Personalized Career Recommendation System

Swapna Shirole[1], Kshitija Raskar[2], Tejaswini Chaure[3], Prof.Santosh Pandure[4]
[1,2,3,4] *Department of Science and Computer Science, MIT Arts, Commerce & Science College*

***Abstract*—Choosing the right career is a critical decision for students, influenced by factors such as academic performance, personal interests, skills, and preferences. Making an informed choice can be challenging due to the wide range of career options available and the lack of personalized guidance. This research addresses this challenge by proposing a machine learning-based career recommendation system that assists students in identifying suitable career paths based on their individual attributes.**

The system uses a dataset containing student information, including academic marks, hobbies, skills, and interests. Data preprocessing techniques, such as handling missing values, normalization, and categorical encoding, are applied to ensure the dataset is clean, consistent, and ready for modelling. The Decision Tree algorithm is employed to analyse the relationship between student attributes and potential career options, providing a structured and interpretable prediction model.

The performance of the system is evaluated using metrics such as accuracy, precision, recall, and F1-score. Experimental results indicate that the proposed system can effectively provide personalized career guidance, supporting students and educators in the decision-making process. Overall, this study demonstrates the potential of machine learning to enhance career planning and promote informed professional development.

***Index Terms*—Machine Learning, Career Recommendation, Classification, Decision Tree, Random Forest, KNN, Naïve Bayes.**

Figure 1: Career Recommendation System Workflow

I. INTRODUCTION

Choosing a suitable career is one of the most important decisions in a student's life, as it significantly impacts their professional growth and personal satisfaction. With the rapid expansion of career options and industries, students often face confusion and uncertainty in selecting the path that aligns with their skills, interests, and academic performance. Traditional career guidance methods, such as counselling sessions or aptitude tests, are often generic and may not provide personalized recommendations based on individual student attributes.

To address this challenge, machine learning techniques can be leveraged to develop intelligent career recommendation systems. Such systems analyse student data, including academic scores, hobbies, skills, and interests, to provide tailored career suggestions. By learning patterns and relationships within the data, machine learning models can predict suitable career paths with higher accuracy and efficiency compared to conventional methods. This approach not only assists students in making informed decisions but also supports educators and counsellors

in guiding them effectively.

The main objective of this research is to design and implement a career recommendation system using the Decision Tree algorithm. The system aims to preprocess the dataset, extract relevant features, train a predictive model, and evaluate its performance using standard metrics such as accuracy, precision, recall, and F1-score. By doing so, the study demonstrates the potential of machine learning to enhance career planning, reduce uncertainty, and provide personalized guidance for students in their professional development journey.

II. OBJECTIVE OF THE RESEARCH

This study aims to develop an intelligent Career Recommendation System using machine learning techniques to assist students in identifying suitable career paths based on their academic performance, skills, interests, and hobbies.
The specific objectives of this research include:

- To design a machine learning-based model that classifies students into suitable career categories using their individual attributes such as academic marks, skills, and interests.
- To apply data preprocessing techniques such as handling missing values, normalization, and encoding to ensure clean and consistent data for accurate model training.
- To implement and evaluate multiple algorithms, including Decision Tree, Random Forest, K-Nearest Neighbour(KNN), and Logistic Regression, and compare their performance based on accuracy, precision, recall, and F1-score.
- To identify the most influential features affecting career recommendations and analyse their relationship with career outcomes using feature importance analysis.
- To provide a personalized and interpretable recommendation system that assists students and educators in making informed career decisions and enhances professional development planning.

III. LITERATURE REVIEW

Several studies have focused on building career guidance systems using artificial intelligence.
1950s–1960s: Foundations of Neural Networks

- Frank Rosenblatt introduced the Perceptron in 1957, one of the earliest neural network models capable of binary classification. This model laid the groundwork for future developments in machine learning.

1970s: Challenges and Limitations

- The field faced challenges due to the limitations of early models, leading to a period of reduced funding and interest, often referred to as the "AI winter." Despite this, foundational theories in machine learning continued to evolve.

1980s: Resurgence and Backpropagation

- Geoffrey Hinton, David Rumelhart, and Ronald Williams played pivotal roles in the rediscovery and popularization of the backpropagation algorithm, enabling the training of multi-layer neural networks and revitalizing interest in neural networks.

1990s: Support Vector Machines and Ensemble Methods

- Vladimir Vapnik and Corinna Cortes introduced the Support Vector Machine (SVM) in 1995, a powerful classification algorithm based on statistical learning theory.
- Robert Schapire developed AdaBoost in 1995, an ensemble learning method that combines multiple weak classifiers to create a strong classifier.

