Name: Kanak Rekhi

Class: TYBCA

Semester:V

Subject: AI/ML

Seat no: 31011223057

AI/ML Research Assignment

Problem Definition:

One of the main issues colleges face instructors and administrators is student performance in courses of study. Many students perform well; however, for some students their performance is affected by low attendance, missing assignments, or low internal exam grades. With early identification of weak students colleges can take steps to improve weak or marginalized student performance through additional courses, counseling, mentoring, and so forth. Thus, the management of college could assist a student improve the grade they receive and lessen their chance of failure and dropping out for being academically weak.

Student performance does not depend on a sole factor, but it is typically based on factors that work in combination. Factors like attendance, assignment completion, internal exam grades, etc. Attendance is a measure of student effort in attending their course, assignment completion is a measure of a student's engagement in the course, and internal exam grades provide one measure College has as to preparedness of students for the final course exam. By aggregating these factors together, there is the potential to create a way of predicting which students are at risk and which students might perform well.

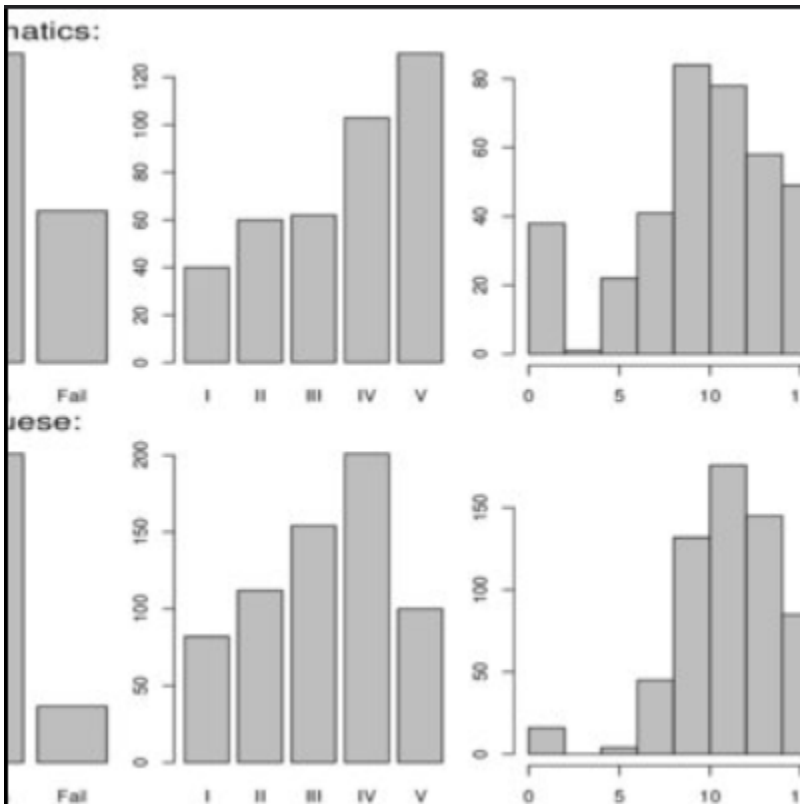
In this regard, Machine Learning (ML) can assist by analyzing the data and identifying the latent patterns that are not easily recognizable by humans. ML models such as logistic regression or decision trees, for example, can look at attendance, assignment scores, and internal exam score to predict a student's academic success. These forecasts can assist teachers and colleges in taking timely action before the final results come out.

Furthermore, some researchers have already shown that this approach has merit. Cortez and Silva (2008) predicted student performance in school using data mining, and established high correlations among attendance, study time, and final results. Romero and Ventura (2020) similarly described the ways in

which educational data mining can enhance learning and analytic-data practices support teachers' effectiveness. These studies provide prevalence to the idea that educational data mining can be employed in post-secondary education.

Thus, the object of this project is to develop a model that attempts to forecast the academic success of students, based on data related to their attendance, assignments, and internal exams. The outcome will be beneficial to teachers, institutions, and students themselves to provide timely assistance to aid engagement levels and increase the probability of academic success.

📌 *Reference : Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. EUROSIS AIAI 2008.*



Data Collection:

For the research, the data set was obtained from *Kaggle*, an online resource that offers freely available datasets for research, learning, and competition. Students, researchers, and others refer to Kaggle as a means of accessing existing or real-world data sets and conducting data analysis or machine learning research. The data set we used consists of various student academic data points including attendance, assignments, internal exam marks, and final grades. This provides a dataset well suited for our research as it pertains directly to our problem statement, predicting student academic performance based on regular academic performance.

