Graduate	Model	Accuracy	F1 Score
Original	ANN	0.86	0.86
Original	Random Forest	0.86	0.88
Original	SVM	0.85	0.85
Original	XGBoost	0.99	0.99

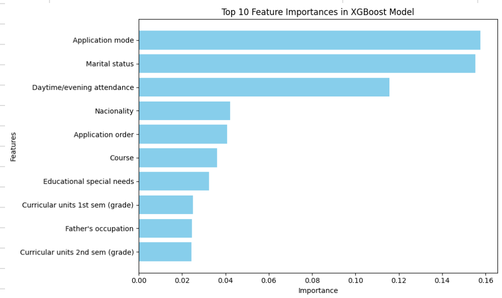
Stage-Drop Out	Model	Accuracy	+/-	F1 Score	+/-					
Original	ANN	0.87		0.78			X_train	X_test	y_pred	
Original	Random Forest	0.86		0.77			X_train_ann_filtered	X_test_ann_filtered	y_pred_ann_filtered	
Original	SVM	0.85		0.77			X_train_ann_f2	X_test_ann_f2	y_pred_ann_f2	Final Best Parameters for ANN: {'neurons_layer1': 16, 'neurons_layer2': 16}
Original	XGBoost	0.85		0.77						
Drop low important(threshold = 0.02, 34->31)	ANN	0.87		0.79						
Drop low important(threshold = 0.01, 34->24)	Random Forest	0.87		0.79			X_train_svm	X_test_svm	y_pred_svm	
Drop low important(threshold = 0.01, 34->31)	SVM	0.88		0.80			X_train_svm	X_test_svm	y_pred_svm	Final Best Parameters: {'C': 10, 'kernel': 'linear'}
Drop low important (threshold = 0.01, 34->30)	XGBoost	0.86		0.79						
Fine-Tuned	ANN	0.87		0.85		{'neurons_layer1': 32, 'neurons_layer2': 8}				
Fine-Tuned	Random Forest	0.87		0.79		Final Best Parameters for Random Forest: {'	X_train_rf	X_test_rf	y_pred_rf	
Fine-Tuned	SVM	0.88		0.80		Final Best Parameters: {'C': 10, 'kernel': 'linear'}	X_train_rf	X_test_rf	y_pred_rf	Final Best Parameters for Random Forest: {'n_estimators': 300, 'max_depth': 3, 'min_samples_split': 10, 'min_samples_leaf': 1}
Fine-Tuned	XGBoost	0.87		0.79		Final Best Parameters: {'learning_rate': 0.1, 'subsample': 0.8, 'colsample_bytree': 0.8, 'tree_method': 'hist', 'early_stopping_rounds': 10, 'eval_metric': 'logloss', 'random_state': 42, 'n_estimators': 100, 'max_depth': 3}				
Feature Engineering	ANN	0.87		0.87						
Feature Engineering	Random Forest	0.87		0.80			X_train_selected	X_test_selected	y_pred_selected	
Feature Engineering	SVM	0.87		0.81			X_train_selected	X_test_selected	y_pred	Final Best Parameters: {'learning_rate': 0.1, 'subsample': 0.8, 'colsample_bytree': 0.8, 'tree_method': 'hist', 'early_stopping_rounds': 10, 'eval_metric': 'logloss', 'random_state': 42, 'n_estimators': 100, 'max_depth': 3}
Feature Engineering	XGBoost	0.87		0.82						
							X_train_final	X_test_final	y_pred	
做到超參數 要查一下為什麼Ann的準確度降低一點							Use the evaluate_and_plot function			
在做特徵工程							evaluate_and_plot(y_test, y_pred_rf, "Random Forest After Feature Selection")			

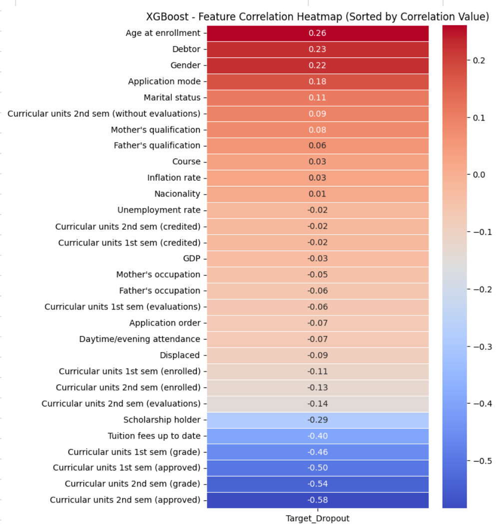
Jasonnwaneti@outlook.com
jutirado2004@gmail.com
isa002014@gmail.com
andrescuervo1024@gmail.com
zoeyhuang1@gmail.com

