

# 城市空气质量监测及归因分析系统

## 项目概览

**项目名称：**城市空气质量监测及归因分析系统

**项目代号：**AirQuality-StatModel-2025

**开发时间：**2025年

**项目仓库地址：**

GitHub: <https://github.com/L-Dramatic/Stat-Modeling>

## 一、问题背景与研究目标

### 1.1 研究背景

城市空气质量问题是当代环境科学领域的核心议题之一。随着城镇化进程的加速和工业活动的增加，大气污染问题日益严峻，直接威胁着居民的身体健康和生活质量。PM2.5（细颗粒物）作为衡量空气质量的关键指标，因其粒径小、能够深入肺泡甚至进入血液循环而备受关注。长期暴露于高浓度PM2.5环境中会增加呼吸系统疾病、心血管疾病等健康风险，因此对PM2.5的监测和归因分析具有重要的公共健康意义。

在实际应用中，PM2.5浓度受多种因素的综合影响，包括气象条件（温度、气压、风速、湿度等）、人类活动（工业排放、交通尾气、生活燃烧等）以及地理环境因素。这些因素之间往往存在复杂的交互作用，使得污染物浓度呈现出明显的时间波动性和空间异质性。传统的定性分析难以揭示这些因素的具体贡献程度，而借助统计建模方法则能够从数据中挖掘出更深层次的规律。

本项目选取北京市2010年至2014年的PM2.5监测数据作为研究对象。北京作为我国首都和人口密集的特大城市，其空气质量问题具有典型性和代表性。这一时期正值我国大气污染防治的关键阶段，研究这段历史数据有助于总结经验教训，为未来的环境治理提供参考依据。

### 1.2 研究目标

本项目的核心目标是构建一套基于统计学方法的城市空气质量分析系统，通过“由浅入深”的分析流程，全面揭示PM2.5浓度的变化规律及其影响因素。具体而言，研究目标包括以下几个层面：

首先，在数据探索层面，我们希望通过可视化手段直观展示PM2.5的时间变化趋势，识别其分布特征和异常模式，为后续建模奠定基础。这一阶段的工作包括绘制时间序列图、分布直方图、相关性热力图等，帮助研究者快速掌握数据的基本面貌。

其次，在统计推断层面，我们利用假设检验方法验证一些常见的经验认知。例如，工作日与周末的PM2.5浓度是否存在显著差异？不同风向条件下的污染水平是否有明显不同？这些问题的回答能够帮助我们理解人类活动和自然因素对空气质量的影响机制。

第三，在回归建模层面，我们将构建多种回归模型来量化各气象因子对PM2.5的贡献程度。从最基础的普通最小二乘回归（OLS）出发，逐步引入Ridge回归、Lasso回归、广义线性模型（GLM）以及贝叶斯回归等方法，通过模型对比分析选择最优的归因分析方案。

第四，在分类预测层面，我们将PM2.5浓度划分为“优良”“轻度污染”“重度污染”三个等级，运用Logistic回归和朴素贝叶斯分类器进行空气质量等级预测，为公众出行和健康防护提供决策支持。

最终，我们将所有分析模块整合到一个交互式Web系统中，使用户能够直观地探索数据、运行模型并获取分析结果。这一系统不仅具有学术研究价值，也为环境监测部门和普通公众提供了实用的分析工具。

## 二、数据收集与处理过程

### 2.1 数据来源与说明

本项目采用的数据来自UCI机器学习数据库（UCI Machine Learning Repository）发布的"Beijing PM2.5 Data Set"。这是一份公开的环境监测数据集，记录了北京市美国大使馆站点2010年1月1日至2014年12月31日期间的逐小时空气质量和气象观测数据。选择这一数据集主要基于以下考量：

数据集的时间跨度长达五年，涵盖了四个完整的年度周期，能够充分反映季节性变化规律和长期趋势。同时，逐小时的采样频率为研究日内变化模式和短期波动提供了丰富的细节信息。数据规模约为1.5MB，符合课程对数据量（小于10MB）的要求，便于在普通计算机上进行处理和分析。

数据集包含的主要字段如下：目标变量为PM2.5浓度值（单位：微克/立方米），特征变量包括露点温度DEWP（摄氏度）、气温TEMP（摄氏度）、气压PRES（百帕）、风向cbwd（分类变量，包括东北NE、西北NW、东南SE、静风cv四个类别）、累计风速lws（米/秒）、累计降雨量lr（毫米）以及累计降雪量ls（毫米）。此外还有年、月、日、小时等时间字段用于构建时间索引。

