A PROJECT REPORT

on

“From Paper to Pixels: A Machine Learning Solution for Sustainable Study Materials”

Submitted to

KIIT Deemed to be University

In Partial Fulfillment of the Requirement for the Award of

BACHELOR’S DEGREE IN

COMPUTER SCIENCE ENGINEERING

BY

AGRASTH NAMAN	2005495
ANSH KAPOOR	2005503
ADITYA NIGAM	20051634
AYUSH ANAND	20051637

UNDER THE GUIDANCE OF

Dr. M. Nazma B. J. Naskar

SCHOOL OF COMPUTER ENGINEERING

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESWAR, ODISHA - 751024

2022 - 2023

KIIT Deemed to be University

School of Computer Engineering
Bhubaneswar, ODISHA 751024

CERTIFICATE

This is certify that the project entitled

From Paper to Pixels: A Machine Learning Solution for Sustainable Study Materials

submitted by

AGRASTH	NAMAN	2005495
ANSH	KAPOOR	2005503
ADITYA	NIGAM	20051634
AYUSH	ANAND	20051637

is a record of bonafide work carried out by them, in the partial fulfillment of the requirement for the award of Degree of Bachelor of Engineering (Computer Science & Engineering OR Information Technology) at KIIT Deemed to be university, Bhubaneswar. This work is done during the year 2022-2023, under our guidance.

Date: 30 / 03 / 2023

(M Nazma BJ Naskar)
Project Guide

Acknowledgements

We are profoundly grateful to **Dr. M. Nazma B. J. Naskar** of **Affiliation** for her expert guidance and continuous encouragement throughout to see that this project rights its target from its commencement to its completion.

AGRASTH NAMAN

ANSH KAPOOR

ADITYA NIGAM

AYUSH ANAND

ABSTRACT

The shift towards electronic media for students and professionals worldwide has led to an increase in the delivery of study materials via courier, resulting in high costs and a significant impact on the environment. To address this issue, we propose a software-based solution that allows users to choose between electronic and hardcopy formats for their study materials based on their preferences. Our solution uses machine learning to analyze past preferences and predict future demands, enabling us to estimate the quantity of material to be delivered in each format. By doing so, we not only reduce transport and delivery costs but also significantly decrease the amount of paper production in institutions on a global scale. Our proposed solution promotes better energy management and water treatment while also promoting the concept of reducing, reusing, and recycling. The advantages of our solution include a reduction in paper consumption, better memory management, and increased satisfaction for users.

Keywords: Sustainability, Electronic media, Machine learning, Study materials, Environmental impact.

Contents

Chapter 1

Introduction

The COVID-19 pandemic has significantly impacted the way we live and work, with the education sector being one of the most affected areas. Online classes have replaced the traditional method of learning, and students have been relying on electronic media for accessing study materials. While this has helped maintain the continuity of education, it has also highlighted the need for a more sustainable and cost-effective approach to providing study materials to students.

The traditional method of providing study materials to students involves the printing of textbooks and other study materials, which are then distributed to the students either in person or through courier services. However, this method has several drawbacks. Firstly, it leads to the large-scale industrial production of paper, which has a significant impact on the environment. Secondly, it adds to the financial burden of the students, as they have to pay for the purchase and delivery of these materials.

To address these issues, we propose a solution that involves providing students with a choice to receive a particular volume of study materials in either electronic format (eBooks - PDFs, ODP, Docs) or hardcopy, depending upon their needs in this online era. This customized mechanism will help a learner optimize his study routine by reading books from a combination of electronic devices (laptops/smartphones/tabs) and paperback (hardcopy). This will not only help save the transport and delivery charges but will also exponentially reduce the production amount of paper in institutions on a global scale.

The proposed solution uses machine learning algorithms to predict the student's preferences for the format of study materials. The recommendation system will be trained using the student's past choices and preferences, and it will predict the format of study materials that the student is likely to prefer in future semesters. The system will also take into account other factors such as the student's course of study, the availability of study materials in different formats, and the cost of the study materials.

The proposed solution has several advantages, including a reduction in paper consumption (and deforestation, as a consequence), better memory management, and increased satisfaction for students. E-books provide the flexibility of adjusting the font style and size according to the reader's preference. E-readers can store thousands of books on a single device, which makes it easier for students to carry their entire library with them. Books on paper are difficult to carry around, especially hardcovers. E-books, on the other hand, are lightweight and easy to carry.

\

Chapter 2

Literature Survey

The proposed software-based solution aims to reduce the impact of paper production on the environment and the financial burden on students by providing a choice between electronic and hardcopy formats for study materials. To achieve this, the solution uses machine learning to analyze past preferences and predict future demands, enabling the estimation of the quantity of material to be delivered in each format. The following literature survey discusses previous research on handling garbage and missing values in machine learning algorithms.

Garbage Value Removal:

Garbage values in data refer to values that are irrelevant, incorrect, or missing. Garbage values can adversely affect the accuracy of machine learning models. Several techniques have been proposed to handle garbage values, including deletion, imputation, and outlier detection. Deletion involves removing records with garbage values, imputation involves replacing missing values with estimated values, and outlier detection involves identifying and removing records with extreme values. Imputation techniques include mean, median, and mode imputation, as well as more advanced techniques such as K-Nearest Neighbors (KNN) and Expectation-Maximization (EM) algorithms. Garbage values can also be handled by using feature selection and feature engineering techniques, which involve selecting or creating new features that are less prone to garbage values.

Handling Missing Values:

Missing values in data can be handled using several techniques, including deletion, imputation, and prediction. Deletion involves removing records with missing values, imputation involves replacing missing values with estimated values, and prediction involves using machine learning algorithms to predict missing values. Imputation techniques include mean, median, and mode imputation, as well as more advanced techniques such as KNN and EM algorithms. Prediction techniques include regression analysis and decision trees. Missing values can also be handled by using feature selection and feature engineering techniques, which involve selecting or creating new features that are less prone to missing values.In the proposed solution, machine learning algorithms are used to analyze past preferences and predict future demands for study materials. Missing values in the data can be handled using imputation techniques such as mean, median, and mode imputation, as well as more advanced techniques such as KNN and EM algorithms. Garbage values in the data can be handled using deletion, imputation, and outlier detection techniques. Feature selection and feature engineering techniques can also be used to reduce the likelihood of garbage and missing values in the data.

Chapter 3

Problem Statement

The traditional method of providing study materials to students involves the printing of textbooks and other study materials, which are then distributed to the students either in person or through courier services. However, this method has several drawbacks. Firstly, it leads to the large-scale industrial production of paper, which has a significant impact on the environment. Secondly, it adds to the financial burden of the students, as they have to pay for the purchase and delivery of these materials.

To address these issues, we propose a solution that involves providing students with a choice to receive a particular volume of study materials in either electronic format (eBooks - PDFs, ODP, Docs) or hardcopy, depending upon their needs in this online era. This customized mechanism will help a learner optimize his study routine by reading books from a combination of electronic devices (laptops/smartphones/tabs) and paperback (hardcopy). This will not only help save the transport and delivery charges but will also exponentially reduce the production amount of paper in institutions on a global scale.

3.1 Project Analysis:

- Identify the current process of providing study materials to students.
- Analyze the impact of the current process on the environment and the financial burden on students.
- Study the feasibility of a software-based solution.
- Identify the requirements for the software solution, including hardware and software components.
- Analyze the data to be collected, including student preferences and choices.
- Identify the required algorithms and methodologies to develop a recommendation system.

3.2 System Design:

- Define the architecture of the software solution.
- Identify the hardware and software components required for the solution.
- Develop the data model for storing student data and material preferences.
- Design the recommendation system and its algorithms.
- Define the user interface for students, teachers, and administrative staff.
- Determine the security requirements and mechanisms for the solution.