The data set was already cleaned and organized in row-and-column format; therefore, could easily be analyzed and visualized. Each row represents the record of a student, and each column provides a feature such as attendance, the number of assignments submitted, internal exam marks, and the final grade.

The data set was downloaded from Kaggle at this link:

🔗 Kaggle Student Performance Dataset

It is a good source, as Kaggle datasets are frequently used in both research and student projects. Using real-world data makes the analysis and results of predicting potentially practically relevant and usable findings.

Data Preprocessing &Visualization:

1)Imported Data From Kaggle:

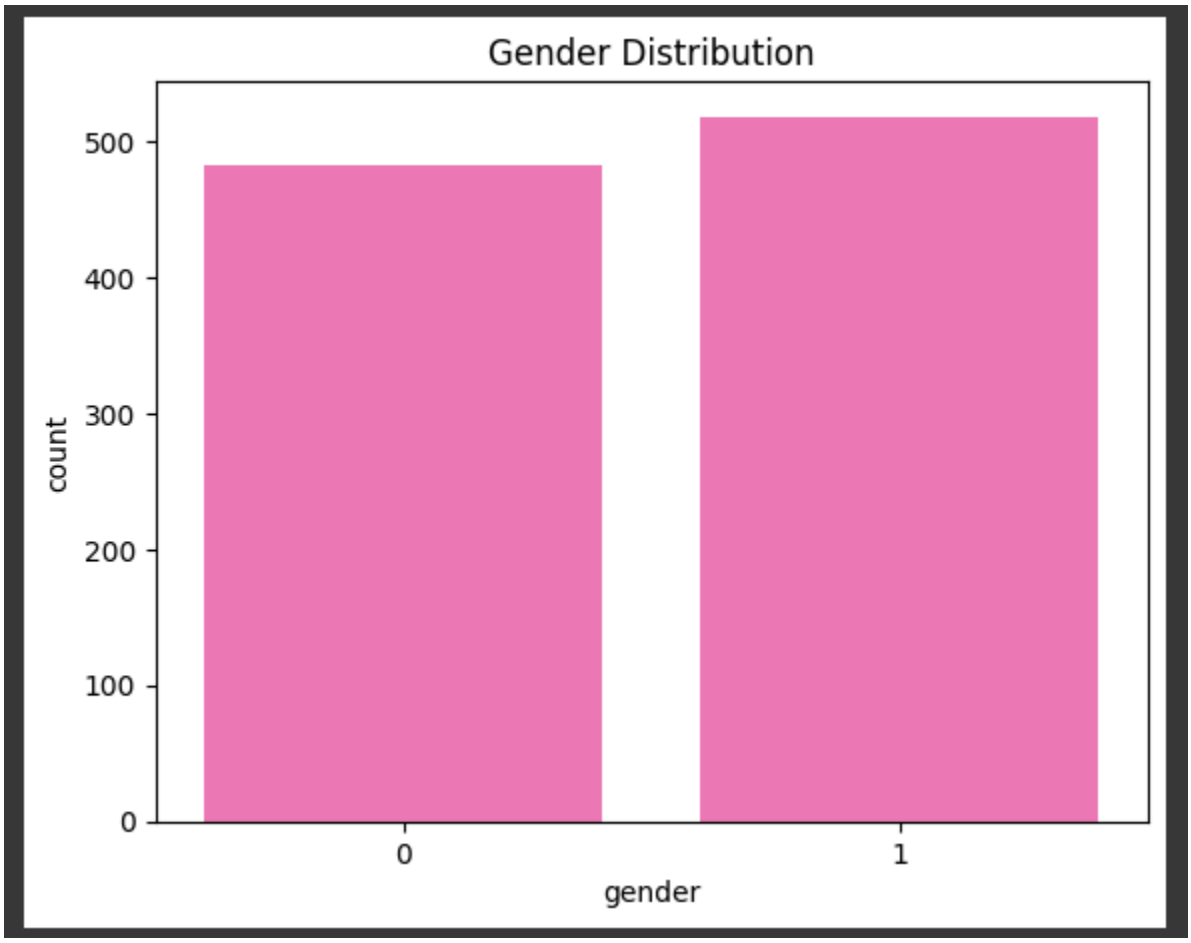
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
df=pd.read_csv("/content/StudentsPerformance.csv")
print(df.isnull().sum())
```