**【插图位置1：数据集基本信息截图】操作说明：运行系统后，在“数据洞察”页面截取显示总记录数、PM2.5均值、峰值、缺失值数量的四个指标卡片区域。**

### 2.2 数据质量评估

在正式建模之前，我们对原始数据进行了全面的质量评估。初步检查发现，数据中存在一定比例的缺失值，主要集中在PM2.5浓度字段上。这些缺失可能源于监测设备故障、数据传输中断或极端天气条件下的采样困难。缺失值的存在会影响后续统计分析的可靠性，必须进行妥善处理。

另一方面，数据中也存在一些明显偏离正常范围的异常值。部分PM2.5记录呈现异常高值，可能对应沙尘暴、重污染事件等极端情况；也有个别极低甚至负值的记录，显然属于仪器误差或数据录入错误。这些异常值如果不加处理，会严重干扰模型的参数估计和预测精度。

从数据分布角度看，PM2.5浓度呈现明显的右偏分布特征。大部分时间段内PM2.5维持在较低水平，但偶发的高污染事件会导致浓度值急剧攀升，形成“长尾”分布形态。这种非正态分布特性对模型选择具有重要启示——传统的基于正态分布假设的方法（如OLS回归）可能并非最优选择，需要考虑采用广义线性模型等更灵活的建模框架。

### 2.3 数据预处理流程

针对上述数据质量问题，我们设计了一套完整的预处理流程，该流程封装在系统的DataPreprocessor类中，用户可以通过侧边栏参数灵活配置预处理策略。

#### （1）缺失值处理

系统提供了两种可选方案：线性插值法（interpolation）根据缺失点前后的有效观测值进行内插，能够保持时间序列连续性；直接删除法（drop）虽会损失数据量，但能保证分析结果的纯净性。默认采用线性插值。

#### （2）异常值检测与剔除

系统实现了两种经典方法：三倍标准差原则（3sigma）基于正态分布假设，将偏离均值超过三个标准差的观测点判定为异常值，适用于近似正态分布的数据；四分位距方法（IQR）更加稳健，将低于 $Q1 - 1.5 \times IQR$ 或高于 $Q3 + 1.5 \times IQR$ 的观测点视为异常，适用于偏态分布数据。

### (3) 数据变换

考虑到PM2.5数据的右偏分布特性，系统提供了对数变换（Log Transform）功能。通过取对数可以将正偏态分布转换为近似正态分布，从而满足许多统计方法对正态性假设的要求。系统会新增log\_PM2.5列存储变换结果，便于对比。

### (4) 分布检验与拟合

为验证预处理效果和数据分布特征，系统实现了正态性检验模块（Shapiro-Wilk检验和D'Agostino检验）以及多种分布拟合功能（正态分布、Gamma分布、对数正态分布），通过K-S检验和AIC准则评估拟合优度。实验结果表明，PM2.5数据更适合采用Gamma分布或对数正态分布来描述，这为后续选择GLM模型提供了理论依据。

【插图位置2：数据分布与正态性检验结果截图】操作说明：在系统中勾选“对PM2.5做Log变换”选项，然后截取相关性热力图和分布对比图区域。

---

## 三、数据可视化分析

### 3.1 时间序列可视化

数据探索的第一步是观察PM2.5浓度随时间的变化趋势。系统在“数据洞察”页面提供了基于Plotly库实现的交互式时间序列图，用户可以通过缩放、平移等操作细致考察任意时间段的污染变化情况。

从宏观趋势来看，五年间PM2.5浓度呈现出明显的季节性波动模式。冬季（11月至次年2月）PM2.5浓度普遍较高，这与北方地区冬季供暖燃煤排放增加、气象条件不利于污染物扩散等因素密切相关。相比之下，夏季（6月至8月）PM2.5浓度相对较低，主要得益于降雨的冲刷作用和较强的大气对流活动。春秋两季则处于过渡状态，偶发的沙尘天气可能导致PM2.5短期内急剧升高。

从日内变化模式来看，PM2.5浓度在一天之内也呈现出规律性波动。早高峰和晚高峰时段由于交通流量增大，尾气排放增加，PM2.5浓度往往出现峰值。夜间由于边界层高度降低，大气稳定性增强，污染物难以扩散，浓度也会维持在较高水平。午后由于太阳辐射增强，大气对流活跃，污染物得以稀释扩散，浓度通常处于日内低点。

