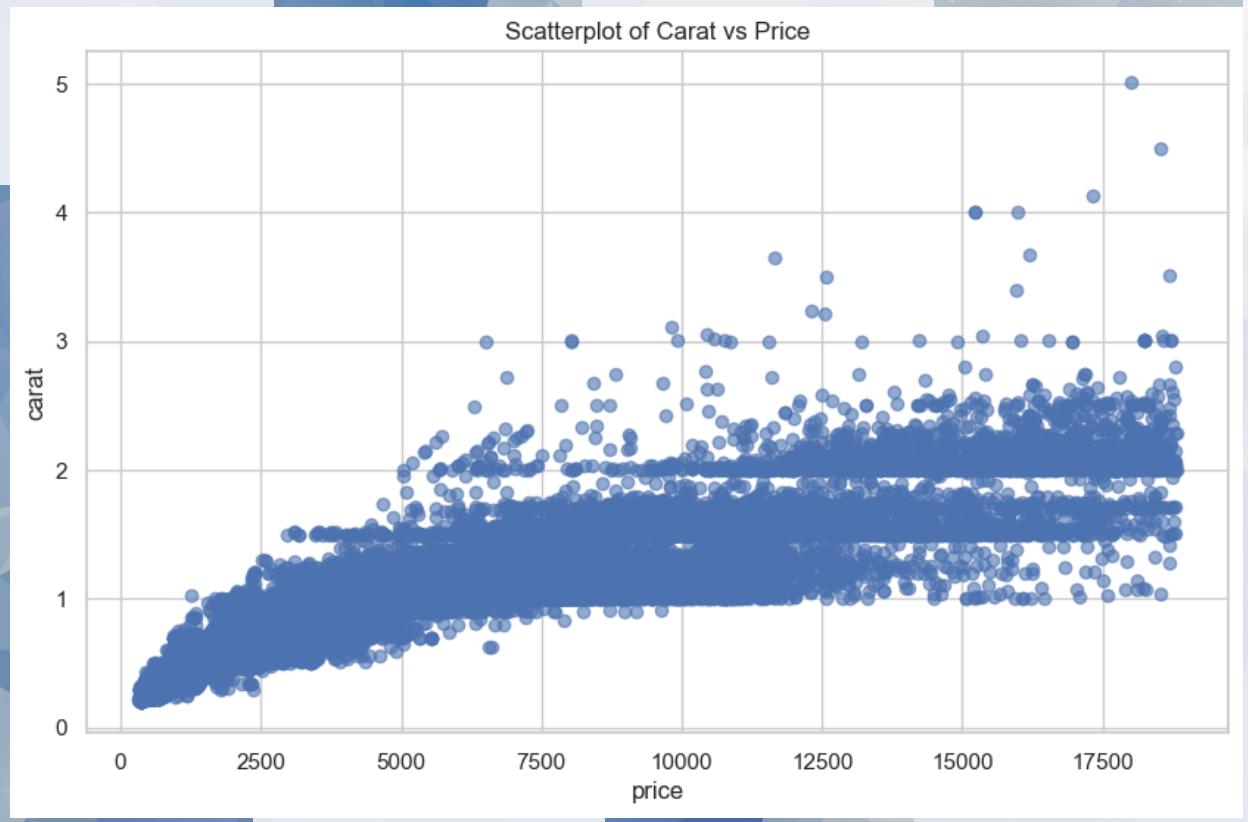
鑽石價格預測之監督式學習模型比較分析

姓名學號： 曾巧庭（5112053014）
所屬系所組：應用數學系大數據

研究動機與論文選擇



傳統鑽石估價依賴 4C（Carat、Cut、Color、Clarity），
從右圖可見，即使 Carat 趨勢與價格成正比，
在高價位區仍存在大量價格離散、波動劇烈的情形。

因此價格可能受多重因素交互影響，
僅靠 4C 難以準確評估
故開始研究機器學習方法來提升預測精度與穩定性。

研究動機與論文選擇

Sharma et al. (2021)

G. Sharma, V. Tripathi, M. Mahajan, and A. K. Srivastava,
"Comparative Analysis of Supervised Models for Diamond Price Prediction,"
in *Proc. 11th Int. Conf. on Cloud Computing, Data Science & Engineering (Confluence)*, 2021, pp. 1019–1022.

Amadavadi et al. (2024)

K. Amadavadi, R. Rane, and R. Patankar,
"Diamond Price Prediction using Machine Learning Techniques,"
in *Proc. 5th Int. Conf. on Smart Electronics and Communication (ICOSEC)*, 2024.

