Titanic

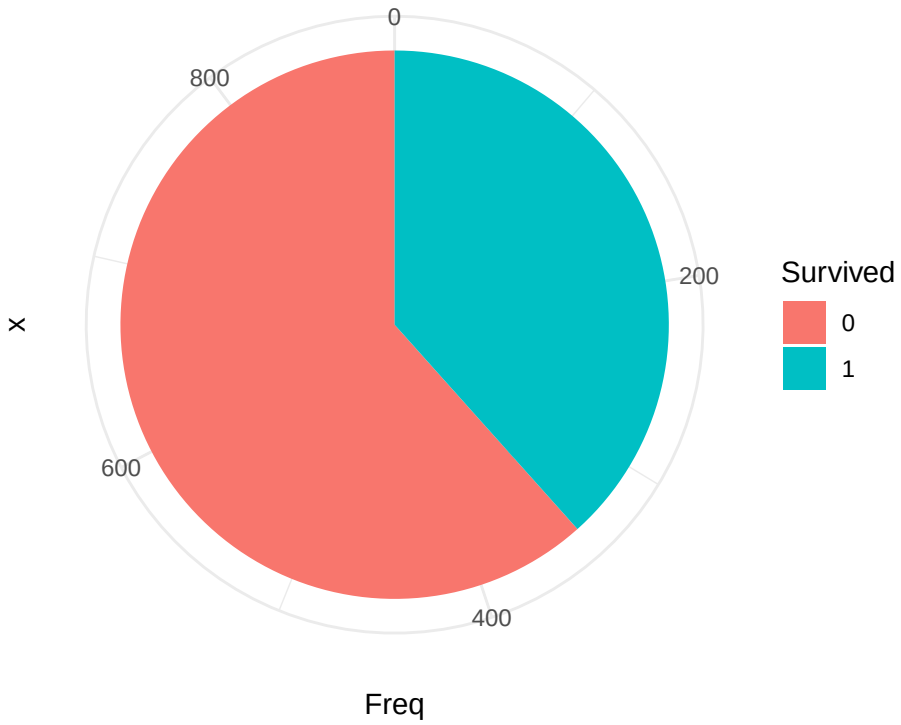
Mateusz Bernart, Patryk Janiak, Patrick Molina, Marek Seget, Chihab Zitouni

2024-04-20

Problem introduction

Introduction

- What is the **Titanic dataset** about?
- The goal: predicting survival on Titanic passengers using classifiers and feature selection