【插图位置3：PM2.5历史趋势图截图】操作说明：在“数据洞察”页面截取“历史趋势回溯”区域的时间序列折线图，建议选择能够展示完整年周期的时间范围。

### 3.2 相关性分析可视化

为了识别影响PM2.5浓度的关键因素，我们绘制了各变量与PM2.5之间的相关系数热力图。热力图采用冷暖色调编码相关性强弱，暖色调（红色）表示正相关，冷色调（蓝色）表示负相关，颜色深浅反映相关系数的绝对值大小。

从热力图可以清晰看出，露点温度（DEWP）与PM2.5呈现较强的正相关关系，这意味着在空气湿度较大的情况下，PM2.5浓度往往也较高。这一现象可能源于以下机制：高湿度环境有利于二次气溶胶的生成，同时也会增强颗粒物的吸湿增长效应，导致PM2.5浓度升高。

风速 (lws) 与PM2.5呈现负相关关系，符合直觉认知——较大的风速有利于污染物的水平扩散和稀释，从而降低局地PM2.5浓度。这一发现为大气污染预警提供了参考：在预报显示持续静风或微风条件时，应提前发布污染预警，提醒公众做好健康防护。

气压 (PRES) 与PM2.5的相关性相对复杂。高气压系统通常伴随下沉气流，抑制污染物的垂直扩散，容易形成污染累积；而低气压系统则有利于大气垂直交换，促进污染物稀释。但气压与其他气象要素之间存在较强的共线性，其独立影响需要通过回归建模进一步剥离。

此外，我们还进行了VIF（方差膨胀因子）共线性诊断。当VIF值超过10时，表明该变量与其他变量存在较强的共线性，可能导致回归系数估计不稳定。诊断结果发现，温度 (TEMP) 和露点 (DEWP) 之间存在一定程度的共线性，在建模时需要考虑采用Ridge回归等正则化方法加以处理。

【插图位置4：相关性热力图截图】**操作说明：**在“数据洞察”页面右侧截取“相关性热力图”区域。

### 3.3 统计检验可视化

为了验证一些经验性假设，我们在系统中实现了假设检验模块，并通过可视化手段直观展示检验结果。

工作日效应检验采用独立样本T检验方法，比较工作日和周末PM2.5浓度的均值差异。从箱线图可以看出，工作日的PM2.5分布与周末存在一定差异，这反映了人类生产活动对空气质量的影响。T检验的p值结果能够告诉我们这种差异在统计学意义上是否显著。如果p值小于0.05，我们可以有95%的把握认为工作日与周末的空气质量确实存在差异，人类活动是影响PM2.5的重要因素。

风向效应检验采用单因素方差分析 (ANOVA) 方法，检验不同风向条件下PM2.5均值是否存在显著差异。初步结果表明，风向对PM2.5浓度有显著影响。西北风 (NW) 条件下PM2.5浓度普遍较低，这与西北方向气团来源洁净、污染物输送少有关；而东南风 (SE) 条件下PM2.5浓度较高，可能与气团携带工业区或城区排放的污染物有关。

【插图位置5：工作日vs周末箱线图及T检验结果截图】**操作说明：**在“数据洞察”页面点击展开“查看更多统计检验”区域，截取工作日vs周末的箱线图和统计显著性检验结果。

---

## 四、数据预处理与建模方法

### 4.1 回归模型体系

本系统实现了完整的回归模型体系，从经典方法到现代技术形成了由浅入深的递进结构。具体包括以下五种模型：

#### (1) OLS普通最小二乘回归（基准模型）

OLS方法通过最小化残差平方和来估计模型参数，在因变量服从正态分布、误差项同方差且相互独立等假设条件下，能够得到最优线性无偏估计 (BLUE)。我们以PM2.5为因变量，以温度、气压、露点、风速等气象因子为自变量构建OLS模型，系数的正负号反映影响方向，绝对值大小反映影响强度，配合t检验的p值可以判断各因子影响是否显著。

#### (2) Ridge回归与Lasso回归（正则化方法）

OLS方法在自变量存在多重共线性时，回归系数估计会变得不稳定。Ridge回归通过L2正则化对过大的系数施加惩罚，稳定参数估计；Lasso回归采用L1正则化，能够将不重要变量的系数压缩至零，自动实现特征选择功能。两种方法均通过交叉验证选择最优正则化参数。