3.2.1 Design Constraints:

- Ensure the software solution meets the needs of all stakeholders, including students, teachers, and administrative staff.
- Ensure the solution is sustainable and cost-effective.
- Ensure the solution is scalable to accommodate future growth.
- Ensure the solution is easy to use and accessible to all students.
- Ensure the solution complies with all relevant regulations and standards.
- Ensure the solution maintains the privacy and security of student data.

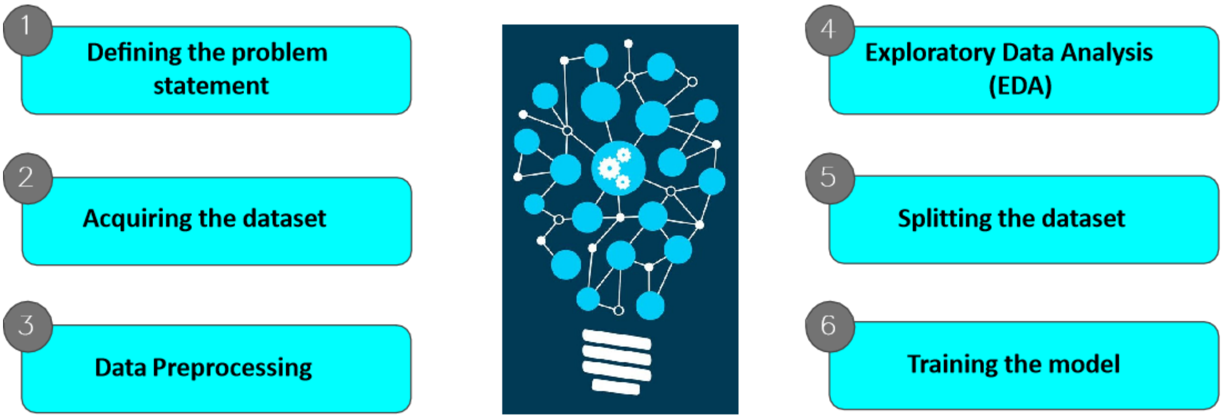
Chapter 4

Implementation

Literature Review

Author	**Title**	**Source**	**Findings**
Chao-Ying Joanne Peng, Kuk Lida Lee, Gary M. Ingersoll (2002)	An Introduction to Logistic Regression Analysis and Reporting	Indiana University Bloomington	Logistic regression is a powerful analytical technique for use when the outcome variable is dichotomous.
Adriana Erthal Abdenur (2020)	How Can Artificial Intelligence Help Curb Deforestation in the Amazon?	The Global Observatory	Tracking deforestation using satellite imagery analysis.
L . Maria Subashini (2015)	Review on Biological Treatment processes of Pulp and Paper Industry Waste Water	IJIRSET	The pulp and paper industrial waste waters are a major environmental concern.

4.1 Methodology OR Proposal



1 Defining the problem statement

2 Acquiring the dataset

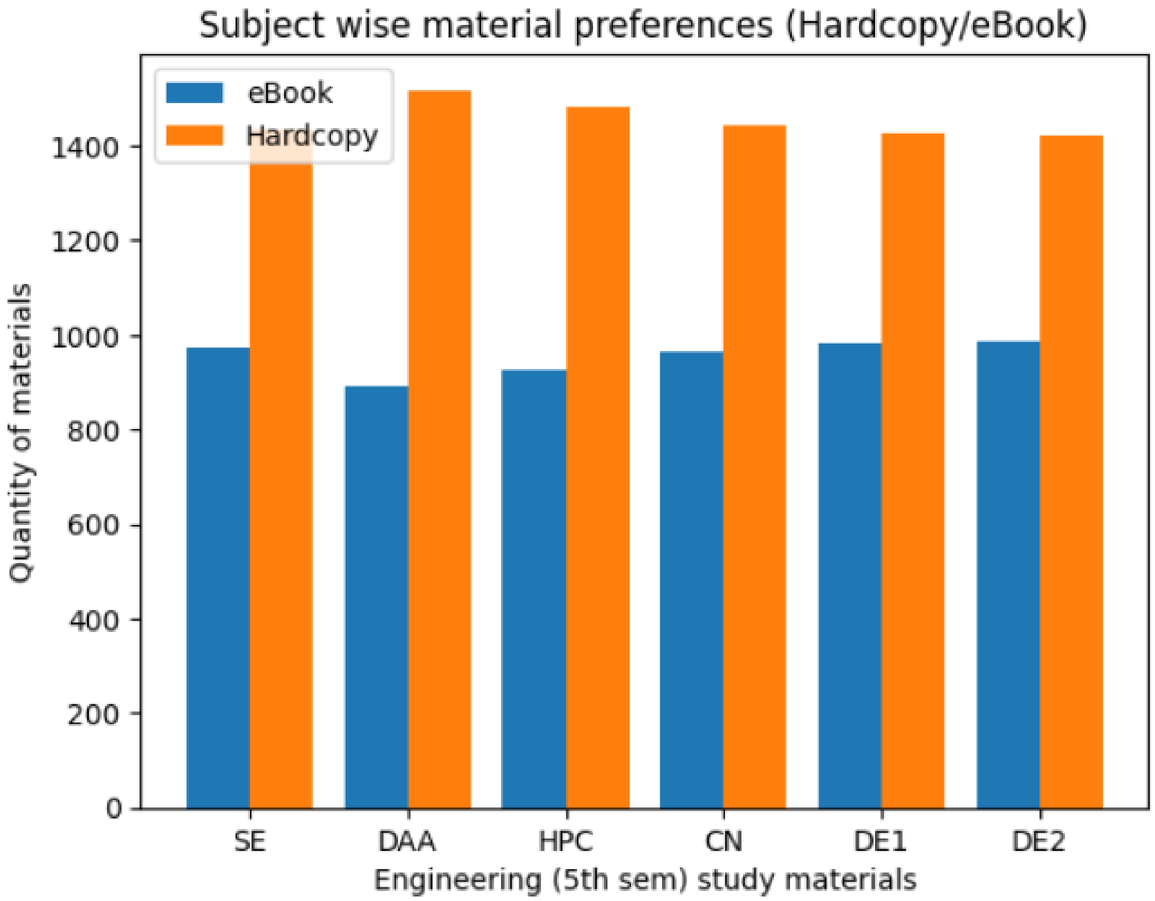
3 Data Preprocessing

4 Exploratory Data Analysis (EDA)

5 Splitting the dataset

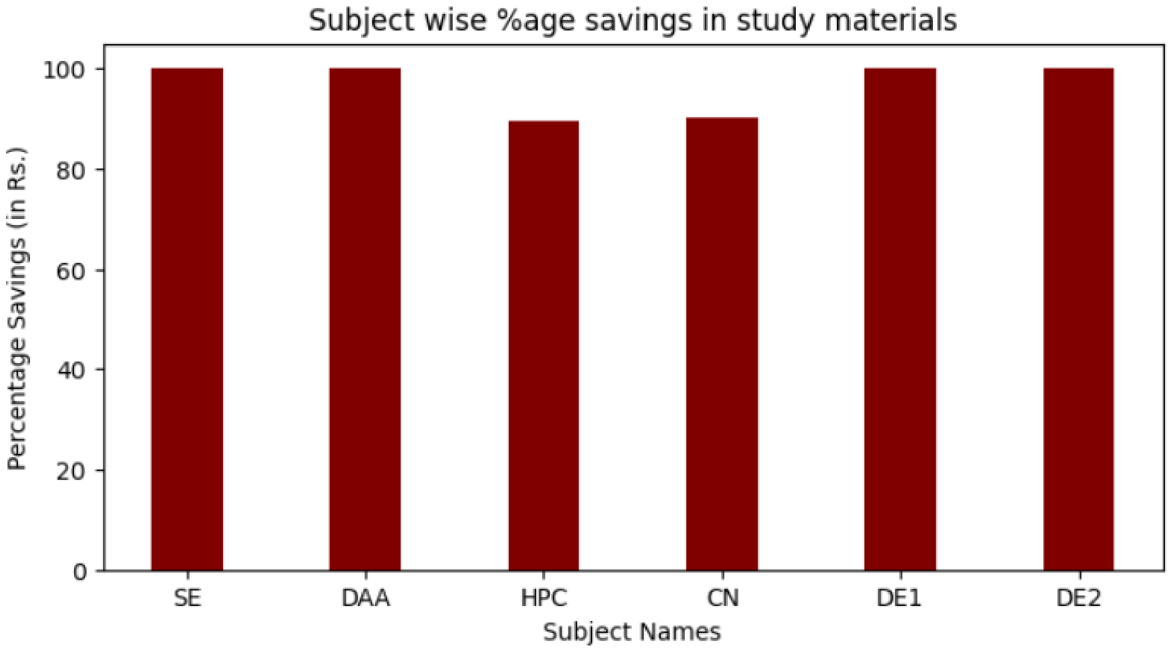
6 Training the model

4.2 Exploratory Data Analysis



Observations:

