第一问：

1. 将势头分为四个影响因素，分别为心理因素、体能因素、技术因素、发球因素。

各个因素的计算，下面以心理因素计算为例：

**p1\_ace、p2\_ace**、**p1\_winner**、**p2\_winner、p1\_double\_fault、p2\_double\_fault、p1\_unf\_err、p2\_unf\_err、p1\_break\_pt\_won、p2\_break\_pt\_won、p1\_break\_pt\_missed、p2\_break\_pt\_missed这些数据是心理因素X。其中p1\_ace、p2\_ace**、**p1\_winner**、**p2\_winner、p1\_break\_pt\_won、p2\_break\_pt\_won这些因素对它所属的选手（p1或者p2）来说是获得胜利的积极因素，相反对另一方选手来说则是消极因素。而p1\_double\_fault、p2\_double\_fault、p1\_unf\_err、p2\_unf\_err、p1\_break\_pt\_missed、p2\_break\_pt\_missed这些因素对它所属的选手（p1或者p2）来说是获得胜利的消极因素，相反对另一方选手来说则是积极因素。X与p1\_ace、p2\_ace**、**p1\_winner**、**p2\_winner、p1\_double\_fault、p2\_double\_fault、p1\_unf\_err、p2\_unf\_err、p1\_break\_pt\_won、p2\_break\_pt\_won、p1\_break\_pt\_missed、p2\_break\_pt\_missed这些因素是线性的关系。**

**即X=w1\*p1\_ace+w2\*p2\_ace+...+err(w1、w2这些参数正负号由积极因素还是消极因素来决定，其中err为误差值，即加入的随机扰动)**

**在其他3个影响因素中要考虑的正负号如下：**

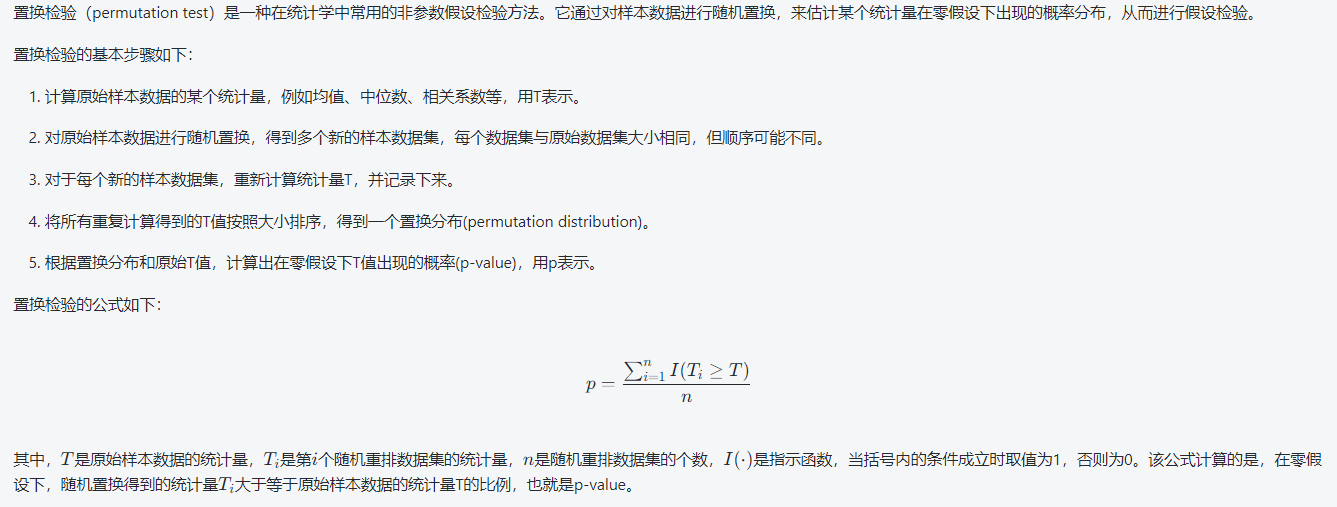
**体能因素（p1\_distance\_run、p2\_distance\_run、rally\_count、speed\_mph，其中p1\_distance\_run、p2\_distance\_run需要考虑正负）：跑动多对自身是消极因素，对对方是积极因素。同样是线性关系，需要加上一个err。**

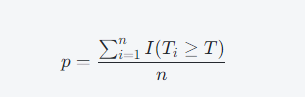
**技能因素（serve\_depth\_score、serve\_width\_score、return\_depth\_score，这些因素都要考虑正负）：若p1为发球方，那么计算其p1\_skill\_factors的线性公式中serve\_depth\_score、serve\_width\_score的权值则为正数，return\_depth\_score的权值为负数。 p1不为发球方则相反。p2也是和p1同理的。同样是线性关系，需要加上一个err。**