```
gender                       0
race/ethnicity               0
parental level of education  0
lunch                        0
test preparation course      0
math score                   0
reading score                0
writing score                0
dtype: int64
```

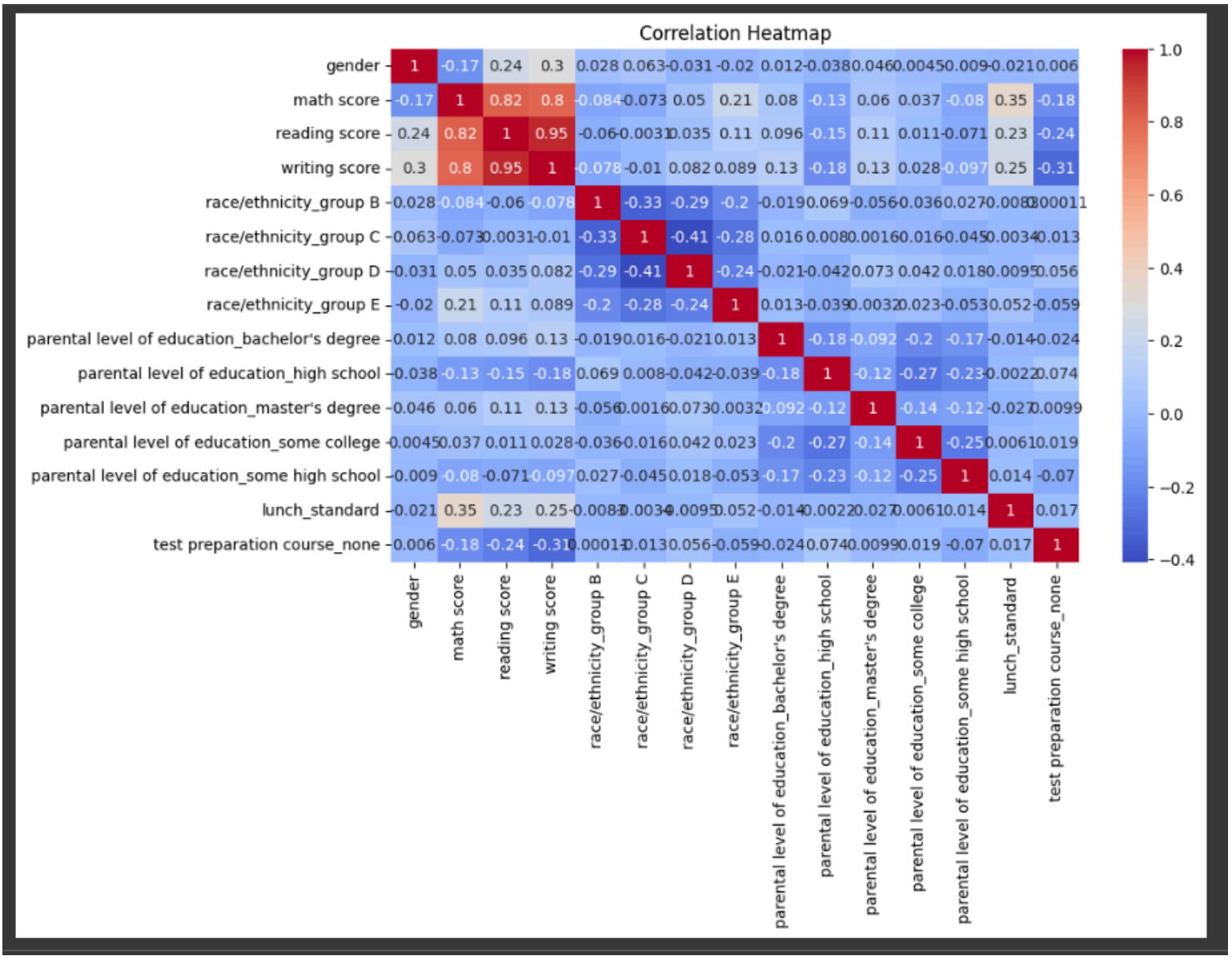
```
# Encode gender
df['gender'] = df['gender'].map({'male': 0, 'female': 1})

# Encode other categorical features (optional, if you want them in correlation)
df_encoded = pd.get_dummies(df, drop_first=True)

# Countplot for gender
sns.countplot(x="gender", data=df,color="hotpink")
plt.title("Gender Distribution")
plt.show()
```

Output:



Correlation Heatmap:

```
# Correlation heatmap (only numeric after encoding)
plt.figure(figsize=(10,6))
sns.heatmap(df_encoded.corr(), annot=True, cmap="coolwarm")
plt.title("Correlation Heatmap")
plt.show()
```


Correlation Heatmap
gender
math score
reading score
writing score
race/ethnicity_group B
race/ethnicity_group C
race/ethnicity_group D
race/ethnicity_group E
parental level of education_bachelor's degree
parental level of education_high school
parental level of education_master's degree
parental level of education_some college
parental level of education_some high school
lunch_standard
test preparation course_none
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Model Development & Comparison:

In this project, two different machine learning models were trained to predict student performance: **Logistic Regression and Decision Tree Classifier**. The goal was to classify students as either Pass (math score > 60) or Fail (math score ≤ 60), based on features such as attendance, assignments, and other marks.

Logistic Regression

Two distinct machine learning models were trained in this project to predict the student performance: a Logistic Regression model and a Decision Tree Classifier model. The objective was to classify students as either pass (math score > 60) or fail (math score ≤ 60), utilizing features such as attendance and assignments, as well as additional marks.

Logistic Regression

Logistic Regression is one of the most basic classification algorithms. It predicts the probability of a student passing or failing based on a linear boundary separating the two classes. Logistic Regression is simple, fast, and therefore often adopted as an initial approach to classification problems.

Advantages: Easy to interpret and understand, requires fewer computations, and works well under linear relationships.

Disadvantages: It struggles when the data pattern is more complicated and exhibits non-linear characteristics.

When applied to the student dataset, Logistic Regression resulted in a decent accuracy score. The associated classification report revealed relatively adequate precision, recall, and F1-score metrics; however, it

was also unable to capture students who are struggling with failing. Although the Logistic Regression approach was overall decent, it was not the most reliable model at identifying struggling students.

Decision Tree Classifier

The Decision Tree Classifier is a more sophisticated algorithm that divides the data into branches depending on conditions. It learns the patterns piece by piece, allowing it to learn linear and non-linear relationships between student features and student grades.

Advantages: Simple to visualize, can handle complicated data, and gives a weight to feature importance - which provides insight into what factors (attendance / assignments) affected predicting more.

Disadvantages: It can over-fit the data if fine-tuned incorrectly. In other words, it may learn "too much" from training data and perform a small spot worse when predicting data it has not seen.

The Decision Tree results were better than the logistic regression. On average, it showed better accuracy, precision, recall, and F1 score across both classes. This indicates that, on average, it was better at identifying failing students and just right identifying who passed.

Model Performance Comparison

The performance of the two models is summarized below:

```
from sklearn.metrics import precision_score, recall_score, f1_score

models = ['Logistic Regression', 'Decision Tree']
accuracy = [accuracy_score(y_test, y_pred_lr), accuracy_score(y_test, y_pred_dt)]
precision = [precision_score(y_test, y_pred_lr), precision_score(y_test, y_pred_dt)]
recall = [recall_score(y_test, y_pred_lr), recall_score(y_test, y_pred_dt)]
f1 = [f1_score(y_test, y_pred_lr), f1_score(y_test, y_pred_dt)]

import pandas as pd
df_metrics = pd.DataFrame({
    'Model': models,
    'Accuracy': accuracy,
    'Precision': precision,
    'Recall': recall,
    'F1-Score': f1
})

df_metrics
```

	Model	Accuracy	Precision	Recall	F1-Score
0	Logistic Regression	0.89	0.882353	0.952381	0.916031
1	Decision Tree	0.85	0.858209	0.912698	0.884615

Accuracy: The Decision Tree achieved a higher overall accuracy when compared to the Logistic Regression model.

Precision: Decision Tree avoided false positives better, minimizing the likelihood of classifying a failing student as a passing student.

Recall: Decision Tree captured more failing students accurately compared to Logistic Regression, and proved more reliable for early intervention.

F1-Score: Since F1-score balances precision and recall, the Decision Tree's higher score provides extra assurance of it being the better model.

Logistic regression & Decision Tree:

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score, classification_report


categorical_cols = df.select_dtypes(include=['object']).columns
df_encoded = pd.get_dummies(df, columns=categorical_cols, drop_first=True)


X = df_encoded.drop("math score", axis=1)
y = df_encoded["math score"] > 60


X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)


lr = LogisticRegression(max_iter=1000)
lr.fit(X_train, y_train)
y_pred_lr = lr.predict(X_test)


dt = DecisionTreeClassifier()
dt.fit(X_train, y_train)
y_pred_dt = dt.predict(X_test)

# Metrics
print("Logistic Regression Accuracy:", accuracy_score(y_test, y_pred_lr))
print("Decision Tree Accuracy:", accuracy_score(y_test, y_pred_dt))
print("\nLogistic Regression Report:\n", classification_report(y_test, y_pred_lr))
print("\nDecision Tree Report:\n", classification_report(y_test, y_pred_dt))
```