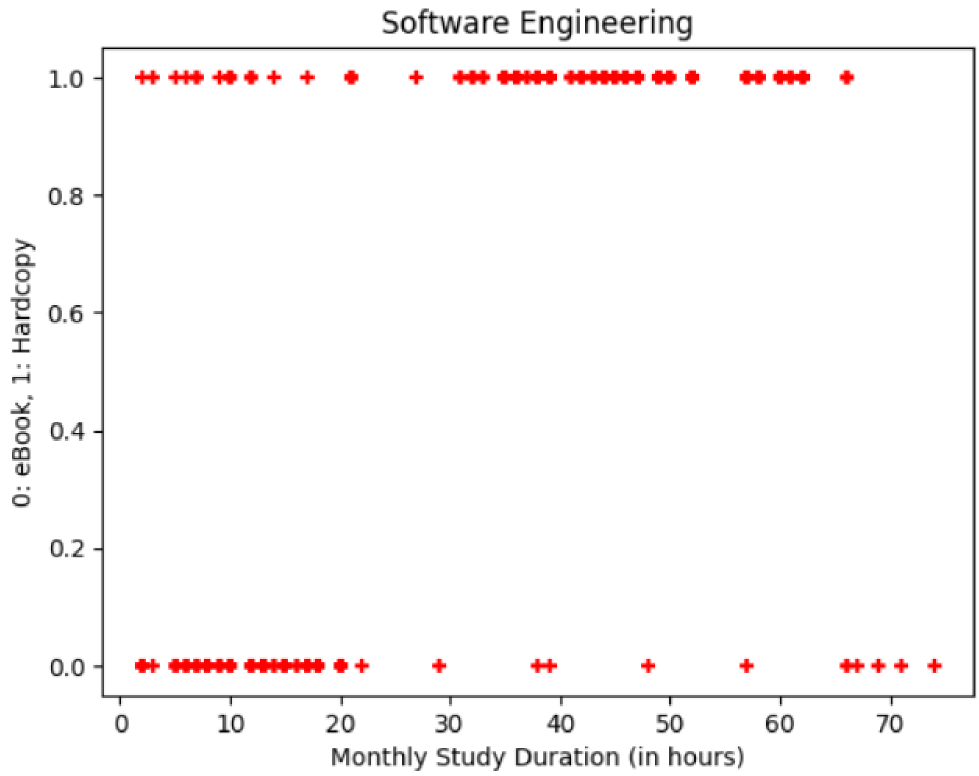
- More no. of students have chosen hardcopy in the subject **Inferential Statistics (IFS)**.
- There is a sharp similarity in the subjects of **Department Elective 1 (DE-3)** and **Open Elective 1 (OE-1)** as both these subjects show an almost identical trend.