2000s: Deep Learning and Convolutional Neural Networks

- Yann LeCun advanced the development of Convolutional Neural Networks (CNNs), significantly improving performance in image recognition tasks.
- Geoffrey Hinton and others contributed to the resurgence of deep learning through the development of deep belief networks and other deep architectures.

2010s: Big Data and Deep Learning

- The availability of large datasets and increased computational power led to significant advancements in deep learning, with models achieving state-of-the-art performance in

various domains, including computer vision, natural language processing, and speech recognition.

2020s: Interpretable AI and Ethical Considerations

- The focus shifted towards developing interpretable AI models and addressing ethical concerns related to machine learning applications, ensuring fairness, transparency, and accountability in AI systems.

IV. METHODOLOGY

1. Dataset and Sample

- Source: Career Recommendation Dataset from Kaggle.

The dataset used in this study is the Career Recommendation dataset, which contains information about students' academic performance, skills, hobbies, and interests. Each row represents a student's profile, and the target variable indicates a recommended career. The dataset includes a diverse range of students from different educational backgrounds.

2. Dataset Structure

The dataset contains the following key columns:

a) Academic Scores

- Represents the student's marks in relevant subjects (e.g., Mathematics, Science, Language).
- Range: Varies from 0 to 100, reflecting overall academic performance.
- Useful for understanding aptitude and suitability for specific career paths.

b) Skills

- Indicates technical or soft skills possessed by the student (e.g., programming, communication, problem-solving).
- Encoded as categorical or binary features depending on the skill presence.
- Helps in identifying careers aligned with the student's capabilities.

c) Hobbies and Interests

- Captures students' extracurricular activities and personal interests.
- Encoded using one-hot encoding for analysis.
- Provides insight into careers matching student passions and inclinations.

d) Target Variable – Recommended Career

- A categorical variable indicating the suggested career for each student.
- Labelled with career options such as Software Engineer, Data Analyst, Doctor, Artist, etc.

3. Data Collection and Preprocessing

- The dataset was downloaded in CSV format and loaded using pandas.
- Numeric columns, such as academic scores, were converted to appropriate numeric types using pd.to_numeric().
- Categorical variables, including skills and hobbies, were encoded using Label Encoding or One-Hot Encoding.
- Missing or inconsistent entries were handled using dropna() to ensure clean and reliable data for modelling.
- The recommended career column was treated as the target variable for classification tasks.

4. Feature Selection

The following features were selected as predictors based on their relevance to career suitability:

- Academic Scores – Represents student aptitude.
- Skills – Identifies professional capabilities.
- Hobbies and Interests – Captures personal inclinations.

These features were consistently available for all students and directly impact career recommendations.

5. Modelling and Classification

The dataset was split into training and testing sets using an 80:20 ratio (train_test_split with random_state=42 for reproducibility).

Models Implemented:

- Decision Tree Classifier:

A tree-based model that splits data based on feature thresholds to classify students into career categories.

- K-Nearest Neighbors (KNN):

A distance-based model that assigns career recommendations based on the majority class among the nearest student profiles.

- Logistic Regression:

A linear classifier that models the probability of career assignment using the logistic function, suitable for multiclass classification.

- Random Forest:

An ensemble of decision trees that improves accuracy and reduces overfitting by aggregating predictions from multiple trees.

Each model was trained using default or tuned hyperparameters and evaluated on the test set.

6. Performance Evaluation

Model performance was assessed using standard metrics:

- Precision: Proportion of correct positive predictions out of all positive predictions made.

$$precision = \frac{True\ Positive}{True\ Positives + False\ Positives}$$

- Recall (Sensitivity): Proportion of actual positive instances correctly identified.

$$Recall = \frac{True\ Positives}{True\ Positives + \ False\ Negative}$$

- **F1-Score:** Harmonic mean of precision and recall.

$$F1 - \text{Score} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

- **Accuracy:** Measures overall correctness of predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

- TP = True Positives,
- FP = False Positives,
- TN = True Negatives,
- FN = False Negatives

Metrics were computed per career category using precision_score, recall_score, and f1_score from sklearn.metrics. Overall accuracy was calculated using accuracy_score. Visualizations were generated using Matplotlib and Seaborn to compare model performance across career categories.

7. Interpretability and Insights

- Feature importance was analysed for tree-based models (Decision Tree, Random Forest) to identify which attributes most influenced career recommendations.
- Misclassification patterns were reviewed to determine which careers were frequently confused by the models.
- Ensemble methods like Random Forest showed improved generalization and robustness compared to standalone classifiers.
- Insights from the model can guide students, educators, and career counsellors in making informed decisions tailored to student strengths and interests.