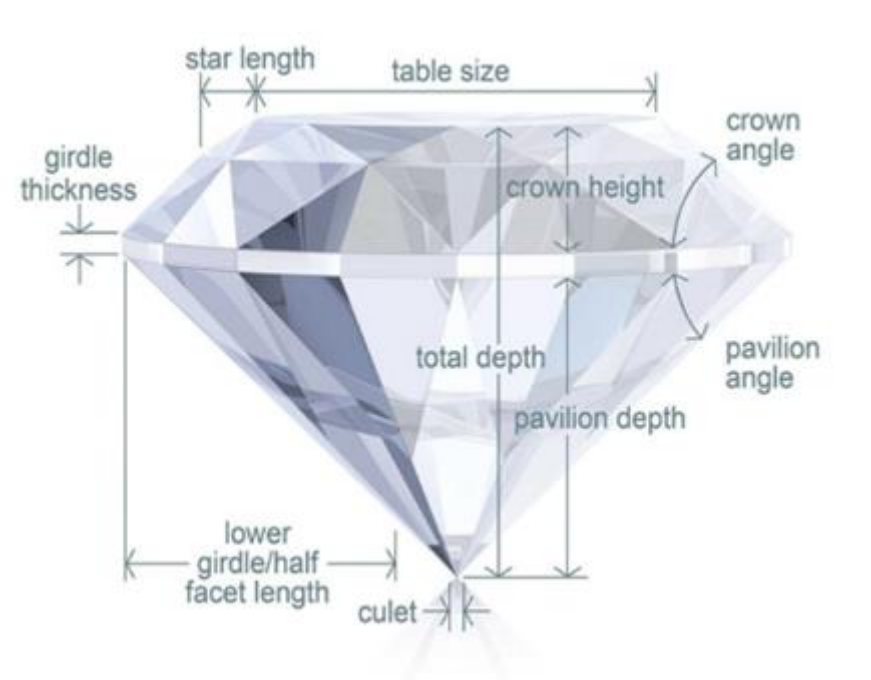
Fitriani et al. (2022)

S. A. Fitriani and I. Surjandari,
"Least Absolute Shrinkage and Selection Operator (LASSO) and k-Nearest Neighbors (k-NN) Algorithm Analysis Based on Feature Selection for Diamond Price Prediction,"
Procedia Computer Science, vol. 197, pp. 457–464, 2022.

數據來源與特徵

使用 Kaggle 鑽石資料集（53940 筆）

特徵包含：Cut、Color、Clarity、Carat、Depth、Table 、 X、Y、Z



Features	*Range*
Cut	(Fair, Good, Very Good, Premium, Ideal)
Color	J (worst) - D (best)
Clarity	(I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
Carat	0.2 - 5.01 ct
Depth	(0 - 31.8) mm
Table	(43 - 95) mm
Price	($326--$18,823)
X(length)	(0 - 10.74) mm
Y(width)	(0 - 58.9) mm
Z(depth)	(0 - 31.8) mm

1

Sharma et al. (2021)

