)

1. Upload CSV File

Choose a file

Drag and drop file here
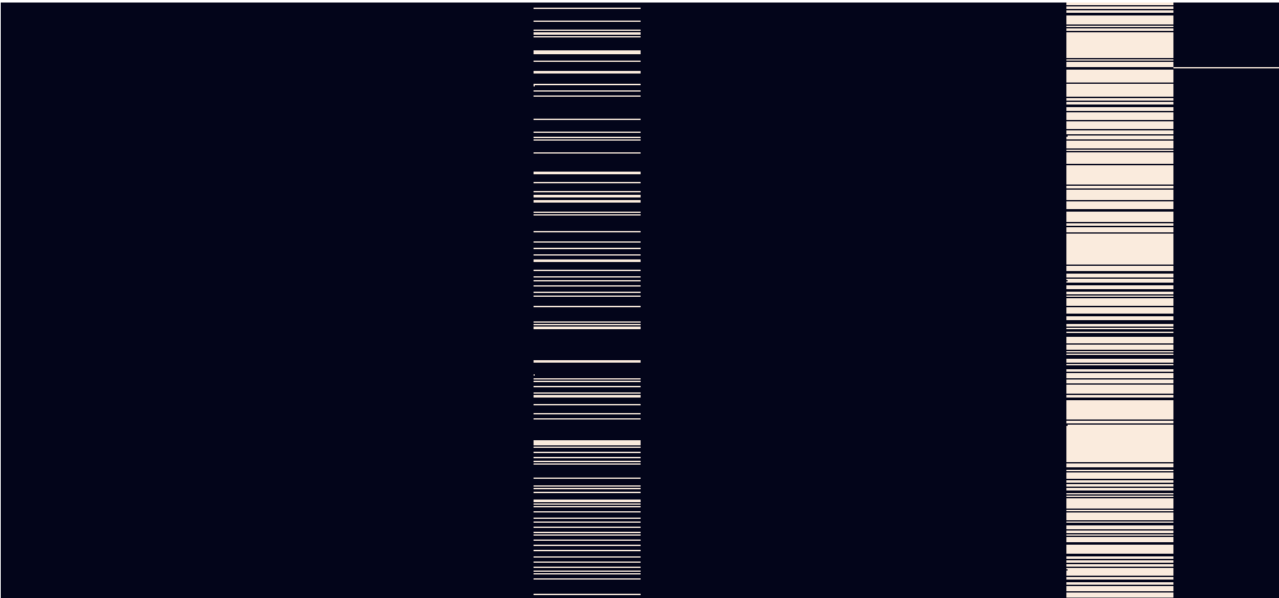Limit 200MB per file
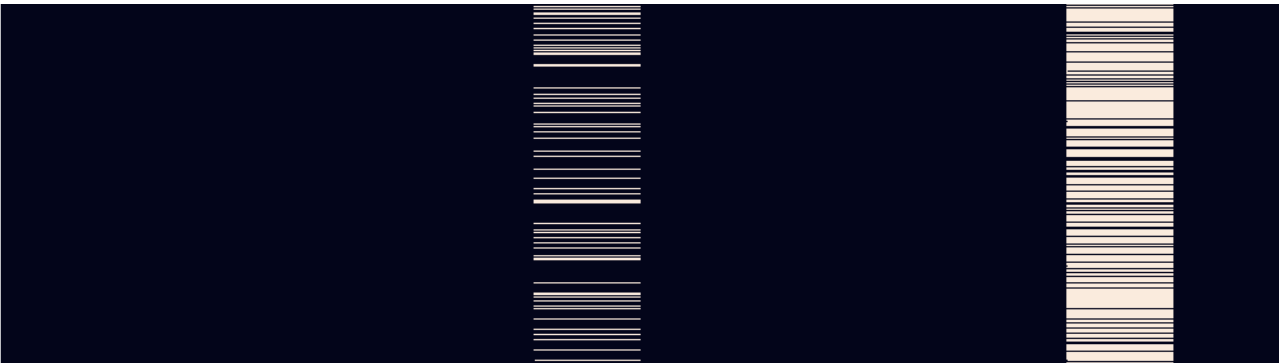
Browse files

