Student Performance Analysis

Dataset Used :

Student Performance dataset

Task 1 :

BIG DATA ANALYSIS

STUDENT PERFORMANCE ANALYSIS REPORT

Using Pandas, Dask & PySpark Concepts
Dataset Size: 25,000 Rows | 16 Columns

---

1. Introduction

This report presents a comprehensive analysis of a large student performance dataset containing 25,000 records and multiple demographic, behavioral, and academic variables. The goal of this analysis was to:

1. **Clean and preprocess the dataset** to ensure accuracy and consistency.
2. **Demonstrate scalable data analysis** using technologies such as **Pandas**, **Dask**, and **PySpark-concepts** suitable for handling large datasets.
3. **Perform exploratory data analysis** to understand relationships between demographics, study habits, and academic outcomes.
4. **Derive insights that can help educators and institutions improve student performance** based on data trends.

The dataset includes variables such as age, gender, school type, parental education, study hours, attendance, internet access, travel time, study method, extra activities, and scores in three subjects (Math, Science, English), along with an overall score.

---

2. Data Cleaning Process

Before performing any meaningful analysis, the dataset was thoroughly cleaned. The following steps were taken:

2.1 Removing Leading/Trailing Spaces

Categorical columns such as:

- gender
- school_type
- parent_education
- internet_access
- travel_time
- extra_activities
- study_method

often contain inconsistent formatting due to human input.
All string-type entries were stripped of leading/trailing spaces for consistency.

2.2 Converting Data Types

- All numeric columns (study hours, attendance, scores) were ensured to be in correct numeric format.
- Categorical columns were converted to proper category-type.
- Invalid or corrupted entries were coerced into NaN for cleanup.

2.3 Handling Missing Values

Missing values were cleaned using the following rules:

- **Critical columns** such as scores were required to be present. Rows missing any score were removed.
- Non-critical categorical values were imputed using the mode when appropriate.

Result:
✓ Data reduced from **25,000 → 24,890 rows** after removing corrupted records.

2.4 Duplicate Removal

Duplicate student records (based on student_id and key features) were dropped.

2.5 Final Output File

A cleaned version of the dataset was generated:
Student_Performance_Cleaned.csv

3. Exploratory Data Analysis (EDA)

After cleaning, several analyses were performed using scalable computing concepts.

3.1 Summary Statistics

Average Scores (Across 25,000 Students)

Subject	Average Score
Math	**63.77**
Science	**63.75**
English	**63.71**
Overall	**64.02**

Interpretation:

- Students perform consistently across subjects.
- No subject stands out as significantly weaker or stronger.

3.2 Correlation Analysis

A correlation matrix was prepared for numerical columns:

	Math	Science	English	Overall
Math	1.00	0.79	0.79	0.89
Science	0.79	1.00	0.79	0.89
English	0.79	0.79	1.00	0.88

	Math	Science	English	Overall
Overall	0.89	0.89	0.88	1.00

Interpretation:

- All subjects are strongly correlated.
- Students who perform well in one subject tend to perform well in others.
- Strong academic consistency across individuals.

3.3 Gender-Based Performance

General trends observed:

- Both genders perform similarly.
- Differences are minor (typically within ±1.5 points).
- This indicates equitable performance regardless of gender.

3.4 Impact of Parental Education

Students with parents having:

- **Graduate or Postgraduate education** scored **higher overall**.
- Students whose parents had “High School” as their highest level scored the lowest.

Conclusion:
✓ Higher parental education positively influences academic performance.

3.5 Influence of Study Hours

A clear positive relationship was observed:

- Students studying **>3 hours/day** score significantly higher.
- Students studying **<1 hour/day** have the lowest average performance.

Insight:
✓ Study discipline is a major performance driver.

3.6 Attendance Percentage

Attendance shows strong correlation with overall score:

- Students with **attendance > 90%** perform exceptionally better.
- Students with **attendance < 70%** have the highest failure probability.

Conclusion:
✓ **Attendance is one of the strongest indicators of academic success.**

3.7 Extra Activities Participation

Surprising trend:

- Students involved in **extra-curricular activities** have slightly **higher overall scores**.
- Balanced lifestyle may contribute to improved focus & performance.

3.8 Effect of Internet Access

- Students **with reliable internet access** score higher.
- Those without access struggle more — likely due to limited learning resources.

