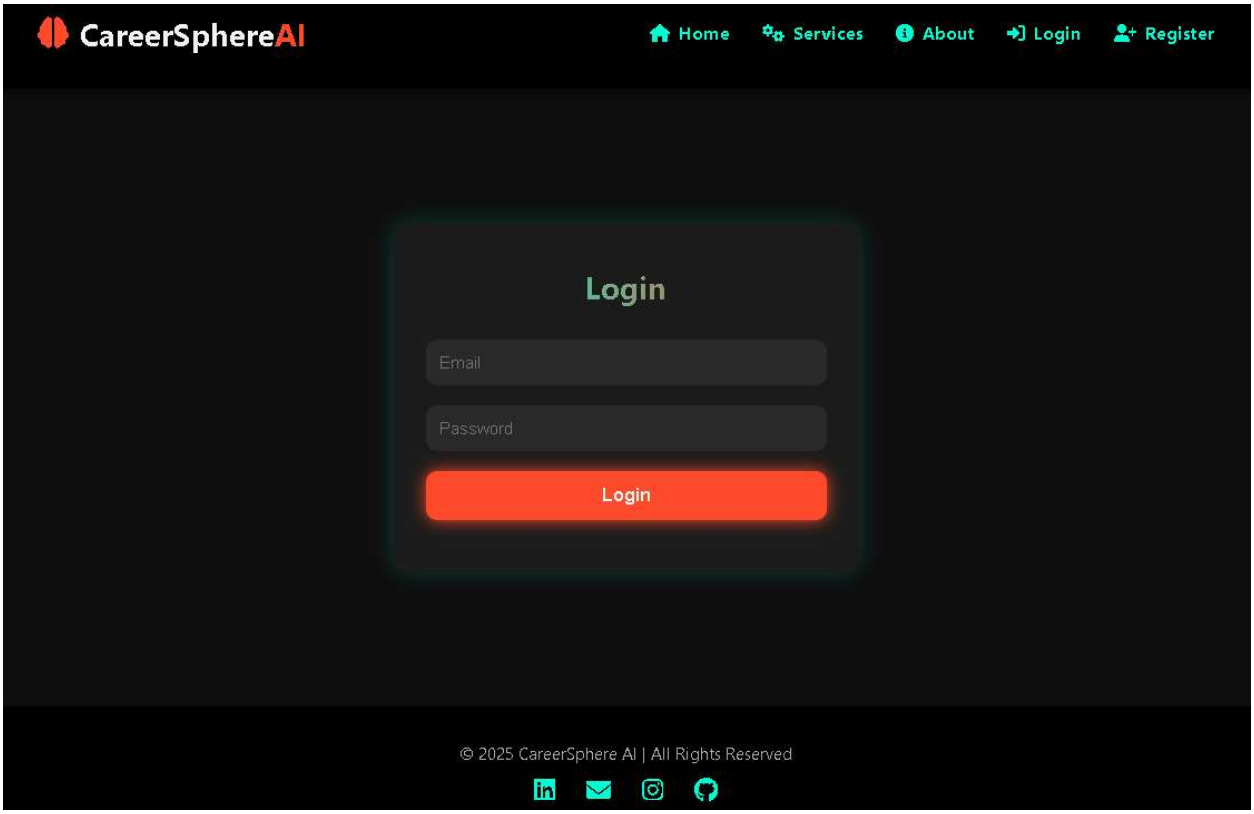
Services Page:



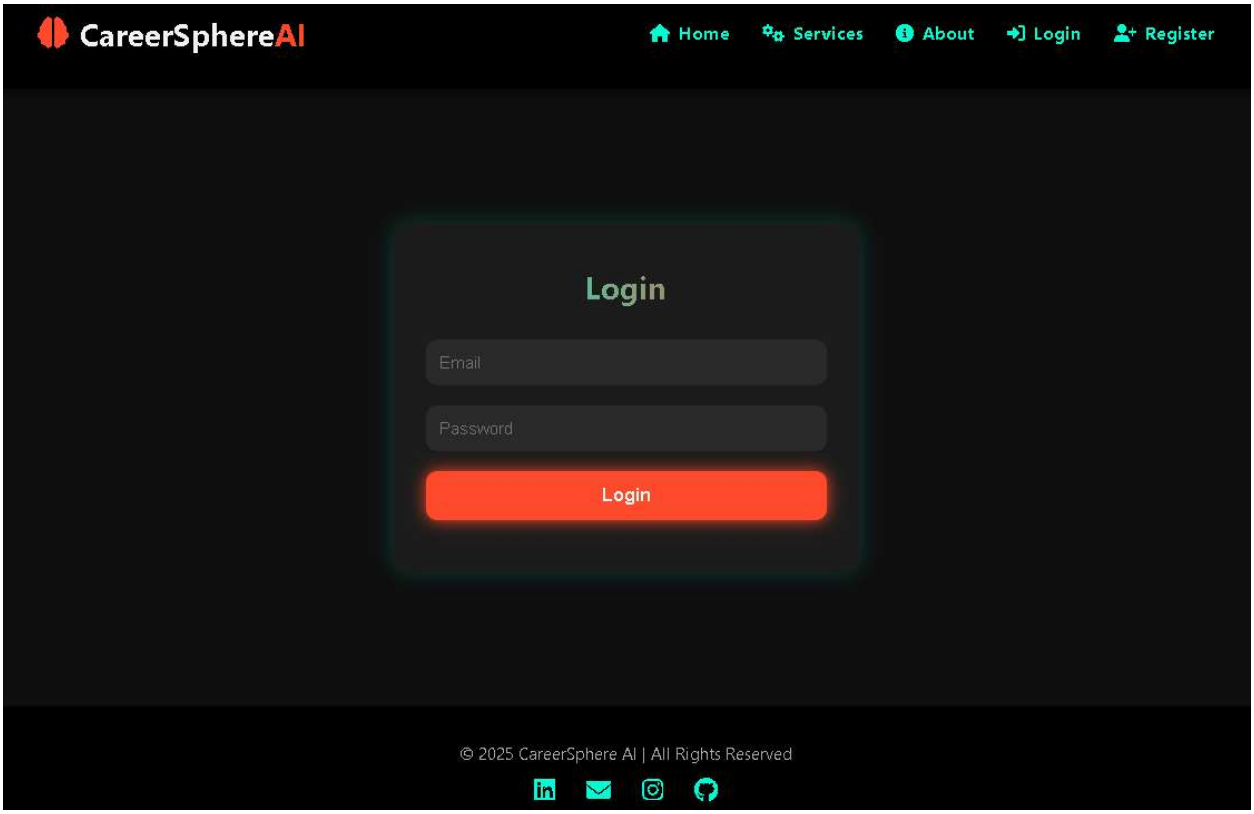
About Page:


CareerSphereAI
Home
Services
About
Login
Register
Login
Email
Password
Login
© 2025 CareerSphere AI | All Rights Reserved

Login Page:


CareerSphereAI
Home
Services
About
Login
Register
Login
Email
Password
Login
© 2025 CareerSphere AI | All Rights Reserved

Register Page:



Admin Panel (Logged In Page):