成人死亡率预测- 程序报告

1 实验介绍

1.1 实验背景

成年人死亡率指的是每一千人中15岁至60岁死亡的概率（数学期望）。这里我们给出了世界卫生组织（WHO）下属的全球卫生观察站（GHO）数据存储库跟踪的所有国家健康状况以及许多其他相关因素。要求利用训练数据建立回归模型，并预测成年人死亡率（**Adult Mortality**）。

1.2 实验要求

1. 训练数据包含`2336`条记录和`22`个字段，对训练数据进行一定的可视化数据分析（章节2.2）
2. 利用训练数据，选择合适的信息作为特征建立回归模型，并预测测试数据成年人死亡率
3. 利用`MO`平台进行模型性能评估

1.3 实验环境

使用基于 Python 的`Pandas`库进行数据相关处理，使用`Sklearn`库进行相关模型构建。

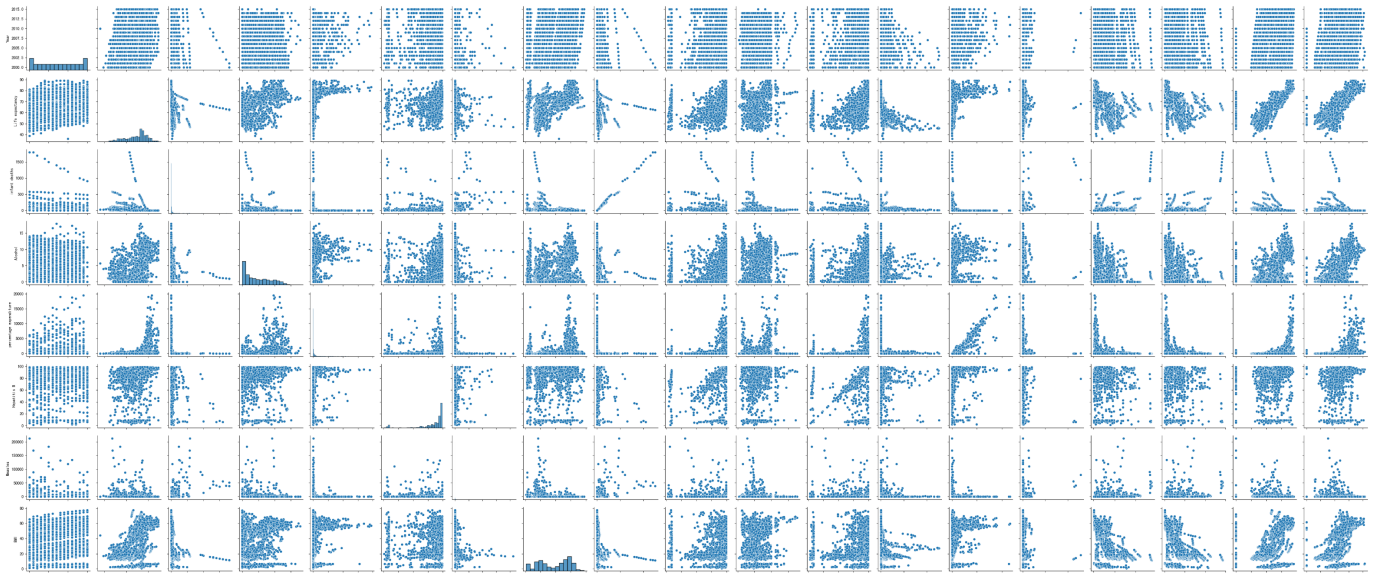
2 实验部分

2.1 数据读取和可视化分析

1. 皮尔森相关性矩阵

	Year	Life expectancy	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	under-five deaths	Polio	Total expenditure	Diphtheria	HIV/AIDS	GDP	Population	thinness 1-19 years	thinness 5-9 years	Income composition of resources	Schooling
Year	1.000000	0.173044	-0.045574	-0.059393	0.027351	0.119861	-0.089535	0.097642	-0.051892	0.099639	0.095594	0.146898	-0.144159	0.094351	0.012174	-0.049637	-0.055414	0.244558	0.219591
Life expectancy	0.173044	1.000000	-0.200307	0.403930	0.385335	0.250026	-0.156621	0.563239	-0.227572	0.470405	0.227194	0.500060	-0.558550	0.466012	-0.028874	-0.481919	-0.476773	0.720200	0.756970
infant deaths	-0.045574	-0.200307	1.000000	-0.117509	-0.088942	-0.233426	0.502258	-0.227356	0.996460	-0.168905	-0.127455	-0.186393	0.027106	-0.112767	0.559472	0.462870	0.466544	-0.148067	-0.197907
Alcohol	-0.059393	0.403930	-0.117509	1.000000	0.351033	0.076217	-0.045272	0.328996	-0.114326	0.223054	0.297324	0.225653	-0.053149	0.367267	-0.032890	-0.437330	-0.425520	0.451557	0.545457
percentage expenditure	0.027351	0.385335	-0.088942	0.351033	1.000000	0.020959	-0.056564	0.243372	-0.091140	0.153229	0.196985	0.149584	-0.102281	0.908855	-0.030910	-0.257604	-0.258956	0.386001	0.399577
Hepatitis B	0.119861	0.250026	-0.233426	0.076217	0.020959	1.000000	-0.122937	0.166267	-0.243389	0.470815	0.054130	0.610947	-0.112097	0.084776	-0.125556	-0.128664	-0.130000	0.196330	0.214588
Measles	-0.089535	-0.156621	0.502258	-0.045272	-0.056564	-0.122937	1.000000	-0.176633	0.510569	-0.143172	-0.106506	-0.151227	0.033809	-0.078137	0.248298	0.224253	0.217004	-0.133857	-0.142477
BMI	0.097642	0.563239	-0.227356	0.328996	0.243372	0.166267	-0.176633	1.000000	-0.237654	0.287353	0.231999	0.304195	-0.243694	0.318139	-0.074910	-0.531130	-0.535272	0.507016	0.562572
under-five deaths	-0.051892	-0.227572	0.996460	-0.114326	-0.091140	-0.243389	0.510569	-0.237654	1.000000	-0.187704	-0.128945	-0.208636	0.040501	-0.116458	0.545763	0.463884	0.467471	-0.167404	-0.214304
Polio	0.099639	0.470405	-0.168905	0.223054	0.153229	0.470815	-0.143172	0.287353	-0.187704	1.000000	0.133347	0.666200	-0.165835	0.219403	-0.036908	-0.237836	-0.237604	0.380321	0.419043