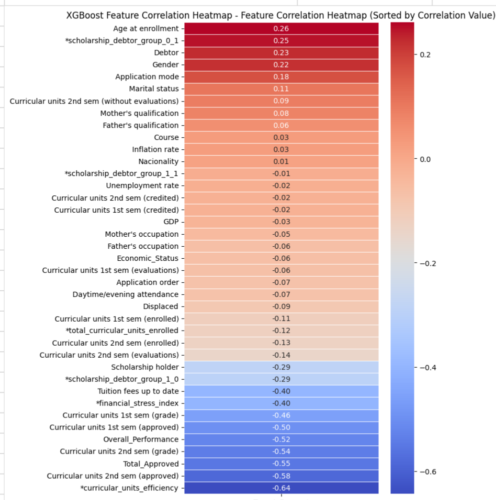
—		
one by one?		
depends		

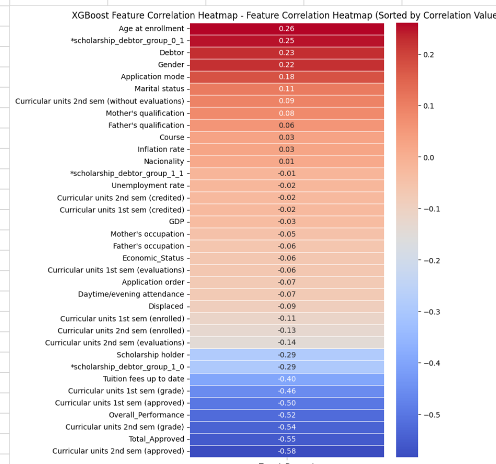
```python

df_factors_encoded['Scholarship Yes Debt Yes' ] = df_factors_encoded['Scholarship holder']& df_factors_encoded['Debtor']

df_factors_encoded['Scholarship No Debt No' ] = ~ df_factors_encoded['Scholarship holder'] & ~ df_factors_encoded['Debtor']

df_factors_encoded['Scholarship No Debt Yes' ] = ~df_factors_encoded['Scholarship holder']& df_factors_encoded['Debtor']

df_factors_encoded['Scholarship Yes Debt No' ] = df_factors_encoded['Scholarship holder'] & ~ df_factors_encoded['Debtor']

grouped_data = df_factors_encoded.groupby('Scholarship Yes Debt Yes')['Target_Graduate'].mean() * 100
plt.figure(figsize=(6, 5))
bars = plt.bar(grouped_data.index, grouped_data.values, color="skyblue")

# Add text labels above bars
for bar, value in zip(bars, grouped_data.values):
    plt.text(
        bar.get_x() + bar.get_width() / 2,
        bar.get_height() +1,
        f"{value:.1f}%",
        ha="center",
        fontsize=12,
    )
```

Base on XGBoost	XGBoost Feature Remove < 0.01	Importance
	Curricular units 2nd sem (approved)	0.255959
	Tuition fees up to date	0.138809
	Curricular units 1st sem (enrolled)	0.057896
	Curricular units 2nd sem (enrolled)	0.043951
	Daytime/evening attendance	0.036293
	Curricular units 1st sem (approved)	0.03179
	Curricular units 2nd sem (credited)	0.031152
	Debtor	0.027588
	Scholarship holder	0.021841
	Course	0.020204
	Curricular units 1st sem (credited)	0.020177
	Application order	0.01929
	Curricular units 2nd sem (without evaluations)	0.019252
	Age at enrollment	0.019017
	Nacionality	0.018756
	Curricular units 2nd sem (grade)	0.017832
	Curricular units 2nd sem (evaluations)	0.017749
	Gender	0.01725
	Mother's occupation	0.017005
	Mother's qualification	0.015271
	Father's qualification	0.014792
	Application mode	0.014656
	Displaced	0.014453
	Unemployment rate	0.014228
	Father's occupation	0.013634
	GDP	0.013287
	Curricular units 1st sem (grade)	0.013131
	Curricular units 1st sem (evaluations)	0.012931
	Inflation rate	0.012662
	Marital status	0.010744
	Previous qualification	0.00885
	Curricular units 1st sem (without evaluations)	0.006383
	Educational special needs	0.003165
	International	0
	scholarship_debtor_group	

