Project: M1 Apple Laptop Purchase Prediction

WISDOM IZUCHUKWU ADIKE

TABLE OF CONTENTS

DATASET
Description

INTRODUCTION

The Apple M1 MacBook is a popular laptop that has gained a lot of attention in recent years due to its impressive performance and energy efficiency. If you are considering purchasing an M1 MacBook, there are several factors that you may want to consider before making your decision. One factor to consider is your budget. The M1 MacBook is available in a range of prices, depending on the specific model and configuration you choose. It's important to carefully consider your budget and choose a model that fits your needs and your financial situation. Another factor to consider is your computing needs.

The M1 MacBook is a powerful machine that is well-suited for a wide range of tasks, including running demanding software, playing games, and handling heavy workloads. However, if you only need a laptop for basic tasks like web browsing and word processing, you may be able to get by with a less powerful and less expensive model. You may also want to consider the design and form factor of the M1 MacBook. The M1 MacBook is available in both 13-inch and 16-inch sizes, and you'll need to decide which size is right for you. Additionally, the M1 MacBook is available in both a standard laptop form factor and a more portable "MacBook Air" form factor.

Finally, you'll want to consider the availability and support options for the M1 MacBook. Apple is known for its strong support network, and the M1 MacBook is no exception. You can find support through Apple's online resources, as well as through authorized Apple service providers.

DATASET DESCRIPTION 01

[243]:

	trust_apple	interest_computers	age_computer	user_pcmac	appleproducts_count	familiarity_m1	f_batterylife	f_pr
0	No	4	8	PC	0	No	5	
1	Yes	2	4	PC	1	No	5	
2	Yes	5	6	PC	0	No	3	
3	Yes	2	6	Apple	4	No	4	
4	Yes	4	4	Apple	7	Yes	5	

5 rows × 22 columns

- trust_apple - Brand trust : (Yes, No)
- Interest_computers - Level of interest in computers : (1 Not interested - 5 Very interested)
- age_computer - Age of your current computer : (0 means less than one year - 6 years or more)
- user_pc or mac - Type of computer : (0 PC , 1 Apple, 2 Hp or Other)
- appleproducts_count - Count of apple products your own : (0 - 10 or more)
- familiarity_m1 - Brand familiarity (Yes, No)

- f_batterylife - Importance of (1 Not important - 5 is very import)
- f_price - Cheaper price (1 Not important - 5 is very import)
- f_size - Thinner of computer (1 Not important - 5 is very import)
- f_multitasking - Improved multitasking power (1 Not important - 5 is very import)
- f_noise - Less noisy (1 Not important - 5 is very import)
- f_performance - Improved performance (1 Not important - 5 is very import)
- f_neural - Neural engine (1 Not important - 5 is very import)
- f_synergy - How important is a seamless experience (1 Not important - 5 is very import)
- f_performanceloss - A small loss in performance (1 Not important - 5 is very import)
- m1_consideration - M1 Chip into account in the selection process of buying a new Apple computer (1 Not important - 5 is very import)
- m1_purchase - Would you buy one of the new Apple M1 Macs (Yes, No)

DATASET DESCRIPTION 02

[249]:

	interest_computers	age_computer	appleproducts_count	f_batterylife	f_price	f_size	f_multitasking
count	133.00000	133.000000	133.000000	133.000000	133.000000	133.000000	133.000000
mean	3.81203	2.827068	2.609023	4.526316	3.872180	3.157895	4.120301
std	0.96256	2.444881	1.898303	0.723826	0.995547	1.166724	0.798081
min	2.00000	0.000000	0.000000	1.000000	1.000000	1.000000	2.000000
25%	3.00000	1.000000	1.000000	4.000000	3.000000	2.000000	4.000000
50%	4.00000	3.000000	3.000000	5.000000	4.000000	3.000000	4.000000
75%	5.00000	5.000000	4.000000	5.000000	5.000000	4.000000	5.000000
max	5.00000	9.000000	8.000000	5.000000	5.000000	5.000000	5.000000

DATASET DESCRIPTION 03

```
[247]: trust_apple            0
       interest_computers     0
       age_computer           0
       user_pcmac             0
       appleproducts_count    0
       familiarity_m1         0
       f_batterylife          0
       f_price                0
       f_size                 0
       f_multitasking         0
       f_noise                0
       f_performance          0
       f_neural               0
       f_synergy              0
       f_performanceloss      0
       m1_consideration       0
       m1_purchase            0
       gender                 0
       age_group              0
       income_group           0
       status                 0
       domain                 0
       dtype: int64
```

After data cleaning, we can see we have no missing value.

DATA
Analysis

Objective of the analysis

In this section, I am demonstrating the connection between various characteristics to identify the features that have the most significant impact on our target variable, 'm1_purchase.' Following that, I am constructing various classification models utilizing advanced methods like GridSearch, ML pipelines, and fine-tuning hyperparameters to attain the optimal predictive model in terms of accuracy. Furthermore, I will address the weaknesses of each model.