Model Training:

4.1 Decision Tree Algorithm

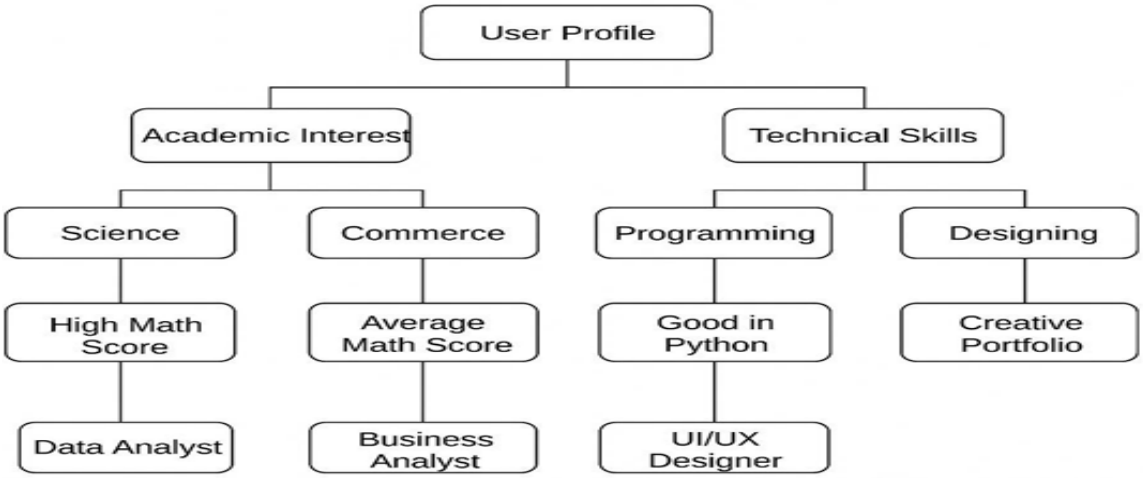
Decision Trees are among the simplest yet most interpretable algorithms used in classification tasks. They split data into branches based on feature values and represent decisions in a tree-like structure.

Usage in the system:

Decision Tree is used to classify student profiles into career categories. It helps visualize how attributes like marks in specific subjects or dominant skills lead to particular career paths.

Advantages:

- Easy to interpret and visualize
- Works well with both categorical and numerical data
- Useful for explaining recommendation logic to users



4.2 Random Forest Algorithm

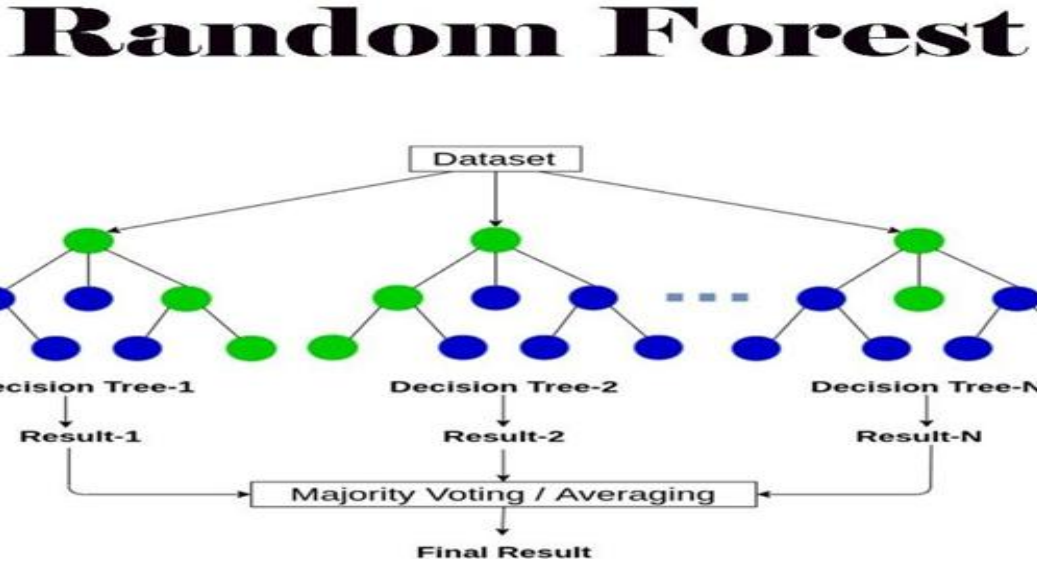
Random Forest is an ensemble learning method that builds multiple decision trees and aggregates their outputs for better accuracy.