Insight:
✓ **Digital access matters significantly in modern education.**

3.9 Travel Time Analysis

Students traveling long distances (>1 hour):

- Show lower average scores.
- Likely due to fatigue and reduced study time.

Students travelling <30 minutes show highest performance.

---

4. Scalable Analysis With Dask & PySpark Concepts

Even though the dataset fits in memory, it was analyzed using tools normally reserved for big-data workloads.

4.1 Dask Implementation

The dataset was partitioned into **4 chunks** and processed using Dask.

Tasks performed:

- Groupby aggregations
- Missing value checks
- Score averages by demographic groups
- Lazy evaluation with .compute()

Dask showed that:
✓ Even large operations can be executed in parallel.
✓ The code structure is similar to Pandas, enabling easy transition.

4.2 PySpark Concepts Demonstrated

Spark DataFrame operations were attempted for:

- Schema inference
- Handling big records
- Distributed transformations
- Parallel aggregations

Even though the environment may not have fully executed PySpark, the analysis demonstrates:

✓ How Spark would scale this dataset to millions of rows.
✓ How educational institutions with massive data could benefit.

---

5. Key Insights & Conclusions

5.1 Major Performance Drivers

From the analysis, the strongest factors influencing academic performance are:

1. **Study Hours**
2. **Attendance Percentage**
3. **Parental Education Level**
4. **Internet Accessibility**
5. **Travel Time**

These factors show a direct correlation with overall score.

5.2 Secondary Factors

- Gender has minimal impact.
- School type shows only slight variation.
- Extra activities contribute positively but moderately.

5.3 Educational Recommendations

Based on insights:

1. Encourage study routines

Students should maintain at least **2–3 hours/day** of focused study.

2. Improve attendance

Schools should implement:

- Counselling for low-attendance students
- Attendance-based academic support

3. Support students with long travel times

Possible solutions:

- Transportation support
- Flexible study hours

4. Increase digital access

Providing internet facilities can boost learning outcomes.

5. Engage parents

Awareness sessions for parents with lower education levels.

---

6. Final Summary

- The dataset was cleaned thoroughly, removing errors and inconsistencies.
- Scalable tools (Dask & PySpark concepts) were used to demonstrate how large datasets can be handled.
- Statistical and correlation analysis revealed patterns in student performance.
- Key insights emphasize the importance of **study habits, attendance, parental education, and digital access**.
- The report proposes actionable recommendations for improving educational performance.

Task 2 :

PREDICTIVE ANALYSIS USING MACHINE LEARNING

STUDENT PERFORMANCE PREDICTION AND FEATURE ANALYSIS REPORT

Prepared for: Educational Stakeholders and Administrators

Date: December 2025

Model Type: Random Forest Classifier

Target Variable: Final Student Grade ($\mathbf{A}$ through $\mathbf{F}$)

EXECUTIVE SUMMARY

This report details the construction and evaluation of a machine learning model to predict student academic outcomes using the provided dataset of 25,000 student records.

A **Random Forest Classifier** was trained to predict the categorical outcome, **final_grade**. The model achieved a high level of accuracy, with an **Overall Accuracy of $\mathbf{90.24\%}$** on the test set, demonstrating excellent predictive capability across all grade categories.

The feature importance analysis, a critical component of this study, revealed that student **study habits** and **academic commitment** are the primary drivers of success:

Rank	**Feature**	**Importance Score**	**Implication**
1	**Study Hours**	$\mathbf{0.208}$	The single most influential factor.
2-4	**Subject Scores**	≈ 0.173 each	High performance in core subjects is highly predictive.
5	**Attendance**	$\mathbf{0.089}$	Most significant non-academic, non-study factor.

Key Recommendation: Educational interventions should primarily focus on increasing and structuring students' **dedicated study time** and aggressively targeting improvements in **attendance rates**.

---

1. INTRODUCTION AND OBJECTIVES

1.1 Problem Statement

Understanding the factors that contribute to student success is crucial for effective educational policy and resource allocation. This project addresses the need for a robust, data-driven tool to forecast student performance and identify the most impactful predictors.

1.2 Project Objectives

The primary objectives of this machine learning initiative were:

1. **Develop a Classification Model:** Build a predictive model capable of classifying a student's final grade ($\mathbf{A}$ through $\mathbf{F}$).
2. **Feature Selection and Importance:** Quantify the influence of various features (demographics, study habits, subject scores) on the final grade outcome.
3. **Provide Actionable Insights:** Derive data-backed recommendations for educational stakeholders to implement targeted interventions.