Dataset

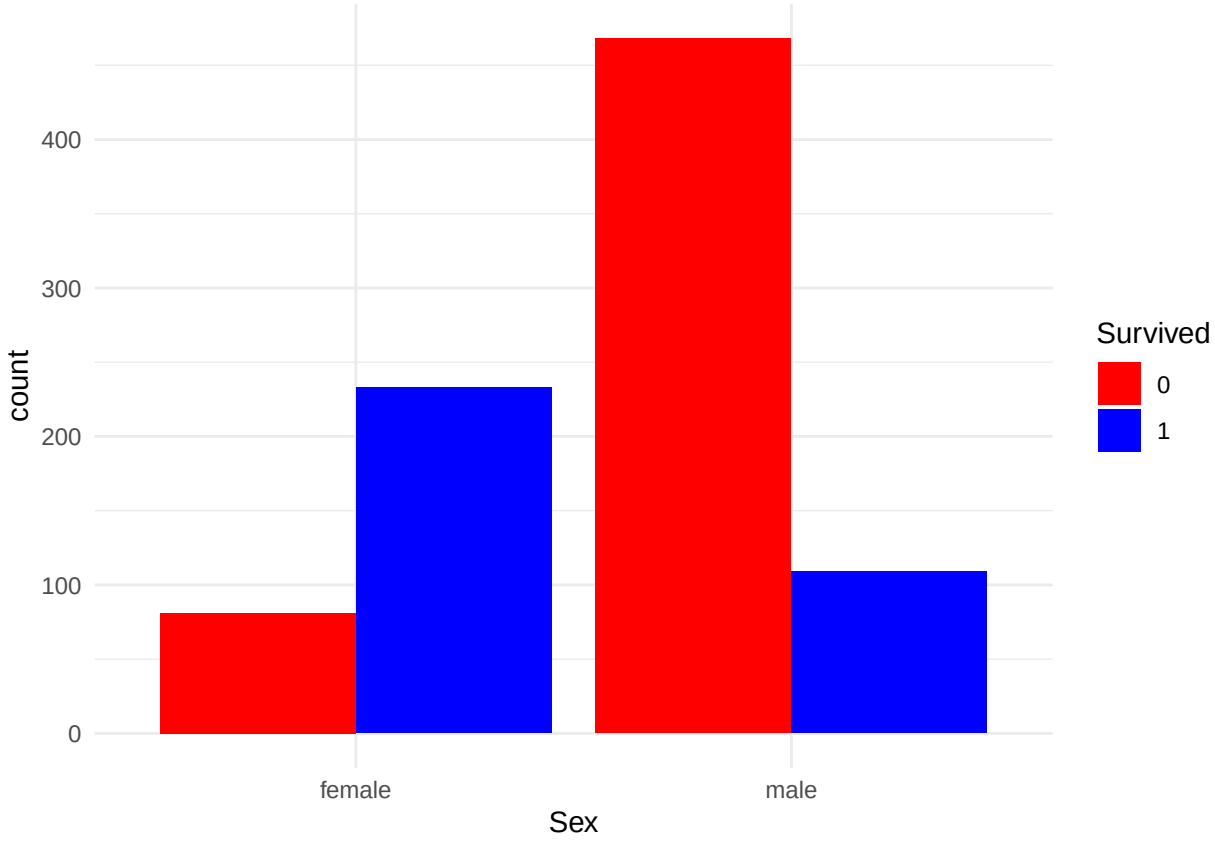
- Dataset structure

PassengerId	Pclass	Sex	Age	SibSp	Parch	Cabin	Fare	Survived
1	3	male	22	1	0		7.2500	0
2	1	female	38	1	0	C85	71.2833	1
3	3	female	26	0	0		7.9250	1
4	1	female	35	1	0	C123	53.1000	1
5	3	male	35	0	0		8.0500	0

*Omited **Name**, **Ticket** and **Embarked** features for visual reason*

Target

- Our target - *Survived?* (binary 0/1)