### (3) GLM广义线性模型 (适应非正态分布)

考虑到PM2.5数据呈现右偏分布特征，我们引入了广义线性模型。与OLS不同，GLM允许因变量服从指数族分布。我们采用Gamma分布族配合对数链接函数，其优势在于保证预测值恒为正值，同时系数具有乘性解释——系数 $\beta$ 意味着自变量每增加一个单位，PM2.5期望值变为原来的 $\exp(\beta)$ 倍。

### (4) 贝叶斯岭回归 (量化不确定性)

贝叶斯方法将模型参数视为随机变量，通过先验分布和似然函数得到参数的后验分布。与频率学派方法仅给出点估计不同，贝叶斯方法能够自然地量化参数不确定性，给出可信区间而非置信区间，适合需要进行风险评估的应用场景。

## 4.2 分类模型体系

除了回归分析，我们还将PM2.5预测问题转化为分类问题。根据中国空气质量标准，我们将PM2.5浓度划分为三个等级：优良 ( $PM2.5 < 75 \mu\text{g}/\text{m}^3$ )、轻度污染 ( $75 \leq PM2.5 < 115 \mu\text{g}/\text{m}^3$ ) 和重度污染 ( $PM2.5 \geq 115 \mu\text{g}/\text{m}^3$ )。

Logistic回归是分类问题的经典方法。对于多分类问题，我们采用多项式Logistic回归 (Multinomial Logistic Regression)，通过softmax函数将线性预测值转换为各类别的概率。Logistic回归的优势在于模型简单、可解释性强，回归系数可以转换为优势比 (Odds Ratio) 进行解读。

朴素贝叶斯分类器 (Naive Bayes) 基于贝叶斯定理和特征条件独立性假设。对于连续型特征，我们假设每个类别下的特征服从高斯分布。尽管条件独立性假设在实际中往往不成立，但朴素贝叶斯分类器在许多实际问题中仍能取得不错的表现，特别是在训练数据较少的情况下具有良好的稳健性。该方法还能自然地输出类别概率，为决策提供更丰富的信息。

## 4.3 模型训练与评估策略

为了确保模型评估结果的可靠性和泛化能力的准确估计，我们采用了训练集-测试集分割策略。默认将数据按照80:20的比例随机划分为训练集和测试集：训练集用于模型参数估计，测试集用于评估模型在未见数据上的预测性能。分类问题中还采用了分层抽样 (Stratified Sampling)，确保训练集和测试集中各类别的样本比例与原始数据一致，避免因类别不平衡导致的评估偏差。

用户可以通过侧边栏灵活调整测试集比例 (10%-40%) 和随机种子，前者影响模型训练的数据量和评估的稳定性，后者确保实验结果的可复现性。

---

## 五、模型性能评估与优化

### 5.1 回归模型评估指标

回归模型的评估采用多维度指标体系，从不同角度全面衡量模型性能：

#### 预测误差指标：

- **RMSE (均方根误差)**：衡量预测值与真实值之间的平均偏差，与因变量具有相同量纲，便于直观理解预测精度。
- **MAE (平均绝对误差)**：计算预测误差绝对值的平均，相比RMSE对极端误差不那么敏感，更能反映典型误差水平。

## 拟合优度指标:

- **R<sup>2</sup> (决定系数)** : 表示模型解释的方差占总方差的比例, R<sup>2</sup>越接近1, 拟合程度越好。但需注意R<sup>2</sup>会随模型复杂度增加而增大, 可能导致过拟合。

## 模型选择指标:

- **AIC (赤池信息量准则)** : 通过对数似然值和参数个数构建惩罚函数, AIC值越小表示模型越优。
- **BIC (贝叶斯信息准则)** : 对参数个数施加更强惩罚, 大样本时倾向于选择更简洁的模型。

## 残差诊断:

残差分析是回归诊断的重要组成部分。我们通过残差分布直方图、残差与预测值散点图以及Q-Q图检验残差是否满足正态性、同方差性和独立性假设。Durbin-Watson检验用于检测残差自相关性, DW统计量接近2表示无显著自相关。

【插图位置6: 回归模型性能对比表格截图】操作说明: 在"模型竞技场"页面运行所有模型后, 截取"模型性能对比 (测试集评估)"表格区域, 包含OLS、Ridge、Lasso、GLM、Bayesian Ridge五个模型的RMSE、MAE、R<sup>2</sup>、AIC、BIC指标。

