# 开发文档

### 1. ****数据理解与探索****

* **数据加载**：先从各个数据集（如幸福指数、训练数据等）加载数据。
* **数据结构分析**：理解各数据集的列、数据类型，以及每个特征的含义，例如收入、教育、宗教信仰、健康等。
* **数据分布可视化**：通过直方图、箱线图、散点图等，探索各个特征的分布和相关性。

### 2. ****数据预处理****

* **缺失值处理**：处理缺失值，常见方法包括：
  + 删除缺失值较多的列或行。
  + 用平均值、中位数或模式等填补缺失值。
* **异常值处理**

异常值指的是那些远离大多数数据点的极端值，比如年收入为1亿或仅为1元的人。 这些异常值可能会让模型产生偏差。

处理思路：

检测异常值：使用统计学方法，如箱线图（Boxplot）或3倍标准差法，来识别异常值。

处理异常值：可以选择删除这些异常值，或者将其限制在某个合理的范围内。

* **数据类型转换**：

**日期数据**：可以将日期列转化为一些有用的时间特征，比如年份、月份，甚至将其表示 为从某个时间点的天数。

**分类数据**：比如“省份”或“城市”这些分类数据不能直接输入到模型中，我们需要将其转 化为数值型数据。常用的方法是 **One-Hot Encoding**，它将每个分类转换为一个独立 的二进制特征。

* **特征工程**：
  + 对收入、教育、宗教等特征进行离散化（例如，将收入分类为低、中、高）。
  + 创建新的特征，例如幸福指数的衍生变量（通过收入、健康等加权组合）。
* **数据标准化或归一化**：  
  不同特征的数值可能有很大的不同，比如收入的数值远大于年龄或教育年限。如果直接使用这些特征，模型可能会过度关注大数值的特征，因此需要对特征进行标准化或归一化。
* 标准化：将数据转化为标准正态分布，均值为0，标准差为1。

归一化：将数据压缩到0-1之间。

### 3. ****探索性数据分析 (EDA)****

* **描述性统计分析**：查看均值、中位数、标准差等统计量，理解数据基本情况。  
  思路：
  + 集中趋势：通过均值、中位数和众数来了解数据的典型值。
  + 离散程度：通过标准差、四分位距等指标来判断数据的波动性。
  + 分布形态：通过查看数据的最大值、最小值，发现数据的范围和是否存在极端值。  
     均值（Mean）：衡量数据的中心点，但容易受到极端值影响。
    - 中位数（Median）：比均值更能代表数据的中间值，适用于有极端值的数据。
    - 标准差（Std）：衡量数据的离散程度，值越大，数据的波动越大。
* **相关性分析**：通过计算特征间的相关系数矩阵，找出与幸福感高度相关的因素。  
  相关性分析可以帮助我们找到哪些变量与幸福感之间有较强的相关性。通过计算两个变量之间的相关系数（如皮尔逊相关系数），我们可以量化它们之间的线性关系。
  + 思路：
  + 相关性矩阵：计算所有变量之间的相关性，用于发现哪些特征与幸福感高度相关。相关性系数的取值范围是 [-1, 1]，其中：
  + 1 表示完全正相关。
  + -1 表示完全负相关。
  + 0 表示无相关性。
  + 热力图：通过热力图（Heatmap）直观展示相关性矩阵，便于快速发现高相关性特征。
* **可视化分析**：使用热力图、散点矩阵等工具进一步分析变量之间的关系。  
  数据分布可视化
  + 数据分布可视化可以帮助我们直观地理解每个变量的分布情况，并且能发现数据中的异常值或分布偏差。
  + 思路：
  + 单变量分布：通过直方图（Histogram）或密度图（Density Plot）查看每个特征的分布形态。
  + 双变量关系：通过散点图（Scatter Plot）查看两个变量之间的关系，分析线性或非线性关联。

### 4. ****数据挖掘方法****

* **分类问题建模**：基于幸福感的离散值，选择合适的分类算法，例如：
  + 决策树
  + 随机森林
  + 支持向量机（SVM）
  + 梯度提升树（GBDT）
* **回归问题建模**：如果幸福感是连续值（如1到10的评分），则可以使用回归算法，如线性回归或随机森林回归。
* **聚类分析**：将相似的人群分组（如根据社会经济状况、健康状况等）以识别幸福感的不同群体。
* **关联规则挖掘**：找出特征之间的隐性关系，例如教育与幸福感之间的规则。

### 5. ****模型评估与优化****

* **交叉验证**：通过K折交叉验证确保模型的泛化能力。
* **评价指标**：根据任务的性质，使用准确率、F1分数、均方误差等评估模型的性能。
* **模型调优**：使用网格搜索或随机搜索优化模型参数。

### 6. ****结果解释与决策支持****

* **特征重要性分析**：使用模型中各特征的权重，分析哪些因素对幸福感影响最大。
* **可视化结果**：通过图表展示关键因素的影响，帮助政策制定者理解模型结果。
* **预测与应用**：根据不同地区和群体的特征，预测他们的幸福感，并为政府提供政策建议。