Analysis of Census Dataset

Predicting income using machine learning models

Design (Idea And Question)

Employing several supervised learning algorithms to accurately model individuals' income using the data collected from 1994 U.S. Census.

The This project to explore whether if demographic characteristics have an effect on an individual's annual income, and can we predict whether if the income is `<=50K` or `>50K`?

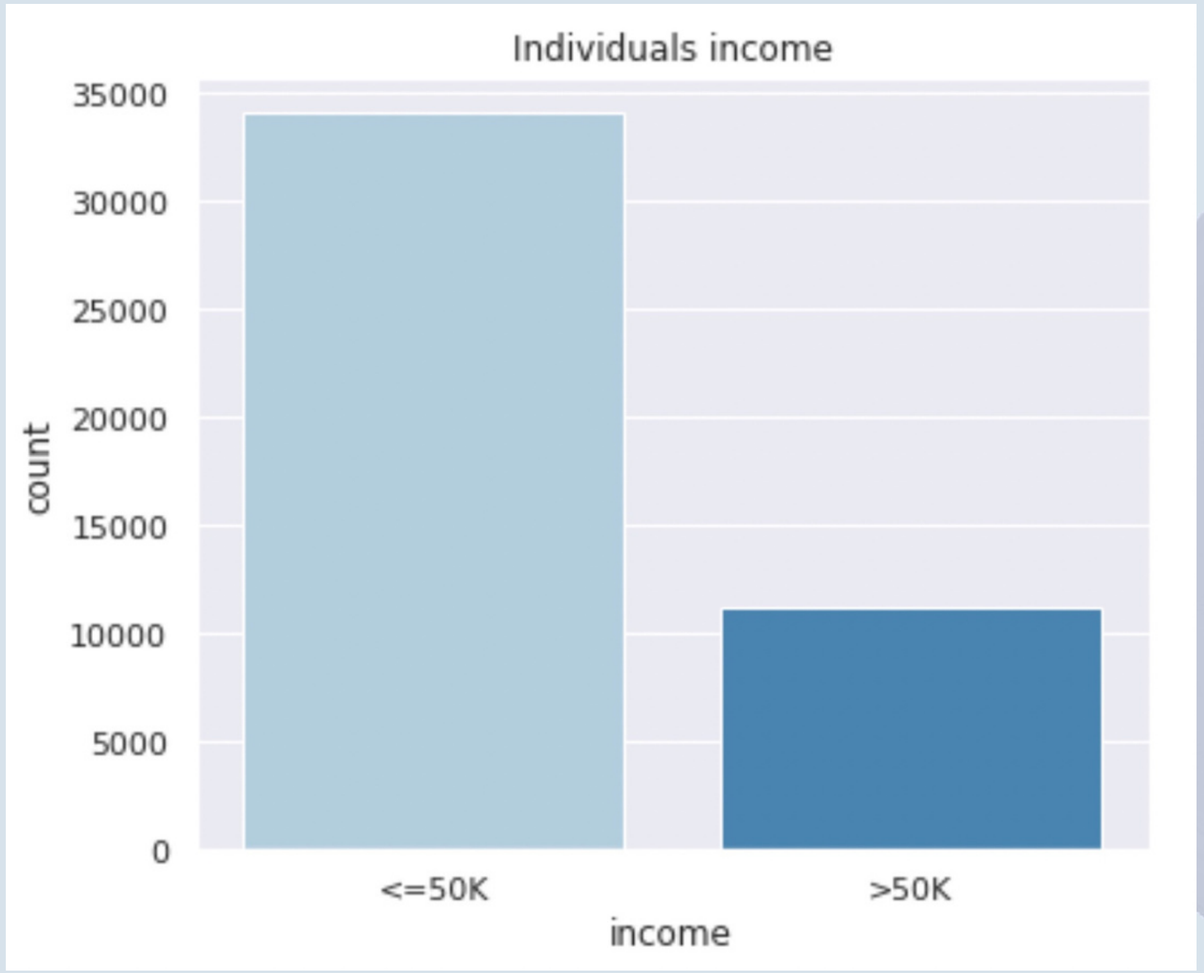
And if so, which models performed better?

Dataset

The dataset contains information about the annual incomes of people from 42 different countries, it contains 45,222 entries with a total of 14 columns representing different attributes of the people.