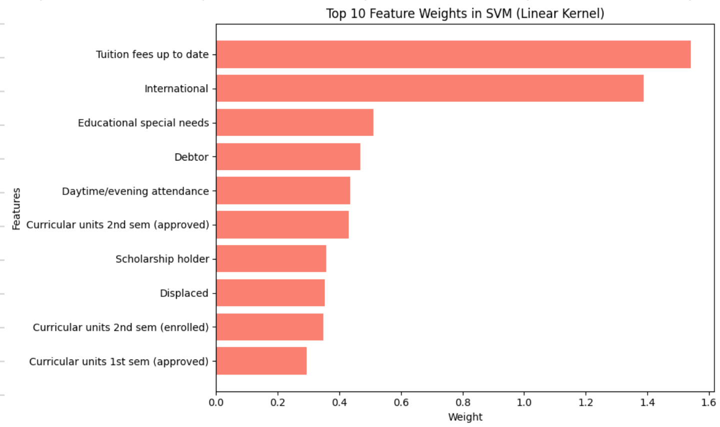






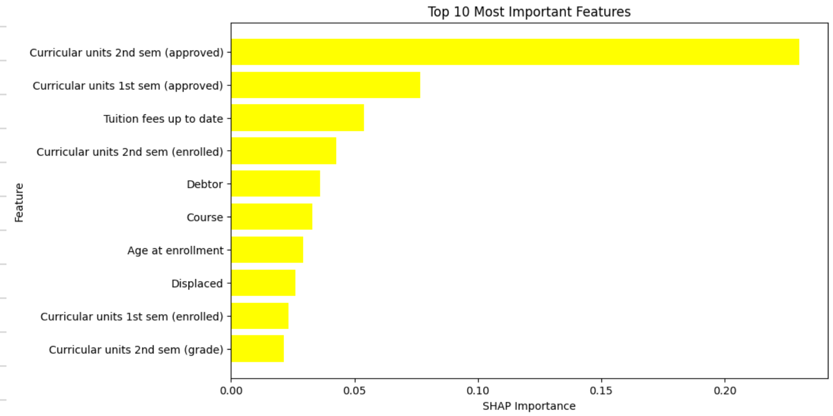


SVM Feature Remove < 0.010	Weight
Tuition fees up to date	1.541711
International	1.386583
Educational special needs	0.510332
Debtor	0.466021
Daytime/evening attendance	0.435421
Curricular units 2nd sem (approved)	0.430958
Scholarship holder	0.356493
Displaced	0.352999
Curricular units 2nd sem (enrolled)	0.34928
Curricular units 1st sem (approved)	0.294194
Curricular units 1st sem (without evaluations)	0.180606
Curricular units 1st sem (credited)	0.126723
Curricular units 2nd sem (credited)	0.126482
Nacionality	0.121448
Gender	0.111448
Curricular units 1st sem (enrolled)	0.085908
Course	0.042924
Unemployment rate	0.039552
Curricular units 1st sem (grade)	0.038162
Curricular units 2nd sem (evaluations)	0.036674
Inflation rate	0.029582
Age at enrollment	0.029292
Father's occupation	0.028219
Marital status	0.025888
Application order	0.025857
Mother's occupation	0.019267
Mother's qualification	0.014491
Curricular units 2nd sem (grade)	0.010583
Curricular units 1st sem (evaluations)	0.008876
Application mode	0.008452
Curricular units 2nd sem (without evaluations)	0.00792
Previous qualification	0.005463
Father's qualification	0.003284
GDP	0.002459