Comparative Analysis of Supervised Models for Diamond Price Prediction

無特徵篩選

RMSE
MAE

Label Encoding

Standardization

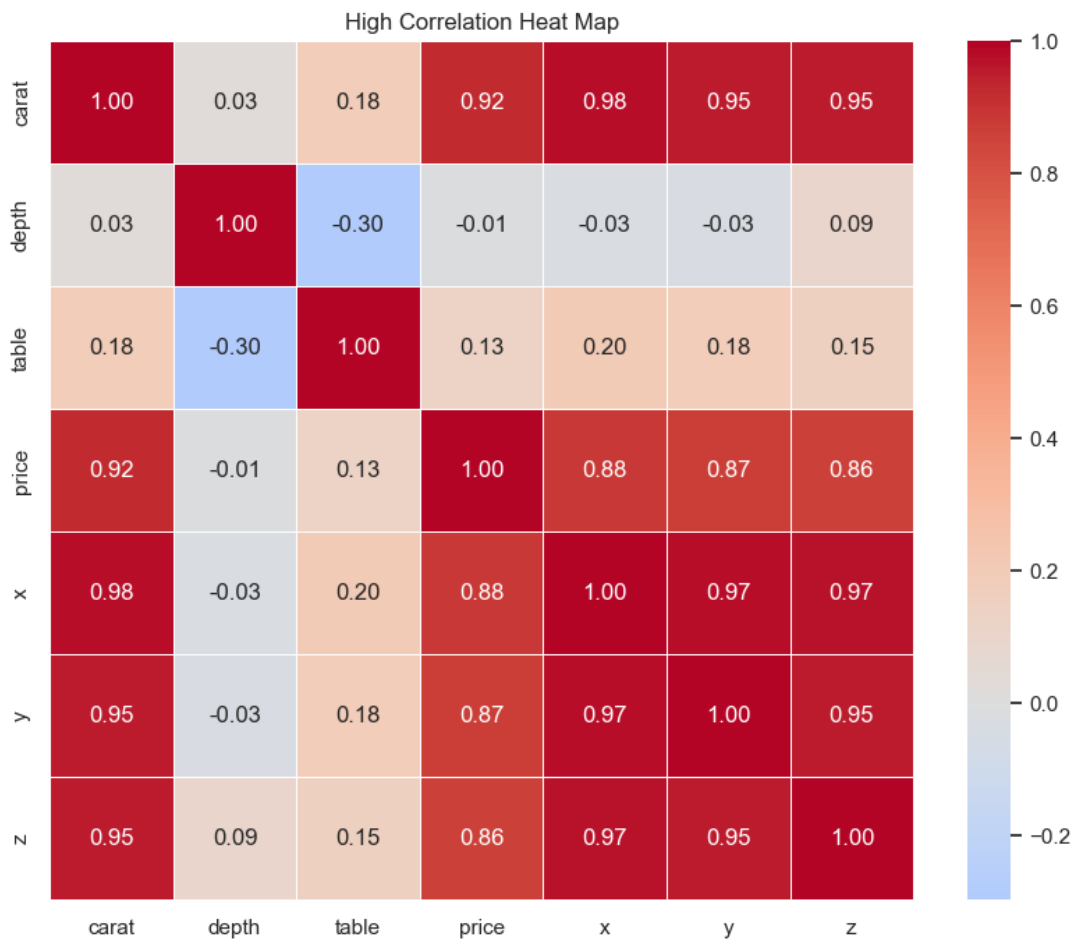
Features	*Range*
Cut	(Fair, Good, Very Good, Premium, Ideal)
Color	J (worst) - D (best)
Clarity	(I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))

Linear Regression
Lasso Regression
Ridge Regression
ElasticNet Regression
Decision Tree Regressor
Random Forest Regressor
AdaBoost Regressor
Gradient Boosting Regressor

1

Sharma et al. (2021)

Comparative Analysis of Supervised Models for Diamond Price Prediction



	Model	RMSE	Cross-Validation RMSE	Accuracy
0	Random Forest	569.237693	582.791666	97.961658
1	Decision Tree	630.164799	665.212760	97.501968
2	Gradient Boosting	655.694773	684.031070	97.295462
3	AdaBoost	1260.560387	1339.595915	90.004213
4	Ridge Regression	1351.262011	1355.212906	88.513999
5	Linear Regression	1351.263480	1355.234769	88.513974
6	Lasso Regression	1351.268941	1355.152417	88.513881
7	ElasticNet	1444.055914	1457.803767	86.882299

Random Forest 模型的 RMSE 為 569.2，
Accuracy 達到 97.96%，
是表現最穩定、誤差最小的模型。

Amadavadi et al. (2024)

Diamond Price Prediction using Machine Learning Techniques

Amadavadi et al. (2024)

Diamond Price Prediction using Machine Learning Techniques

Model	RMSE	Cross-Validation RMSE	Accuracy
Random Forest	569.237693	582.791666	97.961658
Decision Tree	630.164799	665.212760	97.501968
Gradient Boosting	655.694773	684.031070	97.295462
AdaBoost	1260.560387	1339.595915	90.004213
Ridge Regression	1351.262011	1355.212906	88.513999
Linear Regression	1351.263480	1355.234769	88.513974
Lasso Regression	1351.268941	1355.152417	88.513881
ElasticNet	1444.055914	1457.803767	86.882299

Sharma et al. (2021)

Model	MAE	MSE	RMSE	R Squared
LinearRegression	567.916027	697003.796559	834.867532	0.896314
DecisionTree	259.120393	238556.832948	488.422801	0.964512
RandomForest	192.162892	123785.374497	351.831458	0.981586
KNeighbors	343.129666	327935.486541	572.656517	0.951217
XGBRegressor	192.134094	122166.640625	349.523448	0.981827
ExtraTreeRegressor	188.525790	119180.357145	345.225082	0.982271
GradientBoosting	223.726090	139723.044494	373.795458	0.979215
MLP	409.160588	439141.865100	662.677799	0.934674