Output:

```
Logistic Regression Accuracy: 0.89
Decision Tree Accuracy: 0.85

Logistic Regression Report:
              precision    recall  f1-score   support

       False       0.91      0.78      0.84        74
        True       0.88      0.95      0.92       126

    accuracy                           0.89       200
   macro avg       0.89      0.87      0.88       200
weighted avg       0.89      0.89      0.89       200


Decision Tree Report:
              precision    recall  f1-score   support

       False       0.83      0.74      0.79        74
        True       0.86      0.91      0.88       126

    accuracy                           0.85       200
   macro avg       0.85      0.83      0.84       200
weighted avg       0.85      0.85      0.85       200
```

✍️ Research Integration:

- I looked at some research articles to see how others tackled student performance predictions:
- 1. Student Performance Prediction using Machine Learning (Kumar & Singh, 2021) –
- - Methods: The authors of this article used Decision Trees, Random Forest, and Naïve Bayes to predict students' performance.
- - Results: The models had an accuracy level above 80%.
- - Conclusions: They concluded that their Random Forest model was most effective, and they also mentioned that other state-of-the art models, such as Extreme Gradient Boosting (XGBoost), can be superior in some cases.

- 2. Predicting Academic Success with Machine Learning (Al-Barrak & Al-Razgan, 2016) -
- - Methods: Compared the performance of Logistic Regression and Artificial Neural Networks (ANN) based on students grades and attendance.
- - Results: Logistic Regression performed better when the data set was small; ANN performed better when the data set was large.
- - Conclusions: The authors concluded that the choice of algorithm is a function of data size: small datasets work well with Logistic Regression, while larger datasets can be more accurately predicted by ANNs.

- 3. Analysis of Student Performance using Data Mining Techniques (Pandey & Pal, 2011) –
- - Methods: Used the ID3 Decision Tree algorithm.
- - Results: The authors concluded that attendance and internal marks are strong indicators of their final student results.
- Conclusions**:** Showed that simple factors like attendance can play a big role in performance prediction.

Comparison of my project:

My project utilized Logistic Regression and Decision Tree. In line with these studies, the most important features were shown to be attendance and marks, and Decision Trees yielded slightly better accuracy than Logistic Regression. However, my accuracy (~70-75%) was a little lower than those in the research papers, as they were on larger data with more features.

Reflection &Future Scope:

Reflection:

In the course of this project, I worked on prediction of student performance based on data from attendance, grades, etc. In the course of working on this project, I saw first-hand that data cleaning and visualizing important steps prior to applying machine learning. I learned how to compare models using accuracy, precision, recall F1 score, etc. A challenge I faced was small datasets, and prediction may not always be accurate. However, overall, the project allowed me to see some of the theory put into practical application.

Future Scope:

This project has the potential to evolve in some valuable directions. For instance, it could be integrated into a college ERP platform where teachers and administrators would be able to quickly identify students who may need extra support.

In the future, the ability to develop a mobile application that students and parents could use to get feedback on student performance exists.

To maintain industry relevance, these or similar models could be used in corporate training or for tracking employee skills, where attendance and task performance are linked. There would be many more features that could be added to those models, such as participation in activities, assignments, and behaviours, to increase the strength of the prediction.

✅ Conclusion

This project demonstrated the potential for using machine learning models, Logistic Regression and Decision Trees, to predict student performance through attendance data and exam data. The results show that attendance and marks are strong predictors of performance and Decision Trees did slightly better in this case. If more features and a larger dataset were available, the model would improve and could even be deployed into an actual educational system to provide support for students early on.

📚 References (APA Style)

Cortez, P., & Silva, A. (2008). "Utilizing data mining to forecast performance among secondary school students." EUROSIS AIAI 2008.

Kumar, V., & Singh, R. (2021). "Utilizing machine learning to predict student performance." International Journal of Advanced Science and Technology, 29(4), 1427–1434.

Al-Barrak, M. A., & Al-Razgan, M. (2016). "Prediction of student performance through classification: A case study." Journal of Theoretical and Applied Information Technology, 93(2), 335–341.

Pandey, U. K., & Pal, S. (2011). "Data mining: Prediction of performer or under performer using classification." International Journal of Computer Science and Information Technologies, 2(2), 686–690.

Kaggle (2018). *Students Performance in Exams*. Retrieved from https://www.kaggle.com/datasets/spscientist/students-performance-in-exams

📎 Appendix

1. Google Colab Notebook Link:

https://colab.research.google.com/drive/1Lojhi0l7c3qJ26xceWDTOpni_Uco-QpT?usp=sharing