

# Лабораторная работа №2

ТЕМА: «КЛАССИФИКАЦИЯ»

Магомедов Эдуард Хасбулаевич, Таланкин Кирилл, Сидоров Дмитрий  
М8О-309Б-23

# Описание датасета

- Датасет: Titanic\_dataset
  - Источник: Titanic\_dataset.csv
- Размер: 889 строк
- Ключевые поля:
  - Pclass, gender, embarked, age

```
display(DATASET.isna().sum())
```

Executed at 2025.10.25 14:19:01 in 756ms

PassengerId	0
Survived	0
Pclass	0
Name	0
Sex	0
Age	177
SibSp	0

```
1 import ...
5 DATASET = pd.read_csv('Titanic-Dataset.csv')
6 display(DATASET.head(-1))
```

Executed at 2025.12.15 11:28:32 in 230ms

	PassengerId	Survived	Pclass	Name
...	...	...	...	...
886	887	0	2	Montvila, Rev. Ju
887	888	1	1	Graham, Miss. Mar
888	889	0	3	Johnston, Miss. C
889	890	1	1	Behr, Mr. Karl Ho

# Валидация данных

```
target_col = 'Survived'  
y = DATASET[target_col].astype(int)  
X = DATASET.drop(columns=[target_col])  
cat_cols = ["Sex", "Embarked", "deck"]  
num_cols = [ "Pclass", "Age", "SibSp", "Parch", "Fare",  
    "cabin_present", "cabin_count", "cabin_num_exists"]  
  
cat_pipeline = Pipeline([  
    ('imp', SimpleImputer(strategy='most_frequent')),  
    ('ohe', OneHotEncoder(handle_unknown='ignore',  
        sparse_output=False))  
])  
  
prep = ColumnTransformer(  
    transformers=[  
        ('num', num_pipeline, num_cols),  
        ('cat', cat_pipeline, cat_cols)  
    ],  
    remainder='drop'  
)
```

# Сравнение LogReg, SVM и KNN

LogReg 5-fold CV

accuracy: 0.789 ± 0.035

f1 : 0.720 ± 0.047

roc\_auc : 0.854 ± 0.035

SVM (RBF) 5-fold CV

accuracy: 0.816 ± 0.028

f1 : 0.748 ± 0.045

roc\_auc : 0.853 ± 0.035

KNN (k=5) 5-fold CV

accuracy: 0.805 ± 0.023

f1 : 0.739 ± 0.024

roc\_auc : 0.825 ± 0.027

# Результаты

## ► GridSearchCV

```
GRID SEARCH
Лучшие параметры GridSearch:
{'model__max_depth': 10, 'model__min_samples_split':
5, 'model__n_estimators': 300}
MSE GridSearch: 1.5077210873228846e+16
```

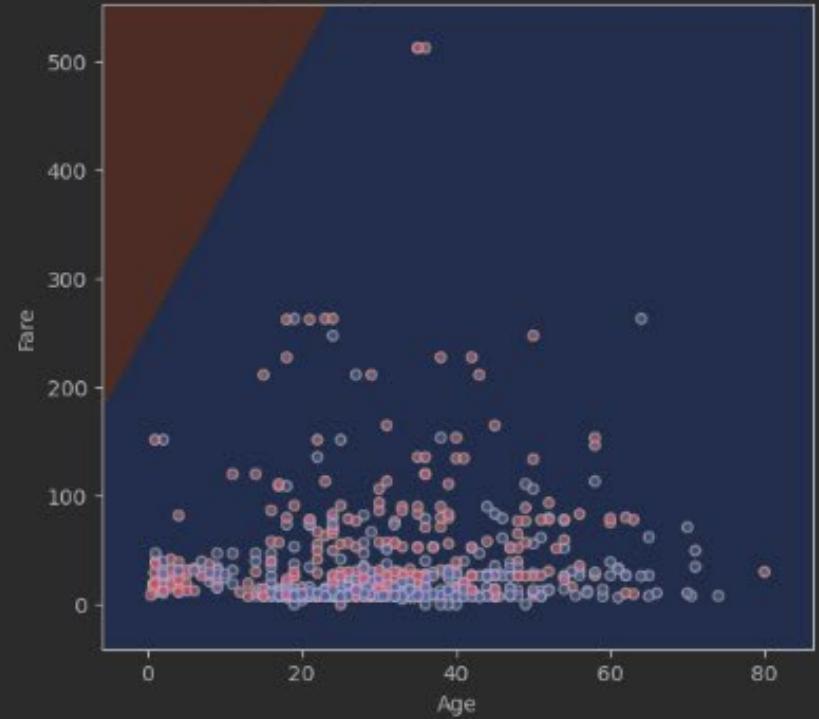
## ► RandomizedSearchCV

```
Лучшие параметры RandomSearch:
{'model__n_estimators': 400,
'model__min_samples_split': 10, 'model__max_depth' :
None}
MSE RandomSearch: 1.5478169879160896e+16
```