📄 Classification-Titanic-train.csv   ✕
59.8KB

	PassengerId	Survived	Pclass	Name	Sex	Age	S
0	1	0	3	Braund, Mr. Owen Harris	male	22	
1	2	1	1	Cumings, Mrs. John Bra…	female	38	
2	3	1	3	Heikkinen, Miss. Laina	female	26	
3	4	1	1	Futrelle, Mrs. Jacques…	female	35	
4	5	0	3	Allen, Mr. William Hen…	male	35	
5	6	0	3	Moran, Mr. James	male	NaN	
6	7	0	1	McCarthy, Mr. Timothy J	male	54	
7	8	0	3	Palsson, Master. Gosta…	male	2	
8	9	1	3	Johnson, Mrs. Oscar W …	female	27	
9	10	1	2	Nasser, Mrs. Nicholas …	female	14	
10	11	1	3	Sandstrom, Miss. Margu…	female	4	

Check the missing values

The missing Proportion for each column

PassengerId: 0.000000

Survived: 0.000000

Pclass: 0.000000

Name: 0.000000

Sex: 0.000000

Age: 0.198653

SibSp: 0.000000

Parch: 0.000000

Ticket: 0.000000

Fare: 0.000000

Cabin: 0.771044

Embarked: 0.002245

Cabin is dropped because its missing proportion is > 50%

2. Data Pre-processing

Which data pre-processing method do you want to apply?

- ☐ Convert Data Type

- ☐ Extract Numbers From A String

- ☐ Extract Letters From A String

- ☐ String Clean-up

- ☒ Drop Columns

Enter the column name for drop column, e.g. Ticket PassengerId

```
Ticket PassengerId
```

	Survived	Pclass	Name	Sex	Age	SibSp	Parch	F;
0	0	3	Braund, Mr. Owen Harris	male	22	1	0	7.2
1	1	1	Cumings, Mrs. John Bra…	female	38	1	0	71.2;
2	1	3	Heikkinen, Miss. Laina	female	26	0	0	7.9;
3	1	1	Futrelle, Mrs. Jacques…	female	35	1	0	53.1(
4	0	3	Allen, Mr. William Hen…	male	35	0	0	8.0

3. Feature transformation

Which feature transformation method do you want to apply?

- ☐ fourier transformation

- ☒ get dummies

Enter the column name for get dummy

```
Sex Embarked
```

	Survived	Pclass	Name	Age	SibSp	Parch	Fare	Sex
0	0	3	Braund, Mr. Owen Harris	22	1	0	7.2500	
1	1	1	Cumings, Mrs. John Bra…	38	1	0	71.2833	
2	1	3	Heikkinen, Miss. Laina	26	0	0	7.9250	
3	1	1	Futrelle, Mrs. Jacques…	35	1	0	53.1000	
4	0	3	Allen, Mr. William Hen…	35	0	0	8.0500	

- ☐ label binarizer

- ☐ ordinal encoder

- ☐ log transformation

☐ A bag of word

Enter the column name for a bag of word

Name

	Survived	Pclass	Name	Age	SibSp	Parch	Fare	Sex
0	0	3	braund mr owen harris	22	1	0	7.2500	
1	1	1	cumings mr john bradle…	38	1	0	71.2833	
2	1	3	heikkinen miss laina	26	0	0	7.9250	
3	1	1	futrelle mr jacques he…	35	1	0	53.1000	
4	0	3	allen mr william henry	35	0	0	8.0500	

