# AutoML - Praca Domowa 2

Mieszko Mirgos, Anna Rutkiewicz

# Modele Manualne - proces

1. Ręczne testowanie pojedynczych modeli
2. Stworzenie pierwszego pipeline'u
3. Rozbudowane pipeline'y
4. Zmniejszanie pipeline'ów
5. Alternatywne metody ekstrakcji feature'ów

# Modele Manualne - rezultaty

| Skład ensemble | Ensemble | Train-Test split | Preprocessing | Test score | Optymalizacja hiper-parametrów |
|---|---|---|---|---|---|
| KNeighborsClassifier x2<br>ExtraTreesClassifier x2<br>BaggingClassifier(KNeighborsClassifier) x2 | VotingClassifier | 80-20 | PCA | 0.9025 | RandomizedSearchCV |
| KNeighborsClassifier x2<br>ExtraTreesClassifier x2<br>BaggingClassifier(KNeighborsClassifier) x2 | VotingClassifier | 80-20 | PCA<br>XGBoost | 0.9200 | RandomizedSearchCV |
| KNeighborsClassifier x3<br>RandomForestClassifier x3<br>SVC x3<br>ExtraTreesClassifier x3<br>XGBClassifier x2 | VotingClassifier | 67-33 | PCA<br>SelectFpr<br>SelectFdr | 0.8621 | RandomizedSearchCV |
| KNeighborsClassifier x3<br>RandomForestClassifier x3<br>SVC x3<br>ExtraTreesClassifier x3<br>XGBClassifier x2 | VotingClassifier | 80-20 | PCA<br>SelectFpr<br>SelectFdr | 0.8925 | RandomizedSearchCV |
| KNeighborsClassifier x3<br>RandomForestClassifier x3<br>SVC x3<br>ExtraTreesClassifier x3<br>BaggingClassifier(KNeighborsClassifier) x3<br>DecisionTreeClassifier x3 | VotingClassifier | 67-33 | PCA<br>SelectFpr<br>SelectFdr | 0.8667 | BayesSearchCV |
| KNeighborsClassifier x2<br>RandomForestClassifier x2<br>SVC x2<br>ExtraTreesClassifier x2<br>XGBClassifier x2<br>final estimator: VotingClassifier | StackingClassifier | 80-20 | PCA<br>SelectFpr<br>SelectFdr | 0.8675 | BayesSearchCV |

# Modele Manualne - zwycięzca

```
pipe = Pipeline([
    ('reduce_dim', 'passthrough'),
    ('clf', VotingClassifier(
            voting='soft',
            verbose=True,
            estimators=[
            ('knn1', KNeighborsClassifier()),
            ('knn2', KNeighborsClassifier()),
            ('et1', ExtraTreesClassifier()),
            ('et2', ExtraTreesClassifier()),
            ('bc1', BaggingClassifier(KNeighborsClassifier(), random_state=4)),
            ('bc2', BaggingClassifier(KNeighborsClassifier(), random_state=4))
            ]
    ))
])
```

```
dimgrid=[
    {
        'reduce_dim': [PCA(svd_solver='full')],
        'reduce_dim__n_components': [4,5,6,7,8,499],
    },
    {
        'reduce_dim': [XGBFeatureExtractor()],
        'reduce_dim__n_components': [4, 6, 7, 20, 50, 100],
        'reduce_dim__xgb_params': [{'n_estimators': 100, 'max_depth': 3, 'learning_rate': 0.1},
        {'n_estimators': 200, 'max_depth': 5, 'learning_rate': 0.2}],
    }
]
```

```
model_distributions=[
    {
        'clf__knn1__n_neighbors': [4,5,6,7],
        'clf__knn2__n_neighbors': [4,5,6,7],

        'clf__et1__n_estimators': [100, 150, 200, 250, 300, 350, 400, 450, 500],
        'clf__et2__n_estimators':  [100, 150, 200, 250, 300, 350, 400, 450, 500],
        'clf__et1__max_depth': list(range(5, 13))+ [None],
        'clf__et2__max_depth': list(range(5, 13))+ [None],

        'clf__bc1__n_estimators': [9, 20, 50, 75, 100, 150, 200],
        'clf__bc2__n_estimators': [9, 20, 50, 75, 100, 150, 200],
        'clf__bc1__bootstrap': [True, False],
        'clf__bc2__bootstrap': [True, False],
    }
]
```

# Modele Automatyczne - proces

1. Testowanie 4 frameworków: AutoGluon, AutoSklearn, MlJar oraz TabPFN na kilku preprocessingach ze średnimi limitami czasowymi.
2. Odrzucenie AutoSklearna i MlJara
3. Wydłużanie czasu dla AutoGluona i eksperymentacja z preprocessingiem dla obu.

# Modele Automatyczne - rezultaty

| Framework | Preprocessor | Train-Test split | Test score (balanced accuracy) | Learning time Limit (hours) |
|---|---|---|---|---|
| AutoGluon | PCA(5) | 80-20 | 0.9025 | 4 |
| AutoGluon | PCA(7) | 80-20 | 0.8825 | 4 |
| AutoGluon | PCA(20) | 80-20 | 0.855 | 4 |
| AutoGluon | PCA(50) | 80-20 | 0.875 | 4 |
| AutoGluon | XGBoost | 80-20 | 0.89 | 4 |
| AutoGluon | PCA(7) | 67-33 | 0.87 | 4 |
| AutoGluon | PCA(20) | 67-33 | 0.86 | 4 |
| AutoGluon | PCA(50) | 67-33 | 0.83 | 4 |
| TabPFN | PCA(5) | 80-20 | 0.9025 | None |
| TabPFN | PCA(5) | 67-33 | 0.89 | None |
| TabPFN | XGBoost | 80-20 | 0.6725 | None |

# Modele Automatyczne - zwycięzca

| | | | | |
|---|---|---|---|---|
| AutoGluon | PCA(5) | 80-20 | 0.9025 | 4 |
| TabPFN | PCA(5) | 80-20 | 0.9025 | None |

# Wnioski

- Modele manualne > Modele automatyczne
- Podział zbioru treningowego ma znaczenie
- Ensemble - czasem mniej modeli znaczy lepiej
- Dobry preprocessing jest ważny
- Od pewnego momentu podnoszenie limitu czasu jest bezcelowe

# Dziękujemy za uwagę