# 数据分析报告

## 1. 引言

本报告对ISCX网络流量数据集进行了全面的分析，包括数据清洗、特征分析、相关性分析和标签分布的可视化。通过这些步骤，我们对数据的特征和分布有了深入的了解，为后续的模型分析奠定了基础。

## 2. 数据集基本信息

数据来源：ISCX网络流量数据集

数据集大小：包含214,099条记录

特征数量：包含91个特征，包括数值特征和分类特征

目标变量：Label，表示网络流量的类型（BENIGN或PortScan）

## 3. 数据清洗

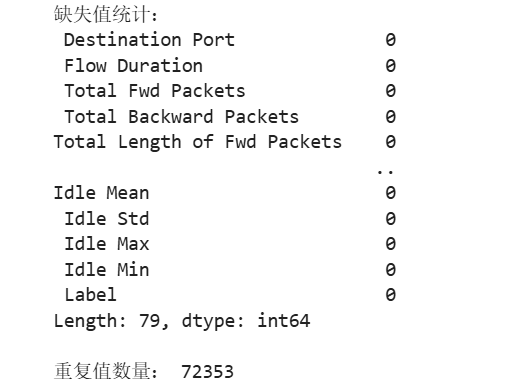
### 3.1 缺失值处理

数据集中存在缺失值，已删除包含缺失值的行。

缺失值统计显示所有特征均无缺失值。

### 3.2 重复值处理

数据集中不存在重复值，无需进行重复值处理。



### 3.3 异常值处理

数据中存在无穷大值和非常大的值，已将其替换为NaN并删除。