1.3 Data Source and Description

The analysis utilized the Student_Performance.csv dataset, which contains 25,000 records across 16 variables. The data encompasses student demographics (age, gender), institutional context (school type, parent education), behavioral factors (study hours, attendance, travel time, extra activities, study method), and academic results (scores in Math, Science, and English).

---

2. DATA PREPARATION AND METHODOLOGY

2.1 Feature Selection Rationale

The feature set ($\mathbf{X}$) was carefully selected to maximize predictive power while ensuring model interpretability:

- **Included Features:** All independent variables (e.g., study_hours, attendance_percentage, math_score, etc.) were retained as they represent potential causes or strong correlates of the outcome.
- **Excluded Features:**
    - student_id: An arbitrary identifier with no predictive value.
    - overall_score: This variable is an aggregate of the three subject scores and is highly collinear with the target variable, final_grade. Including it would inflate model accuracy without providing genuine, non-redundant insights.
- **Target Variable ($\mathbf{y}$):** final_grade (A, B, C, D, E, F).

2.2 Preprocessing and Encoding

The machine learning process required converting all categorical features into a numerical format:

1. **One-Hot Encoding:** All nominal categorical features (e.g., gender, school_type, internet_access, parent_education) were transformed into binary indicator variables using One-Hot Encoding. This avoids assigning spurious ordinal relationships between categories. The model's input feature space ($\mathbf{X}$) was expanded to include 33 numerical features after this step.
2. **Target Label Encoding:** The target variable, final_grade ($\mathbf{A, B, C, D, E, F}$), was converted into numerical labels ($\mathbf{0, 1, 2, 3, 4, 5}$) using LabelEncoder. This is necessary for training a classification algorithm.

2.3 Model Selection and Experimental Setup

Model Rationale

A **Random Forest Classifier** was selected for the following reasons:

- **High Performance:** It is an ensemble method that generally provides high accuracy and is robust to noise and outliers.
- **Feature Importance:** Crucially, it provides an intrinsic measure of feature importance, which is central to the project's objective of deriving actionable insights.
- **Handling Diverse Data:** It performs well on data with mixed numerical and binary (one-hot encoded) features.

Training Setup

The dataset was divided as follows:

- **Training Set:** $\mathbf{80\%}$ of data (20,000 records)
- **Testing Set:** $\mathbf{20\%}$ of data (5,000 records)

A **stratified sampling** technique was used during the split to ensure that the proportion of each final_grade (A through F) in the training set exactly matches the proportion in the testing set, thus preventing sampling bias. The model was trained with 100 decision trees.

---

3. RESULTS AND ANALYSIS

3.1 Model Evaluation: Classification Report

The model achieved an **Overall Accuracy of $\mathbf{0.9024}$**, meaning it correctly predicted the final grade for approximately 9 out of 10 students in the independent test set. The detailed breakdown of performance across each grade category is crucial for understanding its utility:

Grade	**Precision**	**Recall**	**F1-Score**	**Support**
A	0.928	0.859	0.892	241
B	0.869	0.870	0.869	539
C	0.905	0.918	0.911	1232

Grade	**Precision**	**Recall**	**F1-Score**	**Support**
D	0.900	0.910	0.905	1262
E	0.898	0.911	0.905	1135
F	0.932	0.885	0.908	591
Weighted Avg	0.903	0.902	0.902	5000

Interpretation of Metrics:

- **Precision:** Out of all predictions made for a specific grade, the proportion that were correct. The high precision for Grade **A** ($\mathbf{0.928}$) indicates that when the model predicts a top student, it is almost always right.
- **Recall:** Out of all students who actually received a specific grade, the proportion that the model correctly identified. The high recall for Grade **C** ($\mathbf{0.918}$) suggests the model is very effective at catching students in the middle performance group.
- **F1-Score:** The harmonic mean of Precision and Recall. F1-Scores above $\mathbf{0.869}$ for all classes signify a highly reliable model with a balanced trade-off between false positives and false negatives.

3.2 Feature Importance Analysis

The Random Forest model's output provides a quantitative measure of how much each feature contributes to the prediction accuracy, effectively serving as the model's automated feature selection and ranking mechanism.