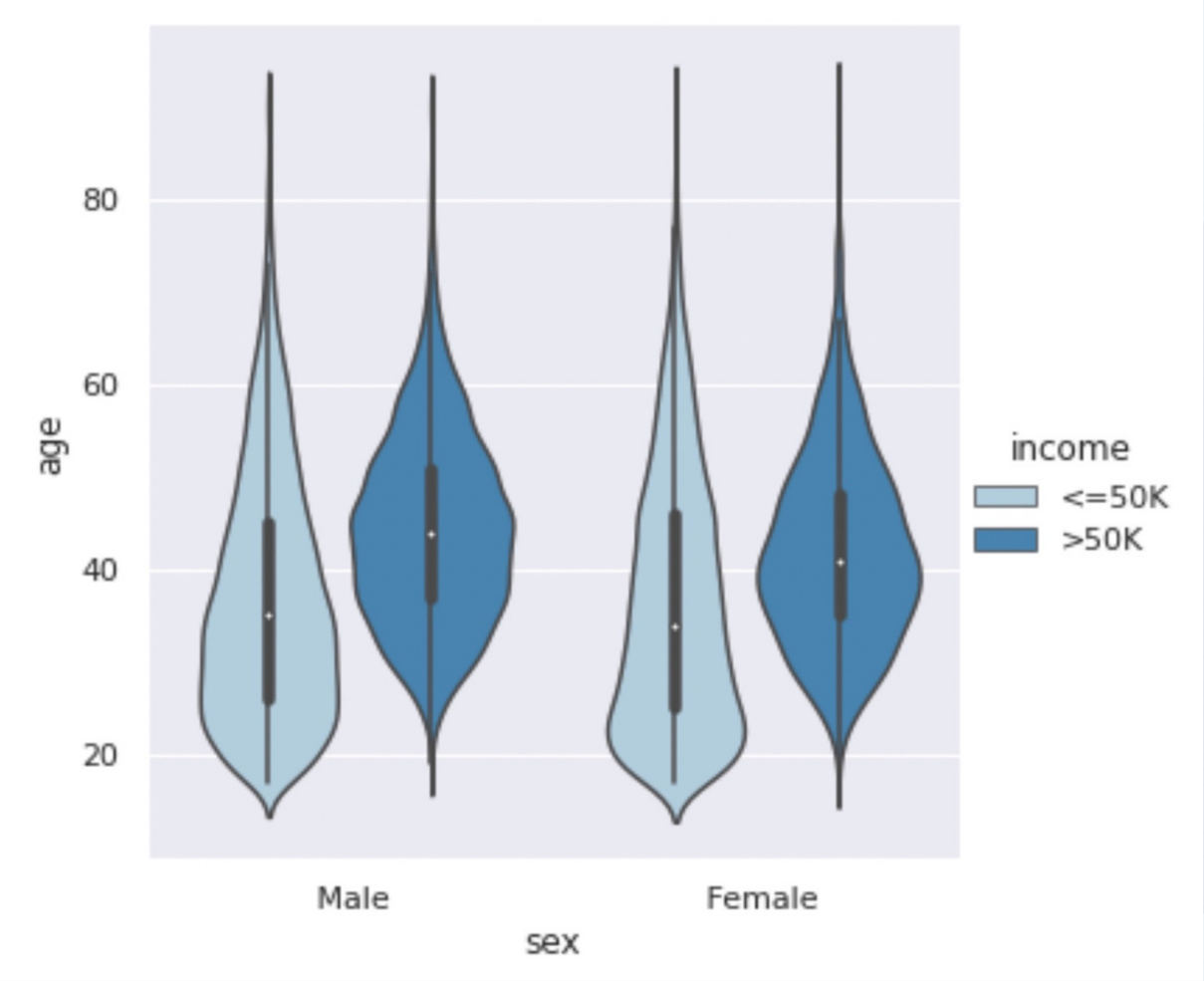
age	workclass	education_level	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	income
39	State-gov	Bachelors	13.0	Never-married	Adm-clerical	Not-in-family	White	Male	2174.0	0.0	40.0	United-States	<=50K
50	Self-emp-not-inc	Bachelors	13.0	Married-civ-spouse	Exec-managerial	Husband	White	Male	0.0	0.0	13.0	United-States	<=50K
38	Private	HS-grad	9.0	Divorced	Handlers-cleaners	Not-in-family	White	Male	0.0	0.0	40.0	United-States	<=50K