## 5.2 分类模型评估指标

分类模型的评估同样采用多指标综合评价体系:

### 基础分类指标:

- **Accuracy (准确率)** : 正确分类样本占总样本的比例, 是最直观的评价指标。但在类别不平衡情况下可能产生误导。
- **Precision (精确率)** : 预测为正类的样本中实际为正类的比例, 反映"查准"能力。
- **Recall (召回率)** : 实际为正类的样本中被正确识别的比例, 反映"查全"能力。在污染预警中, 召回率尤为重要——宁可虚警也不愿漏报。
- **F1-Score**: 精确率和召回率的调和平均, 在两者之间取得平衡。

### 多分类汇总指标:

对于多分类问题, 我们计算宏平均 (Macro Average) 和微平均 (Micro Average) 两种汇总方式。宏平均对各类别赋予相同权重, 适合关注少数类表现; 微平均按样本量加权, 反映整体分类效果。

### 可视化评估工具:

- **混淆矩阵**: 以表格形式展示各类别的分类结果, 对角线元素为正确分类数量, 可细致分析模型的分类偏好和错误模式。
- **ROC曲线与AUC**: 以假正率为横轴、真正率为纵轴绘制曲线, 曲线下面积AUC衡量整体分类性能。AUC越接近1, 分类能力越强。

【插图位置7: 分类模型性能对比表格和混淆矩阵截图】操作说明: 在"分类与状态"页面运行分类模型后, 截取"分类模型性能对比 (测试集评估)"表格以及Logistic Regression和Naive Bayes的混淆矩阵热力图。

## 5.3 模型选择与优化

基于上述评估结果, 我们对各模型进行综合比较和选择。回归模型中, GLM模型因其能够适应PM2.5数据的非正态分布特性, 通常在AIC/BIC指标上表现优异。Ridge和Lasso回归通过正则化有效缓解了多重共线性问题, 在

存在高度相关特征的情况下表现稳健。贝叶斯回归则提供了参数不确定性的量化信息，适合需要进行风险评估的应用场景。

分类模型中，Logistic回归和朴素贝叶斯各有优势。Logistic回归的可解释性更强，系数具有明确的统计含义；朴素贝叶斯在训练效率和小样本表现上具有优势。实际应用中应根据具体需求选择合适的模型。

【插图位置8：ROC曲线对比图截图】操作说明：在“分类与状态”页面或“评估中心”页面截取ROC曲线对比图区域。

---

## 六、模型解读与实际应用分析

### 6.1 回归系数解读

回归模型的核心价值在于量化各因素对PM2.5的贡献程度。以GLM模型（Gamma分布族+对数链接函数）为例，我们可以将回归系数转换为实际含义丰富的解读。

对于对数链接函数，系数 $\beta$ 的解读需要通过指数变换。具体而言，当某气象因子增加一个单位时，PM2.5的期望值将变为原来的 $\exp(\beta)$ 倍。如果系数为正， $\exp(\beta) > 1$ ，表示该因子增加会导致PM2.5浓度上升；如果系数为负， $\exp(\beta) < 1$ ，表示该因子增加会导致PM2.5浓度下降。相对变化百分比可以计算为 $(\exp(\beta)-1) \times 100\%$ 。

以温度为例，假设其回归系数为0.015，则 $\exp(0.015)-1 \approx 0.0151$ ，意味着温度每升高1摄氏度，PM2.5浓度的期望值约增加1.5%。这一结果与气象学知识一致：较高的温度有利于光化学反应进行，促进二次气溶胶的生成，从而导致PM2.5浓度升高。

风速的系数通常为负值，反映了风的稀释扩散作用。假设风速系数为-0.08，则 $\exp(-0.08)-1 \approx -0.077$ ，意味着风速每增加1米/秒，PM2.5浓度的期望值约下降7.7%。这一发现具有明确的政策含义：在预报显示持续静风条件时，应加强污染源管控，降低排放强度。

系数解读时还需关注显著性水平。 $p$ 值小于0.05的系数在统计学意义上显著，其影响可以较为确定地归因于该因子； $p$ 值较大的系数则可能是由随机波动导致，解读时需谨慎。系统在可视化时采用颜色编码区分显著与不显著的系数，帮助用户快速识别关键因子。