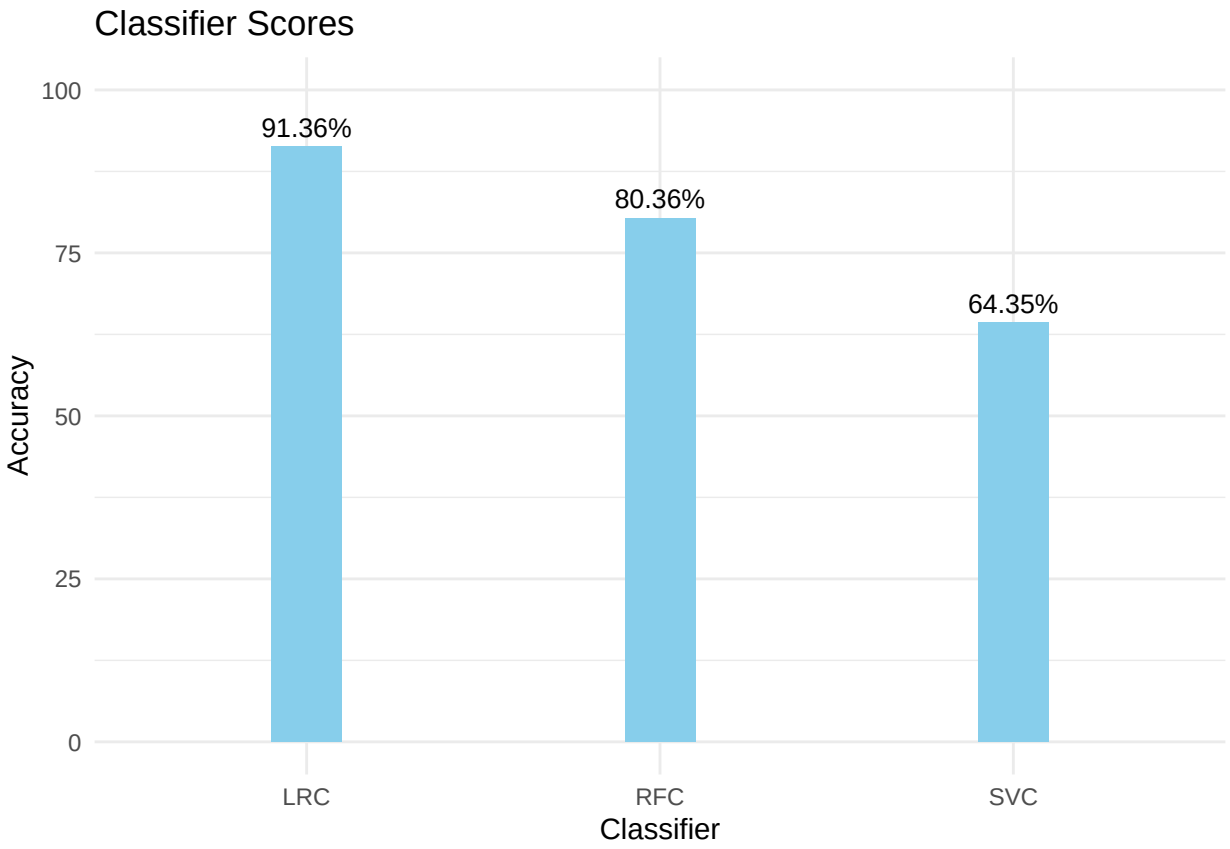


Solving the problem

Data preprocessing

- Missing values:
  - Dropped irrelevant columns
- Encoding - Sex + Embarked (categorical -> numeric)

Comparison of classifiers before preprocessing



Preprocessing techniques

- Normalization: `MinMaxScaler`
- Standardization: `StandardScaler`

Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
1	-1.482983	1.322511	0.577094	0.522511	-0.506787	0.694046	-2.049487
1	-1.482983	1.322511	0.577094	0.522511	-0.506787	0.694046	-2.049487
1	0.908600	1.322511	-0.251478	0.552714	-0.506787	-0.503620	0.519588
1	-1.482983	1.322511	0.369951	0.522511	-0.506787	0.350326	0.519588
0	0.908600	-0.756138	0.369951	-0.552714	-0.506787	-0.501257	0.519588

Feature Selection

- RFE (Recursive Feature Elimination) with LR, RFC
- Selected features:
  - for LR: **Pclass**, **Sex**, **Age**, **SibSp**
  - for RFC: **Pclass**, **Sex**, **Age**, **Fare**

Feature extraction

- PCA (Principal Component Analysis)