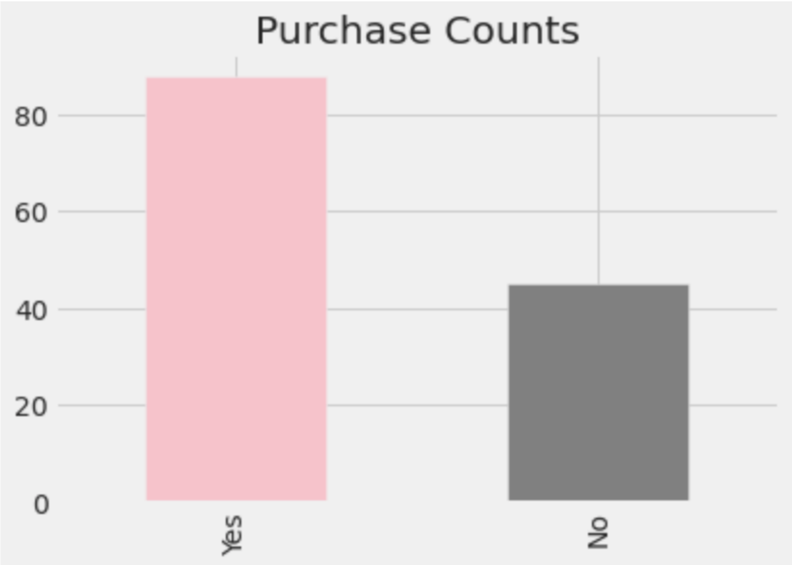
DATA ANALYSIS 01

```
Categorical Features : ['trust_apple', 'interest_computers', 'age_computer', 'user_pcmac', 'appleproducts
_count', 'familiarity_m1', 'f_batterylife', 'f_price', 'f_size', 'f_multitasking', 'f_noise', 'f_performa
nce', 'f_neural', 'f_synergy', 'f_performanceloss', 'm1_consideration', 'm1_purchase', 'gender', 'age_gro
up', 'income_group', 'status']
Continuous Features : ['domain']
```

After data processing, we identified the categorical and continuous features so we can know their relationships.

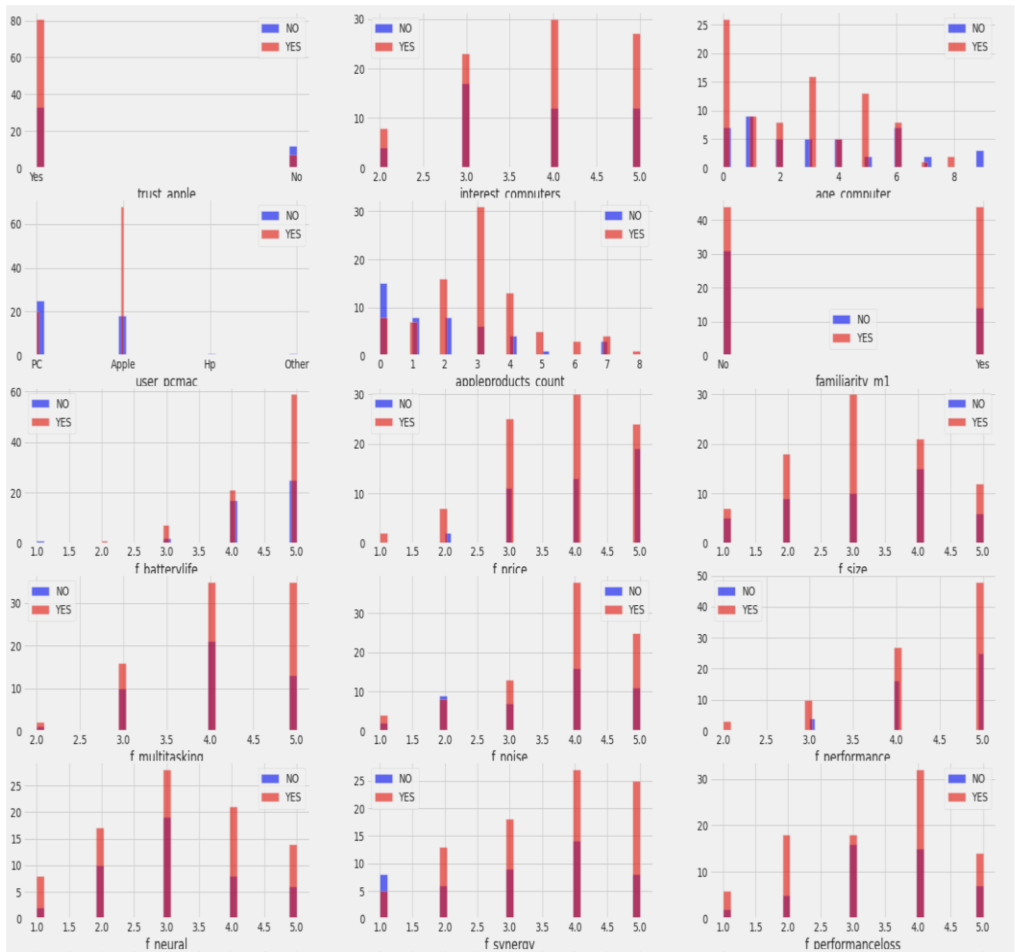
DATA ANALYSIS 02

```
Yes    88
No     45
Name: m1_purchase, dtype: int64
[253]: <AxesSubplot:title={'center':'Purchase Counts'}>
```



From the chart, we have 88 purchases of the M1 Apple Laptop and 45 no purchase.