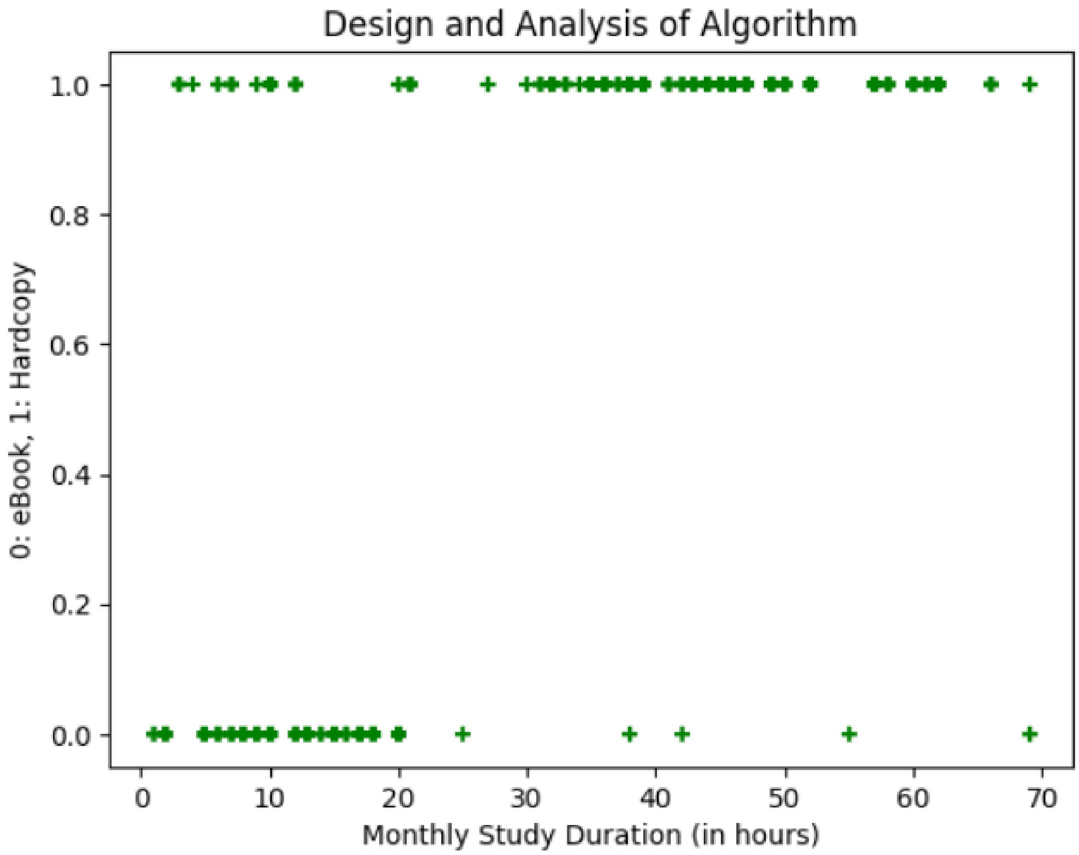


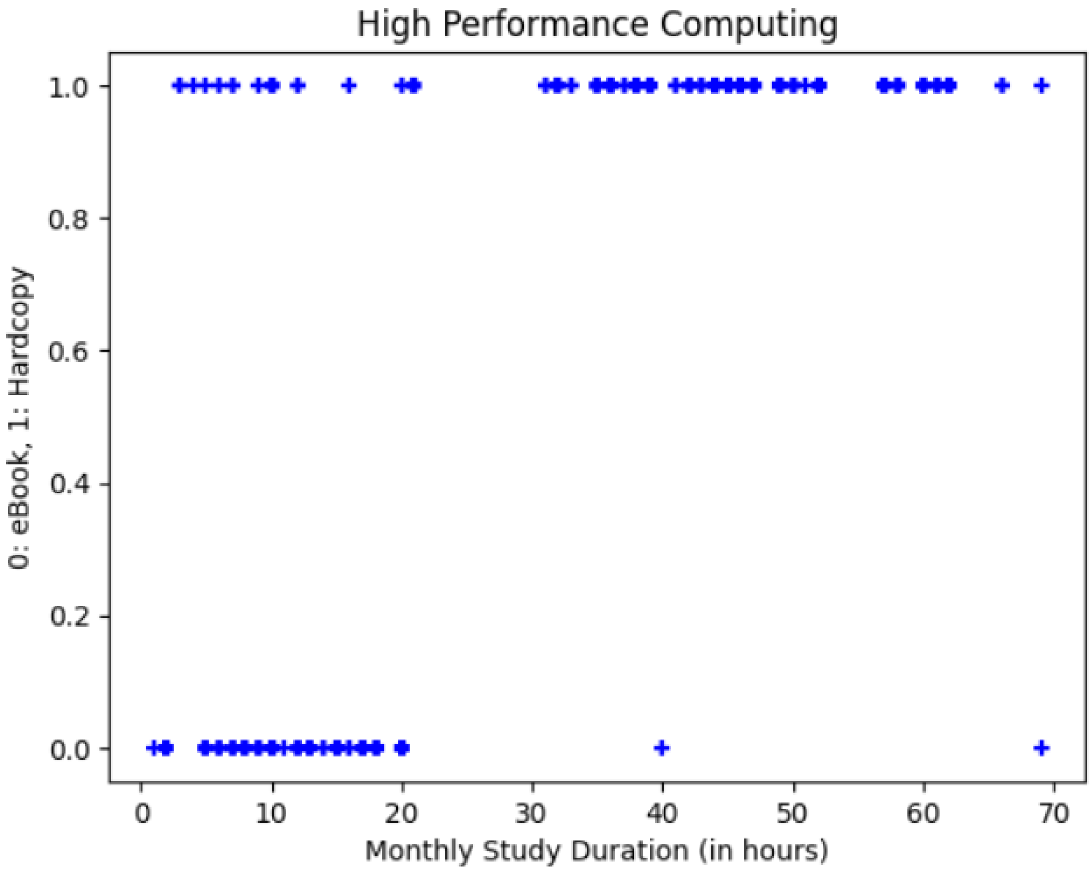
Observations:

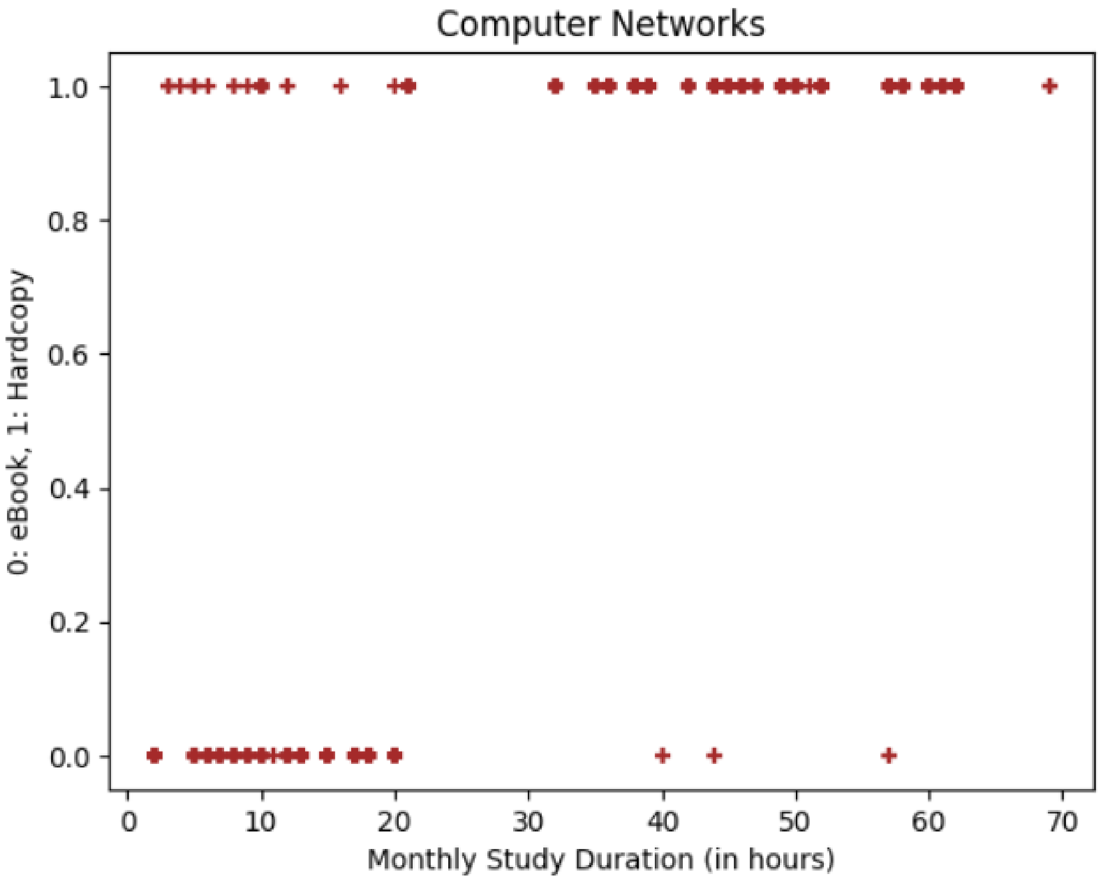
- It can be seen that 100% of the ‘total amount’ of Cellular Communication **(CC)**, Inferential Statistics **(IFS)** and Open Elective-1 **(OE-1)** is saved.
- The subject of Department Elective-3 **(DE-3)** stands lowest with around 30% savings.

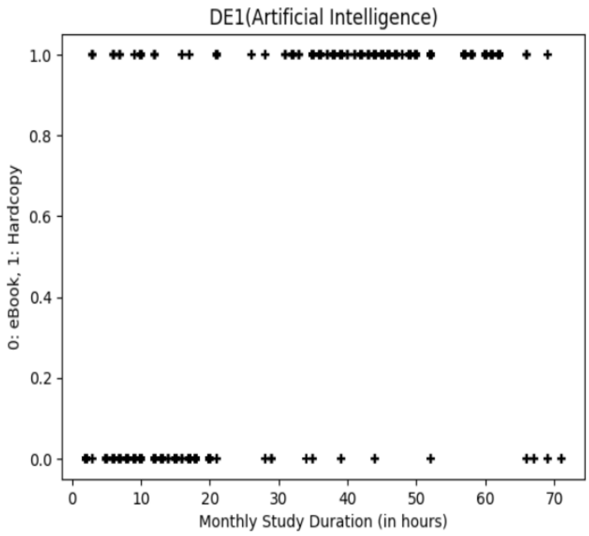
4.5 Scatter Plots

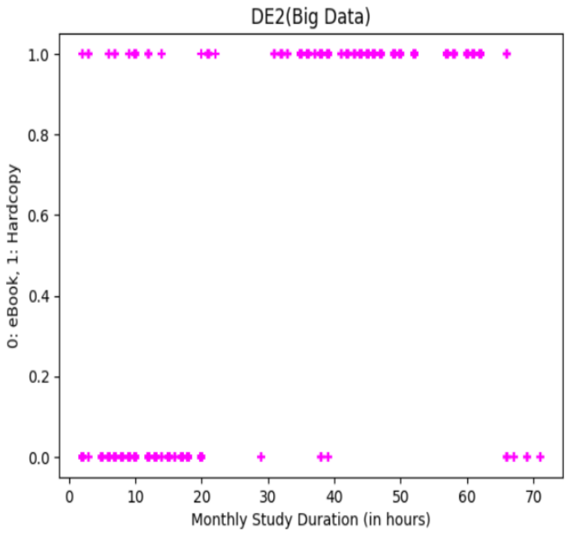





High Performance Computing
0: eBook, 1: Hardcopy
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
60
70
Monthly Study Duration (in hours)


Computer Networks
0: eBook, 1: Hardcopy
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
60
70
Monthly Study Duration (in hours)