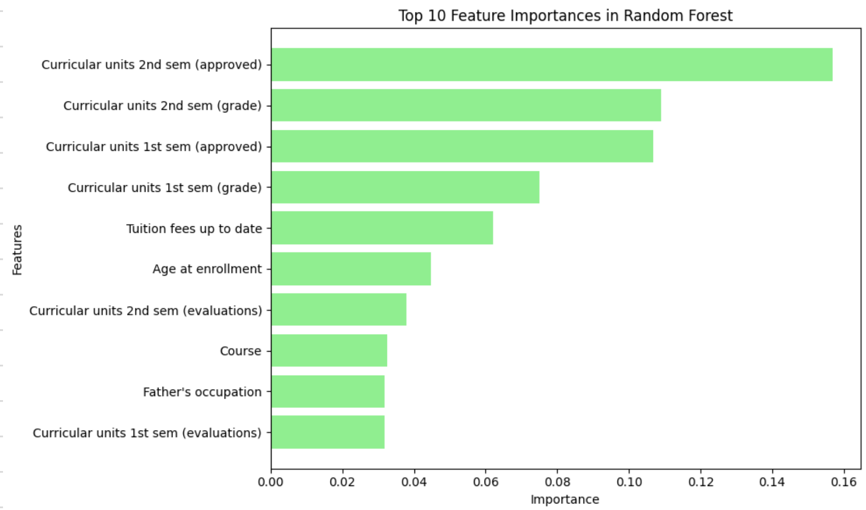


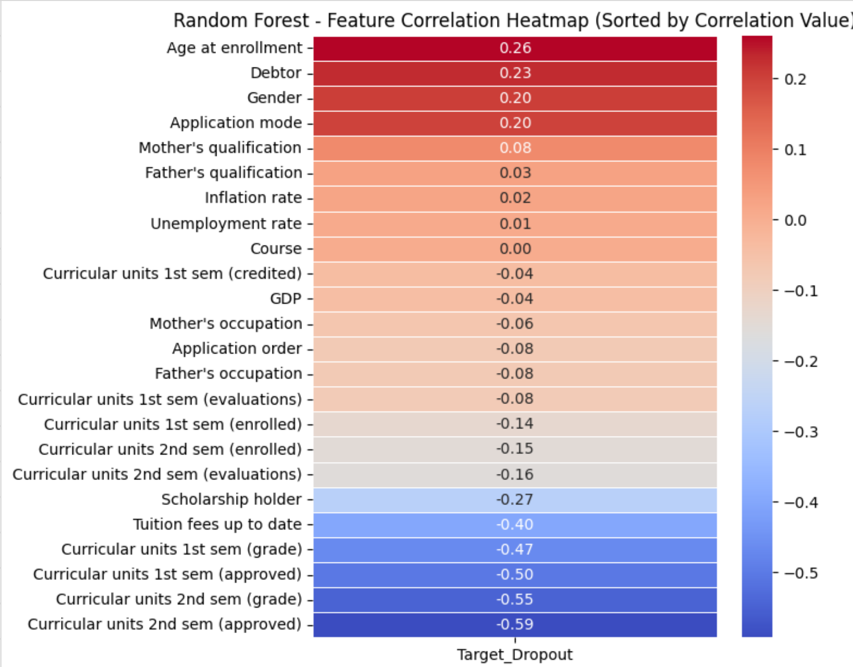
Feature,Importance	Importance
Curricular units 2nd sem (approved)	0.230968
Curricular units 1st sem (approved)	0.076557
Tuition fees up to date	0.054343
Curricular units 2nd sem (enrolled)	0.044345
Debtor	0.03641
Course	0.032716
Age at enrollment	0.029964
Displaced	0.025933
Curricular units 1st sem (enrolled)	0.023776
Curricular units 2nd sem (grade)	0.021577
Mother's qualification	0.020278
Scholarship holder	0.019598
Curricular units 2nd sem (credited)	0.016207
Mother's occupation	0.014565
Nacionality	0.014021
Unemployment rate	0.013562
Gender	0.013441
Daytime/evening attendance	0.013382
Curricular units 1st sem (evaluations)	0.013162
Curricular units 2nd sem (evaluations)	0.010569
Curricular units 1st sem (credited)	0.010349
Father's qualification	0.009964
International	0.009891
GDP	0.00574
Application mode	0.004474
Educational special needs	0.003716
Inflation rate	0.002799
Application order	0.002613
Father's occupation	0.00238
Curricular units 1st sem (grade)	0.00207
Previous qualification	0.001595
Curricular units 2nd sem (without evaluations)	0.000828
Marital status	0.000609
Curricular units 1st sem (without evaluations)	0.000377