☐ Word2Vec

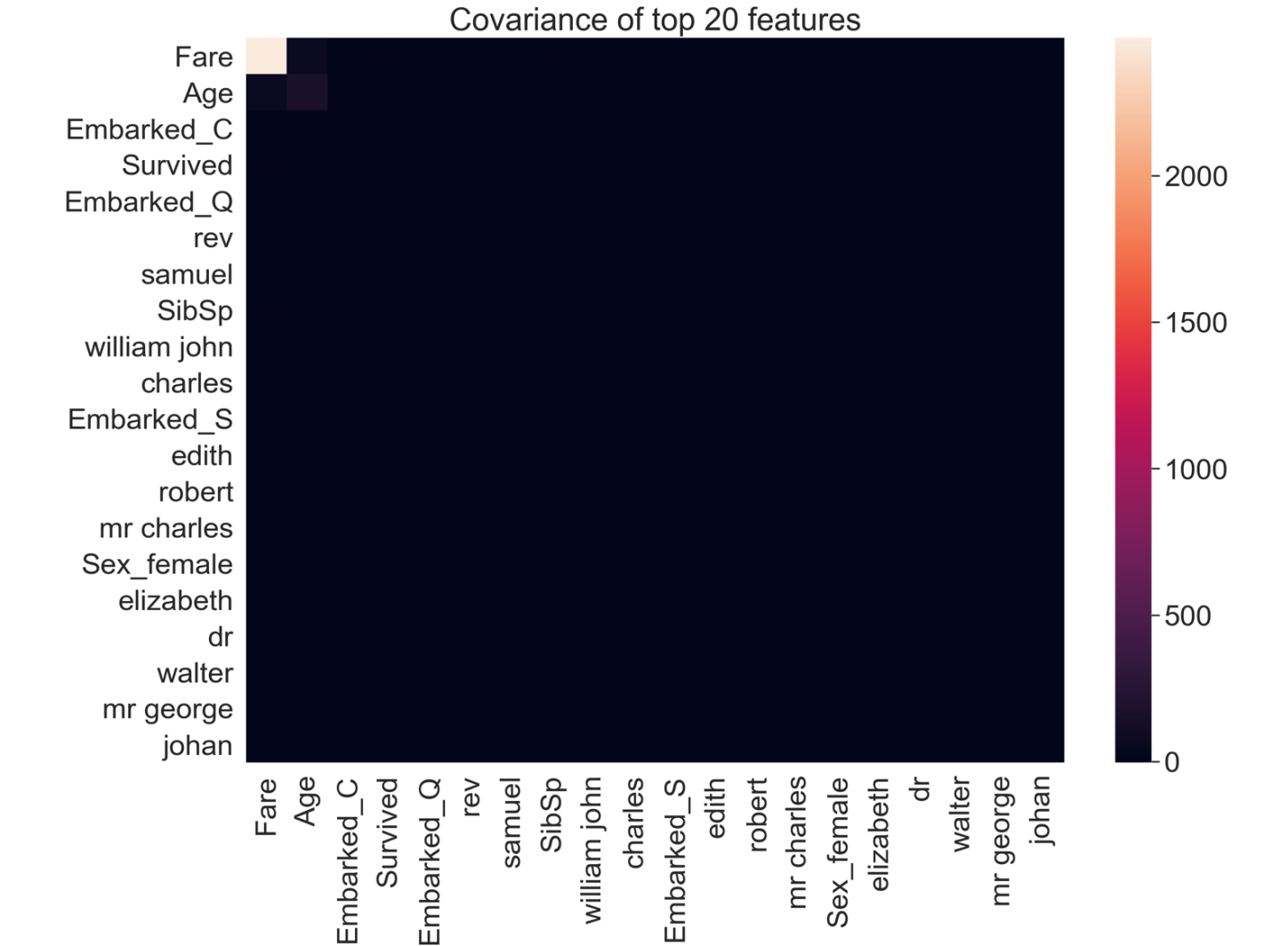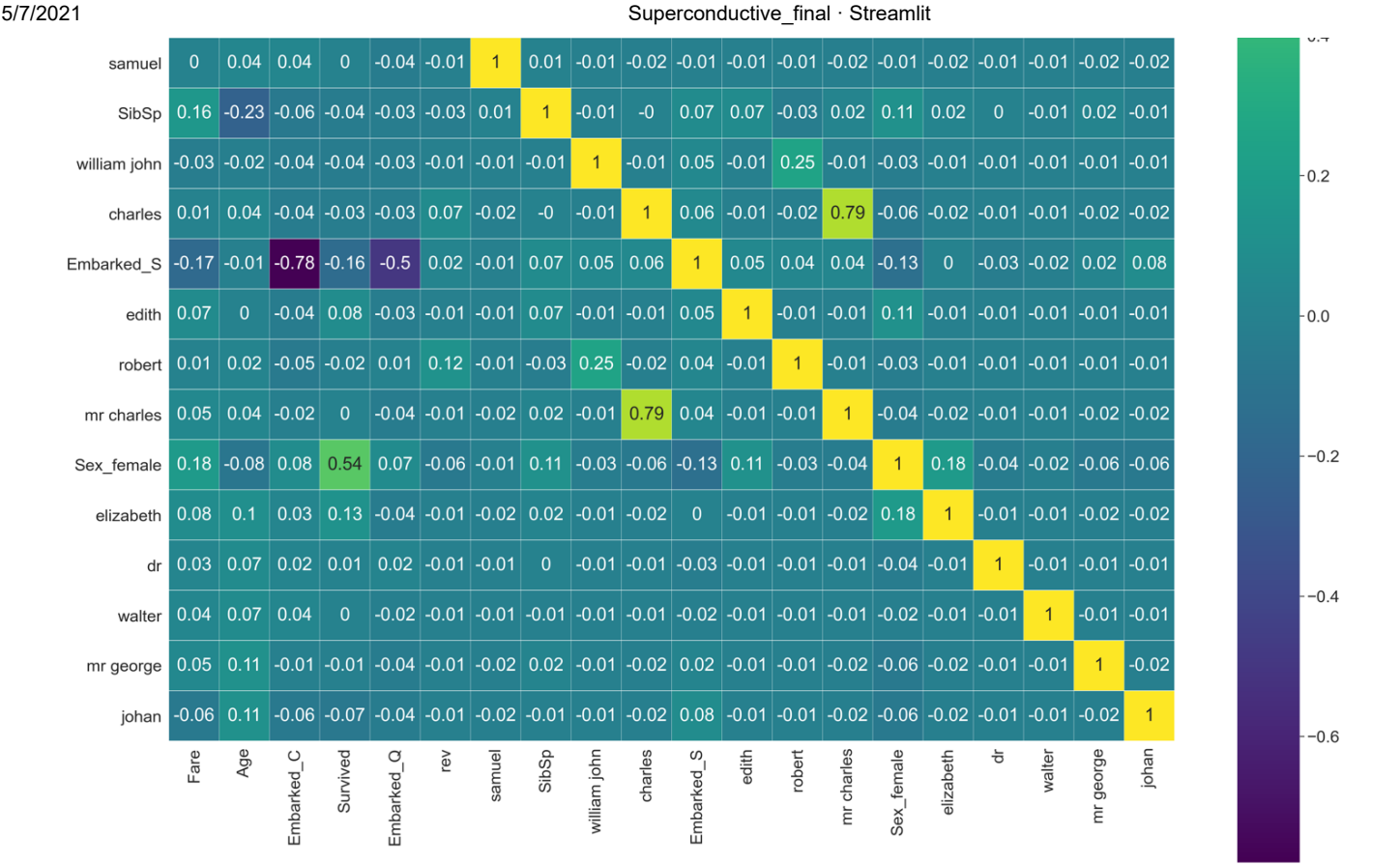# 4. Feature Engineering and selection