DATA ANALYSIS 03



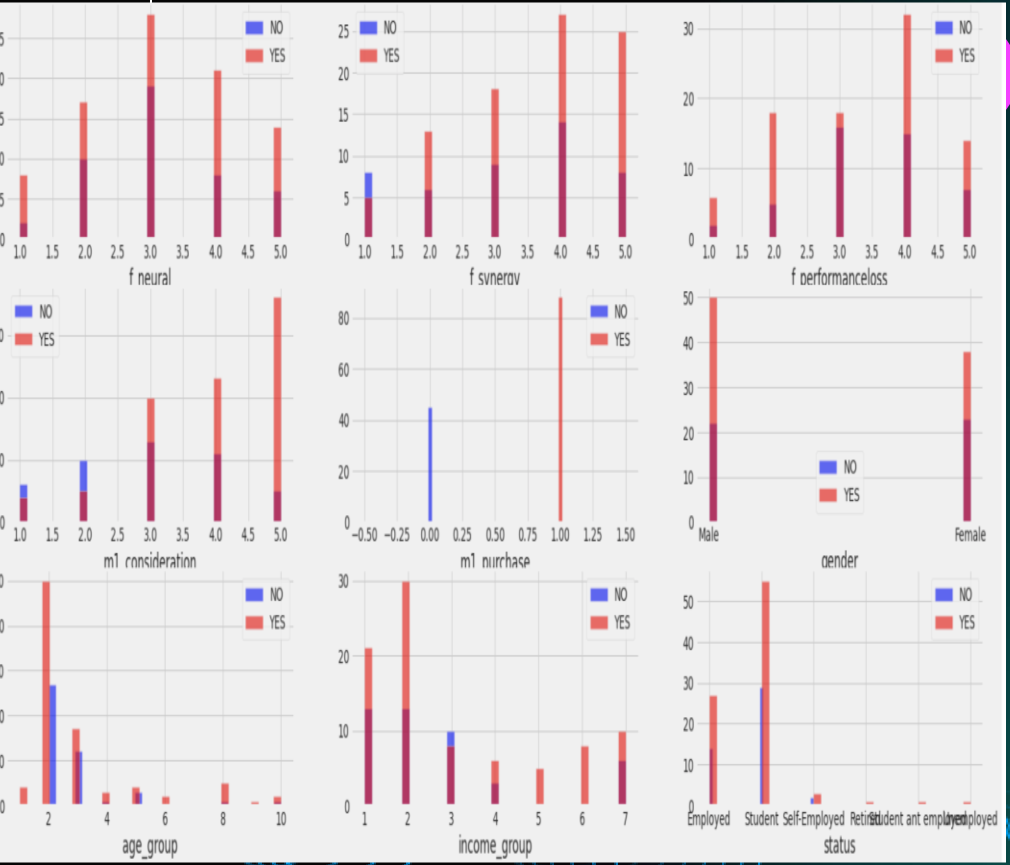
Study of the relationship of categorical features and the m1_purchase:

trust_apple: Looking at the chart, we can see based on the brand trust, people who tends to trust the Apple brand have more tendency to purchase the M1 Apple Laptop.

age_computer: On the chart, 0 means the computer is less than 1 year old and we have more of it which influences the purchase outcome, followed by 3 years old computers.

f_performance: The performance of the M1 Apple Laptop product greatly influenced the purchase outcome. On the chart, 5 means its a high performance product. Having more of 5 on the bar chart convinces people to purchase more of the high performance M1 Laptop.