Exploring Data



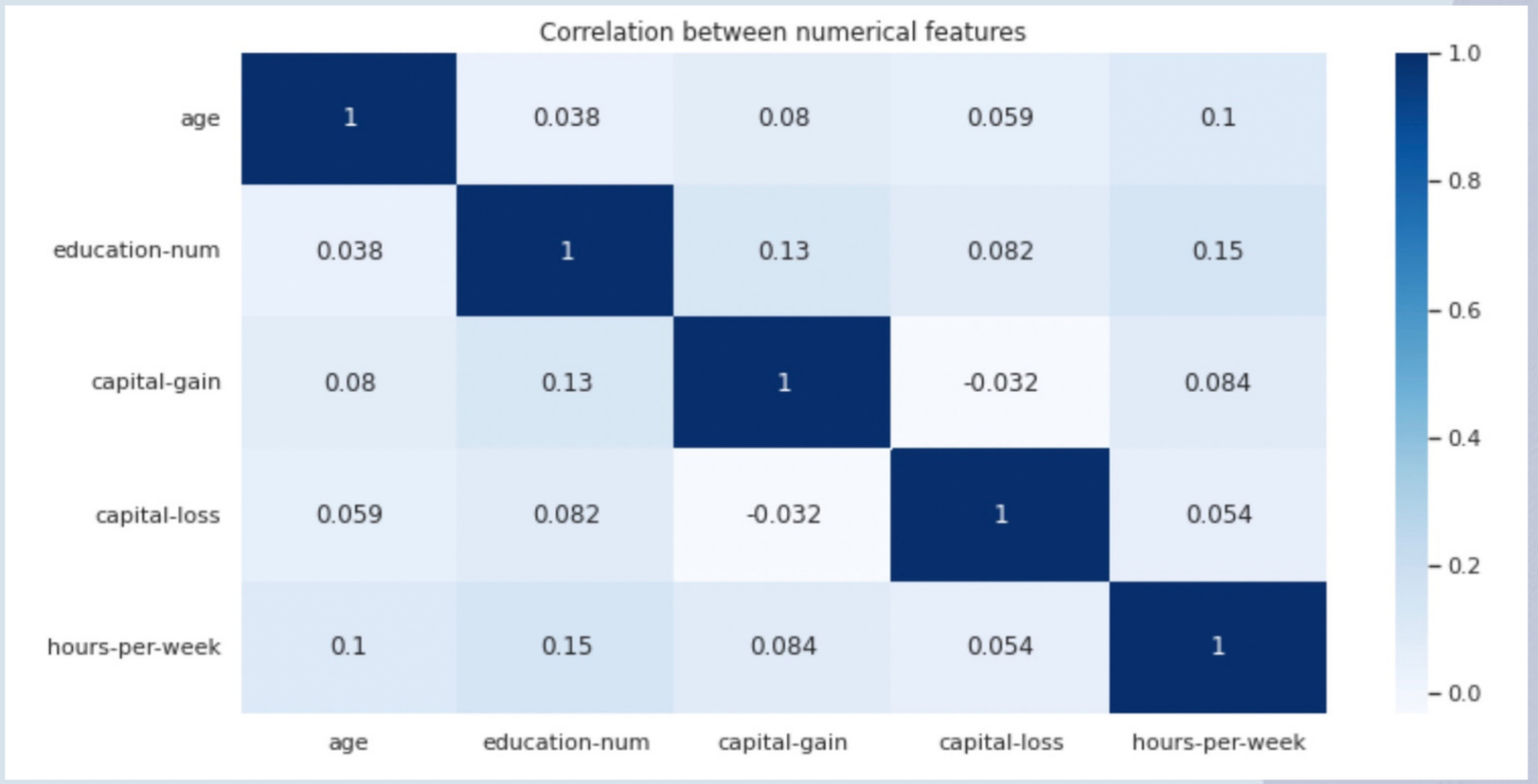
Bar plot to show the number of individuals making more than $50,000 or less (check Imbalance)

