Rain in Australia

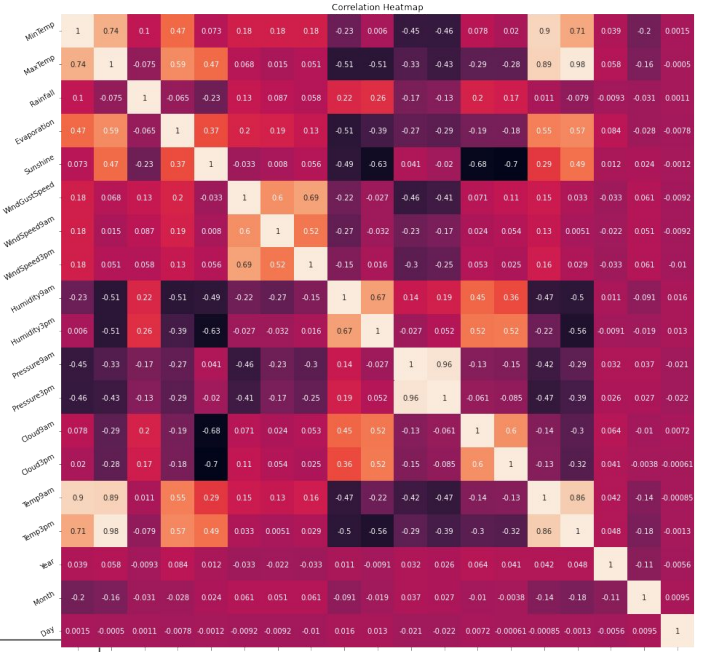
Kaggle Project
MatCAD

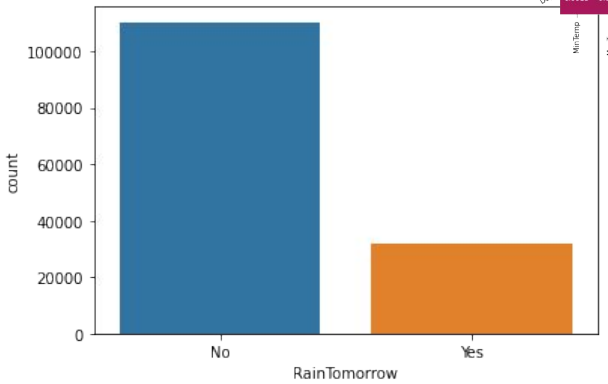
Maria Pallejà: 1570129

https://github.com/mpvt2001/Rain_in_Australia

Exploratory Data Analysis

- Tipus d'atributs
- 23 atributs
- Atribut objectiu: RainTomorrow
- Correlacions
- Dades no balancejades





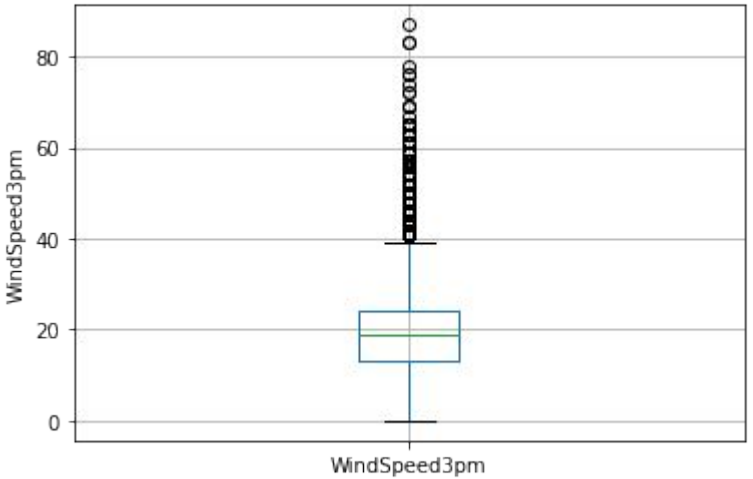
Preprocessing

· Tractament dels Nans.