Amadavadi et al. (2024)

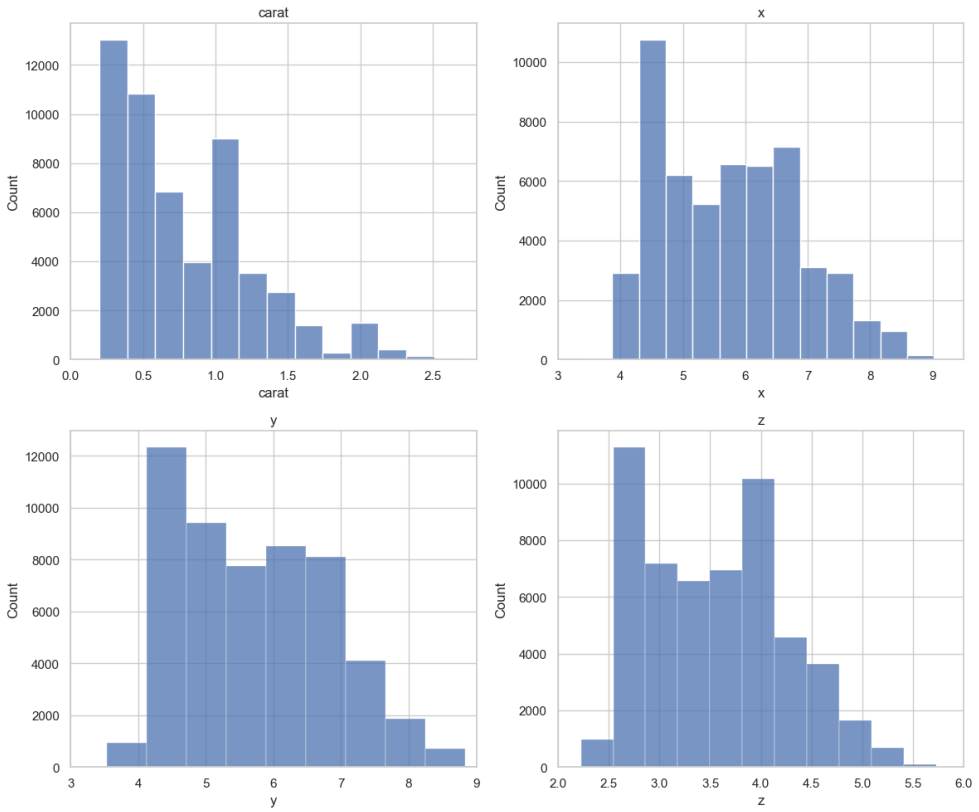
Extra Trees Regressor 在 Amadavadi 的模型中表現最優，R^2 高達 0.9823，略勝 Sharma 文獻中最佳的 Random Forest（$R^2 = 0.9796$）

Fitriani et al. (2022)

Least Absolute Shrinkage and Selection Operator (LASSO) and k-Nearest Neighbors (k-NN) Algorithm Analysis Based on Feature Selection for Diamond Price Prediction

Amadavadi et al. (2024)