Total expenditure	0.095594	0.227194	-0.127455	0.297324	0.196985	0.054130	-0.106506	0.231999	-0.128945	0.133347	1.000000	0.149584	-0.010628	0.171697	-0.078890	-0.273844	-0.282508	0.164861	0.256631
Diphtheria	0.146898	0.500060	-0.186393	0.225653	0.149584	0.610947	-0.151227	0.304195	-0.208636	0.666200	0.149584	1.000000	-0.182804	0.210663	-0.034762	-0.253593	-0.245601	0.415212	0.439829
HIV/AIDS	-0.144159	-0.558550	0.027106	-0.053149	-0.102281	-0.112097	0.033809	-0.243694	0.040501	-0.165835	-0.010628	-0.182804	1.000000	-0.141261	-0.025819	0.209907	0.214169	-0.249723	-0.228113
GDP	0.094351	0.466012	-0.112767	0.367267	0.908855	0.084776	-0.078137	0.318139	-0.116458	0.219403	0.171697	0.210663	-0.141261	1.000000	-0.036506	-0.294106	-0.299203	0.464682	0.457766
Population	0.012174	-0.028874	0.559472	-0.032890	-0.030910	-0.125556	0.248298	-0.074910	0.545763	-0.036908	-0.078890	-0.034762	-0.025819	-0.036506	1.000000	0.262519	0.261089	-0.013043	-0.035931
thinness 1-19 years	-0.049637	-0.481919	0.462870	-0.437330	-0.257604	-0.128664	0.224253	-0.531130	0.463884	-0.237836	-0.273844	-0.253593	0.209907	-0.294106	0.262519	1.000000	0.937984	-0.424126	-0.482056
thinness 5-9 years	-0.055414	-0.476773	0.466544	-0.425520	-0.258956	-0.130000	0.217004	-0.535272	0.467471	-0.237604	-0.282508	-0.245601	0.214169	-0.299203	0.261089	0.937984	1.000000	-0.414132	-0.471809
Income composition of resources	0.244558	0.720200	-0.148067	0.451557	0.386001	0.196330	-0.133857	0.507016	-0.167404	0.380321	0.164861	0.415212	-0.249723	0.464682	-0.013043	-0.424126	-0.414132	1.000000	0.796696
Schooling	0.219591	0.756970	-0.197907	0.545457	0.399577	0.214588	-0.142477	0.562572	-0.214304	0.419043	0.256631	0.439829	-0.228113	0.457766	-0.035931	-0.482056	-0.471809	0.796696	1.000000

2. 数据依赖关系(完整结果见文件)



2.2 模型拟合与成人死亡率预测

我测试了以下10种模型：

1. LinerRegression:
2. Ridge
3. LassorLars
4. BayesianRidge
5. SVR
6. GradientBoostingRegressor
7. KernelRidge
8. XGBRegressor
9. ElasticNet
10. SGDRegressor

结果如图所示：

```
0th regressor, MSE=7451.44129449904, R2=0.5207060487947698
1th regressor, MSE=7486.412310976235, R2=0.518456632070868
2th regressor, MSE=7504.45733556351, R2=0.5172959343223109
3th regressor, MSE=7485.569668840523, R2=0.5185108327634207
4th regressor, MSE=10134.962969701097, R2=0.348095723353167 63
5th regressor, MSE=4028.7416201308424, R2=0.7408620140478084
6th regressor, MSE=11126.399294964815, R2=0.2843242441287597
7th regressor, MSE=146.11832919181336, R2=0.9906013011932373
8th regressor, MSE=13239.635177816112, R2=0.1483960208374876
9th regressor, MSE=7600.925729947412, R2=0.5110908639093132
```

可以看到，5th GradientBoostingRegressor和7th XGBRegressor效果较好，但经过验证集发现，XGBRegressor实际上过拟合了，因此选择了GradientBoostingRegressor