根据KPSS检验1久远的年份后，我们对表格5（programs.xlsx）进一步处理：按Discipline去除从1932年以来未能再次举办的项目（这里还对应前面一个假设，未举办过的运动在后面2028年也不会举办），随后剔除这些数据后还剩下53个discipline（项目）

第一问首先我们聚类后，将类别也标在表中，随后构建53维向量（分别代表某个国家在某次奥运会参加53个events的次数），并于是否东道主国家增广成为54维向量作为特征向量，将得到的金银铜牌个数作为三维的目标向量。

就是根据运动员表格，把年内某国统计出参加了哪些events（运行项目表格（表格中是项目简写）的53个events没参加就是0，参加多少次就是几，类似于生物生态学的生态位，占位子），这些都写到统计国家在某一年金银奖牌个数的表格，（其中Is\_host是与处理后的数据代表是否东道主）

第0类：若表现国家（大量，得分低甚至0金牌）我们测试了支持向量机、随机森林对抗生成神经网络以及SHAP+随机森林。我们发现支持向量机和随机森林由于样本标签分布范围比较小所以准确率比较低（给数据），这也许是模型对差异较小数据的分辨率不足的问题。观察对抗生成神经网络损失曲线发现当训练到一定次数的时候，鉴别器会明显强于生成器，精确率（precision)、召回率（recall）、F1值急剧下降，说明梯度爆炸造成模型崩溃。

第1类：强表现国家（只有几个）运用支持向量机模型并进行3-折交叉验证（因为样本少所以随机森林效果不好）

第2类：一般国家（数量适中）采取随机森林预测合适，将54维向量与3维目标向量进行训练得到模型，

（1）

第一小问给出国家和参加的项目后可以快速选出类别，不考虑国家的体育文化偏好（比如中国乒乓球很强），然后分别对三类数据根据相应的模型训练的模型预测目标向量（金银铜-->总），并获得预测的整数范围及其概率分布

预测目标向量与2024届向量相减，先看金牌再看奖牌的顺序确定强弱（题干也是这么做的），就是变强变弱（也可以让（x1,x2,x3）与）

（2）

我们通过对运动员表格的筛选并剔除已经不存在的国家，我们发现有将近70个国家从未获得过奖牌，这些国家在下次奥运会中每个国家最多产生一个人获得首枚奖牌该国有没有这个人获得首枚奖牌与该国最终结束奥的时候是否获得过奖牌是充要的，因此有多少人获得首枚，也就等同于夏季奥运会结束的时候，这些国家有多少个国家获得了奖牌。我们截取1932年以来的数据，（这里我们做了一个假设：每一个国家我们都用他最近一次参与奥运会的参赛情况，作为下次参赛奥运会的参赛情况（人次））通过观察发现这些国家都只参加一项运动而且只参加1人次且从未作为过东道主国家，因此根据聚类特征分布这些国家是第（1）小问里面的第0类里面的特例（由于每个国家只参1人次的项目且Is\_host始终为0，对这些国家的参赛情况所对应的特征向量类似于读热编码向量），运用对应训练好的模型，模拟输入符合该类国家特征的特征向量得到预测值，预测区间及其概率分布（我们将该概率分布转化为该国最终是否获得过奖牌的概率分布，是每个国家分别得奖牌和不得奖牌的概率，在对这将近70个国家进行概率上的预测）也就能估计出下届奥运会有多少人获得首枚奖牌并根据这个概率得到其可能性。

（3）

我们前面的问题运用的特征向量前53维中的内容是每个项目参加多少人次，这仅仅是一个参与度方面的问题，并不能体现国家的能力层面上的问题。对此我们根据动员表格，将上面54维向量内前53维关于这个国家参加多少人次、获得奖牌的成绩通过Topsis模型获得三种奖牌分别相对于获得三种奖牌，分别相对于参与人次的评分构建出一种评分体系：

Gold Medal TOPSIS Score: 1.5178245702589412

Silver Medal TOPSIS Score: 1.356259715599741

Bronze Medal TOPSIS Score: 1.0031494248878818

Events Medal TOPSIS Score: 1.0

值得注意的是如果一个运动员获得了一枚奖牌。他肯定同时获得了参与这次项目的参与分数，比如一个获得金牌的运动员，在获得金牌之后，该运动员分数应该为1.517+1.0=2.517。

我们再次对运动员表格中的奖牌儿和项目进行统计,得到在一定的年份下某个国家在。某个项目下的总评分值得注意的是该评分是这个国家在这一年在人次上的总评分，得到**基于人次人次的积分表\***。我们应该注意到有些项目是多人参与的，比如一场足球赛每支队伍有11人上场如果足球赛举办两场，那么该国就参与该项目22人次，因此我们收集了[en.wikipedia.org](https://en.wikipedia.org/wiki/Olympic_sports)下对应项目超链接内的每场比赛上场人数，得到**基于参赛场次的积分表**。将复合数值重新载入向量前53维，得到修正后的特征向量，这些获奖情况能反映一个国家对某些运动的偏好，获得的新向量自然能够反映国家体育文化倾向，第54维仍为If\_host。

对各个国家来说，什么运动最重要？（综合评分最高的）为什么？

显然每一个向量中除了If\_host外数值最大的维度所对应的运动项目就是这个国家在该年度最重要的运动项目。我们以2024年各参赛国的向量为例，做了一个什么图，该图颜色越深的位置数值越大，对应的该项目对这个国家越重要。

母国选择的项目如何影响结果？

~~对于东道主的国家，我们只需要训练第54维为1的向量与其对应的目标向量即可，这样可以将54维向量简化为53维（若有全为0的某一维度可也以更简化）。~~

要分析主办国选择的项目如何影响比赛结果，检查主办国在特定项目上的表现是否优于其他国家。这可以通过比较主办国在其选择的项目与其他项目上的奖牌数来实现，我们以2024年法国巴黎奥运会为例：