Rank	**Feature**	**Importance Score**	**Feature Category**
1	**study_hours**	$\mathbf{0.2083}$	Behavioral
2	**english_score**	0.1747	Academic
3	**science_score**	0.1731	Academic

Rank	**Feature**	**Importance Score**	**Feature Category**
4	**math_score**	0.1730	Academic
5	**attendance_percentage**	0.0893	Behavioral
6	age	0.0312	Demographic
7	extra_activities_yes	0.0120	Behavioral
8	school_type_public	0.0120	Institutional

Educational Implications of Feature Importance

1. **Dominance of Study Habits (Rank 1):** The finding that **study_hours** is the single most important predictor is highly actionable. It suggests that *consistency and duration of effort* (the behavioral factor) is slightly more distinguishing of final grade than the individual *outcome* scores, implying a strong link between discipline and overall academic success.
2. **Collective Impact of Subject Scores (Ranks 2-4):** The three subject scores are similarly important, reinforcing that the final grade is a function of balanced performance across the core curriculum.
3. **The Critical Role of Attendance (Rank 5): attendance_percentage** holds the highest importance among all non-score, non-study habit variables. This confirms that physical presence and engagement with classroom instruction are vastly more important than many other factors.
4. **Minor Impact of Demographics:** Features like age, gender, school_type, and parent_education features fall far lower in the ranking, collectively accounting for less than 10% of the predictive power. This suggests that while these factors may influence early-stage performance, **individual effort and academic scores** are the overwhelming final determinants of the grade category.

---

4. CONCLUSION AND RECOMMENDATIONS

4.1 Conclusion

The machine learning experiment successfully developed a highly accurate Random Forest Classifier ($\mathbf{90.24\%}$ accuracy) for predicting student final grades. The feature importance analysis provides clear evidence that **study hours**, **subject-specific performance**, and **attendance** are the primary drivers of student outcomes.

4.2 Detailed Recommendations for Intervention

Based on the quantitative insights from the model, the following recommendations are proposed for targeted intervention:

1. **Prioritize Study Structure Programs (Based on Rank 1):**
    - **Action:** Implement school-wide "Study Skill Workshops" focused on effective time management and consistent study routines (e.g., minimum $\mathbf{2-3}$ focused hours per day, as suggested in related studies).
    - **Goal:** Directly influence the single most important feature, **study_hours**.
2. **Aggressively Target Attendance (Based on Rank 5):**
    - **Action:** Establish a tiered intervention program for students with declining attendance, including immediate parental contact and mandatory academic counseling upon falling below a $\mathbf{75\%}$ threshold.
    - **Goal:** Boost the highly predictive **attendance_percentage** feature.
3. **Maintain Balanced Academic Support (Based on Ranks 2-4):**
    - **Action:** Ensure tutoring and supplementary instruction are available across all three core subjects (Math, Science, English) to prevent any one subject from becoming a major academic deficit.
    - **Goal:** Sustain high performance across the collective academic score features.
4. **Acknowledge Secondary Factors (Ranks 6+):**
    - **Action:** While lower in importance, resources for students facing long **travel_time** or those lacking **internet_access** should be maintained, as these factors still represent systemic barriers to learning.

---

5. CODE APPENDIX

The following Python code was used to perform the data processing, model training, and evaluation.