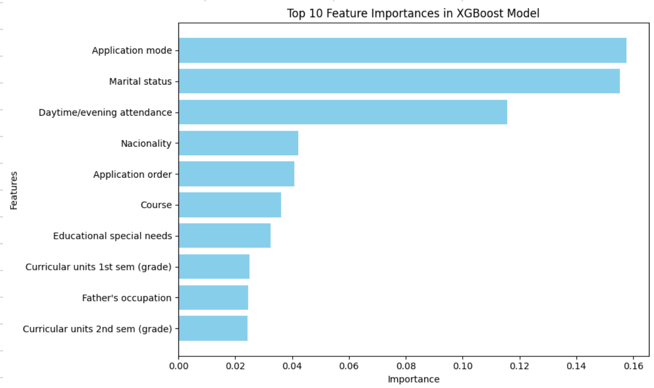
meaningless

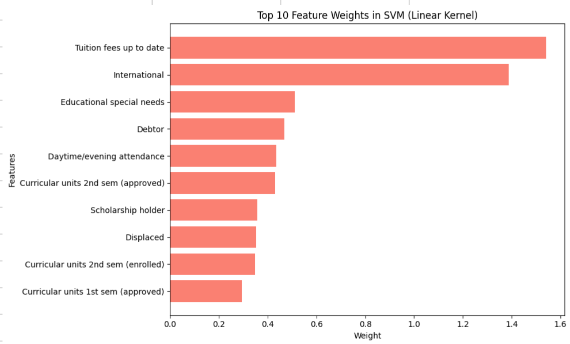


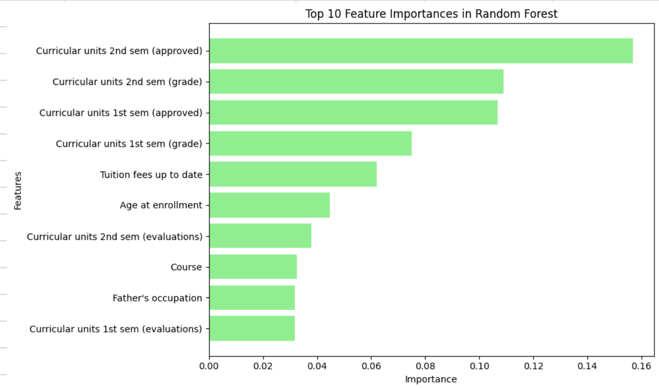
Random Forest Feature Remove < 0.02	Importance
Curricular units 2nd sem (approved)	0.156829
Curricular units 2nd sem (grade)	0.109001
Curricular units 1st sem (approved)	0.106847
Curricular units 1st sem (grade)	0.075058
Tuition fees up to date	0.062178
Age at enrollment	0.044679
Curricular units 2nd sem (evaluations)	0.037951
Course	0.032564
Father's occupation	0.03191
Curricular units 1st sem (evaluations)	0.031858
Mother's occupation	0.02902
Father's qualification	0.023198
Application mode	0.022737
Mother's qualification	0.022084
GDP	0.021915
Curricular units 2nd sem (enrolled)	0.021109
Curricular units 1st sem (enrolled)	0.020501
Unemployment rate	0.020319
Inflation rate	0.019943
Debtor	0.016775
Scholarship holder	0.014936
Application order	0.014498
Gender	0.011465
Curricular units 1st sem (credited)	0.008272
Previous qualification	0.007997
Displaced	0.007605
Curricular units 2nd sem (credited)	0.006756
Curricular units 2nd sem (without evaluations)	0.004908
Curricular units 1st sem (without evaluations)	0.004454
Marital status	0.003529
Daytime/evening attendance	0.002973
Nacionality	0.00268
International	0.001751
Educational special needs	0.001696











XGBoost Feature Remove < 0.0	Importance
Curricular units 2nd sem (approved)	0.255959
Tuition fees up to date	0.138809
Curricular units 1st sem (enrolled)	0.057896
Curricular units 2nd sem (enrolled)	0.043951
Daytime/evening attendance	0.036293
Curricular units 1st sem (approved)	0.03179
Curricular units 2nd sem (credited)	0.031152
Debtor	0.027588
Scholarship holder	0.021841
Course	0.020204

SVM Feature Remove < 0.	Weight
Tuition fees up to date	1.541711
International	1.386583
Educational special needs	0.510332
Debtor	0.466021
Daytime/evening attendance	0.435421
Curricular units 2nd sem (app	0.430958
Scholarship holder	0.356493
Displaced	0.352999
Curricular units 2nd sem (enr	0.34928
Curricular units 1st sem (appr	0.294194

Feature,Importance	Importance
Curricular units 2nd sem (appr	0.230968
Curricular units 1st sem (appro	0.076557
Tuition fees up to date	0.054343
Curricular units 2nd sem (enro	0.044345
Debtor	0.03641
Course	0.032716
Age at enrollment	0.029964
Displaced	0.025933
Curricular units 1st sem (enroll	0.023776
Curricular units 2nd sem (grad	0.021577