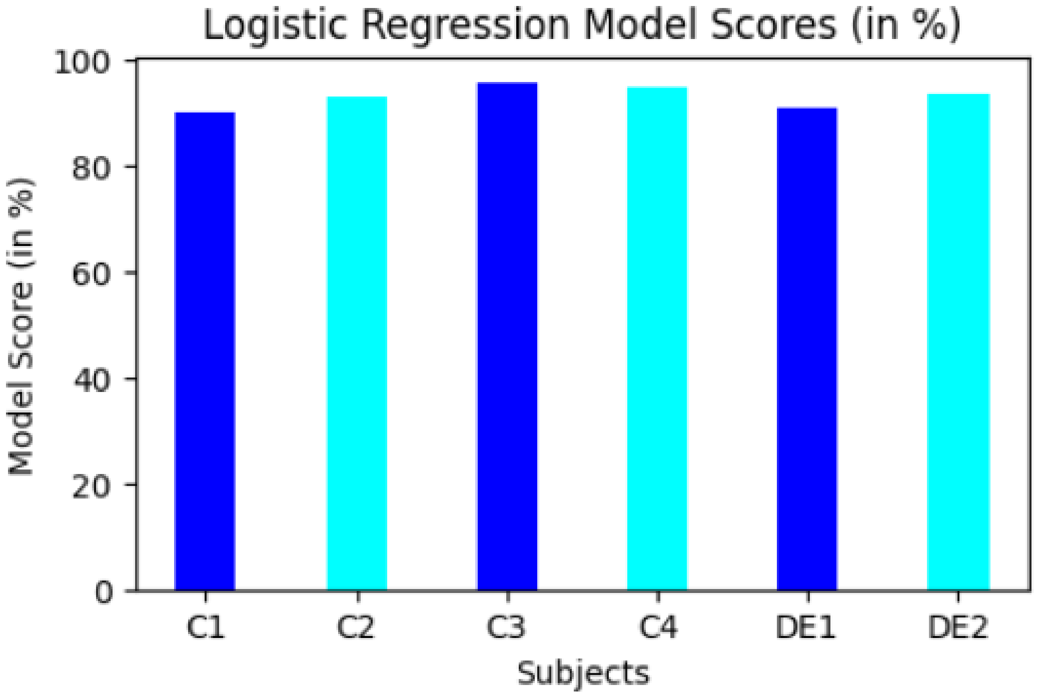




4.4 Subjectwise Model Scores

Observations:

- It can be inferred that the model score for **OE-1** stands highest with **95.63%** and **IFS** holds the second place at **93.89%**.
- **Cellular Communication** stands lowest with a regression model score of 92.15%.



Chapter 5

Standards Adopted

5.1 Design Standards

- User interface design should be intuitive and user-friendly, with a clear and simple navigation structure.

- The software should be designed to be platform-agnostic, allowing it to be used on different devices, operating systems, and web browsers.

- The system should be scalable to handle large volumes of data, including student records, course materials, and preferences.

- Data privacy and security standards should be maintained throughout the system, with secure authentication and access control mechanisms.

- The software should be designed to be modular, with separate modules for data processing, analysis, and presentation.

- The system should be designed to be robust and reliable, with built-in error handling and exception management mechanisms.

- The software should be tested thoroughly to ensure that it meets the functional requirements specified in the design, and that it is free of errors and bugs.

- The software should be designed to be easily maintainable, with clear and concise documentation, and easy-to-understand code.

- The system should be designed to be compatible with other systems and applications, allowing for seamless integration with existing systems.

- The design should be based on best practices and industry standards, including software engineering principles and coding standards.

5.2 Coding Standards

- Follow a consistent coding style throughout the project to improve readability and maintainability. Use an established coding standard, such as the Google Python Style Guide, to ensure consistency.

- Use version control systems such as Git to keep track of changes in the codebase and collaborate with other developers. Use branches to isolate different features or bug fixes, and use pull requests to review and merge changes.

- Write modular code that is easy to understand and maintain. Use functions and classes to encapsulate logic and improve code reuse.

- Follow best practices for machine learning, such as splitting the data into training and validation sets, using appropriate evaluation metrics, and avoiding overfitting.

- Document the code thoroughly to make it easier for other developers to understand and use. Use descriptive variable and function names, add comments to explain complex logic, and provide documentation strings for functions and classes.

- Write automated tests to ensure that the code behaves as expected and catches errors early in the development process. Use a testing framework such as pytest or unittest, and aim for high test coverage.

- Use continuous integration and continuous deployment (CI/CD) to automate the testing and deployment process. Use tools such as Travis CI or Jenkins to run automated tests and deploy the code to a staging or production environment.

- Use security best practices to protect sensitive data and prevent unauthorized access to the system. Use secure coding practices, such as parameterized queries and input validation, to prevent SQL injection and other attacks.

- Use logging to track errors and events in the system. Use a logging framework such as Python's built-in logging module to log events to a file or a centralized logging service.

- Follow software development life cycle (SDLC) processes to ensure that the project is delivered on time and meets the requirements. Use agile development methodologies, such as Scrum or Kanban, to manage the project and prioritize tasks.

5.3 Testing Standards

- Unit Testing: Each module of the software solution should be tested individually to ensure that it functions correctly. This involves testing the code, input/output parameters, and expected output for each module.

- Integration Testing: Once the individual modules have been tested, they need to be integrated to ensure that they function correctly together. This involves testing the interaction between different modules, and verifying that the data is being passed correctly between them.

- System Testing: Once the integration testing has been completed, the entire system should be tested to ensure that it meets the requirements of the problem statement. This involves testing the system as a whole, and verifying that it meets the needs of the users and the environment.

- Acceptance Testing: This stage involves testing the system with actual users to ensure that it meets their requirements and expectations. The users should be asked to perform various tasks and provide feedback on the system's usability, reliability, and performance.

- Regression Testing: After any changes or updates to the system, regression testing should be performed to ensure that the system still works correctly and that no new errors have been introduced.

- Security Testing: This involves testing the system's security features to ensure that they are robust and reliable. It includes testing for vulnerabilities, authentication, access control, and data protection.

- Performance Testing: This involves testing the system's performance under different conditions, including heavy usage, multiple users, and peak loads. It includes testing for speed, scalability, and reliability.

Chapter 6

Conclusion and Future Scope

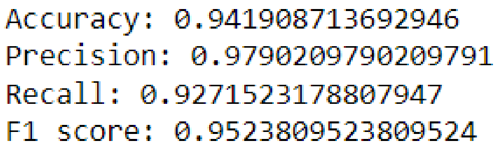
Conclusion:

The COVID-19 pandemic has resulted in a shift towards electronic media for students and professionals worldwide, leading to an increase in the delivery of study materials via courier. However, this method has significant drawbacks, including high costs and a significant impact on the environment due to the large-scale industrial production of paper. This paper proposes a software-based solution that allows students to choose between electronic and hardcopy formats for their study materials based on their preferences. By using machine learning algorithms to analyze past preferences and predict future demands, the solution enables institutions to estimate the quantity of material to be delivered in each format, reducing transport and delivery costs and significantly decreasing the amount of paper production in institutions on a global scale.

The proposed solution has several advantages, including a reduction in paper consumption, font style and size flexibility, better memory management, efficient carrying, and the satisfaction of having an entire library at your fingertips. These advantages not only help to reduce the financial burden on students but also promote better energy management and water treatment, while also encouraging the concept of reducing, reusing, and recycling.