Which feature engineering and selection do you want to apply?

☐ Choosing top features and visualize

Here is your top 20 feature scores

	Specs
26	charles
9	Embarked_S
31	edith
98	robert
74	mr charles
5	Sex_female
34	elizabeth
30	dr
106	walter
77	mr george
50	johan

Pearson Correlation of top 20 Features

	Fare	Age																		
Fare	1	0.1	0.27	0.26	-0.12	-0.02	0	0.16	-0.03	0.01	-0.17	0.07	0.01	0.05	0.18	0.08	0.03	0.04	0.05	-0.06
Age	0.1	1	0.03	-0.06	-0.03	0.09	0.04	-0.23	-0.02	0.04	-0.01	0	0.02	0.04	-0.08	0.1	0.07	0.07	0.11	0.11
Embarked_C	0.27	0.03	1	0.17	-0.15	-0.04	0.04	-0.06	-0.04	-0.04	-0.78	-0.04	-0.05	-0.02	0.08	0.03	0.02	0.04	-0.01	-0.06
Survived	0.26	-0.06	0.17	1	0	-0.06	0	-0.04	-0.04	-0.03	-0.16	0.08	-0.02	0	0.54	0.13	0.01	0	-0.01	-0.07
Embarked_Q	-0.12	-0.03	-0.15	0	1	0.02	-0.04	-0.03	-0.03	-0.03	-0.5	-0.03	0.01	-0.04	0.07	-0.04	0.02	-0.02	-0.04	-0.04
rev	-0.02	0.09	-0.04	-0.06	0.02	1	-0.01	-0.03	-0.01	0.07	0.02	-0.01	0.12	-0.01	-0.06	-0.01	-0.01	-0.01	-0.01	-0.01

Covariance of top 20 features

☐ Plot Distribution

☐ Remove Outliers

☐ Top Interaction

☐ Drop Column

Enter a column name for drop column, e.g. Name

```
Name
```

	Survived	Pclass	Age	SibSp	Parch	Fare	Sex_female	Sex_male	Emba
0	0	3	22	1	0	7.2500	0	1	
1	1	1	38	1	0	71.2833	1	0	
2	1	3	26	0	0	7.9250	1	0	
3	1	1	35	1	0	53.1000	1	0	
4	0	3	35	0	0	8.0500	0	1	