DATA ANALYSIS 04



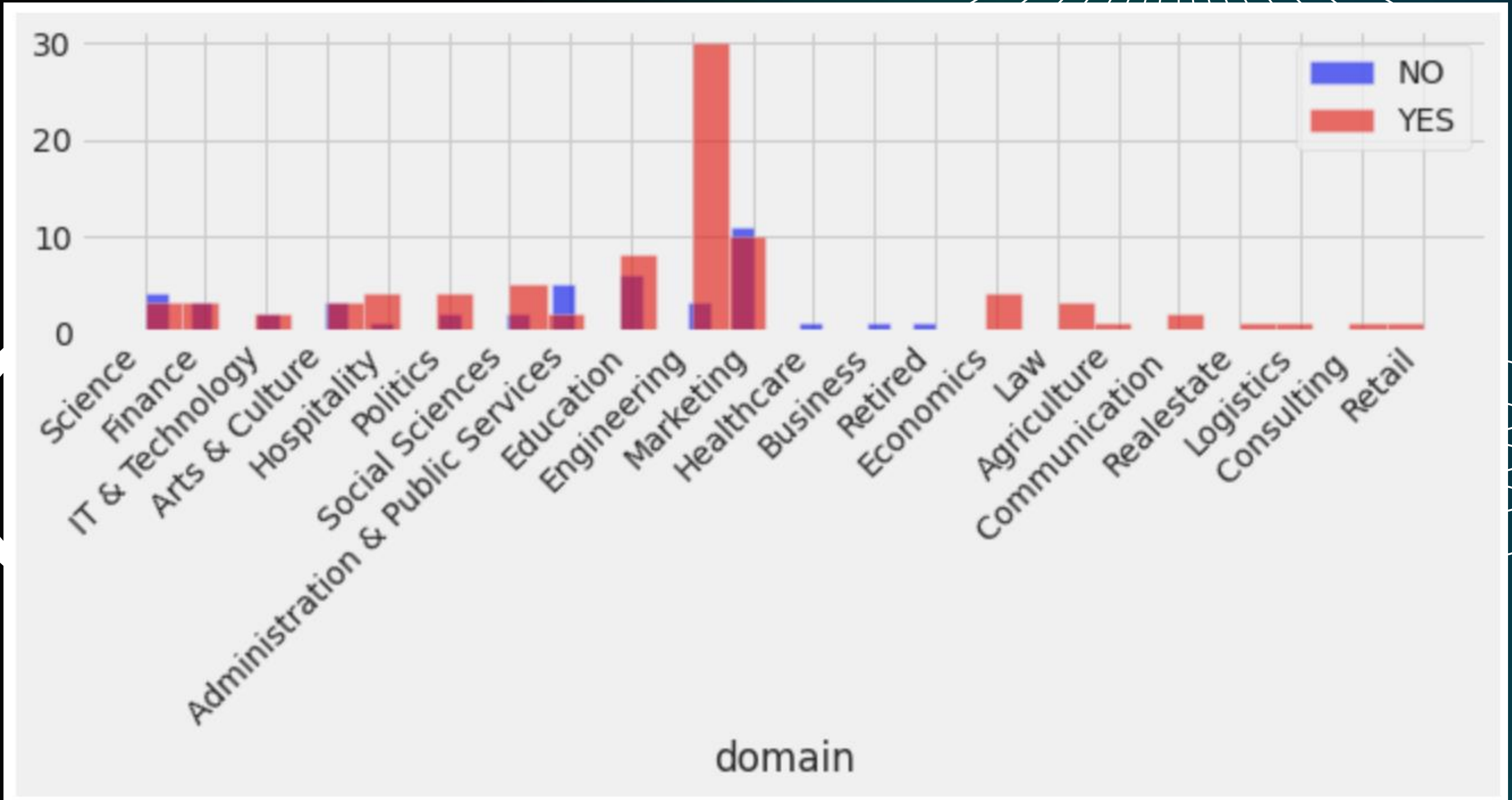
Study of the relationship of categorical features and the m1_purchase:

f_synergy: Looking at the chart, we can see based on the importance of the seamless experience, the M1 Laptop has more of value 4 which shows the product can be used without hassle and that influences the purchase outcome

f_neural: On the chart, 5 means the computer has an important neaural engine which influences the purchase outcome, we can notice value 3 is more, followed by 4 and it's one of the reason it got more purchases.

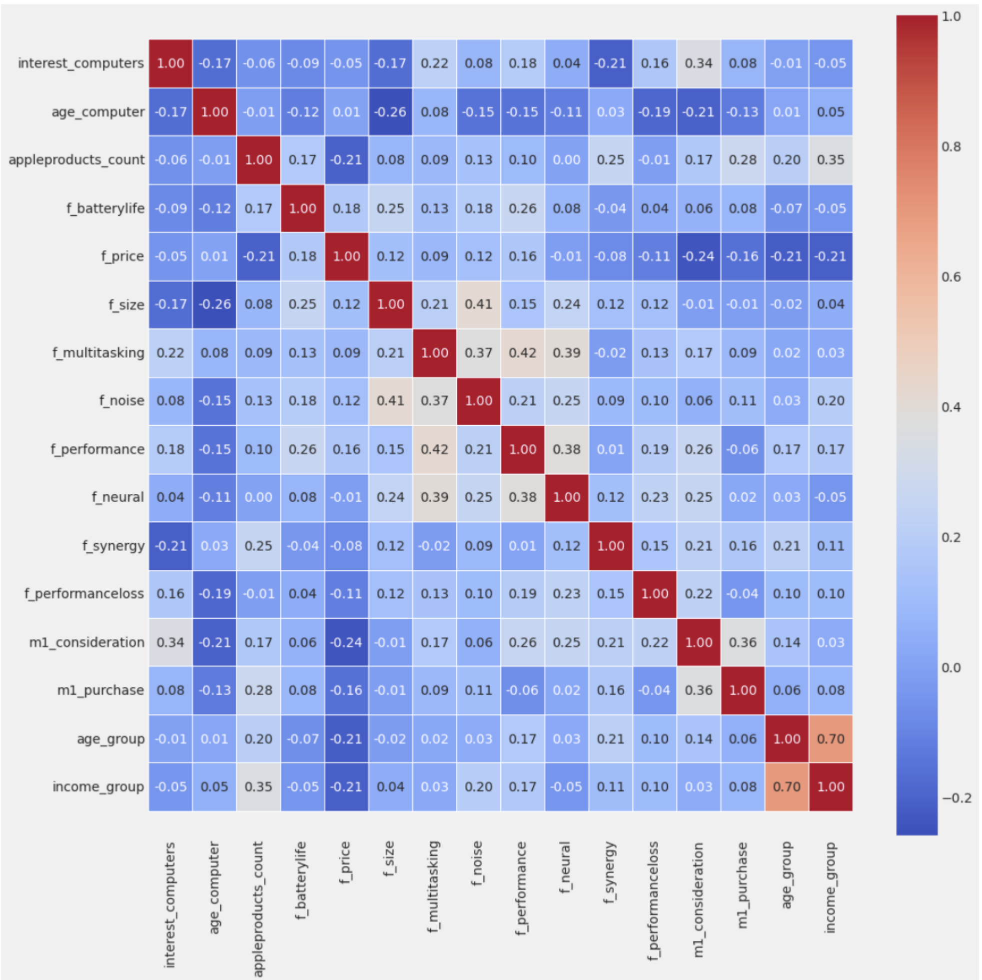
gender: What do we say here? Gender might also influence the purchasing outcome. On the chart, males are more in possession of the M1 laptop than females. This can be an issue for the ladies who are more into girly things.