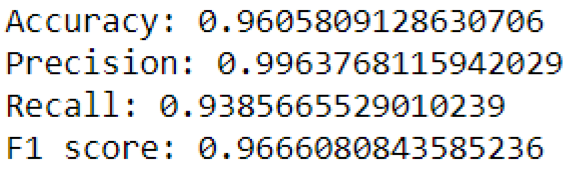
Implementing Random Forest:

Random Forest classification for Software Engineering:



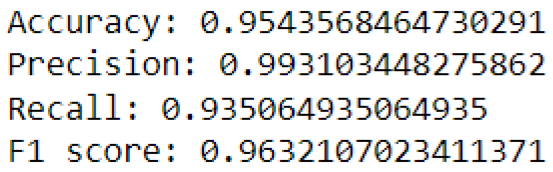
```
Accuracy: 0.941908713692946
Precision: 0.9790209790209791
Recall: 0.9271523178807947
F1 score: 0.9523809523809524
```

Random Forest classification for Design and Analysis of Algorithm:



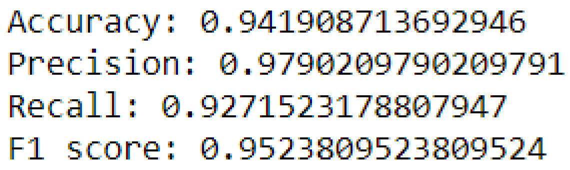
```
Accuracy: 0.9605809128630706
Precision: 0.9963768115942029
Recall: 0.9385665529010239
F1 score: 0.9666080843585236
```

Random Forest classification for High-Performance Computing:



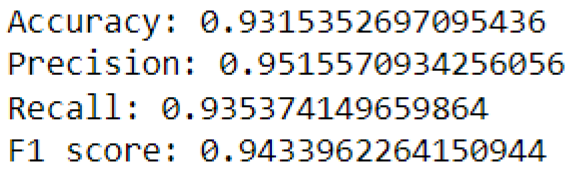
```
Accuracy: 0.9543568464730291
Precision: 0.993103448275862
Recall: 0.935064935064935
F1 score: 0.9632107023411371
```

Random Forest classification for Computer Networks:



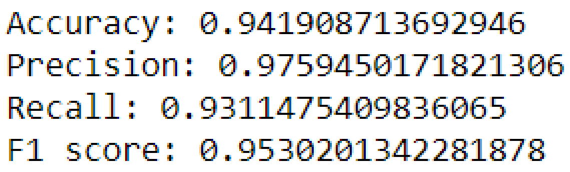
```
Accuracy: 0.941908713692946
Precision: 0.9790209790209791
Recall: 0.9271523178807947
F1 score: 0.9523809523809524
```

Random Forest classification for DE1(Artificial Intelligence):



```
Accuracy: 0.9315352697095436
Precision: 0.9515570934256056
Recall: 0.935374149659864
F1 score: 0.9433962264150944
```

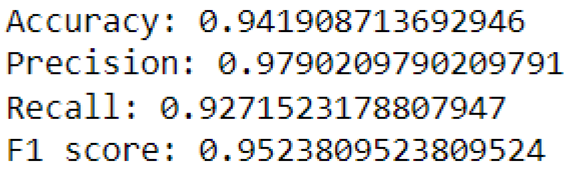
Random Forest classification for DE2(Big Data):



```
Accuracy: 0.941908713692946
Precision: 0.9759450171821306
Recall: 0.9311475409836065
F1 score: 0.9530201342281878
```

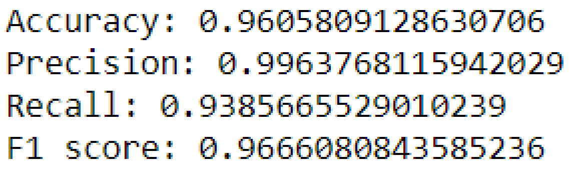
Decision Tree:

Decision Tree classification for Software Engineering:



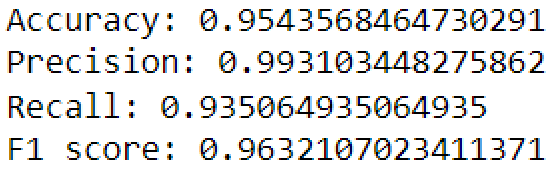
```
Accuracy: 0.941908713692946
Precision: 0.9790209790209791
Recall: 0.9271523178807947
F1 score: 0.9523809523809524
```

Decision Tree classification for Design and Analysis of Algorithm:



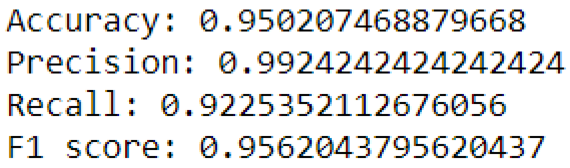
```
Accuracy: 0.9605809128630706
Precision: 0.9963768115942029
Recall: 0.9385665529010239
F1 score: 0.9666080843585236
```

Decision Tree classification for High Performance Computing:



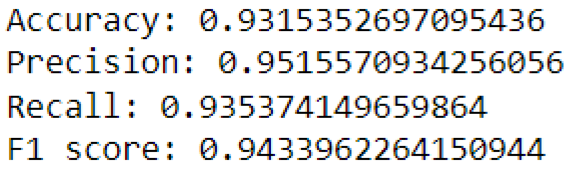
```
Accuracy: 0.9543568464730291
Precision: 0.993103448275862
Recall: 0.935064935064935
F1 score: 0.9632107023411371
```

Decision Tree classification for Computer Networks:



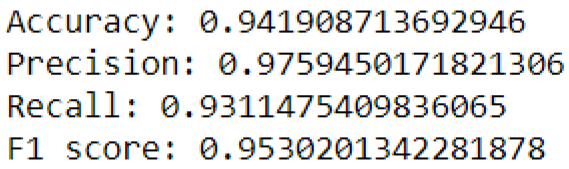
```
Accuracy: 0.950207468879668
Precision: 0.9924242424242424
Recall: 0.9225352112676056
F1 score: 0.9562043795620437
```

Decision Tree classification for DE1(Artificial Intelligence):



```
Accuracy: 0.9315352697095436
Precision: 0.9515570934256056
Recall: 0.935374149659864
F1 score: 0.9433962264150944
```

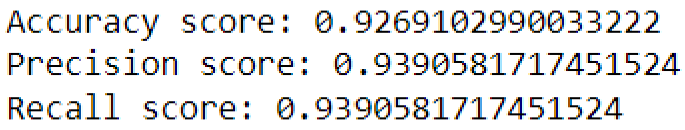
Decision Tree classification for DE2(Big Data):



```
Accuracy: 0.941908713692946
Precision: 0.9759450171821306
Recall: 0.9311475409836065
F1 score: 0.9530201342281878
```

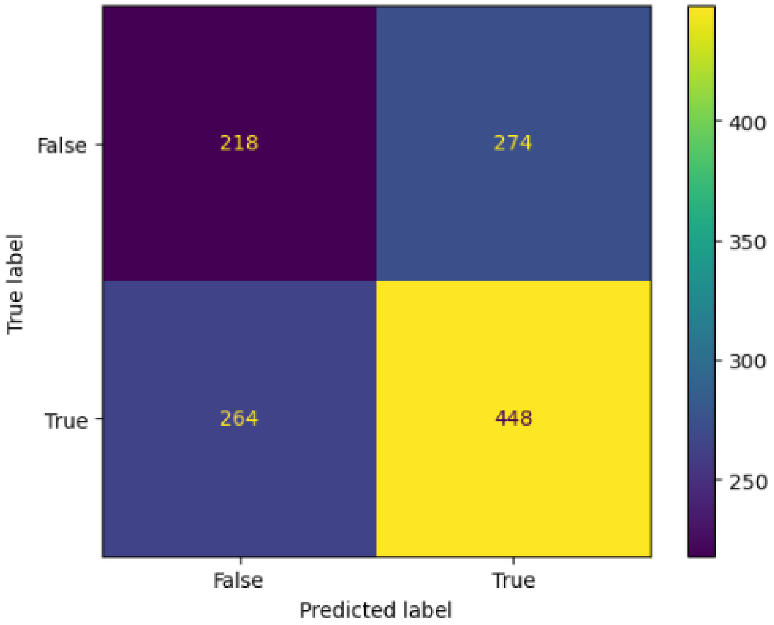
K-Nearest Neighbor

Applying k-Nearest Neighbor for Software Engineering:



```
Accuracy score: 0.9269102990033222
Precision score: 0.9390581717451524
Recall score: 0.9390581717451524
```