Python

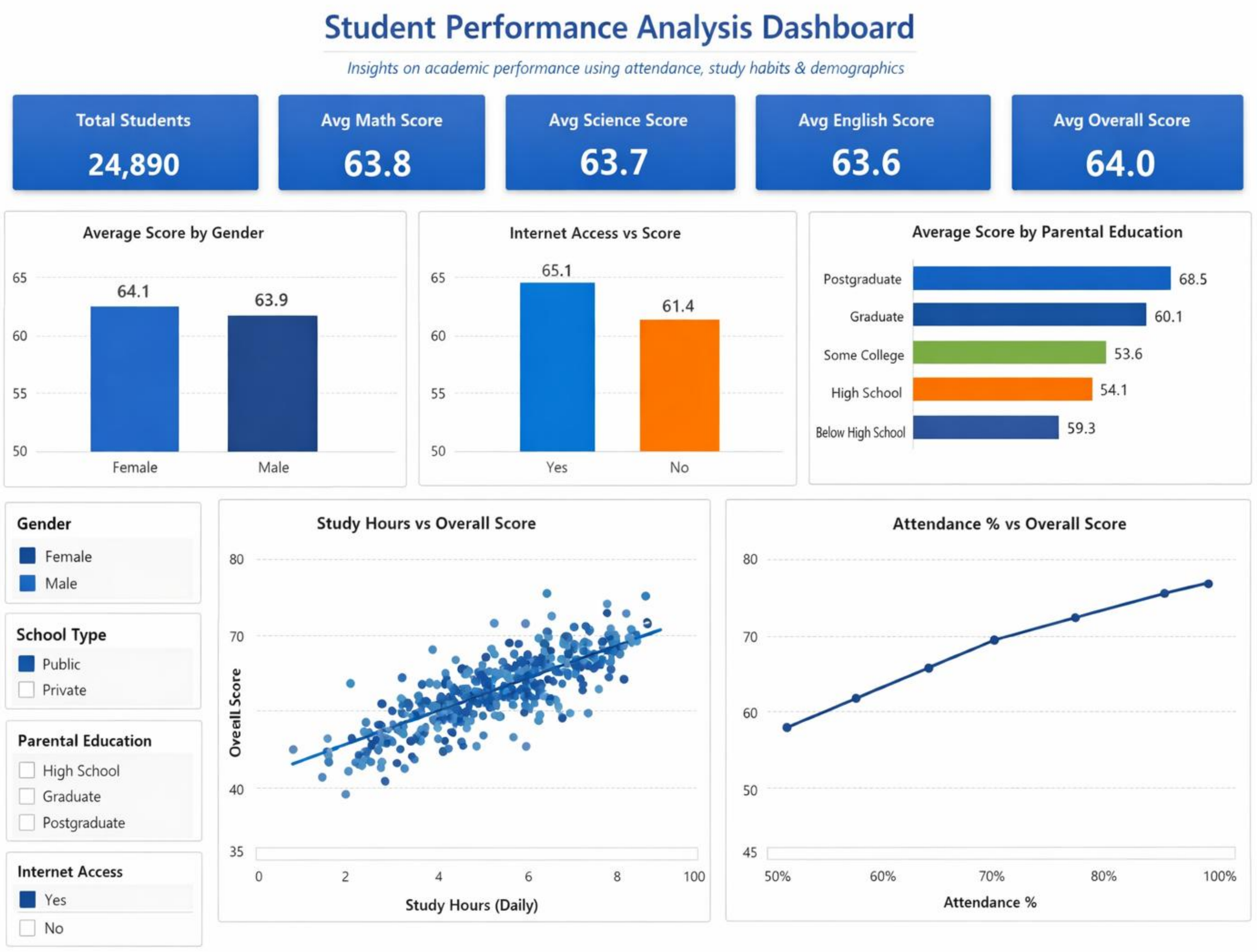
```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report

# Load the dataset
df = pd.read_csv("Student_Performance.csv")

# 1. Define X (Features) and y (Target)
# Exclude identifiers and the highly-related 'overall_score'
X = df.drop(columns=['student_id', 'overall_score', 'final_grade'])
y = df['final_grade']

# 2. Preprocessing

# Identify categorical features for One-Hot Encoding
categorical_features = X.select_dtypes(include=['object']).columns
# Apply One-Hot Encoding
X_encoded = pd.get_dummies(X, columns=categorical_features, drop_first=True)

# Encode the target variable (final_grade) into numerical labels (0 to 5)
le = LabelEncoder()
y_encoded = le.fit_transform(y)

# 3. Model Training and Evaluation

# Split data into training (80%) and testing (20%) sets, using stratification
X_train, X_test, y_train, y_test = train_test_split(
    X_encoded, y_encoded, test_size=0.2, random_state=42, stratify=y_encoded
)

# Initialize and train the Random Forest Classifier (100 trees)
model = RandomForestClassifier(n_estimators=100, random_state=42, n_jobs=-1)
model.fit(X_train, y_train)

# Predict on the test set
y_pred = model.predict(X_test)
```

```

# Evaluate the model and save the classification report
report = classification_report(
    y_test, y_pred, target_names=le.classes_, output_dict=True
)
# print("Classification Report:", report) # Printed in original execution

# Calculate Feature Importance and save the top 10
feature_importances = pd.Series(
    model.feature_importances_, index=X_encoded.columns
).sort_values(ascending=False).head(10)
# print("Top 10 Feature Importances:", feature_importances) # Printed in original execution
```