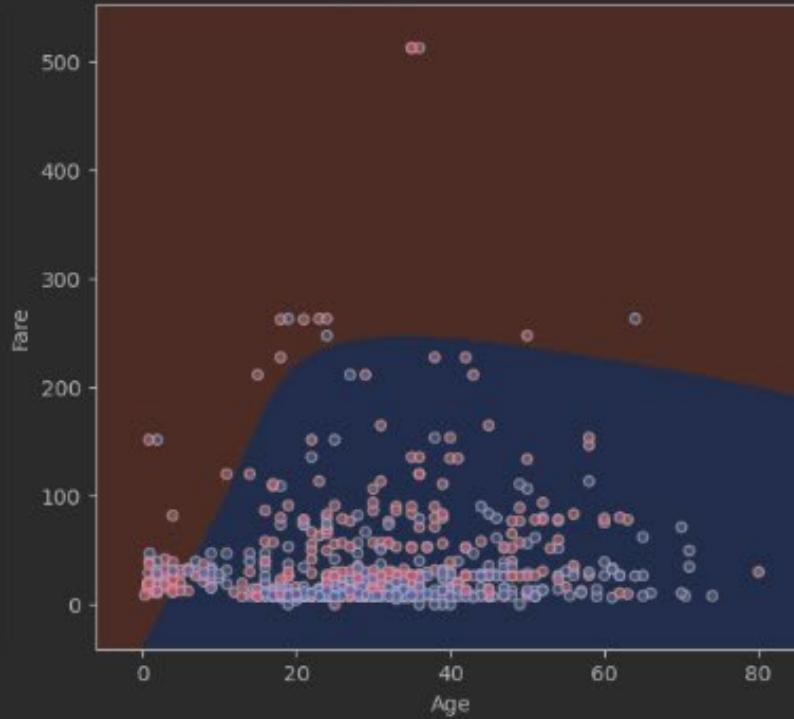
## ► Optuna

```
{'n_estimators': 146, 'max_depth': 9,
'min_samples_split': 4}. Best is trial 18 with value:
1.4863087578715278e+16.
```