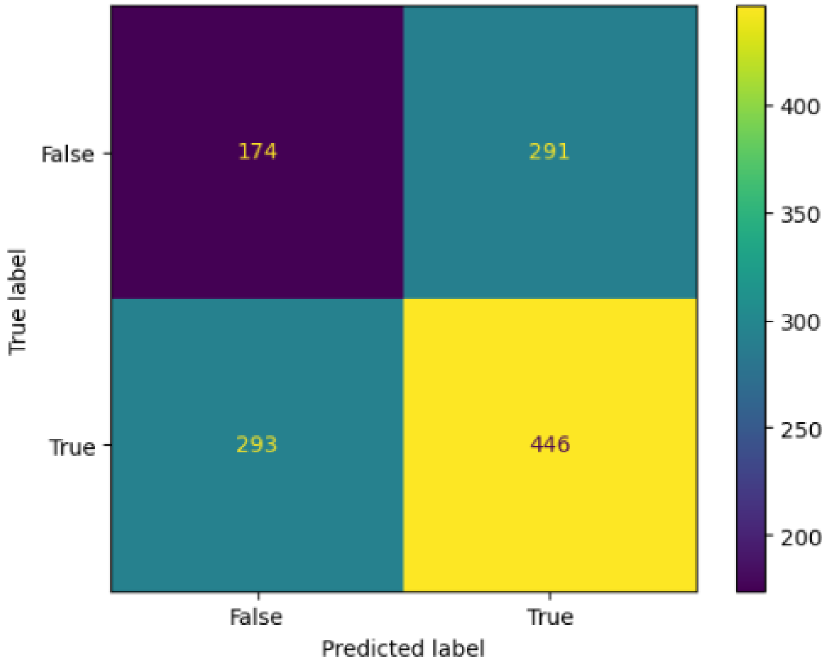
```
[[218 274]
 [264 448]]
```



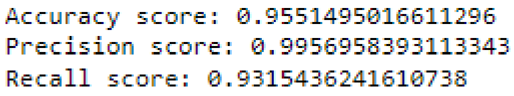
Applying k-Nearest Neighbor for Design and Analysis of Algorithm:

```
Accuracy score: 0.9551495016611296
Precision score: 0.9918588873812755
Recall score: 0.938382541720154
```

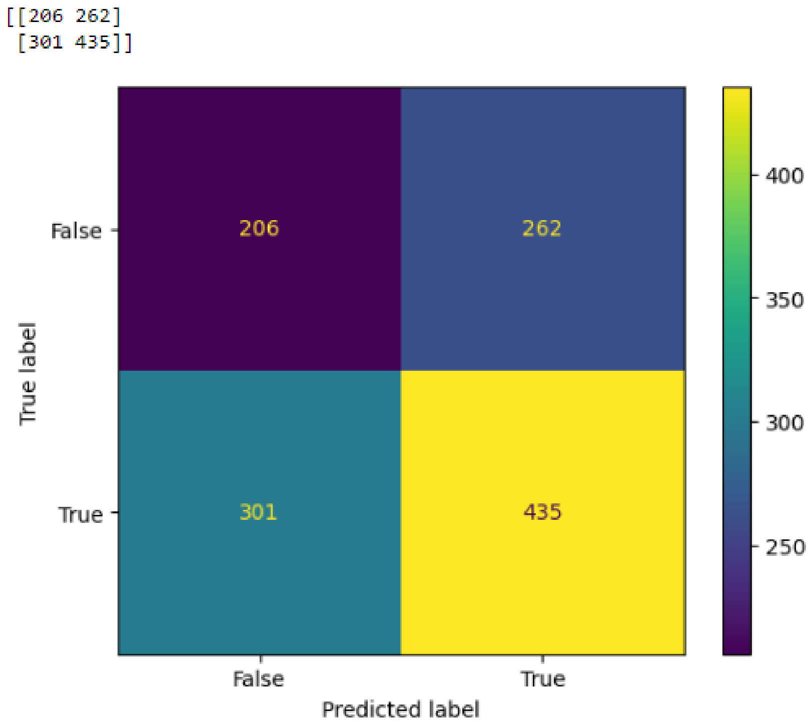
```
[[174 291]
 [293 446]]
```



Applying k-Nearest Neighbor for High Performance Computing:

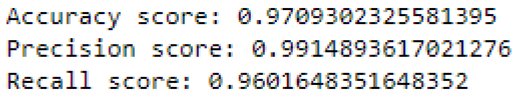


```
Accuracy score: 0.9551495016611296
Precision score: 0.9956958393113343
Recall score: 0.9315436241610738
```

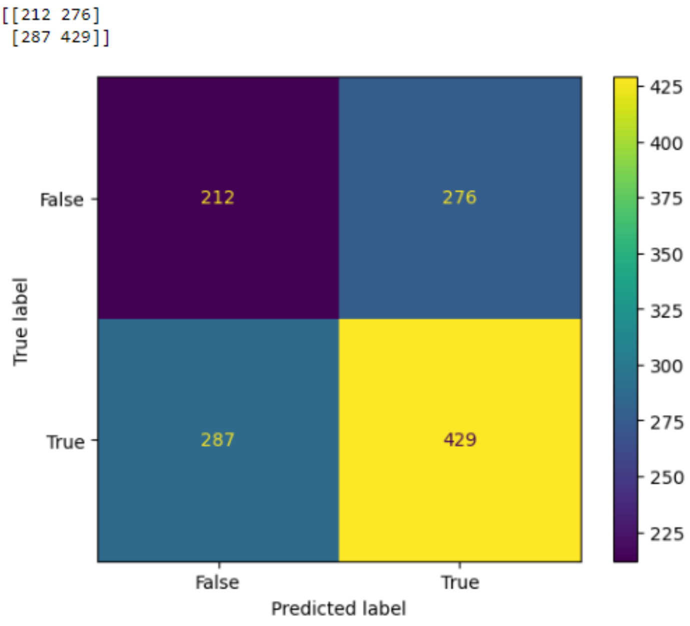


```
[[206 262]
 [301 435]]
```

Applying k-Nearest Neighbor for Computer Networks:

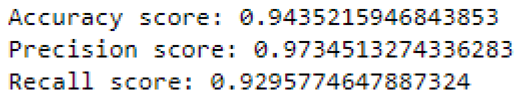


```
Accuracy score: 0.9709302325581395
Precision score: 0.9914893617021276
Recall score: 0.9601648351648352
```

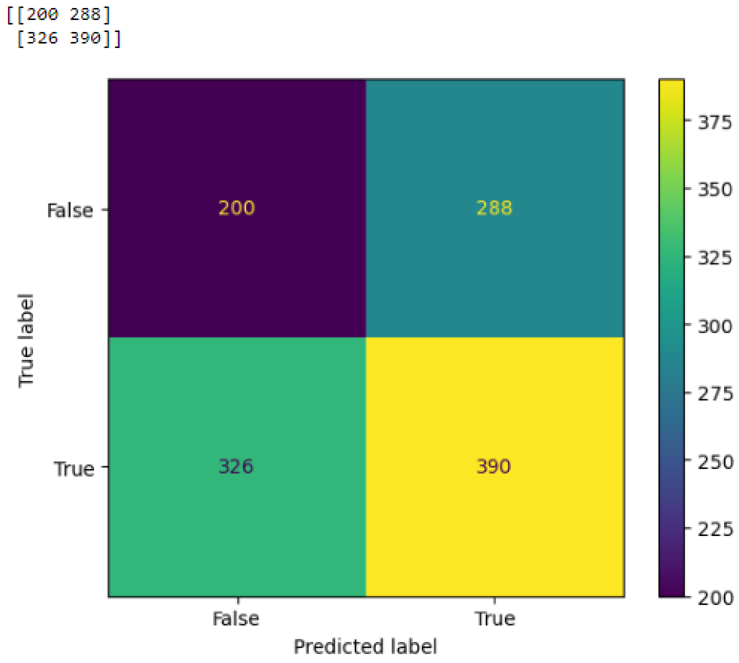


```
[[212 276]
 [287 429]]
```

Applying k-Nearest Neighbor for DE1(Artificial Intelligence):