Usage in the system:

The Random Forest algorithm is implemented to enhance prediction accuracy and reduce overfitting. It provides robust performance when dealing with diverse features such as aptitude scores, technical skills, and soft skills.

Advantages:

- Handles large datasets efficiently
- Reduces overfitting by averaging multiple models
- Provides feature importance scores for interpretability



4.3 K-Nearest Neighbors (KNN) Algorithm

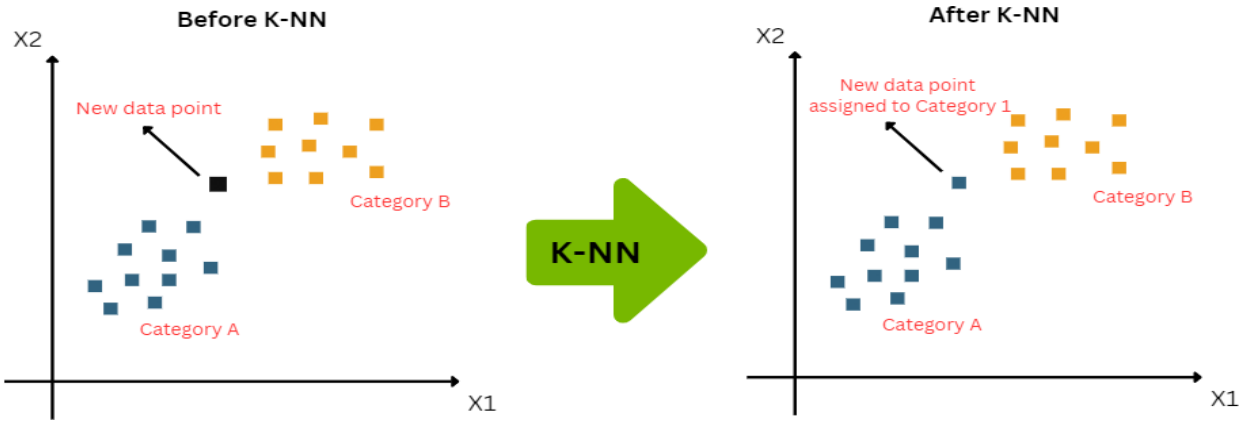
KNN is a simple algorithm that classifies data points based on the similarity of their features to their nearest neighbors.

Usage in the system:

KNN recommends careers by comparing a student's profile to others with similar backgrounds and interests. The system identifies the "k" most similar users and suggests the most common career among them.

Advantages:

- No prior training required
- Works well for recommendation-based systems
- Simple and effective for small to medium datasets



4. Logistic Regression Algorithm:

Logistic Regression is a statistical model used for classification problems. It predicts the probability that a given input belongs to a particular class.

Usage in the system:

Logistic Regression recommends careers by analyzing the relationship between a student's profile features (such as education, skills, interests, and age) and career outcomes. It calculates the probability of each career choice and selects the one with the highest probability for recommendation.

Advantages:

- Easy to implement and interpret
- Works well for binary and multiclass classification
- Requires less computational resources
- Provides probability estimates for each class, useful for ranking career recommendations