## 4. 数据预处理

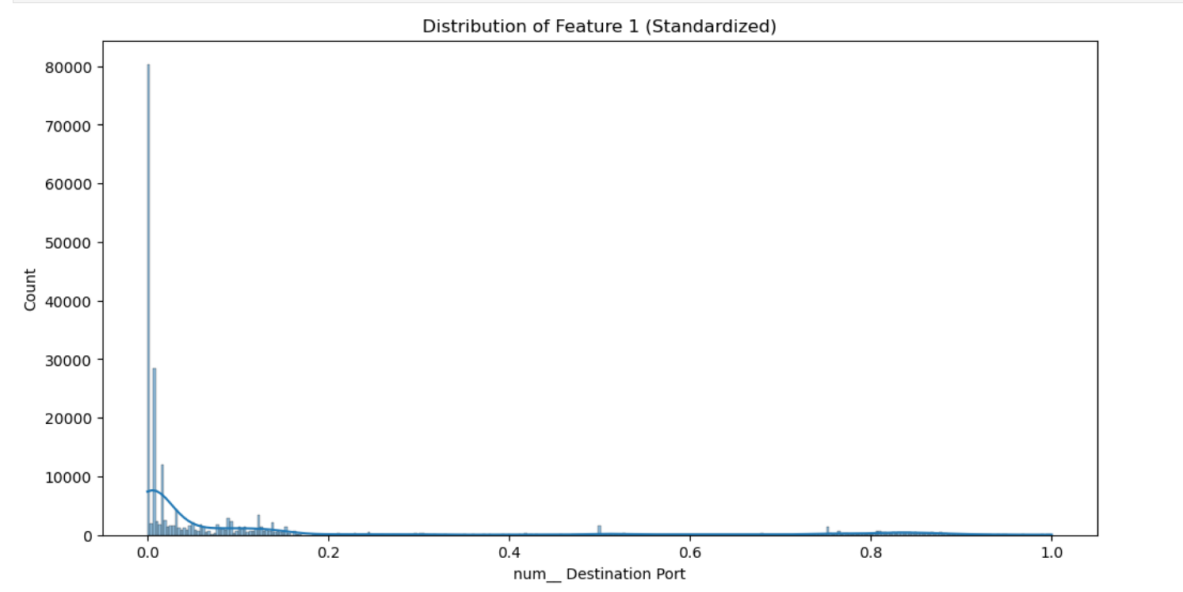
### 4.1 特征标准化

使用MinMaxScaler对所有数值特征进行标准化，将数据缩放到[0, 1]范围内。

## 5. 数据分析与可视化

### 5.1 特征分布

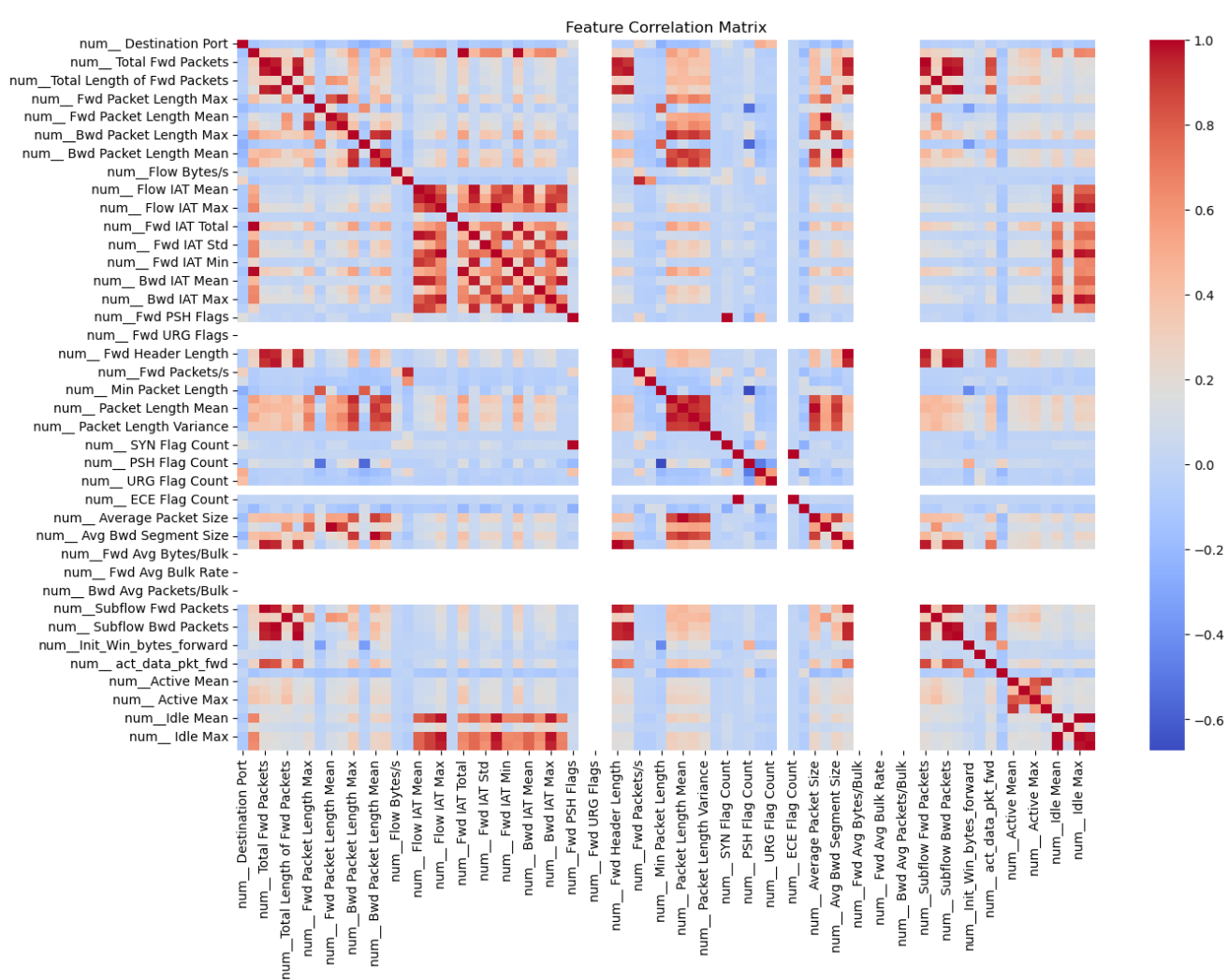
特征分布图：展示了标准化后的特征分布，大多数特征集中在较低的值范围内。



该图展示了“num\_\_Destination Port”特征的分布情况。横轴表示该特征的取值范围，纵轴表示对应取值的样本数量。从图中可以看出，大多数样本的“num\_\_Destination Port”取值集中在较低的范围内，尤其是在0附近有一个明显的峰值。这表明在数据集中，大部分网络流量的目的端口号相对较小，可能集中在一些常见的端口上，如HTTP（80）、HTTPS（443）等。这种分布模式对于后续的模型分析是有帮助的，因为它可以反映出网络流量的常见模式和异常情况。