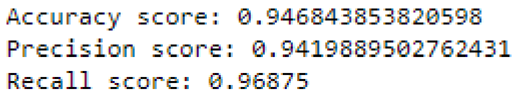


```
Accuracy score: 0.9435215946843853
Precision score: 0.9734513274336283
Recall score: 0.9295774647887324
```

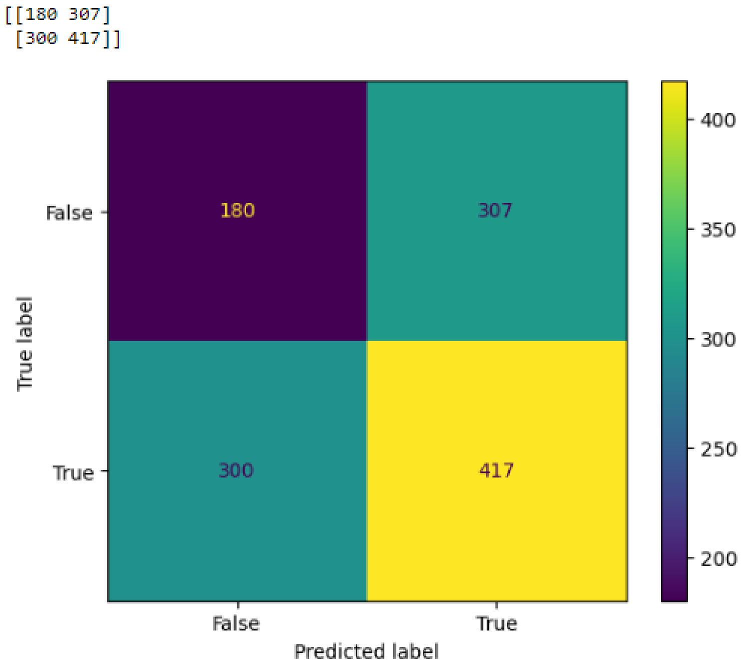


```
[[200 288]
 [326 390]]
```

Applying k-Nearest Neighbor for DE2(Big Data):



```
Accuracy score: 0.946843853820598
Precision score: 0.9419889502762431
Recall score: 0.96875
```



```
[[180 307]
 [300 417]]
```

Advantages

1. The proposed solution has several advantages, which are as follows:

2. Reduction in paper consumption (and deforestation, as a consequence): By providing students with a choice to receive study materials in electronic format, the solution will help in reducing the production amount of paper in institutions on a global scale, which will, in turn, reduce deforestation.

3. E-books come with font style and size flexibility: E-books provide the flexibility of adjusting the font style and size according to the reader's preference.

4. Better Memory Management: E-readers can store thousands of books on a single device, which makes it easier for students to carry their entire library with them.

5. Efficient in carrying: Books on paper are difficult to carry around, especially hardcovers. E-books, on the other hand, are lightweight and easy to carry.

6. Satisfaction of having an entire library at your fingertips, an infinite supply just a click away, ready to download instantly

Facts

1. The IT industry once predicted that its emergence is to make paperless offices but it's amazing that about 95% of office work worldwide is still done on paper.

2. 42% of all global wood harvest is used to make paper. Is it really worth it to cut down our life-saving trees for this product? Also, the bleaching reagents without proper treatment that get released into river bodies prove to be graveyards for hydrological ecology.

3. According to a survey, an average tree gets converted into 12000 A4 sheets. As per our calculations, we require almost 300 trees to run an engineering semester (6-month duration) and a tree requires a minimum of 3 years to grow and become mature.

Proposed Solution:
Software-Based Solution for Study Materials

The COVID-19 pandemic has resulted in a significant shift toward electronic media for students and professionals worldwide. To address the impact of this shift on the environment and the high cost of study material delivery, we propose a software-based solution that allows students to choose between electronic and hardcopy formats based on their preferences. Our solution uses machine learning to analyze past preferences and predict future demands, enabling us to estimate the quantity of material to be delivered in each format. By doing so, we reduce transport and delivery costs, as well as the amount of paper production in institutions on a global scale.

The software solution will be based on a recommendation system that uses machine learning algorithms to predict the student's preferences for the format of study materials. The recommendation system will be trained using the student's past choices and preferences, and it will predict the format of study materials that the student is likely to prefer in future semesters. The system will also take into account other factors such as the student's course of study, the availability of study materials in different formats, and the cost of the study materials.

Our solution addresses the problem of traditional study material distribution, which involves the printing of textbooks and other study materials that are then distributed to students. This method has several drawbacks, including the large-scale industrial production of paper and the financial burden on students. With our solution, students can choose to receive a particular volume of study materials in either electronic or hardcopy format, depending on their needs.

Our solution not only reduces the impact on the environment but also provides several benefits to students. E-books provide font style and size flexibility, making it easier for students to read according to their preferences. Additionally, e-readers can store thousands of books on a single device, making it easier for students to carry their entire library with them. Books on paper are difficult to carry around, especially hardcovers, while e-books are lightweight and easy to carry. Furthermore, students can have an entire library at their fingertips, with an infinite supply just a click away, ready to download instantly.

Future Scope

The proposed software-based solution has the potential to make a significant impact on the environment and the delivery of study materials. However, there is still room for improvement and future development.

One possible future scope is to expand the recommendation system to include other factors that may impact a student's preferences, such as their learning style or the time of day they prefer to study. This could lead to a more personalized approach to study material distribution, further improving the student's learning experience. Another future scope is to integrate the solution with existing learning management systems. This would enable seamless integration of study materials and delivery, making it easier for students to access their study materials in their preferred format. Furthermore, there is potential for the solution to expand beyond study materials and into other areas, such as office work or government paperwork. By adopting a more sustainable and cost-effective approach to document distribution, we can significantly reduce the environmental impact of paper production while reducing costs for businesses and individuals alike.

In conclusion, the proposed software-based solution offers a sustainable and cost-effective solution to the traditional delivery of study materials. By using machine learning to analyze past preferences and predict future demands, we reduce transport and delivery costs, as well as the amount of paper production in institutions on a global scale. With the potential for future expansion and development, our solution has the potential to make a significant impact on the environment and document distribution.

References

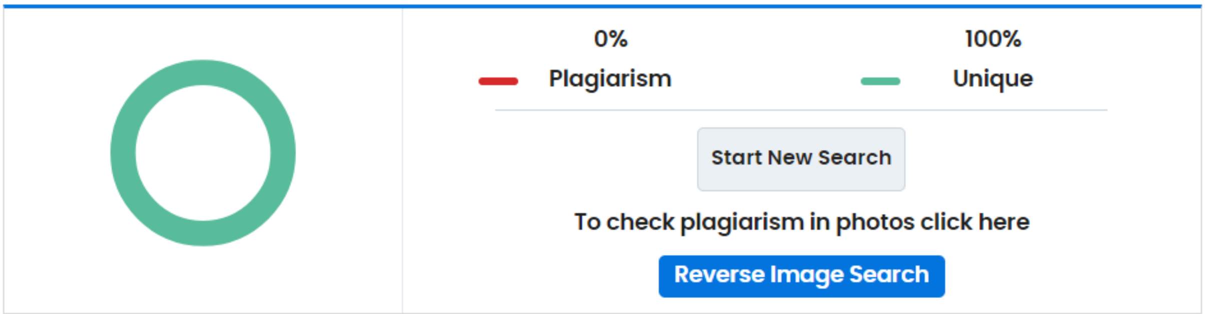
Chao-Ying Joanne Peng, Kuk Lida Lee, Gary M. Ingersoll (2002). "An Introduction to Logistic Regression Analysis and Reporting" Indiana University-Bloomington.

Adriana Erthal Abdenur (2020). "How Can Artificial Intelligence Help Curb Deforestation in the Amazon?" The Global Observatory.

L . Maria Subashini (2015). "Review on Biological Treatment processes of Pulp and Paper Industry Waste Water" IJIRSET.

TURNITIN PLAGIARISM REPORT

(This report is mandatory for all the projects and plagiarism must be below 25%)



To check plagiarism in photos click here