【插图位置9：GLM系数可视化条形图截图】操作说明：在“归因分析”页面选择GLM模型运行后，截取系数可视化条形图（绿色表示显著，灰色表示不显著）。

### 6.2 贝叶斯方法的参数不确定性分析

贝叶斯回归相比频率学派方法的独特优势在于能够量化参数的不确定性。系统展示了各回归系数的后验分布信息，包括后验均值和后验标准差。

后验均值可以理解为参数的“最佳估计”，其含义与OLS系数类似。后验标准差则反映了对该估计的确信程度——标准差越小，表示对估计结果越有信心；标准差越大，表示存在较大的不确定性。

基于后验分布，我们可以构建可信区间（Credible Interval）。95%可信区间意味着参数落在该区间内的概率为95%，这与频率学派95%置信区间的解释有本质区别。可信区间的直接概率解释更符合决策者的直觉需求，在风险评估和不确定性决策中具有优势。

通过比较贝叶斯估计与频率学派估计的差异，我们还可以评估正则化效应的强度。贝叶斯岭回归隐含地引入了参数的先验分布，其效果类似于Ridge回归的L2正则化。当两种方法的估计结果差异较大时，说明数据中存在共线性或过拟合风险，正则化起到了重要的稳定作用。

【插图位置10：贝叶斯后验分布可视化截图】操作说明：在“模型竞技场”页面运行模型后，截取“贝叶斯方法：参数后验分布”区域的系数均值和重要性条形图。

## 6.3 分类预测的实际应用

分类模型的输出可以直接应用于空气质量预警决策。根据气象预报数据输入模型，可以预测未来一段时间内空 气质量等级的概率分布，为公众出行和健康防护提供参考。

当模型预测“重度污染”的概率超过某一阈值（如30%）时，可以触发预警机制，提醒敏感人群减少户外活动、佩戴防护口罩。当预测“优良”的概率较高时，则可以推荐适宜的户外运动时段。

分类模型还可以辅助环境管理部门的决策。当预测显示污染风险较高时，可以提前启动应急减排措施，如限制工业排放、实施机动车限行等，将污染遏制在发生之前。这种“预防性管控”策略相比事后应急更加经济有效。

---

# 七、系统创新性与改进方向

## 7.1 系统创新特点

本项目在技术实现和应用设计上体现了若干创新特点。

首先，系统构建了完整的统计建模流程，从数据预处理、探索性分析、统计推断、回归建模到分类预测，形成了“由浅入深”的递进式分析框架。这种设计不仅符合统计分析的一般规范，也便于用户理解各环节的逻辑关系和方法原理。

其次，系统实现了多模型对比分析功能。用户可以在同一界面下运行OLS、Ridge、Lasso、GLM、贝叶斯回归五种回归模型，以及Logistic回归、朴素贝叶斯两种分类模型，通过统一的评估指标体系进行横向比较。这种设计有助于深入理解各方法的优劣势和适用场景，培养模型选择的实践能力。

第三，系统强调了Train/Test分割策略的重要性。所有模型评估指标均基于测试集计算，真实反映模型的泛化能力，避免了在训练集上评估导致的过度乐观估计。这一设计体现了机器学习和统计建模的最佳实践。

第四，系统采用Streamlit框架构建交互式Web界面，用户无需编写代码即可完成全流程分析。这种设计降低了统计分析的技术门槛，使非专业背景的用户也能借助系统获取数据洞见。

## 7.2 潜在改进方向

尽管系统已经实现了较为完整的功能，但仍存在若干可以进一步完善的方向。

在时间序列建模方面，当前系统主要关注静态回归分析，未充分利用数据的时序结构。未来可以引入ARIMA、季节性分解、状态空间模型等时间序列方法，对PM2.5的动态演化规律进行建模和预测。这将使系统具备短期预报功能，进一步提升实用价值。

在特征工程方面，当前系统仅使用原始气象观测变量作为特征。未来可以构建更多衍生特征，如气象变量的滞后项、移动平均、波动率等，以及交互项和多项式特征，捕捉变量之间的非线性关系和动态效应。

在模型集成方面，可以探索将多个单一模型的预测结果进行加权组合，通过模型平均或Stacking等技术提升预测精度和稳健性。集成方法往往能够综合各模型的优势，获得比单一模型更好的表现。

在系统交互设计方面，可以增加更多的用户定制选项，如自定义污染等级阈值、选择性包含/排除特征变量、调整可视化样式等，提升系统的灵活性和适应性。

## 八、团队分工与协作

本项目由五位成员共同完成，采用模块化分工与协作开发的模式。在项目启动阶段，团队根据系统功能架构进行任务拆解，每位成员负责一到两个核心模块的开发工作。分工的基本原则是“功能独立、接口清晰”，确保各模块可以并行开发而不产生代码冲突。