Random Forest Feature I	Importance
Curricular units 2nd sem (ap	0.156829
Curricular units 2nd sem (gra	0.109001
Curricular units 1st sem (app	0.106847
Curricular units 1st sem (grad	0.075058
Tuition fees up to date	0.062178
Age at enrollment	0.044679
Curricular units 2nd sem (eva	0.037951
Course	0.032564
Father's occupation	0.03191
Curricular units 1st sem (eval	0.031858

	4 Methods have three factor overlap
	3 Methods have three factor overlap
	2 Methods have six factor overlap

Why is XGBoost so effective for predicting the Graduate target?		為什麼 XGBoost 在 Graduate 目標的預測上效果這麼好？
Strong data-fitting capability:		對於數據的擬合能力強：XGBoost 是一種基於梯度提升的樹模型，擅長處理數 值型和分類型混合的數據，對於高維度、非線性關係的數據具有很強的表現力。
XGBoost, as a gradient boosting tree model, excels at handling mixed numerical and categorical data. Its ability to capture non-linear relationships and man		我們的數據中，影響畢業與否的主要特徵(如 Curricular units approved、Tuition fees up to date 等)具有很強的判別力，且數據之間的模式對 XGBoost 來說更容易捕捉。
Highly predictive features in the data:		
In your dataset, key features influencing graduation status, such as Curricular units approved and Tuition fees up to date, possess strong discriminative power. These patterns are well-captured by XGBoost, contributing to its excellent performance.		
		為什麼特徵工程能顯著提升 Drop Out 的 F1 Score？
Why does feature engineering significantly improve the F1 Score for Drop Out?		
Reasons for the F1 Score improvement:		F1 Score 提升的原因：
		新特徵捕捉了關鍵信息：
New features captured critical information:		你提到新增了 Scholarship 和 Debt Ratio，這些可能是與學生輟學密切相關的因素，補充了數據中原本缺失的重要信號。
The inclusion of Scholarship and Debt Ratio added vital signals closely associated with student dropout, compensating for previously missing data.		解決數據分布不均問題：
Addressed data imbalance issues:		原始數據可能存在不平衡或噪聲特徵，通過特徵工程過濾掉低相關特徵，可以提高模型對輟學學生的辨別能力。
Original data may have contained imbalanced or noisy features. Feature engineering filtered out less relevant features, improving the model's ability to ident		準確度沒有提升的原因：
Why accuracy did not improve:		準確度受所有樣本的預測影響，但新增特徵主要對輟學這一類別的預測有幫助，因此主要反映在 F1 Score 而非準確度上。
Accuracy reflects predictions for all samples, but the newly added features primarily benefited dropout predictions, which impacted F1 Score rather than overall accuracy.		
Which model performs best for Drop Out, and why?		哪個模型在 Drop Out 表現最好，為什麼？
Best-performing model:		
		表現最好的模型：
XGBoost: Performed the best due to its high flexibility and responsiveness to feature engineering.		XGBoost 表現最佳：由於其高靈活性和對特徵工程的良好響應， XGBoost 在 Graduate 和 Dropout 目標上都表現出色。
Random Forest: Demonstrated high stability, making it a reliable choice for interpretability and consistent performance.		Random Forest 穩定性高：在解釋性和穩定性方面， Random Forest 是一個可靠的選擇
Why XGBoost and Random Forest outperform others:	分析一下各模型優缺點	XGBoost 的優勢：
	跟我們的資料？	高靈活性：XGBoost 能夠處理複雜的非線性關係，並且通過正則化防止過擬合。
XGBoost's strengths:		特徵重要性：XGBoost 能夠自動選擇重要特徵，這在特徵工程後尤其有用。
High flexibility: Handles complex non-linear relationships effectively.		Random Forest 的優勢：
Feature importance: Automatically selects the most critical features, which is especially useful after feature engineering.		穩定性：Random Forest 通過集成多棵樹來減少方差，表現通常較為穩定。
Random Forest's strengths:		解釋性：Random Forest 的特徵重要性易於解釋，這在分析輟學因素時非常有用。
Stability: Reduces variance through the ensemble of trees.		為什麼比其他模型好？
Interpretability: Feature importance in Random Forest is easy to understand, aiding dropout analysis.		集成學習的優勢：XGBoost 和 Random Forest 都是集成學習方法，能 夠通過組合多個弱學習器來提高預測能力。
Why these models are better than others:		對特徵工程的響應：這些模型能 夠充分利用特徵工程後的新特徵，從而提升性能
Ensemble learning advantages: Both XGBoost and Random Forest leverage multiple weak learners to enhance prediction power.		相同的特徵：
Response to feature engineering: These models effectively utilize new features, leading to superior performance.		
Common and unique top features		Curricular units 2nd sem (approved)
Common features across models:		Curricular units 1st sem (approved)
		Tuition fees up to date
Curricular units 2nd sem (approved)		Age at enrollment
Curricular units 1st sem (approved)		International
Tuition fees up to date		
Age at enrollment		不同的特徵：
International		
Unique features:		Scholarship holder 和 Debt Ratio 是新增的特徵，對於輟學模型的 F1 Score 提升有顯著幫助。
Newly added features like Scholarship holder and Debt Ratio significantly improved F1 Score for dropout prediction.		
Why do the new features, Scholarship and Debt Ratio, improve the F1 Score?		
Scholarship:		為什麼 Scholarship 和 Debt Ratio 新特徵對 F1 Score 有幫助？
Students receiving scholarships typically exhibit stable academic performance, reducing their risk of dropout. This feature helps the model better identify low-risk groups.		
		Scholarship：
Debt Ratio:		
Debt ratio is an indicator of financial pressure, directly linked to dropout likelihood. This feature enables the model to distinguish high-risk groups impacted		獲得獎學金的學生通常具有穩定的學業表現，輟學風險更低。這一特徵能幫助模型更準確地辨別出低風險群體。
Impact on accuracy:		Debt Ratio：
The new features primarily improved the model's ability to predict dropout (a minority class). Overall accuracy is influenced by the majority class, so the improvement is not as noticeable.		
		債務比例是經濟壓力的指標，與輟學的可能性有直接關聯。該特徵幫助模型區分出受經濟壓力影響的高風險群體。
Key Takeaways		
Feature engineering is critical for improving model performance:		影響準確度的原因：
Adding Scholarship and Debt Ratio significantly enhanced the F1 Score, highlighting the importance of effective feature creation for predicting key categories.		
		新特徵主要改善了模型對輟學學生(小比例類別)的預測，而整體準確度受大比例類別影響更大，因此對準確度的影響不明顯。