Diamond Price Prediction using Machine Learning Techniques



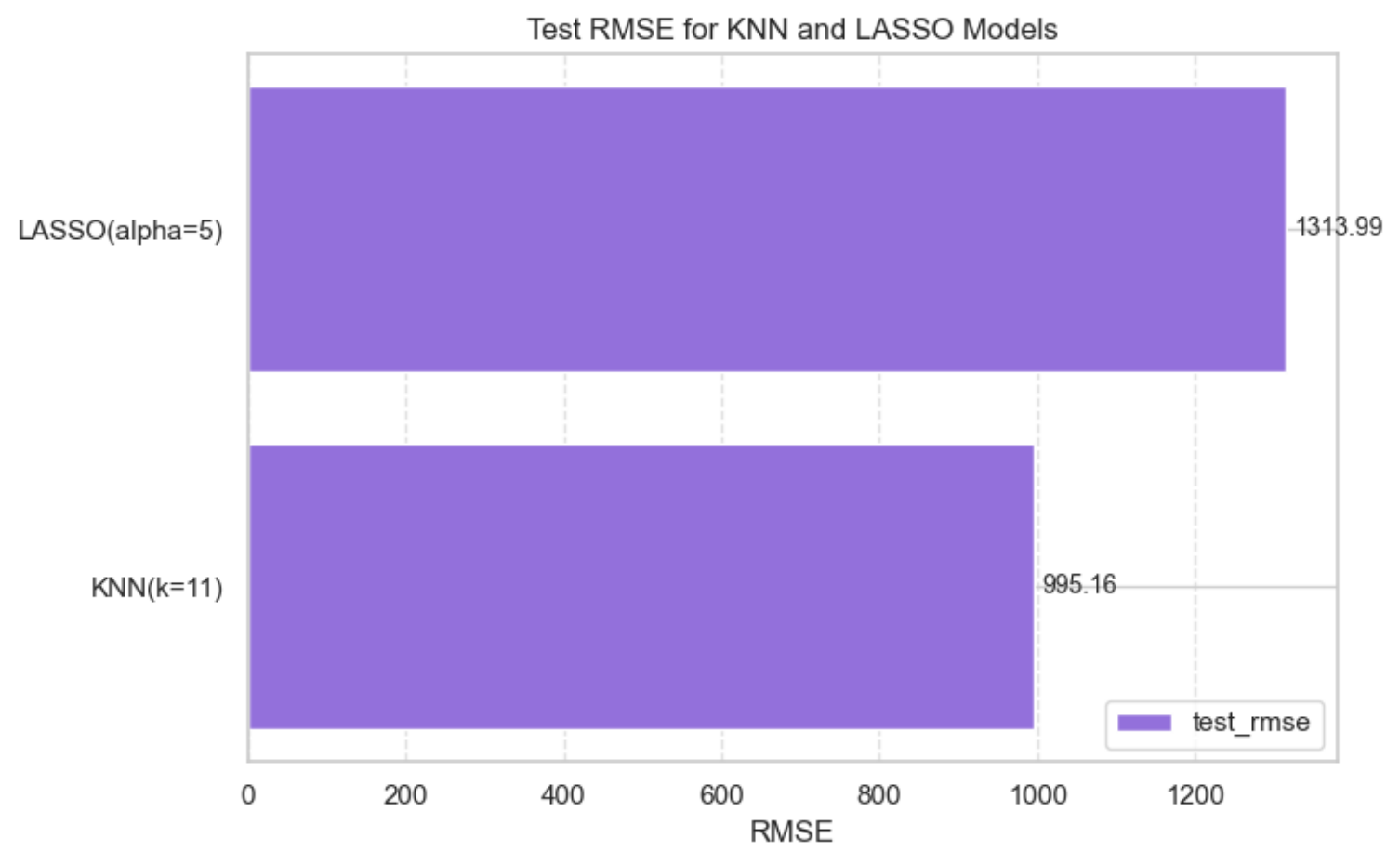
carat 呈現偏態分布，
而其他幾個幾何特徵也略有分散，
在建模前使用了標準化。

	alpha	RMSE	R2_Score
0	5	1313.99	0.8867
1	10	1317.79	0.8861
2	20	1330.06	0.8839
3	30	1347.07	0.8809
4	40	1349.53	0.8805
5	45	1350.99	0.8803
6	50	1352.59	0.8800
7	55	1354.35	0.8797
8	100	1376.05	0.8758

	k	RMSE	R^2
0	1	1287.632419	0.891222
1	2	1131.287504	0.916034
2	3	1078.388610	0.923703
3	4	1047.003890	0.928080
4	5	1029.347280	0.930485
5	6	1022.208706	0.931446
6	7	1010.699664	0.932981
7	8	1005.370024	0.933686
8	9	999.773503	0.934422
9	10	996.623793	0.934834
10	11	995.159619	0.935026
11	12	995.642522	0.934963
12	13	995.408070	0.934993
13	14	996.575584	0.934841
14	15	996.807165	0.934810

Amadavadi et al. (2024)

Diamond Price Prediction using Machine Learning Techniques



在特徵簡化後， k-NN 在 k = 11 時表現最佳。
RMSE（995.16）明顯優於 LASSO（1313.99）。

Model Comparison Summary

文獻	最佳模型	準確率	特徵處理
Sharma et al.(2021)	Random Forest	0.9796	無特徵篩選
Amadavadi et al.(2024)	Extra Trees	0.9823	離群值處理
Fitriani et al.(2022)	k-NN (k=11)	0.935	特徵篩選後效果佳

Q & A

THANK YOU