### 8.1 功能模块分工

#### 张洛梧 —— 原型开发与系统架构

张洛梧负责项目的整体架构设计和原型开发工作。在项目初期，他搭建了基于Streamlit的Web应用框架，确定了系统的页面结构和导航逻辑。主要工作包括：设计系统的五大功能页面布局，实现侧边栏参数配置机制，编写CSS样式美化界面，以及构建数据加载和缓存机制。他还负责将其他成员开发的功能模块整合到主程序app.py中，确保各模块之间的接口对接顺畅。

#### 许奕 —— 数据预处理与数据洞察

许奕负责数据处理的前端工作，主要开发了“数据洞察”页面的全部功能。具体包括：实现data\_preprocessing.py模块中的缺失值插值、异常值检测（ $3\sigma$ 原则和IQR方法）、对数变换和正态性检验等功能；开发statistical\_inference.py模块中的T检验和ANOVA分析；设计并实现数据概览指标卡片、PM2.5历史趋势图、相关性热力图、工作日与周末对比箱线图等可视化组件。这些工作为后续的建模分析提供了高质量的数据基础。

#### 王天一、肖相宇 —— 归因分析与模型竞技场

王天一和肖相宇共同负责回归建模相关模块的开发。两人合作完成了“归因分析”和“模型竞技场”两个核心页面。具体分工如下：王天一主要负责glm\_model.py广义线性模型的实现，包括Gamma分布族拟合、对数链接函数、系数解释等功能；肖相宇主要负责regression\_models.py模块，实现了OLS、Ridge、Lasso三种回归模型以及模型对比功能。两人共同开发了bayesian\_models.py贝叶斯回归模块和feature\_selection.py特征选择模块，并合作设计了模型性能对比表格、残差分析图、VIF共线性诊断等可视化功能。

#### 李星烁 —— 分类与状态模块

李星烁负责分类建模相关功能的开发，主要工作集中在“分类与状态”页面。他开发了classification\_models.py模块，实现了Logistic回归和朴素贝叶斯两种分类算法，设计了PM2.5浓度到空气质量等级的转换逻辑（优良/轻度污染/重度污染三分类）。此外，他还负责model\_evaluation.py评估模块中分类相关指标的计算，包括混淆矩阵、ROC曲线、AUC值、精确率、召回率、F1-Score等，并实现了对应的可视化功能。

### 8.2 协作与版本控制

团队使用Git进行版本控制，通过GitHub仓库实现代码共享和协作。

---

## 九、项目总结与反思

### 9.1 项目成果总结

通过本项目的开发与实践，我们完成了以下主要成果：

在技术实现方面，我们构建了一个功能完整的城市空气质量分析系统，涵盖数据预处理、可视化探索、统计推断、回归建模、分类预测和模型评估六大核心功能模块。系统代码结构清晰，模块化程度高，共计实现10余个

Python模块，代码量超过2000行。

在课程覆盖方面，系统实现了统计分析与建模课程的全部核心章节要求，包括数据预处理、分布拟合、假设检验、方差分析、相关分析、回归分析、广义线性模型、贝叶斯方法、分类模型等，形成了理论与实践相结合的完整学习案例。

在应用价值方面，系统以北京PM2.5数据为案例，深入分析了气象因素对空气质量的影响机制，得出了若干有意义的研究结论。系统的交互式设计使其不仅可用于教学演示，也具备一定的实际应用潜力。

## 9.2 经验与反思

项目开发过程中，我们积累了若干宝贵经验，也发现了一些值得反思的问题。

在技术选型方面，Streamlit框架为快速构建数据应用提供了极大便利，但其交互模式也带来了状态管理的挑战。页面刷新会导致计算结果丢失，需要借助session\_state机制进行状态持久化。这一问题在开发初期造成了一些困扰，后通过重构代码得以解决。

在模型选择方面，我们深刻体会到“没有免费的午餐”定理的含义——没有哪个模型在所有情况下都是最优的。不同模型基于不同的假设和优化目标，适用于不同的数据特征和应用场景。通过多模型对比分析，我们能够更全面地理解数据特性，做出更合理的模型选择。

在团队协作方面，清晰的接口规范和充分的前期设计是并行开发成功的关键。开发初期我们在接口定义上投入了较多时间，但这些投入在后续集成阶段得到了回报——模块之间能够顺利对接，减少了返工成本。