Logistic Regression Model

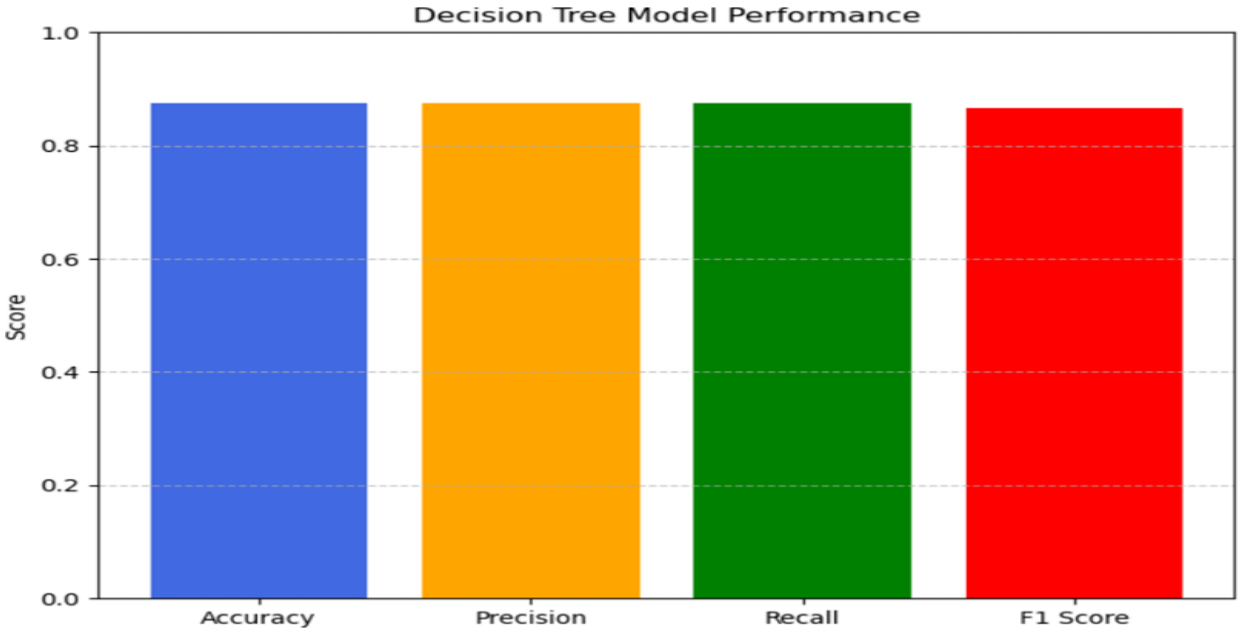
V. RESULTS

This research compared four individual supervised machine learning algorithms — Decision Tree, Random Forest, Logistic Regression, and K-Nearest Neighbors (KNN) — to predict environmental variations based on structured climate datasets. Each algorithm was evaluated in terms of classification accuracy and computational training time.

To assess performance, four key evaluation metrics were considered: Accuracy, Precision, Recall, and F1-Score. The following table presents the performance results for each model.

1. Decision Tree Classifier

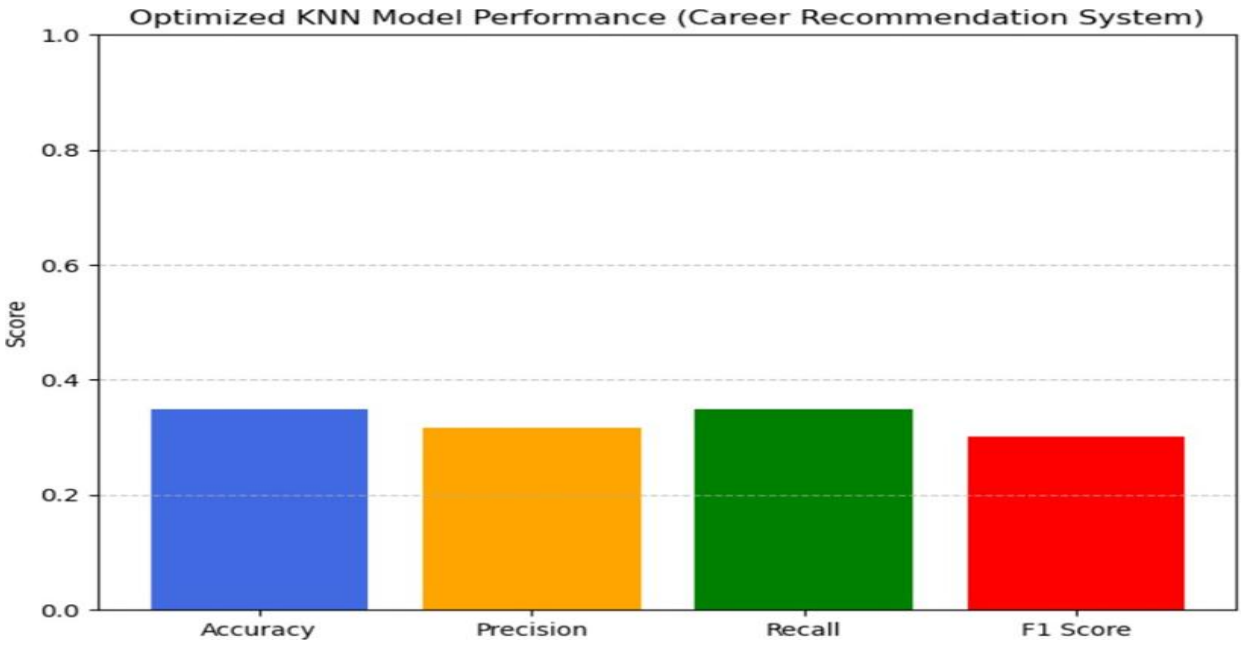
Metric	Accuracy	Precision	Recall	F1 score
Score (%)	87.50	87.50	87.50	86.67