DATA ANALYSIS 05



As clearly seen in our bar chart, the M1 Laptop is more used in the Engineering sector and it also influences the purchase outcome in the case where Engineers tends to buy more which is followed by the marketers.

DATA ANALYSIS 06



Studying the correlations between features using Heat Map:

The purpose of this matrix is to illustrate the connections among the features. While this is valuable for feature engineering techniques, our primary concern in this lesson is the association between the target variable (identifying whether they purchased the M1 Laptop or not) and the remaining features. In essence, our attention will be primarily directed towards the last row of the matrix.

f_price, f_batterylife are amongst the features least related to the target variable while the rest features has high correlation to the target m1_purchase including m1_consideration as the strongest correlation.

FEATURE ENGINEERING 01

[273]:

f_noise	f_performance	f_neural	...	domain_IT & Technology	domain_Law	domain_Logistics	domain_Marketing	domain_Politic
4	2	2	...	0	0	0	0	
4	5	2	...	0	0	0	0	
1	4	2	...	1	0	0	0	
4	4	4	...	0	0	0	0	
4	5	3	...	0	0	0	0	

We converted the categorical continuous column using the pd.get_dummies method while we also converted the ordinal categorical values using the LabelEncoder library from the sklearn.preprocessing.

FEATURE ENGINEERING 02

	interest_computers	age_computer	appleproducts_count	f_batterylife	f_price	f_size	f_multitasking
0	0.196020	2.123821	-1.379594	0.656892	0.128877	-0.135843	-0.151307
1	-1.889629	0.481564	-0.850816	0.656892	1.137147	1.584840	-1.409050
2	1.238844	1.302693	-1.379594	-2.116651	0.128877	-0.996185	-0.151307
3	-1.889629	1.302693	0.735518	-0.729880	-0.879394	-0.135843	-0.151307
4	0.196020	0.481564	2.321852	0.656892	-0.879394	-0.135843	-0.151307
...	...	...	...	...	...	...	...
128	1.238844	-1.160693	-0.850816	0.656892	-0.879394	-0.135843	1.106436
129	1.238844	2.123821	1.264296	-0.729880	-0.879394	-0.996185	-0.151307
130	0.196020	-1.160693	2.850630	0.656892	0.128877	-0.135843	1.106436
131	1.238844	0.892128	0.735518	0.656892	-0.879394	0.724498	-0.151307
132	1.238844	-0.339564	2.321852	-0.729880	-0.879394	0.724498	-0.151307

133 rows × 54 columns

Also scaled the existing numerical values in the dataset with StandardScaler so as to keep every values in the same scale for our Machine Learning Algorithm.

Machine Learning Analysis

Objective of this section

In the upcoming analysis, we'll compare five classification models (Logistic Regression, KNN, SVM, Decision Tree, and Random Forest) in predicting purchasing outcomes. To build robust models, we'll use techniques like standard scaling, cross-validation, grid search for hyperparameters, and various metrics (e.g., accuracy, precision, F1 Score). Our aim is to determine the best model for accurate predictions.

