AN INITIATIVE OF:

science, technology & innovation
Department:
Science, Technology and Innovation
REPUBLIC OF SOUTH AFRICA

CSIR 80th anniversary
Touching lives through innovation

NICIS
NATIONAL INTEGRATED CYBERINFRASTRUCTURE SYSTEM
DIRISA

Train Test Split

DIRISA Datathon | Instructor: Kgaugelo Mmakola

Train-Test Split Exploring

Student	Study Hours	Passed (Yes=1 / No=0)
A	2	0
B	4	0
C	5	1
D	7	1
E	1	0
F	8	1
G	3	0
H	6	1
I	9	1
J	2	0

Train-Test Split Exploring

- Python Function:
  - train_test_split(...)
- What are these 4 variables?

Variable	Meaning
X_train	The **study hours** used for training
X_test	The **study hours** used for testing
y_train	The **pass/fail labels** for training
y_test	The **pass/fail labels** for testing

Train-Test Split Exploring

- **Test_size** – That means **20% of your data** is used for testing.
- **Random_state** – This is like **locking the shuffle**.

Without it, the computer would randomly choose different training/testing sets **every time you run the code**.

Train-Test Split Exploring

- X = [2, 4, 5, 7, 1, 8, 3, 6, 9, 2] # Study Hours
- y = [0, 0, 1, 1, 0, 1, 0, 1, 1, 0] # Passed or not

- X_train = X[:8] # First 8 students → training
- X_test = X[8:] # Last 2 students → testing

- y_train = y[:8] # Training labels
- y_test = y[8:] # Testing labels

Final Summary

- You **split your data** to teach your robot and test it properly.
- **train_test_split()** is the function that helps with that.
- **test_size**=0.2 means 20% for testing, 80% for learning.
- **random_state**=42 keeps the split the same every time.
- **X_train, y_train** = learning materials
- **X_test, y_test** = exam paper

AN INITIATIVE OF:

science, technology & innovation
Department:
Science, Technology and Innovation
REPUBLIC OF SOUTH AFRICA

CSIR 80th anniversary
Touching lives through innovation

nicis
NATIONAL INTEGRATED CYBERINFRASTRUCTURE SYSTEM
DIRISA

Exploring Different random state values

DIRISA Datathon | Instructor: Kgaugelo Mmakola

Different random_state values

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
```

Model	R^2 Score (CV)	MAE (Test)	MSE (Test)
Linear Regression	0.71	3.23	21.45
Ridge Regression	0.71	3.22	21.40
Lasso Regression	0.67	3.52	24.17
Decision Tree	0.79	2.67	14.32
Random Forest	0.86	2.15	10.89
Gradient Boosting	0.87	2.10	10.02

Different random_state values

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

Model	R^2 Score (CV)	MAE (Test)	MSE (Test)
Linear Regression	0.72	3.11	20.34
Ridge Regression	0.72	3.11	20.34
Lasso Regression	0.68	3.39	23.01
Decision Tree	0.82	2.45	12.67
Random Forest	0.88	2.01	9.85
Gradient Boosting	0.89	1.98	9.12

Different random_state values

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=100)
```

Model	R^2 Score (CV)	MAE (Test)	MSE (Test)
Linear Regression	0.73	3.05	19.87
Ridge Regression	0.73	3.05	19.87
Lasso Regression	0.69	3.31	22.45
Decision Tree	0.81	2.52	13.01
Random Forest	0.87	2.08	10.21
Gradient Boosting	0.88	2.03	9.45


Thank you!!
NATIONAL INTEGRATED
CYBERINFRASTRUCTURE SYSTEM
DIRISA