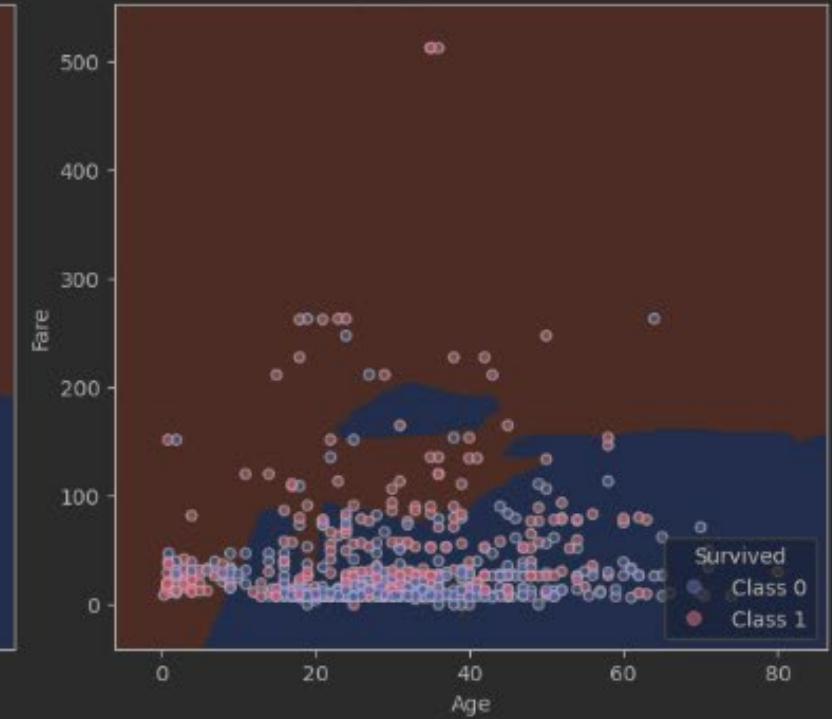
Logistic Regression: decision surface



SVM (RBF): decision surface



KNN (k=5): decision surface



Survived  
Class 0  
Class 1

# Библиотека Catboost: обучение и результаты

```
pipe_cat = Pipeline([
    ("cabin_feat", CabinFeaturizer(cabin_col="Cabin",
        add_hash=False, drop_original=True)),
    ("select", select_tf),
    ("fill_cats", fill_tf),
    ("clf", CatBoostClassifier(
        iterations=100,
        learning_rate=0.1,
        depth=5,
        loss_function="Logloss",
        eval_metric="AUC",
        random_seed=42,
        verbose=0
    ))
])
```

Model	Acc_mean	Acc_std	F1_mean	F1_std
	AUC_mean	AUC_std		
GradientBoosting	0.841	0.029	0.774	0.043
	0.875	0.028		
CatBoost	0.824	0.022	0.755	0.030
	0.872	0.021		