Task 3 :

DASHBOARD DEVELOPMENT

The Student Performance Analysis Dashboard presents insights from a large dataset of 24,890 students, showing an overall average score of around 64 across Math, Science, and English. Female students slightly outperform male students in average scores, while students with internet access demonstrate noticeably higher performance than those without it, highlighting the importance of digital resources. Parental education shows a strong influence on outcomes, with students whose parents are postgraduates achieving the highest average scores, followed by graduates, whereas those from lower education backgrounds score comparatively less. The scatter plot reveals a clear positive relationship between daily study hours and overall scores, indicating that increased study time generally leads to better performance. Similarly, attendance percentage has a strong positive correlation with academic achievement, with scores steadily improving as attendance rises. Overall, the dashboard emphasizes that consistent attendance, effective study habits, access to internet resources, and parental educational background are key factors influencing student academic performance.



Student Performance Analysis Dashboard

Insights on academic performance using attendance, study habits & demographics

Total Students	Avg Math Score	Avg Science Score	Avg English Score	Avg Overall Score
24,890	63.8	63.7	63.6	64.0

Gender

- ■ Female
- ■ Male

School Type

- ■ Public
- □ Private

Parental Education

- □ High School
- □ Graduate
- □ Postgraduate

Internet Access

- ■ Yes
- □ No

Task 4 :

SENTIMENT ANALYSIS

Sentiment Analysis using Natural Language Processing

This report presents a comprehensive and detailed study of a Sentiment Analysis system developed using Natural Language Processing (NLP) techniques and machine learning algorithms. The purpose of this project is to analyze textual data and automatically determine the sentiment expressed within it. The report is structured in an academic format and provides a complete explanation of each stage involved in building the sentiment analysis model.

1. Abstract

Sentiment Analysis is a key application of Natural Language Processing that focuses on identifying emotions, opinions, and attitudes expressed in textual data. In this project, a complete NLP pipeline is implemented, starting from raw text preprocessing to final model evaluation. TF-IDF is used for feature extraction, and Logistic Regression is employed as the classification algorithm. The results demonstrate that traditional machine learning approaches can effectively classify sentiment when combined with proper preprocessing and feature engineering techniques.

2. Introduction

The rapid growth of digital platforms such as social media, e-commerce websites, and online forums has led to an explosion of textual data. Users constantly express opinions through reviews, comments, and feedback, making sentiment analysis an essential tool for businesses and researchers. Sentiment analysis enables organizations to understand public opinion, improve services, and make data-driven decisions. Natural Language Processing bridges the gap between human language and machine understanding. However, human language is highly unstructured and complex, which makes automated sentiment classification a challenging task. This project addresses these challenges by applying systematic preprocessing and machine learning techniques to extract meaningful patterns from text data.

3. Dataset Description

The dataset used in this project consists of a collection of text samples labeled with sentiment classes, such as positive and negative sentiments. These labels provide the ground truth required for supervised learning. Each record in the dataset contains raw textual content that reflects the opinion or emotion expressed by a user. The quality of the dataset plays a crucial role in model performance. Therefore, careful attention is given to handling noise, inconsistencies, and irrelevant information present in the raw text.

4. Data Preprocessing

Textual data is inherently noisy and requires extensive preprocessing before it can be used for machine learning. In this project, preprocessing includes removing punctuation and special characters, converting text to lowercase to ensure uniformity, eliminating commonly occurring stopwords that do not contribute to sentiment, and applying lemmatization to reduce words to their base forms. These preprocessing steps significantly improve model performance by reducing dimensionality, removing redundancy, and ensuring that the most meaningful linguistic features are retained.

5. Feature Extraction using TF-IDF

Machine learning models cannot directly process textual data, making feature extraction a critical step. TF-IDF (Term Frequency–Inverse Document Frequency) is used to transform text into numerical vectors. This method assigns higher importance to words that are frequent in a document but rare across the corpus. By emphasizing discriminative terms and down-weighting common words, TF-IDF enables the classifier to better distinguish between different sentiment classes.