## 9.3 改进方向展望

展望未来，本项目可以在以下几个方向进一步拓展：

可以引入更丰富的数据源，如多站点监测数据、卫星遥感数据、污染源清单数据等，构建空间分析功能，揭示污染的区域传输规律。

可以开发移动端应用或微信小程序，将空气质量预测服务推送给更广泛的公众用户，提升系统的社会影响力。

可以与气象部门或环保部门合作，获取实时监测数据流，将系统升级为在线预警平台，为空气质量管理决策提供技术支撑。

---

# 十、AI工具使用情况介绍

## 10.1 AI辅助开发

在本项目的开发过程中，我们适度使用了AI辅助工具来提升开发效率。

在代码编写阶段，AI工具帮助我们快速生成了一些重复性较高的代码框架，如数据处理的基础函数、可视化图表的样板代码等。但核心的算法逻辑和业务设计仍由团队成员独立完成，AI工具主要起到加速作用而非替代作用。

在文档撰写阶段，AI工具协助进行了语句润色和格式规范化，提升了文档的可读性。但文档的内容架构、核心观点和技术描述均由团队成员根据项目实际情况撰写，确保文档内容的准确性和原创性。

在Debug阶段，AI工具帮助分析了一些难以定位的错误信息，提供了可能的修复方向。但最终的问题诊断和修复仍需要团队成员理解代码逻辑后手动完成。

## 10.2 使用原则与反思

我们在使用AI工具时遵循以下原则：

**辅助而非替代**: AI工具是提升效率的辅助手段，而非替代人类思考和创造的工具。核心的设计决策、算法选择、结果解读等仍由团队成员完成。

**理解而非照搬**: 对于AI生成的代码或建议，我们会先理解其原理和逻辑，确认正确性后才纳入项目，而非盲目复制粘贴。

通过本项目的实践，我们认识到AI工具是一把双刃剑。合理使用可以显著提升开发效率，但过度依赖则可能削弱自身的理解和思考能力。在未来的学习和工作中，我们将继续探索人机协作的最佳模式，充分发挥AI工具的优势，同时保持独立思考和创新能力。

---

## 附录

### 附录A：系统运行说明

**环境要求**: Python 3.8+

**依赖安装**:

```
pip install -r requirements.txt
```

**启动系统**:

```
streamlit run Code/app.py
```

或使用启动脚本:

- Windows: 双击 `run_app.bat`
- Linux/Mac: 执行 `./run_app.sh`

**访问地址**: 浏览器打开 `http://localhost:8501`

### 附录B：主要依赖库

库名	版本	用途
streamlit	$\geq 1.28.0$	Web应用框架
pandas	$\geq 2.0.0$	数据处理
numpy	$\geq 1.24.0$	数值计算
statsmodels	$\geq 0.14.0$	统计建模
scipy	$\geq 1.10.0$	统计检验

库名	版本	用途
scikit-learn	≥1.3.0	机器学习
matplotlib	≥3.7.0	静态可视化
seaborn	≥0.12.0	统计可视化
plotly	≥5.14.0	交互式可视化

## 附录C：项目文件结构

```
Stat-Modeling/
├── Code/
│   ├── __init__.py
│   ├── app.py
│   ├── data_preprocessing.py
│   ├── statistical_inference.py
│   ├── feature_selection.py
│   ├── regression_models.py
│   ├── bayesian_models.py
│   ├── classification_models.py
│   ├── model_evaluation.py
│   └── glm_model.py
├── Data/
│   ├── beijing+pm2+5+data.zip
│   └── README.md
└── Docs/
    ├── 统计分析文档.md
    ├── 项目增强说明.md
    └── 项目提案-终极增强版.md
├── README.md
└── requirements.txt
└── run_app.bat
└── run_app.sh
```

# 核心代码模块  
# 包初始化文件  
# Streamlit主应用  
# 数据预处理模块  
# 统计推断模块  
# 特征选择模块  
# 回归模型模块  
# 贝叶斯模型模块  
# 分类模型模块  
# 模型评估模块  
# 广义线性模型模块  
# 数据目录  
# UCI北京PM2.5数据集  
# 数据说明文档  
# 文档目录  
# 本文档  
# 功能增强说明  
# 项目提案  
# 项目说明  
# 依赖清单  
# Windows启动脚本  
# Linux/Mac启动脚本