### 5.2 特征相关性

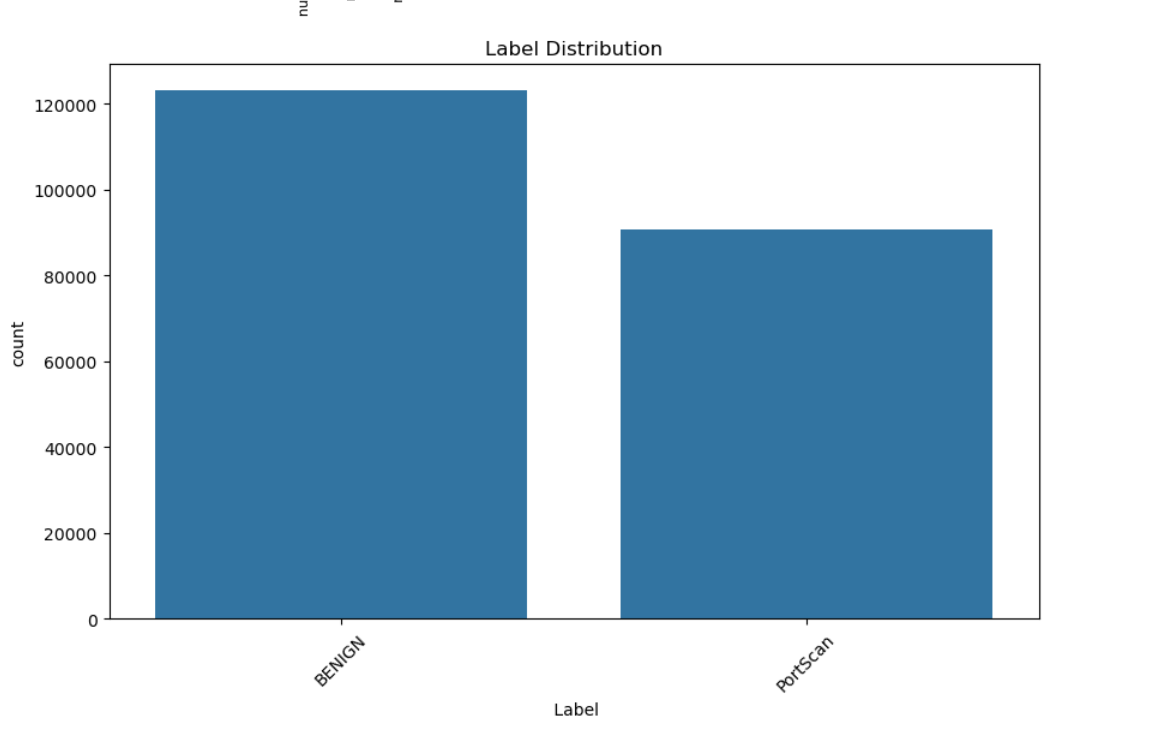
特征相关性热力图：显示了特征之间的相关性，部分特征之间存在较高的相关性。



该图展示了数据集中各特征之间的相关性。颜色越深，表示相关性越强。从图中可以看出，某些特征之间存在较高的相关性，例如“num\_\_Total Fwd Packets”和“num\_\_Flow Packets/s”。这表明这些特征可能在某种程度上反映了相似的网络行为或模式。在后续的模型分析中，可以考虑对这些高度相关的特征进行进一步的分析和处理，以避免多重共线性问题，提高模型的性能和稳定性。

### 5.3 标签分布

标签分布图：显示了两类标签（BENIGN和PortScan）的数量分布，BENIGN样本数量明显多于PortScan样本。



该图展示了数据集中不同标签（BENIGN和PortScan）的样本数量分布。从图中可以看出，BENIGN样本的数量明显多于PortScan样本。这种类别不平衡的情况在网络安全数据集中较为常见，因为正常流量通常远多于攻击流量。在后续的模型分析中，需要特别注意这种类别不平衡问题，可以采用过采样、欠采样或使用适合处理不平衡数据的算法等方法来解决。