· Tractament outliers

· Tractament de les dades categòriques
  - get_dummies

· Escalatge de dades
  - *MinMaxScaler*
  - *StandardScaler*



Model Selection

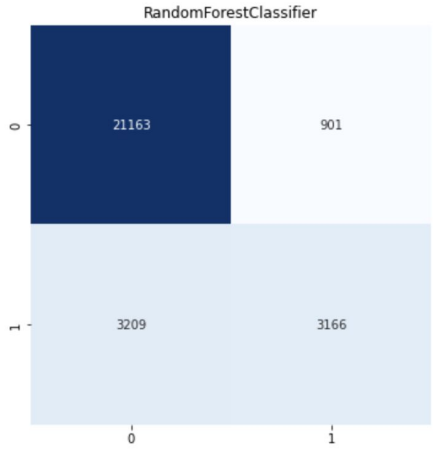
· Testeig de models

· Matrius de confusió

· Mètodes d'Ensemble

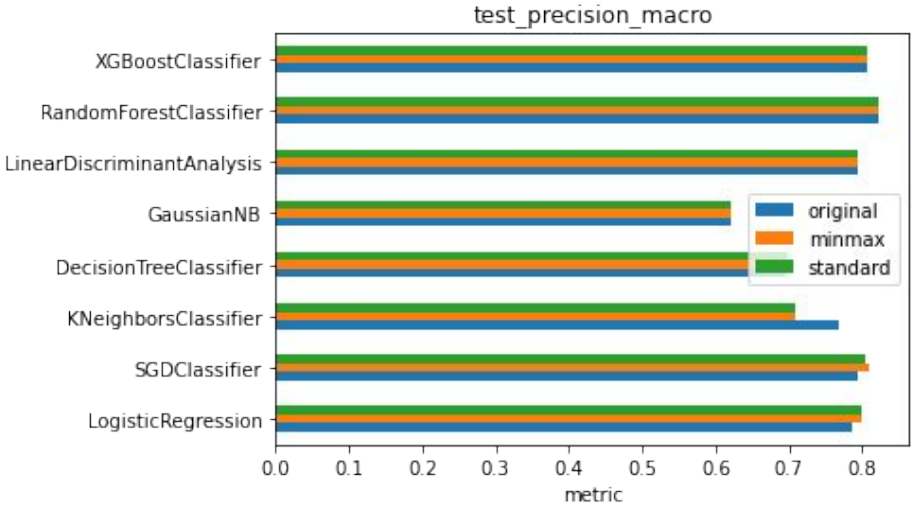
- RandomForestClassifier: 0.8410984915081402
- BaggingClassifier: 0.8338197545623967
- GradientBoostingClassifier: 0.845071908294947

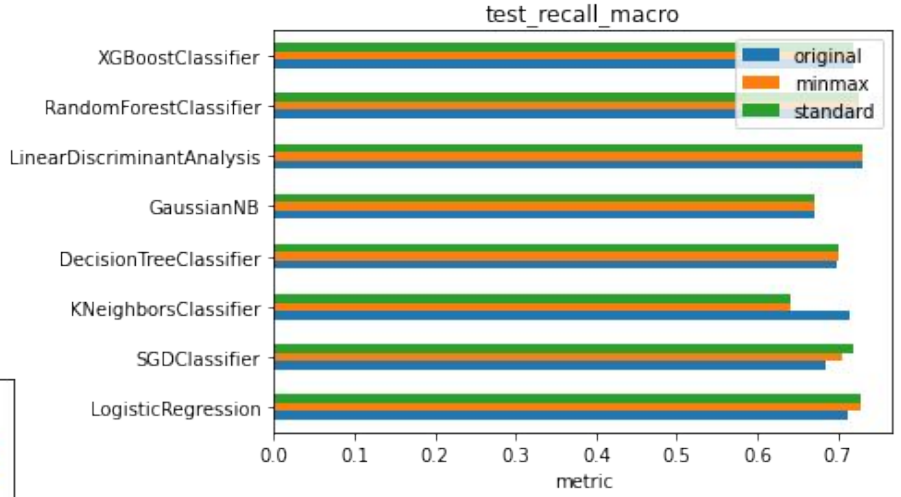
	original	minmax	standard
LogisticRegression	0.840677	0.845916	0.845916
SGDClassifier	0.822849	0.844650	0.845248
KNeighborsClassifier	0.832308	0.801329	0.801329
DecisionTreeClassifier	0.784767	0.775801	0.775906
GaussianNB	0.641338	0.637329	0.637329
LinearDiscriminantAnalysis	0.846338	0.845599	0.845599
RandomForestClassifier	0.855480	0.854249	0.854601
XGBoostClassifier	0.847744	0.847147	0.847147