Exploring Data



Violin plot to show the distribution of age based on the sex and income

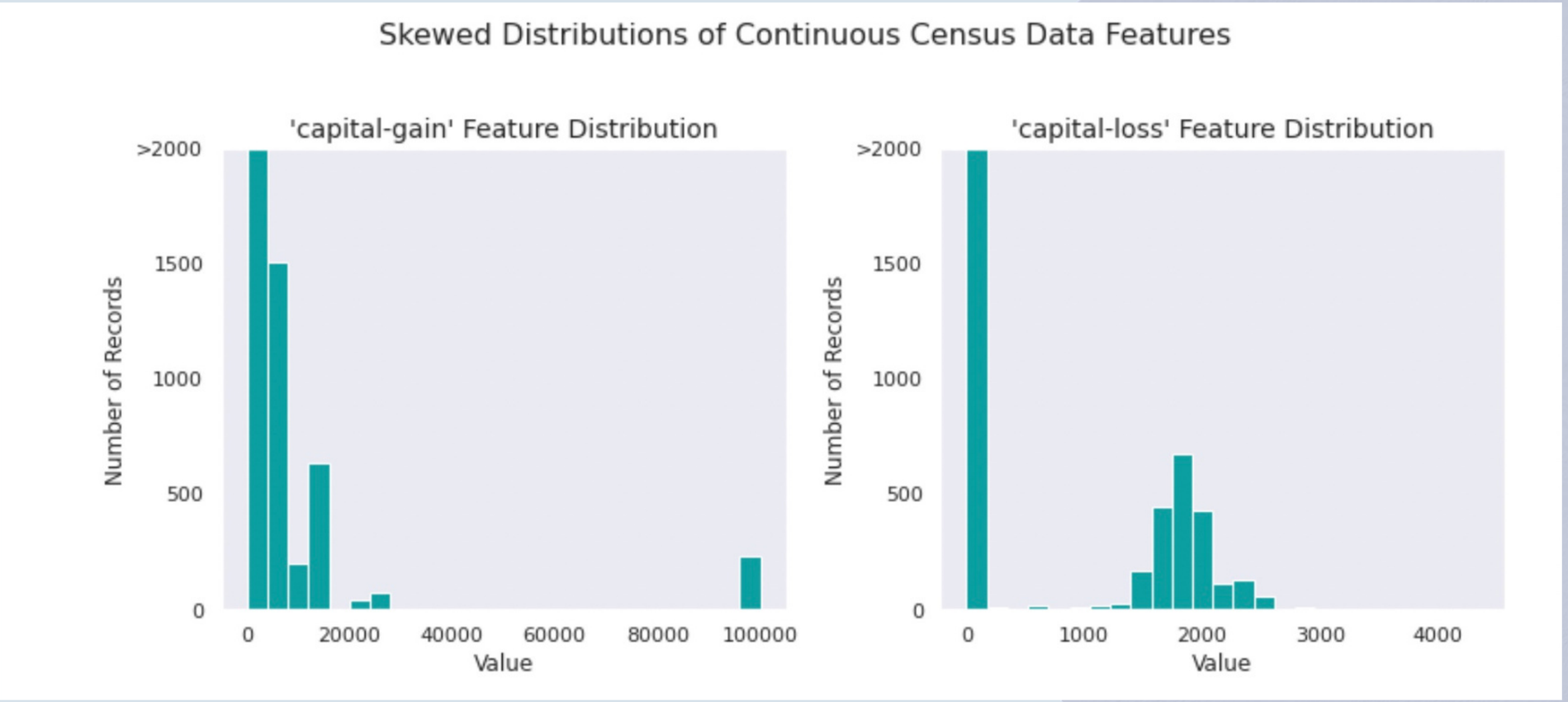
Exploring Data



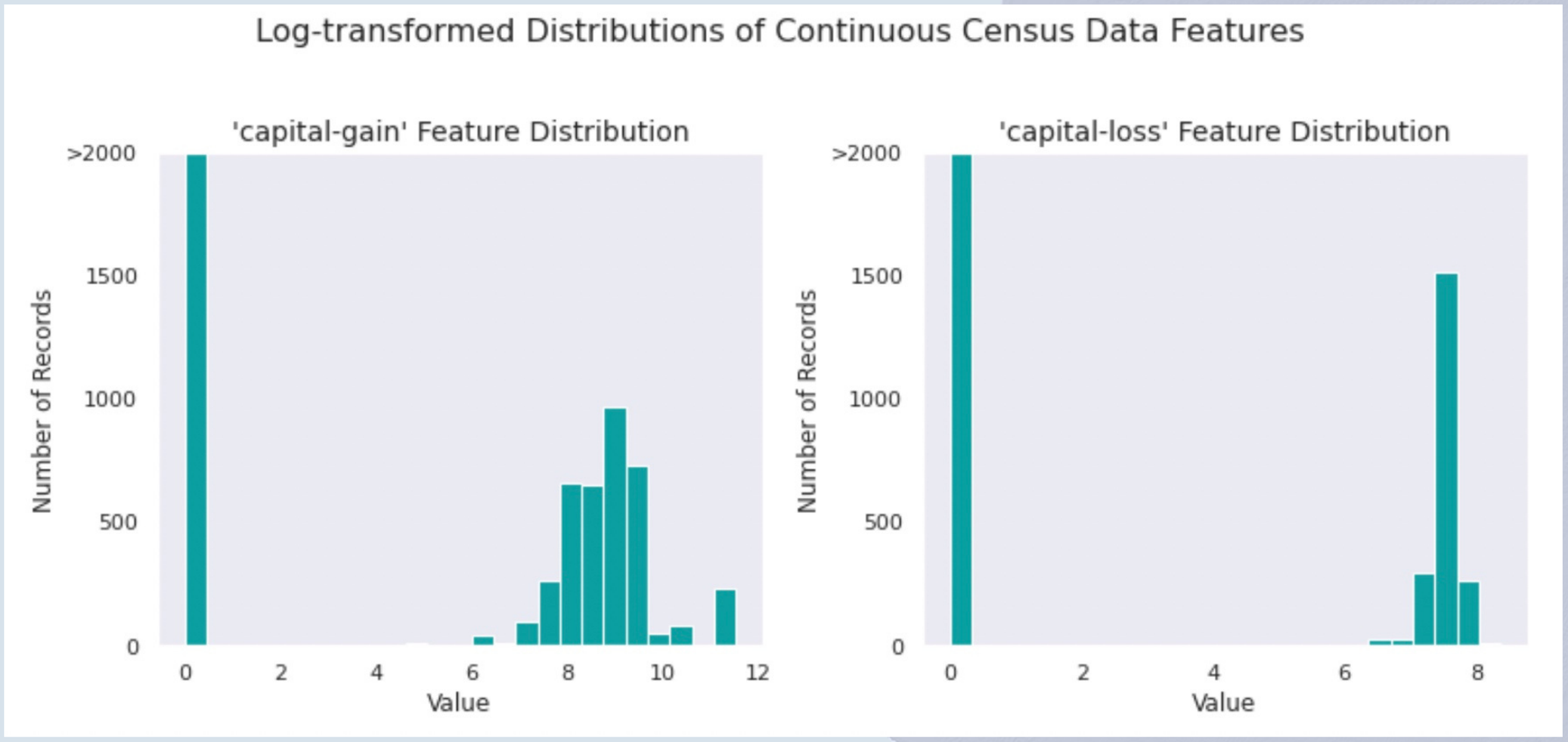
Heatmap of correlation between only the numerical features

Preparing the Data

1.Dealing with Numerical Data



Preparing the Data



Preparing the Data

After scaling

	age	workclass	education_level	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week
0	0.30137	State-gov	Bachelors	0.8	Never-married	Adm-clerical	Not-in-family	White	Male	0.02174	0.0	0.397959

Preparing the Data

2.Dealing with categorical Data

```
103 number of features after one-hot encoding.
```

Output

Modeling

Naive Predict

```
Naive Predictor: [Accuracy score: 0.2478, F-score: 0.2917]
```