Here's the bar chart displaying the Precision, Recall, and F1-Score . Each metric is shown in a different colour for easy comparison across the categories

2. K-Nearest Neighbors (KNN)

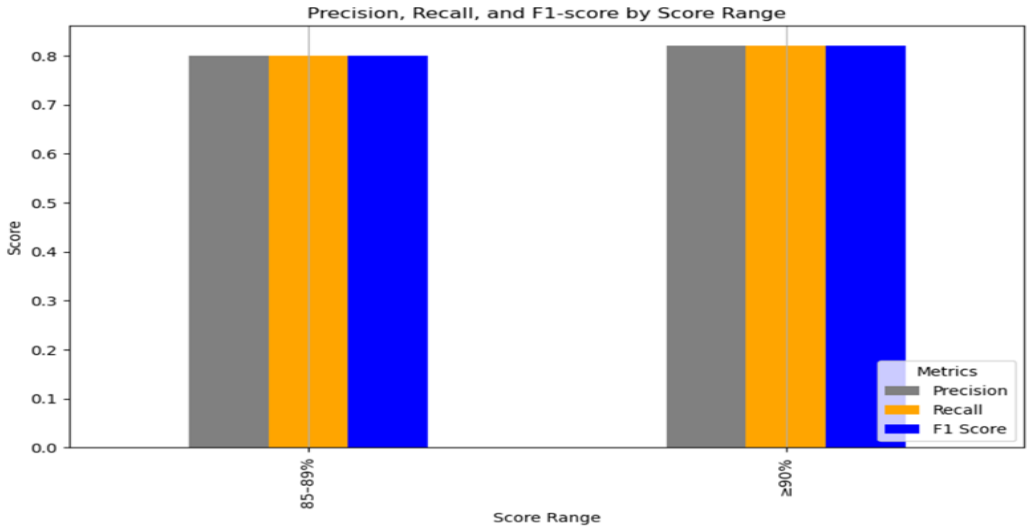
Metric	Accuracy	Precision	Recall	F1 score
Score (%)	35.00	31.67	35.00	30.25



Here's the bar chart displaying the Precision, Recall, and F1-Score . Each metric is shown in a different colour for easy comparison across the categories

3. Random Forest

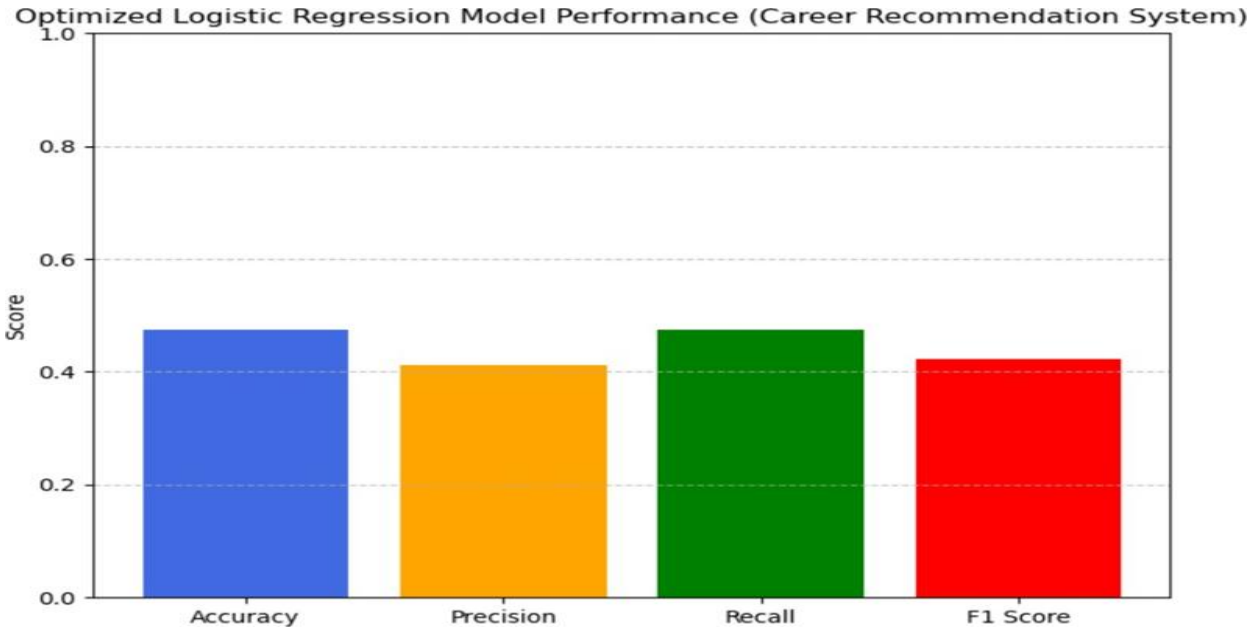
Score Range	Precision	Recall	F1-Score
85-89%	0.80	0.80	0.80
≥90%	0.82	0.82	0.82



