Выполнил: ЧжаоЛян

Группа: ИУ5И-22М

Random Forest Classifier

Complement Naive Bayes

```
In [1]: import numpy as np
         import pandas as pd
         from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
         from sklearn.model_selection import train_test_split
         from sklearn.metrics import classification_report
         from sklearn.ensemble import RandomForestClassifier
         from sklearn.naive_bayes import ComplementNB
In [6]: df = pd.read_csv(r'C:\Users\80667\Desktop\文件\MY5\研一下\MMO\lab\lab\\数据集\StudentsPerformance.csv')
Out[6]:
               gender race/ethnicity
                                    bachelor's degree
                           group C
                                           some college
                                                                             completed
                          group B
                                         master's degree
                                                           standard
                                                                                                                     93
                                                                                                         57
                           group A
                                        associate's degree free/reduced
                male
                          group C
                                       some college
                                                                                             76
                                                                                                         78
                                                                                                                     75
                                                           standard
                                                                                none
          995 female
                          group E
                                                                                             88
                                                                                                         99
                                                                                                                     95
                                          master's degree
                                                          standard
                                                                             completed
                                                                                                         55
                                                                                                                     55
          996
                male
                           group C
                                             high school free/reduced
                                                                                none
                                                                             completed
                                                                                             59
                                                                                                         71
                                                                                                                     65
          997 female
                          group C
                                            high school free/reduced
          998 female
                                                                                                         78
                                                                                                                     77
                           group D
                                            some college
                                                           standard
                                                                             completed
          999 female
                                           some college free/reduced
                                                                                                         86
         1000 rows × 8 columns
```

Предобработка признаков

TFIDF

```
In [7]: tfidfy = TfidfVectorizer()
    tfidf_ngram_features = tfidfv.fit_transform(df['parental level of education'])
    tfidf_ngram_features

Out[7]: <1000x8 sparse matrix of type '<class' numpy.float64'>'
    with 2179 stored elements in Compressed Sparse Row format>
```

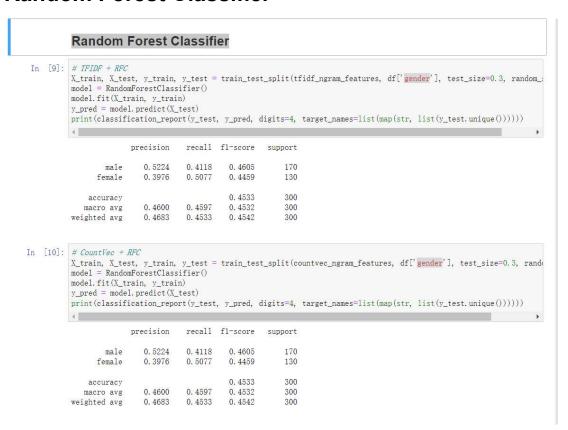
CountVectoriser

CountVectoriser

```
In [8]: countvec = CountVectorizer()
countvec_ngram_features = countvec.fit_transform(df['parental level of education'])
countvec_ngram_features

Out[8]: <1000x8 sparse matrix of type '<class 'numpy.int64'>'
with 2179 stored elements in Compressed Sparse Row format>
```

Random Forest Classifier



Complement Naive Bayes

```
Complement Naive Bayes
In [11]: # TFIDF + CNB
          X_train, X_test, y_train, y_test = train_test_split(tfidf_ngram_features, df['gender'], test_size=0.3, random_smodel = ComplementNB()
           model.fit(X_train, y_train)
           y_pred = model.predict(X_test)
           print(classification_report(y_test, y_pred, digits=4, target_names=list(map(str, list(y_test.unique())))))
                       precision recall f1-score support
                             0. 5500 0. 2588 0. 3520
0. 4273 0. 7231 0. 5371
                   male
                 female
                                                  0.4600
                                                                 300
               accuracy
                          0. 4886 0. 4910 0. 4446
0. 4968 0. 4600 0. 4322
                                                                 300
           weighted avg
                                                                 300
In [12]: # CountVec + CNB
          X_train, X_test, y_train, y_test = train_test_split(countvec_ngram_features, df['gender'], test_size=0.3, randomodel = ComplementNB()
           model.fit(X_train, y_train)
           y_pred = model.predict(X_test)
           print(classification_report(y_test, y_pred, digits=4, target_names=list(map(str, list(y_test.unique())))))
                         precision recall fl-score support
                            0. 5500 0. 2588 0. 3520
0. 4273 0. 7231 0. 5371
                   male
                 female
                                                  0.4600
                                                                 300
               accuracy
          macro avg 0.4886 0.4910 0.4446
weighted avg 0.4968 0.4600 0.4322
                                                                 300
```

Выводы:

CountVectorizer с Random Forest Classifier показал лучшие результаты, чем TFIDF, а с Complement Naive Bayes оба векторизатора показали одинаковые результаты Random Forest Classifier показал лучшие, чем Complement Naive Bayes результаты