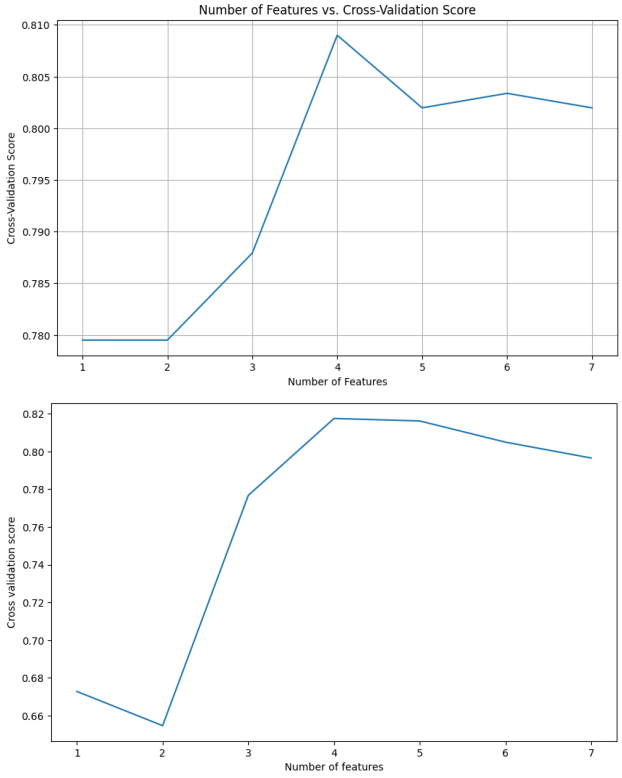


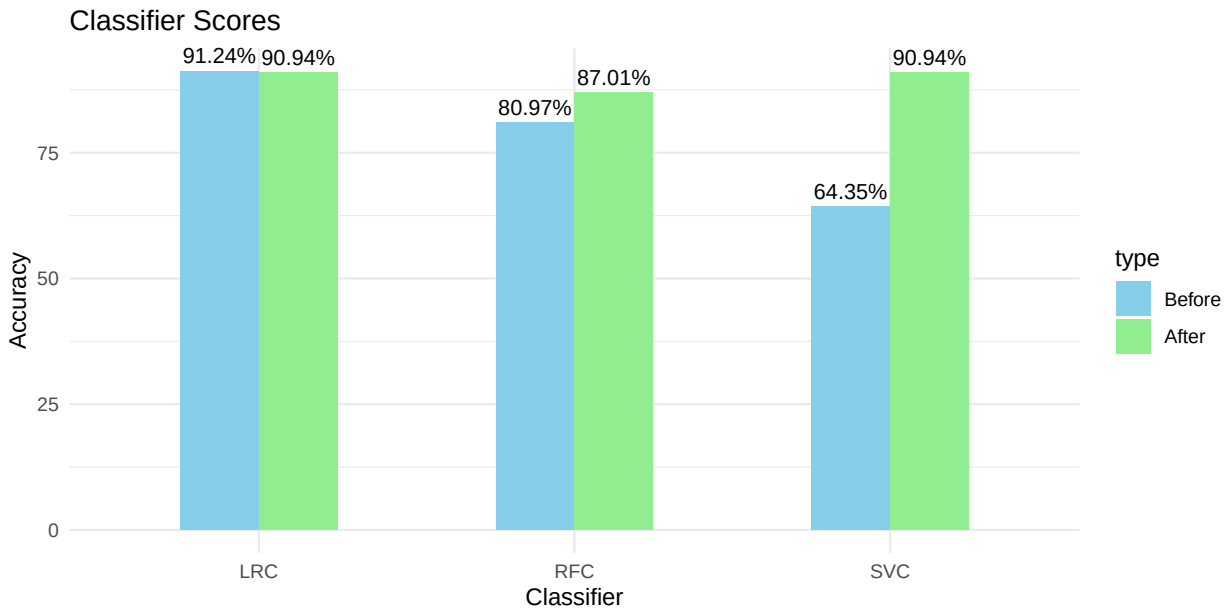
Figure 1:

```
[[ 1.03979617, -0.8338755 ,  0.39574407, -0.124945  ],
    [-0.8159031 ,  1.9567542 ,  0.23809646, -0.0552459 ],
    [ 0.48443562,  0.35598023, -1.52130571, -0.51725331],
    ...,
    [-0.44638046,  1.59476638, -0.76656293,  1.20977618],
    [-0.94330883, -0.02518358,  0.51553682,  1.40835751],
    [ 0.03490097, -1.26169953, -0.25152961, -0.27004702]]
```

- Selected: **Pclass**, **Sex**, **Age**, **SibSp**

Classifiers comparison after selecting and extracting features

- SVC, RFC, LR after actions



Conclusion

- Data preprocessing significantly improved classifier accuracy.
- Feature selection techniques identified key predictors: Pclass, Sex, Age, and SibSp.
- Feature extraction via PCA simplified models while retaining essential information.
- Post-preprocessing, classifiers (SVC, RFC, LR) demonstrated improved performance.