MACHINE LEARNING 01

```
X = data.drop(columns=['m1_purchase'])
y = data['m1_purchase']

# Split data into train and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

DATA SPLITTING

MACHINE LEARNING 02

```python
from sklearn.linear_model import LogisticRegression

lr = LogisticRegression(solver='liblinear').fit(X_train, y_train)
y_pred_0 = lr.predict(X_test)
clf_report = pd.DataFrame(classification_report(y_test, y_pred_0, output_dict=True))
clf_report
```

	0	1	accuracy	macro avg	weighted avg
precision	0.625	0.842105	0.777778	0.733553	0.777778
recall	0.625	0.842105	0.777778	0.733553	0.777778
f1-score	0.625	0.842105	0.777778	0.733553	0.777778
support	8.000	19.000000	0.777778	27.000000	27.000000

LOGISTIC REGRESSION

MACHINE LEARNING 03

```
from sklearn.linear_model import LogisticRegressionCV

# L1 regularized logistic regression
lr_l1 = LogisticRegressionCV(Cs=10, cv=4, penalty='l1', solver='liblinear').fit(X_train, y_train)
y_pred_1 = lr_l1.predict(X_test)
clf_report = pd.DataFrame(classification_report(y_test, y_pred_1, output_dict=True))
clf_report
```

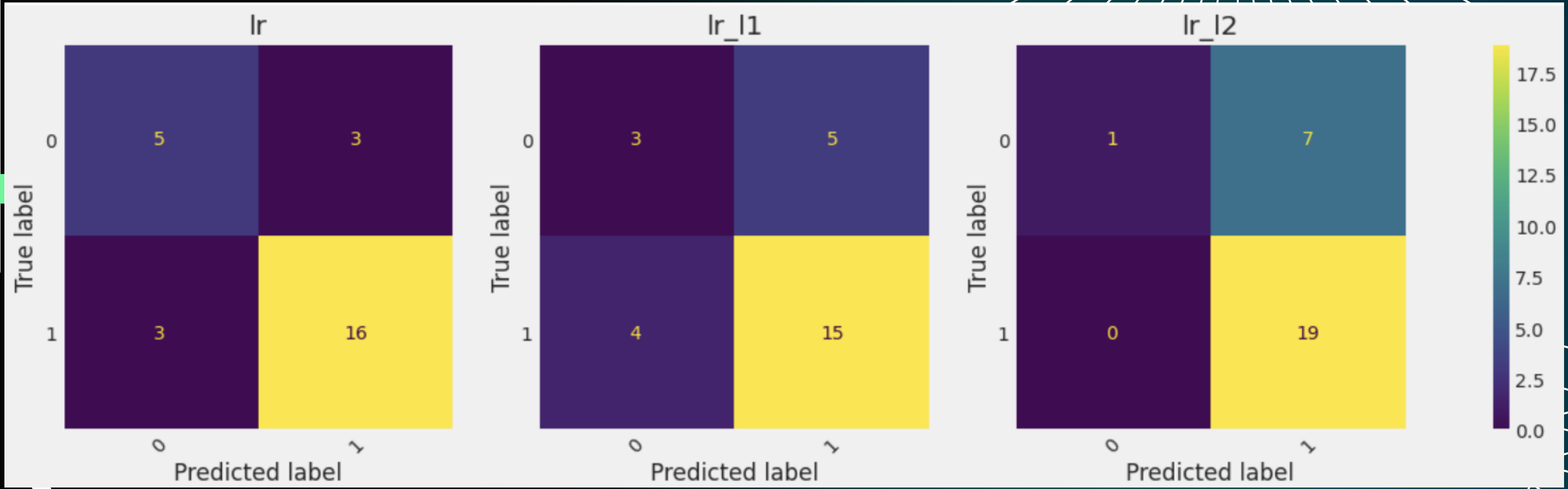
	0	1	accuracy	macro avg	weighted avg
precision	0.428571	0.750000	0.666667	0.589286	0.654762
recall	0.375000	0.789474	0.666667	0.582237	0.666667
f1-score	0.400000	0.769231	0.666667	0.584615	0.659829
support	8.000000	19.000000	0.666667	27.000000	27.000000

