01. train_val_test

October 8, 2024

1 Training, Validation, and Tesing Datasets

[17]:
```
import pandas as pd

from pathlib import Path
from sklearn.model_selection import train_test_split
```

1.1 Import Final Dataset

[19]:
```
# Import final dataset
datasets = Path("../datasets")
df = pd.read_csv(datasets / "school_final_dataset.csv")
df.head()
```

[19]:
```
   Undergrad_Degree Work_Experience  Employability_Before      Status  \
0          Business              No                 252.0      Placed
1          Business              No                 423.0  Not Placed
2  Computer Science             Yes                 101.0      Placed
3       Engineering              No                 288.0  Not Placed
4           Finance              No                 248.0  Not Placed

   Status_enc
0           1
1           0
2           1
3           0
4           0
```

1.2 Split Dataset

[21]:
```
# Split dataset for 20% test_data
train_val_data, test_data = train_test_split(df, test_size=0.2, random_state=42)
# Split remaining dataset into 60% training and 40% validation
train_data, val_data = train_test_split(train_val_data, test_size=0.4,␣
 ↪random_state=42)

print(f"Training dataset: {train_data.shape[0]}")
```

```
print(f"Validation dataset: {val_data.shape[0]}")
print(f"Testing dataset: {test_data.shape[0]}")
```

```
Training dataset: 576
Validation dataset: 384
Testing dataset: 240
```

1.3 Export and Save Datasets

[16]:
```
train_data.to_csv(datasets / "training_data.csv", index=False)
val_data.to_csv(datasets / "validation_data.csv", index=False)
test_data.to_csv(datasets / "testing_data.csv", index=False)
```

[ ]:
```

```