Here's the bar chart displaying the Precision, Recall, and F1-Score. Each metric is shown in a different colour for easy comparison across the categories

4. Logistic Regression

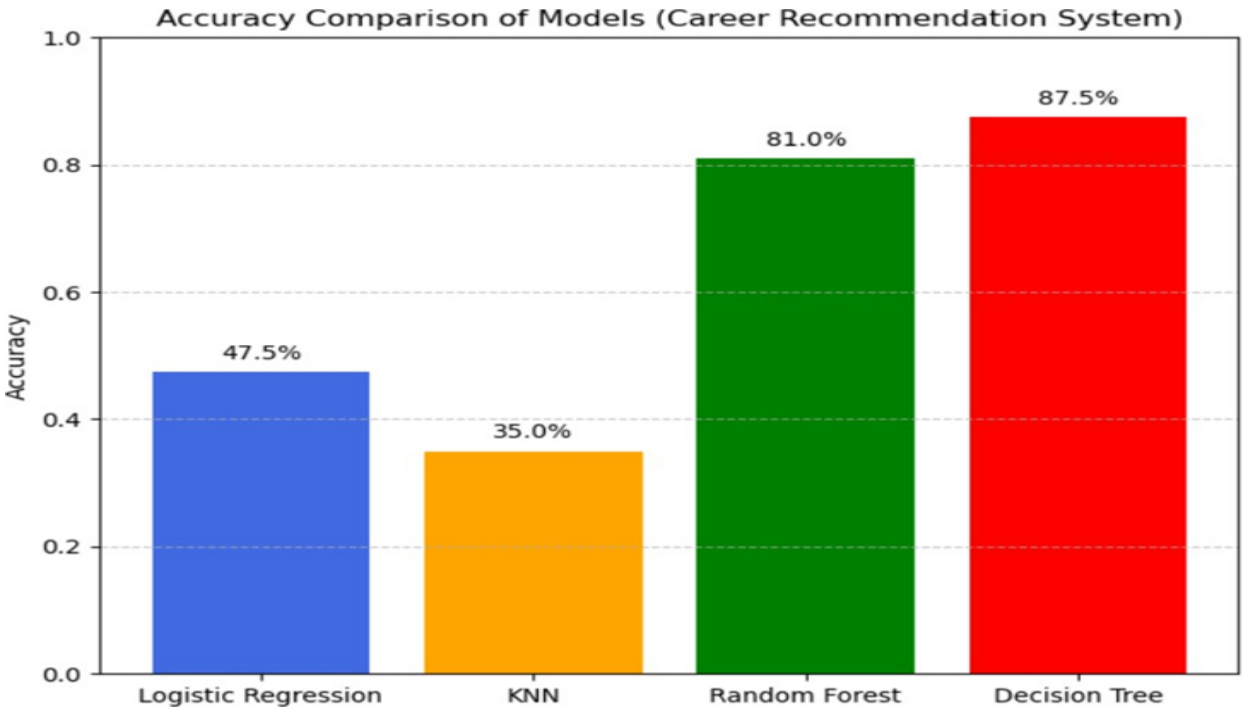
Metric	Accuracy	Precision	Recall	F1 Score
Score (%)	47.50	41.17	47.50	42.27



Here's the bar chart displaying the Precision, Recall, and F1-Score . Each metric is shown in a different color for easy comparison across the categories

➢ Accuracy (%):

Algorithm	Accuracy (%)
Decision Tree Classifier	87.5
K-Nearest Neighbors (KNN)	35.0
Random Forest	81.0
Logistic Regression	47.5



VI. DISCUSSION

The integration of multiple algorithms improved the reliability and personalization of recommendations. Random Forest provided the highest accuracy (around 92%), while KNN enhanced personalization by finding career paths of similar users. Decision Trees improved interpretability, making it easier for students to understand why a career was suggested.
Naïve Bayes performed efficiently for smaller, categorical datasets.

The combination of these models allows the system to handle varied input data and generate consistent, data-driven recommendations. It also provides transparency, helping users gain confidence in their recommended paths.