```
# L2 regularized logistic regression
lr_l2 = LogisticRegressionCV(Cs=10, cv=4, penalty='l2', solver='liblinear').fit(X_train, y_train)
y_pred_2 = lr_l2.predict(X_test)
clf_report = pd.DataFrame(classification_report(y_test, y_pred_2, output_dict=True))
clf_report
```

	0	1	accuracy	macro avg	weighted avg
precision	1.000000	0.730769	0.740741	0.865385	0.810541
recall	0.125000	1.000000	0.740741	0.562500	0.740741
f1-score	0.222222	0.844444	0.740741	0.533333	0.660082
support	8.000000	19.000000	0.740741	27.000000	27.000000

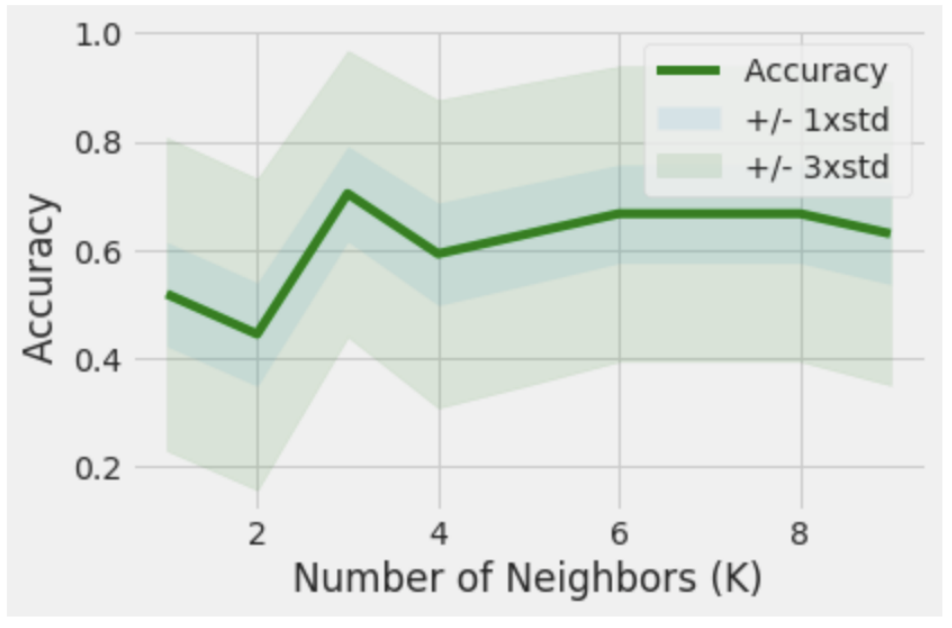
LOGISTIC REGRESSION WITH L1 AND L2 PENALTY

MACHINE LEARNING 04



The best model if we want to consider the Logistic Regression is the model without penalty as it comes with an accuracy of 77% while the L1 and L2 penalties contain 66%b and 74% respectively.

MACHINE LEARNING 05

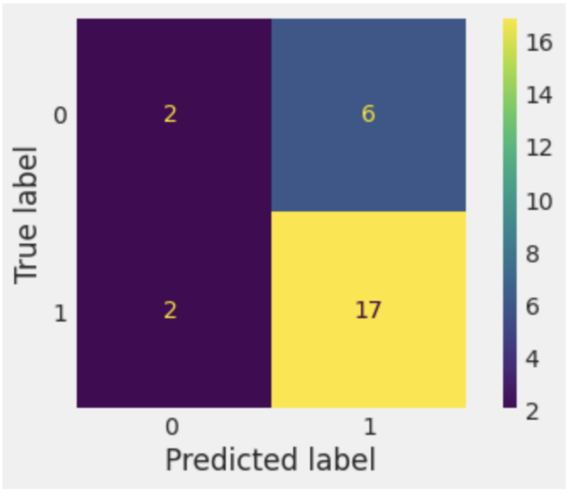


	0	1	accuracy	macro avg	weighted avg
precision	0.500000	0.823529	0.703704	0.661765	0.727669
recall	0.625000	0.736842	0.703704	0.680921	0.703704
f1-score	0.555556	0.777778	0.703704	0.666667	0.711934
support	8.000000	19.000000	0.703704	27.000000	27.000000

K Nearest Neighbors

MACHINE LEARNING 06

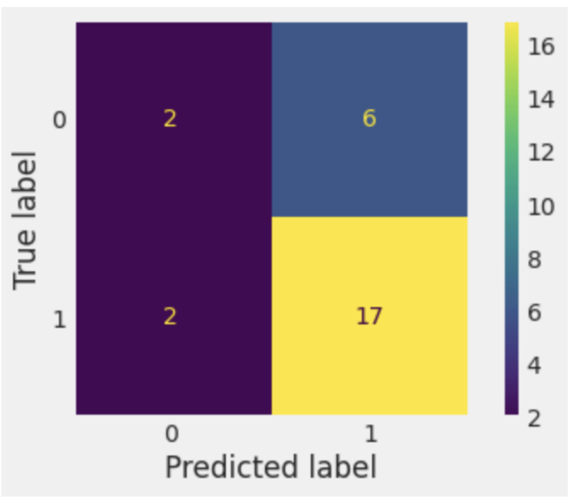
	0	1	accuracy	macro avg	weighted avg
precision	0.500000	0.739130	0.703704	0.619565	0.668277
recall	0.250000	0.894737	0.703704	0.572368	0.703704
f1-score	0.333333	0.809524	0.703704	0.571429	0.668430
support	8.000000	19.000000	0.703704	27.000000	27.000000



Simple Vector Machine

MACHINE LEARNING 07

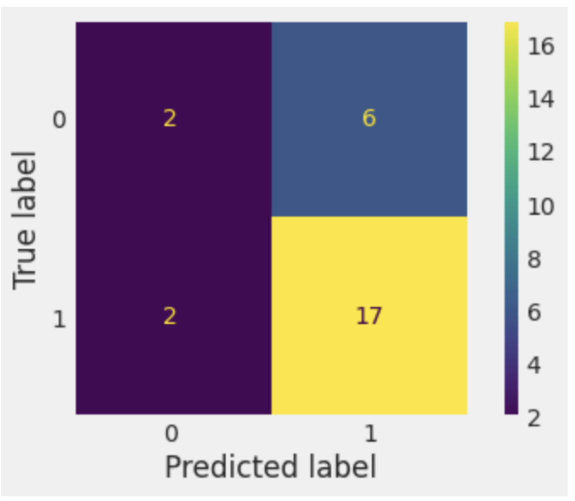
	0	1	accuracy	macro avg	weighted avg
precision	0.636364	0.937500	0.814815	0.786932	0.848274
recall	0.875000	0.789474	0.814815	0.832237	0.814815
f1-score	0.736842	0.857143	0.814815	0.796992	0.821498
support	8.000000	19.000000	0.814815	27.000000	27.000000



Decision Tree