# Выбор лучшего метода

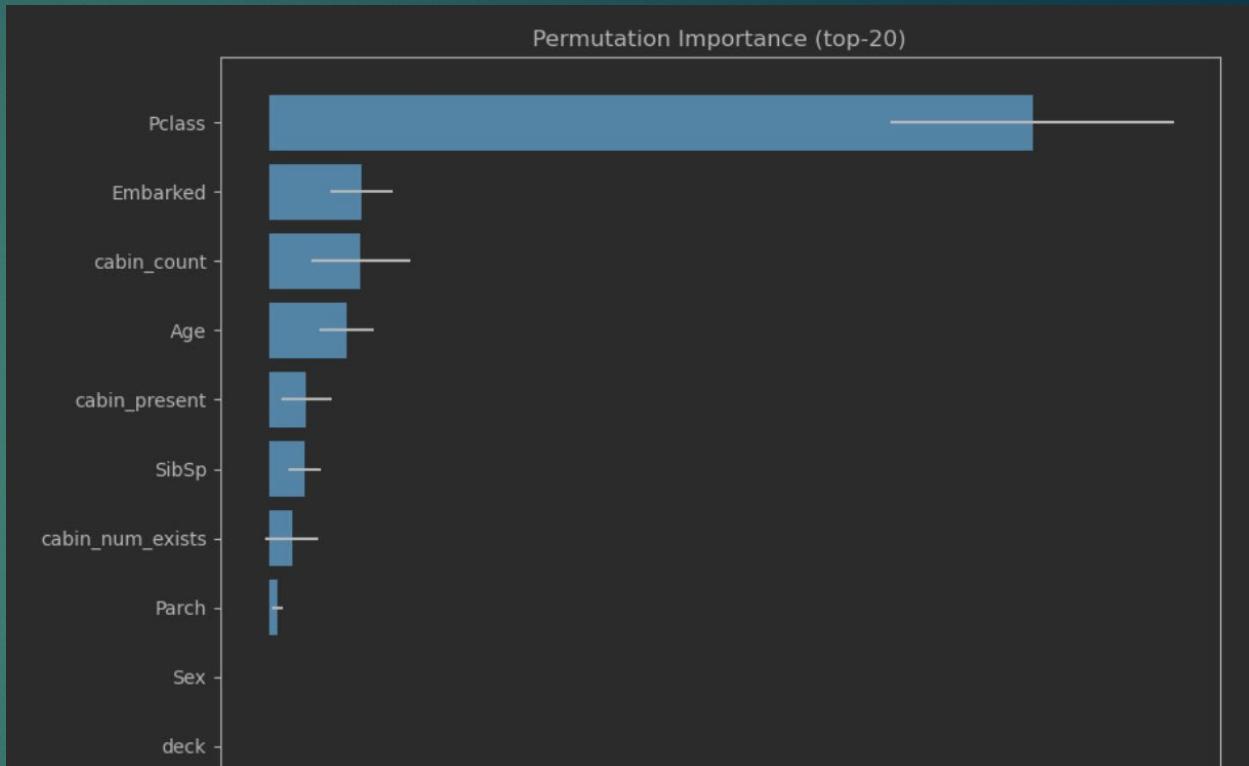
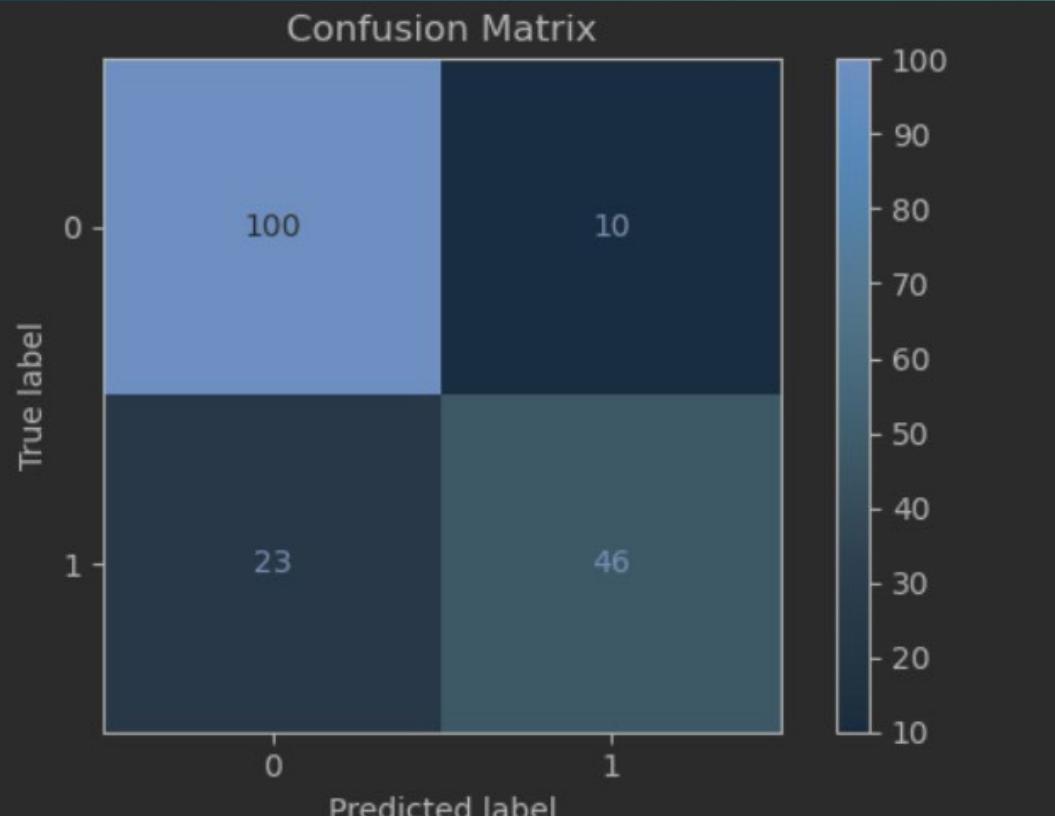
- ▶ Сравнив три метода видно, что среднеквадратичная ошибка меньше всего у метода Optuna со значением

```
Лучшие параметры: {'n_estimators': 131, 'max_depth':  
12, 'min_samples_split': 5}
```

```
Лучшее MSE: 1.4863087578715278e+16
```

# Confusion Matrix и топ-признаки

▼



# Глобальная интерпретация