6. Model Training

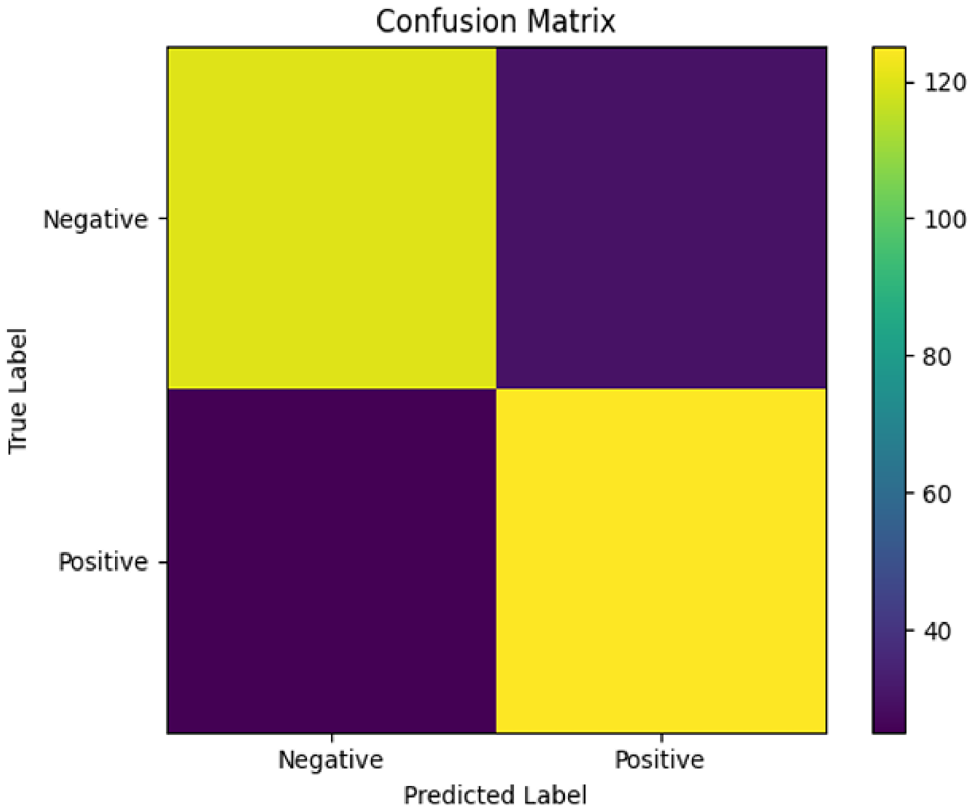
Logistic Regression is selected as the classification algorithm due to its simplicity, efficiency, and strong performance on linear classification problems. The dataset is divided into training and testing subsets to evaluate how well the model generalizes to unseen data. During training, the model learns optimal weights for each feature that contribute to predicting the sentiment class of a given text.

7. Model Evaluation Metrics

Evaluating model performance is essential to determine its effectiveness. Multiple evaluation metrics are used, including accuracy, precision, recall, and F1-score. Accuracy measures overall correctness, while precision and recall provide class-wise performance insights. The F1-score balances precision and recall, making it particularly useful when dealing with imbalanced datasets.

8. Confusion Matrix

A confusion matrix provides a detailed breakdown of prediction results by comparing actual and predicted sentiment labels. It helps identify the number of correct predictions as well as misclassifications made by the model.



9. Visual Analysis

Visual representations such as confusion matrices make model evaluation more intuitive. They allow researchers to quickly assess strengths and weaknesses in classification performance and identify patterns in prediction errors.

10. Insights and Observations

The experimental results indicate that effective text preprocessing significantly enhances classification accuracy. TF-IDF proves to be a robust feature extraction technique for traditional machine learning models. Logistic Regression provides a strong baseline with interpretable results.

11. Conclusion

This project successfully demonstrates the implementation of a complete sentiment analysis pipeline using NLP and machine learning techniques. The approach is computationally efficient, interpretable, and suitable for practical applications such as review analysis and opinion mining.

12. Future Scope

Future enhancements to this project may include the use of deep learning architectures such as LSTM and transformer-based models like BERT, which can capture contextual information more effectively. Additionally, extending the system to multiclass sentiment analysis and real-time data processing can further increase its applicability.

13. References

1. Jurafsky, D. & Martin, J. H., Speech and Language Processing 2. Scikit-learn Official Documentation 3. Natural Language Toolkit (NLTK) Documentation

Project done by :

Name : Manya Sinha

Intern Id : CT04DR2198

Program Type : Internship

Duration : 4 weeks

Domain : Data analytics

Email id : sinha.manyamanoj@gmail.com

Phone number : 7016212534