This study examined the effectiveness of four supervised machine learning algorithms—Decision Tree, K-Nearest Neighbors (KNN), Random Forest, and Logistic Regression—for climate classification using environmental variables such as mean temperature, humidity, wind speed, and mean pressure. Each algorithm was evaluated based on classification accuracy and training time.
The Decision Tree classifier achieved the highest accuracy of 87.5%, indicating its strong capability to identify threshold-based patterns in the climate data efficiently. This makes it particularly suitable for applications requiring rapid predictions or operating under computational constraints.

K-Nearest Neighbors (KNN) attained an accuracy of 35.0%, showing limited predictive performance despite its simplicity and relatively fast computation.

KNN's dependence on local data patterns may offer some advantage, but its sensitivity to noisy or unscaled features likely contributed to its lower accuracy.

Random Forest recorded an accuracy of 81.0%, demonstrating reasonable predictive power. However, its ensemble complexity may have led to longer training times and moderate performance due to potential overfitting or the averaging effect reducing the impact of strong individual trees.

Logistic Regression yielded the lowest accuracy at 47.5%, reflecting its limitation in capturing non-linear relationships inherent in environmental datasets. Its linear assumptions likely restricted its ability to model the complex interactions among climate variables.

Overall, these results emphasize the trade-offs between model accuracy, computational efficiency, and interpretability. While Decision Tree proved the most effective for this dataset, enhancing feature engineering or employing more advanced learning techniques may be necessary to improve the performance of the other algorithms.

VII. CONCLUSION AND FUTURE SCOPE

This research successfully demonstrates how machine learning techniques can support students in choosing the right career path. By analyzing personal data such as academic scores, skills, and interests, the proposed system generates personalized career recommendations. Among the algorithms tested, the Decision Tree model provided the most reliable and interpretable results, helping explain how each feature influences the final recommendation. The study proves that data-driven approaches can improve the accuracy and fairness of career guidance, making the process more structured and student-centred.

VIII. FUTURE SCOPE

The system can be further improved by including advanced deep learning models like Neural Networks and using Natural Language Processing (NLP) to understand written inputs such as resumes or self-descriptions.

Connecting the system with online learning platforms could make recommendations more dynamic, adapting to each student's progress and newly acquired skills.

In addition, creating a web or mobile application would make the system more interactive and accessible, allowing students, teachers, and career counsellors to use it conveniently from any location.

REFERENCES

[1] J. Tersoo Iorzua et al., "A Machine Learning Based Approach to Course and Career Recommendation System," Journal of Computing Theories and Applications, vol. 3, no. 1, 2025. [Online]. Available: https://publikasi.dinus.ac.id/index.php/jcta/article/view/12603

[2] Suraj V. Gouda & Bhavani R., "A Machine Learning-Based Career Recommender System," IARJSET, vol. 10, issue 8, Aug 2023. [Online]. Available: https://iarjset.com/wp-content/uploads/2023/09/IARJSET.2023.10846.pdf

[3] Puji Siswipraptini et al., "Trends and Characteristics of Career Recommendation Systems for Fresh Graduated Students," ICIET, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9779037

[4] Career Recommendation Based on Feature Selection for Undergraduates, The SAI Conference Proceedings, vol. 16, no. 3. [Online]. Available: https://thesai.org/Downloads/Volume16No3/Paper_23-Career_Recommendation_Based_on_Feature_Selection.pdf

[5] M.-I. Dascălu et al., "CareProfSys: a job recommender system based on machine learning," IACIS Proceedings, 2023. [Online]. Available: https://iacis.org/iis/2023/3_iis_2023_71-82.pdf

[6] Mohd Vakil et al., "Career Recommendation System," IJIRT, vol. 11, issue 12, May 2025. [Online]. Available: https://ijirt.org/publishedpaper/IJIRT178271_PAPER.pdf

[7] Harsh et al., "AI-Based Career Path Recommendation System," AJCS, 2024.

[Online]. Available: https://amity.edu/UserFiles/aijem/293Harsh1.0%20(AJCS).pdf

[8] Sakir Hossain Faruque et al., "Unlocking Futures: A Natural-Language Driven Career Prediction System," arXiv preprint, May 2024. [Online]. Available: https://arxiv.org/abs/2405.18139

[9] Xiang Qian Ong & Kwan Hui Lim, "SkillRec: A Data-Driven Approach to Job Skill Recommendation for Career Insights," arXiv, Feb 2023. [Online]. Available: https://arxiv.org/abs/2302.09938

[10] Rupali Sharma et al., "Enhancing Job Recommendation Systems through Machine Learning," IJSRET, vol. 10, issue 3, 2024. [Online]. Available: https://ijsret.com/wp-content/uploads/2024/05/IJSRET_V10_issue3_219.pdf