XGBoost's stability in predicting Graduate:	
XGBoost fully utilized highly predictive features like Curricular units approved, achieving high accuracy in most cases.	
	Key Takeaways
Different models excel in different scenarios:	特徵工程是提升模型性能的關鍵:新增的 Scholarship 和 Debt Ratio 成功提高了 F1 Score，說明有效的特徵能顯著改善模型對關鍵類別的預測。
	XGBoost 對 Graduate 預測的穩定性:該模型充分利用了數據中強相關特徵(如 Curricular units approved)，在多數情況下能達到高準確度。
SVM and Random Forest performed best after feature engineering for Drop Out, demonstrating their ability to leverage new features effectively.	不同模型適合不同場景:
Shared and unique features:	SVM 和 Random Forest 在 Drop Out 的特徵工程後表現最佳，說明這些模型更能利用新增特徵。
While most models shared common top features, newly added features like Scholarship and Debt Ratio provided critical improvements for predicting dropo	共通與差異性:多數模型的前 10 大特徵相似，但新增的特徵(如 Scholarship 和 Debt Ratio)為提升預測輟學率提供了關鍵幫助。
	數據洞察:
Data insights:	經濟因素(獎學金、負債)對學生的學習成果有顯著影響
	學業表現指標(GPA、出席率)是很好的預測指標
Economic factors (e.g., scholarships, debt levels) significantly influence student outcomes.	家庭背景對學生成功有重要影響
Academic performance indicators (e.g., GPA, attendance) are strong predictors.	
Family background plays a crucial role in student success.	
Practical Applications	實際應用價值:
Early identification of at-risk students:	可以早期識別有輟學風險的學生
Enables proactive interventions for students with a high risk of dropping out.	幫助學校更好地分配獎學金和支援資源
	提供具體的干預指標(如經濟支援需求)
Better allocation of resources:	
Assists schools in distributing scholarships and support services more effectively.	
Specific intervention indicators:	
Provides actionable insights (e.g., financial support needs) for targeted assistance programs.	