Output

Modeling

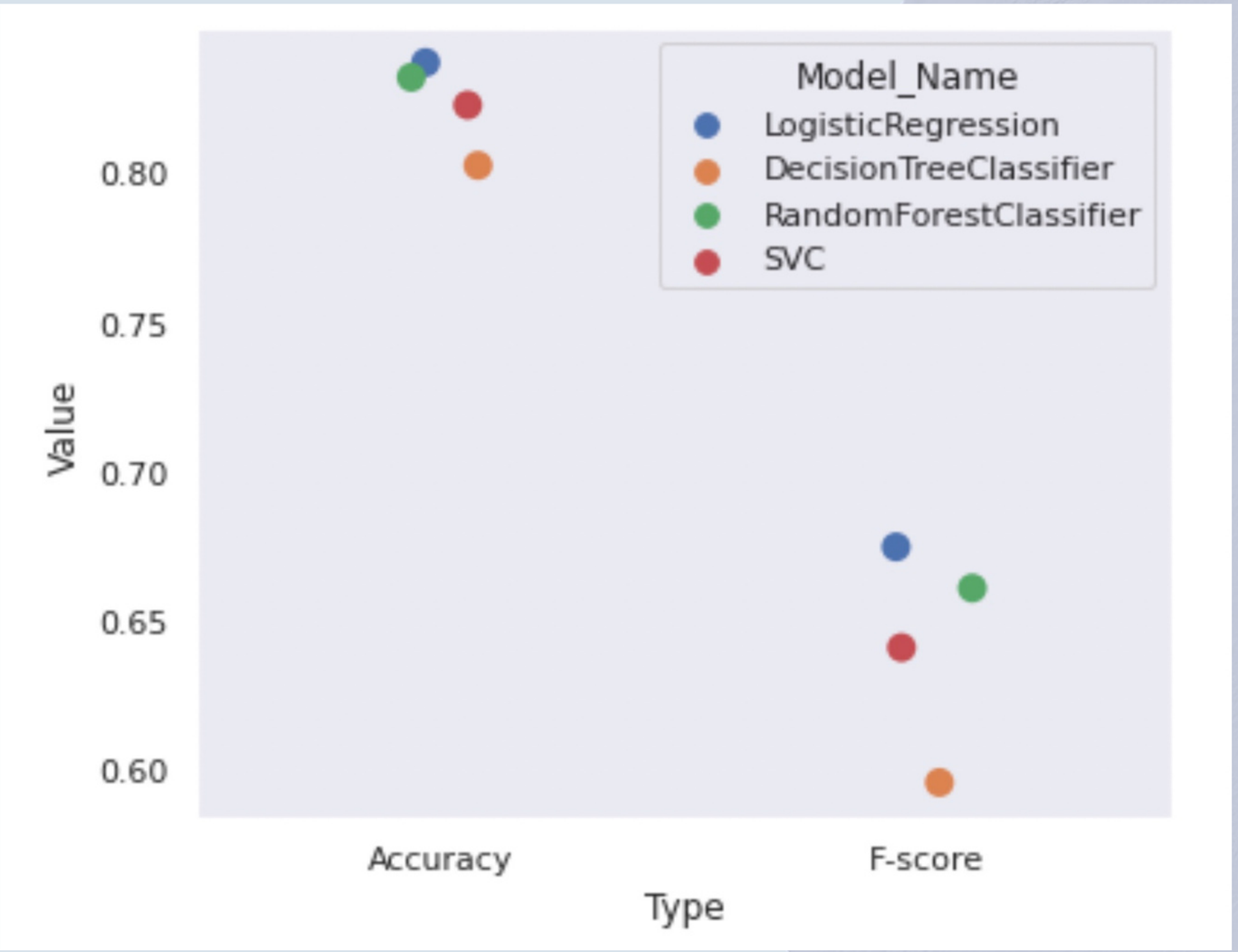
We preformed:

1. Logistic regression
2. Decision Tree Classifier
3. Random Forest Classifier
4. Support Vector Classifier (SVC)

Modeling

	Type	Value	Model_Name
0	Accuracy	0.836926	LogisticRegression
1	F-score	0.674581	LogisticRegression
2	Accuracy	0.802432	DecisionTreeClassifier
3	F-score	0.595626	DecisionTreeClassifier
4	Accuracy	0.831951	RandomForestClassifier
5	F-score	0.660781	RandomForestClassifier
6	Accuracy	0.822664	SVC
7	F-score	0.640846	SVC

Comparison


Model_Name
LogisticRegression
DecisionTreeClassifier
RandomForestClassifier
SVC
0.80
0.75
0.70
0.65
0.60
Value
Accuracy
F-score
Type