# 第八周作业

2024年11月3日　　14:23

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.metrics import f1_score
import numpy as np
file_path = r"C:\Users\ASUS\Documents\WeChat Files\wxid_gtfqqzwcu89o22\FileStorage
\File\2024-10\fraudulent.csv"
df = pd.read_csv(file_path)

missing_percentage = df.isnull().sum() / len(df)
columns_to_drop = missing_percentage[missing_percentage > 0.5].index
df = df.drop(columns=columns_to_drop)
.
for col in df.columns:
    if missing_percentage[col] < 0.5:
        df[col] = df[col].fillna(df[col].mode()[0])
X = df.drop("y", axis=1)
y = df["y"]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=1)
models = [
    ("K-近邻", KNeighborsClassifier(n_neighbors=5)),
    ("决策树", DecisionTreeClassifier()),
    ("逻辑回归", LogisticRegression()),
    ("支持向量机", SVC(kernel="linear"))
]
for name, model in models:
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    f1 = f1_score(y_test, y_pred)
    print(f"{name}模型的F1值为: {f1}")
```

```
PS F:\下载的文件\zuoye> python -u "f:\下载的文件\zuoye\10_28.py"
K-近邻模型的F1值为: 0.8373540856031129
决策树模型的F1值为: 0.8648648648648648
逻辑回归模型的F1值为: 0.8490718321226796
支持向量机模型的F1值为: 0.8490566037735849
PS F:\下载的文件\zuoye>
```

把数据少于一半的剔除，然后多于一半的用众数填充。综合来看决策树模型在此次针对该数据集的测试中表现最佳，能够更有效地对网站是否为钓鱼欺诈网站进行分类预测。