第十一次作业简单报告

10235501461 柯宇

作业要求与目的：

使用不同方法对相同数据集进行机器学习建模，以查验出钓鱼网站，并且以 **F1** 指标为准，比较不同预处理对数据建模的有效性和不同方法所建出的模型的分类能力。

预处理方法的不同：

使用删去过多确实值的列的方法，当某一列超过 **30%**缺失时即删除时，结果如下。

```
KNN F1 Score: 0.8386
Decision Tree F1 Score: 0.8720
Logistic Regression F1 Score: 0.8548
SVM F1 Score: 0.8668
```

而超过 **10%**即删除时，结果如下。

```
KNN F1 Score: 0.8636
Decision Tree F1 Score: 0.8782
Logistic Regression F1 Score: 0.8542
SVM F1 Score: 0.8664
```

难道这完全是因为删除列减少误差的功能吗？事实上，删除列同时导致我们在进行机器学习建模的时候指标减少，能够得出看似更加准确的结论。但对于实际的数据，如果删除了过多的列，其实模型是无法拟合出真实情况并且分类的。

使用众数的方法更加简单粗暴，但是实际上能够更加随机地对模型进行处理，不至于导致某一个指标的缺失。

建立模型方法的不同：

以下分别为训练集比例不同时，模型的 **F1** 指标值。

0.1

```
KNN F1 Score: 0.8731
Decision Tree F1 Score: 0.8689
Logistic Regression F1 Score: 0.8646
SVM F1 Score: 0.8655
```

0.3

```
KNN F1 Score: 0.8353
Decision Tree F1 Score: 0.8731
Logistic Regression F1 Score: 0.8548
SVM F1 Score: 0.8673
```

0.5

```
KNN F1 Score: 0.8338
Decision Tree F1 Score: 0.8554
Logistic Regression F1 Score: 0.8559
SVM F1 Score: 0.8655
```

0.7

```
KNN F1 Score: 0.8344
Decision Tree F1 Score: 0.8420
Logistic Regression F1 Score: 0.8442
SVM F1 Score: 0.8638
```

在所有测试中，可以得到的结论是，**SVM** 向量机模型是对于数据拟合最好的模型。

同时，也可以注意到数据划分的比例会影响整个模型的训练质量。划分训练集时不可以过高也不可以过低。例如在本次实验中，大于 **0.3** 的情况下，**F1** 指标会下降，可能发生了过拟合的情况；而取 **0.1** 时数据集过少，尽管看似效果很好，有可能是忽略了某部分数据的影响，发生欠拟合的情况，在实际运行中仍然不会是最好的模型。

——————————————————报告完毕——————————————————