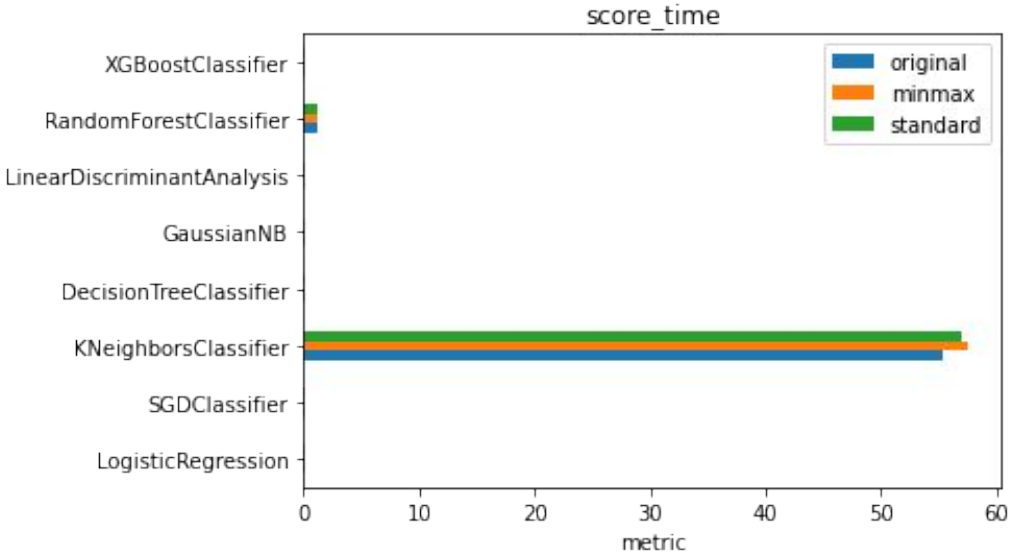
Crossvalidation

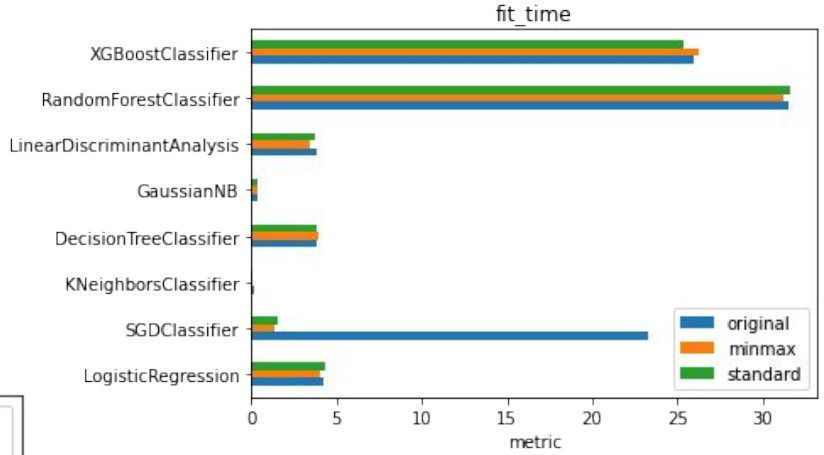
k = 5





Crossvalidation





Hyperparameter Search

· Mètodes
  - RandomizedSearchCV

LogisticRegression
{'penalty': 'l2', 'C': 10}
Tarda 2.585540073394777 per training

RandomForestClassifier
{'n_estimators': 100, 'min_samples_split': 2, 'criterion': 'entropy'}
Tarda 24.920233707427975 per training