5. Train Test Split

☐ train-test split

Please enter the column name that you want as the output(label) eg. Survived

```
Survived
```

(712, 110) (712,)

(179, 110) (179,)

	Pclass	Age	SibSp	Parch	Fare	Sex_female	Sex_male	Embarked_C
140	3	28	0	2	15.2458	1	0	1
439	2	31	0	0	10.5000	0	1	0
817	2	31	1	1	37.0042	0	1	1
378	3	20	0	0	4.0125	0	1	1
491	3	21	0	0	7.2500	0	1	0

	Survived
140	0
439	0
817	0
378	0
491	0

	Pclass	Age	SibSp	Parch	Fare	Sex_female	Sex_male	Embarked_C
495	3	28	0	0	14.4583	0	1	1
648	3	28	0	0	7.5500	0	1	0
278	3	7	4	1	29.1250	0	1	0

31	1	28	1	0	146.5208	1	0	1
255	3	29	0	2	15.2458	1	0	1

	Survived
495	0
648	0
278	0
31	1
255	1

6. Feature Scaling

☐ Feature scaling - Normalization

☐ Feature scaling - Standard Scaler

7. Models & Evaluations

Which models do you want to apply? Select Classification or Regression

Select Regression or Classification

Classification   ▼

Selected Option: 'Classification'

Logistic Regression Running...

```
▼ {
    "C" : 1
    "class_weight" : NULL
    "dual" : false
    "fit_intercept" : true
    "intercept_scaling" : 1
    "l1_ratio" : NULL
    "max_iter" : 10000
    "multi_class" : "auto"
    "n_jobs" : NULL
    "penalty" : "l2"
    "random_state" : NULL
    "solver" : "liblinear"
    "tol" : 0.0001
    "verbose" : 0
    "warm_start" : false
```

}

Logistic Regression accuracy: 81.01 %

Logistic Regression precision score: 75.36 %

Logistic Regression recall score: 75.36 %

Logistic Regression F1 score: 75.36 %

CONFUSION MATRIX:

Predicted

neg pos

Actual

neg

	0
0	93
1	17

pos

	0
0	17
1	52

SVM Running...

Support Vector Machines labeling accuracy: 81.01 %

Support Vector Machines labeling precision score: 76.12 %

Support Vector Machines labeling recall score: 73.91 %

Support Vector Machines labeling F1 score: 75.0 %

CONFUSION MATRIX:

Predicted

neg pos

Actual

neg

	0

	0
0	93
1	17

pos

	0
0	17
1	52

KNN classifier Running...

K-Nearest Neighbors labeling accuracy: 70.95 %

K-Nearest Neighbors labeling precision score: 67.35 %

K-Nearest Neighbors labeling recall score: 47.83 %

K-Nearest Neighbors labeling F1 score: 55.93 %

CONFUSION MATRIX:

Predicted

neg pos

Actual

neg

	0
0	94
1	16

pos

	0
0	36
1	33

Random Forest Classifier Running...

Random Forest Classifier labeling accuracy: 81.01 %

Random Forest Classifier labeling precision score: 75.36 %

Random Forest Classifier recall score: 75.36 %

Random Forest Classifier labeling F1 score: 75.36 %

CONFUSION MATRIX:

Predicted

neg pos

Actual

neg

	0
0	93
1	17

pos

	0
0	17
1	52

Decision Tree Classifier Running...

Decision Tree Classifier labeling accuracy: 80.45 %

Decision Tree Classifier labeling precision score: 74.29 %

Decision Tree Classifier recall score: 75.36 %

Decision Tree Classifier labeling F1 score: 74.82 %

CONFUSION MATRIX:

Predicted

neg pos

Actual

neg

	0
0	92
1	18

pos

	0
0	17
1	52

PyplotGlobalUseWarning: You are calling `st.pyplot()` without any arguments. After

December 1st, 2020, we will remove the ability to do this as it requires the use of Matplotlib's global figure object, which is not thread-safe.

To future-proof this code, you should pass in a figure as shown below:

```
>>> fig, ax = plt.subplots()
>>> ax.scatter([1, 2, 3], [1, 2, 3])
>>>    ... other plotting actions ...
>>> st.pyplot(fig)
```

You can disable this warning by disabling the config option: `deprecation.showPyplotGlobalUse`

```
st.set_option('deprecation.showPyplotGlobalUse', False)
```

or in your `.streamlit/config.toml`

```
[deprecation]
showPyplotGlobalUse = False
```