MACHINE LEARNING 08

	0	1	accuracy	macro avg	weighted avg
precision	0.666667	0.888889	0.814815	0.777778	0.823045
recall	0.750000	0.842105	0.814815	0.796053	0.814815
f1-score	0.705882	0.864865	0.814815	0.785374	0.817759
support	8.000000	19.000000	0.814815	27.000000	27.000000



Random Tree

MACHINE LEARNING 09: model comparison

As demonstrated in the previous analysis, all models yield excellent prediction results, and these results are closely aligned. However, the final model selection hinges on identifying the model with the highest performance score.

- Logistic Regression: accuracy of 77%.
- KNN: accuracy of 70%
- SVM: accuracy of 70%
- Decision Tree: accuracy of 81%
- Random Forests: accuracy of 81%

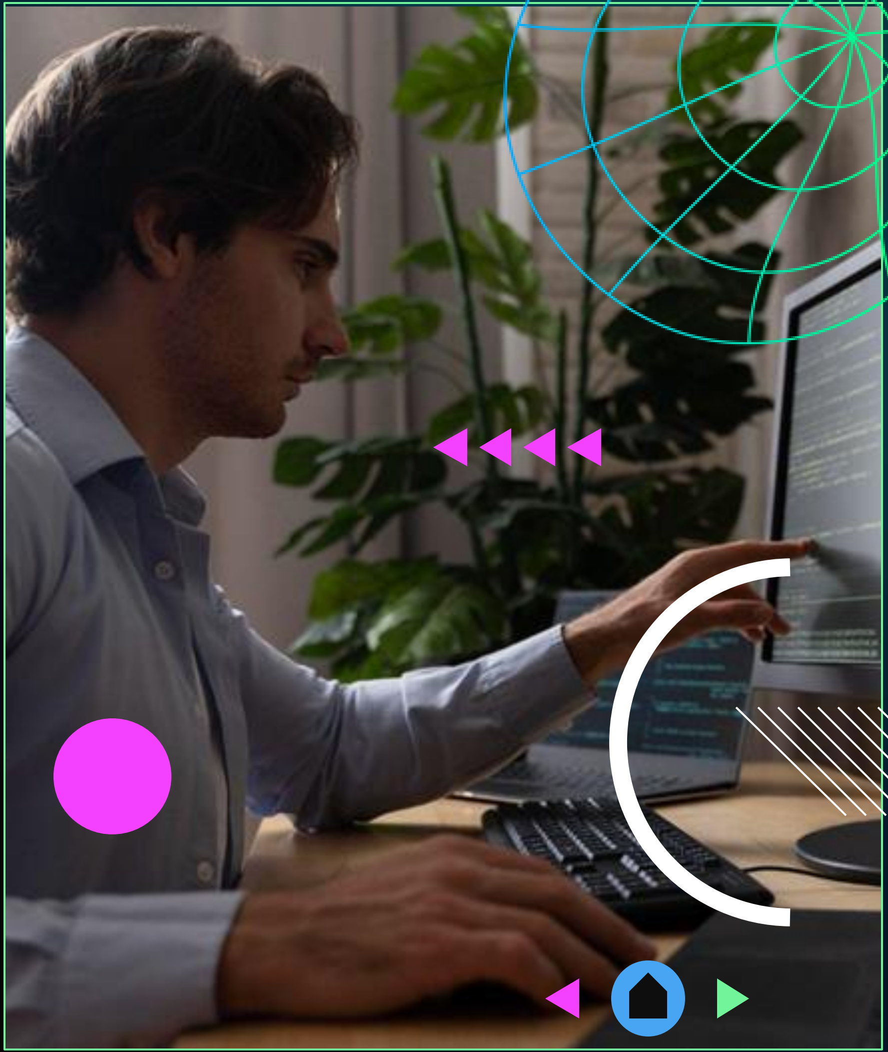
Model Flaws and Strength

MACHINE ANALYSIS 10

In terms of simplicity, the Decision Tree model stands out because it provides strong predictive results while being the easiest and quickest to train due to its fewer parameters. On the contrary, models like K Nearest Neighbors (KNN) achieve optimal results with a K value of 3, but they are slower in the prediction phase because they need to calculate distances between all data points to classify each one. The Random Forest model also performs well, but its training process takes longer, mainly due to the grid search technique used to find the best parameters. This tradeoff implies that with larger datasets, these models may offer improved performance, but the training time will be longer. Ultimately, the model choice depends on your project's specific requirements, taking into account factors like dataset size, training time, and the desired level of predictive accuracy.

THANKS

WISDOM IZUCHUKWU ADIKE
7/09/2023



CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon** and infographics & images by **Freepik**