**发球因素（server和serve\_no）:** **谁是server谁就是积极因素（对应加号），对另一方就是消极因素（对应减号），若该server的serve\_no为2则是二次发球，此时serve\_no的符号就是减号（消极因素），若为1则是第一次发球，此时是加号（积极因素）。不是server就与上面相反。同样是线性关系，需要加上一个err。**

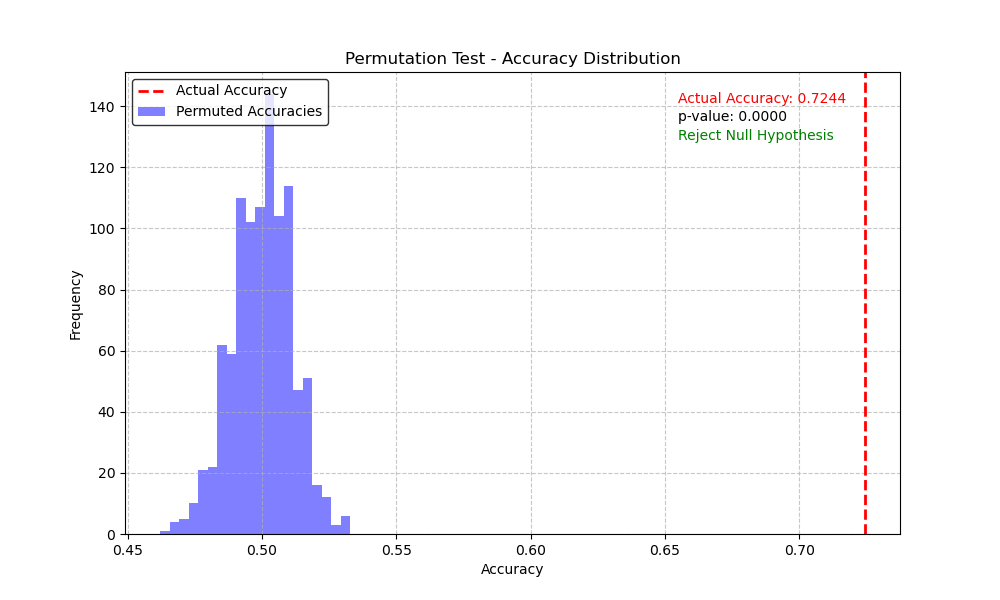
1. **总的势头与心理因素、体能因素、技术因素、发球因素四个方面是线性关系再加上一个其他因素（随机扰动作为干扰）。**

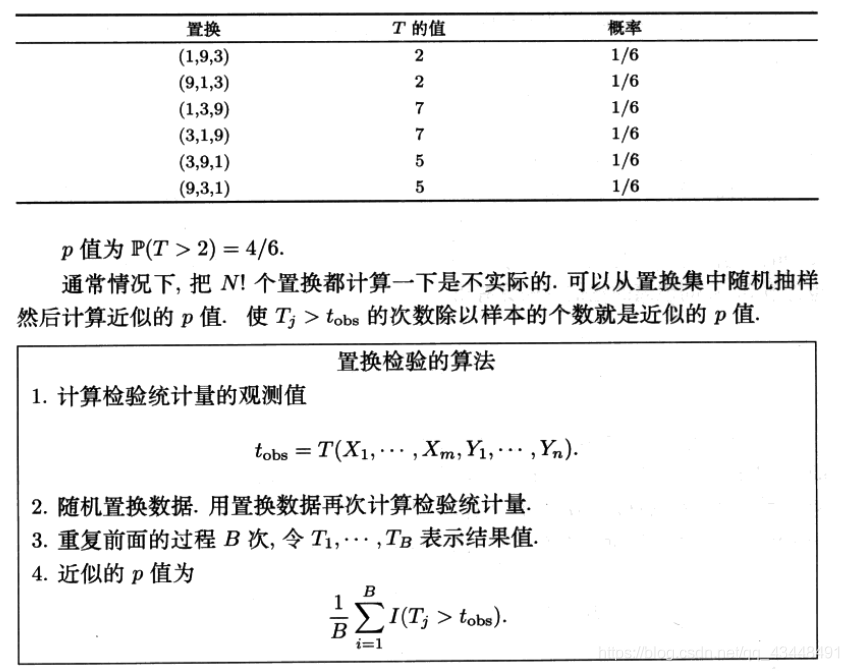
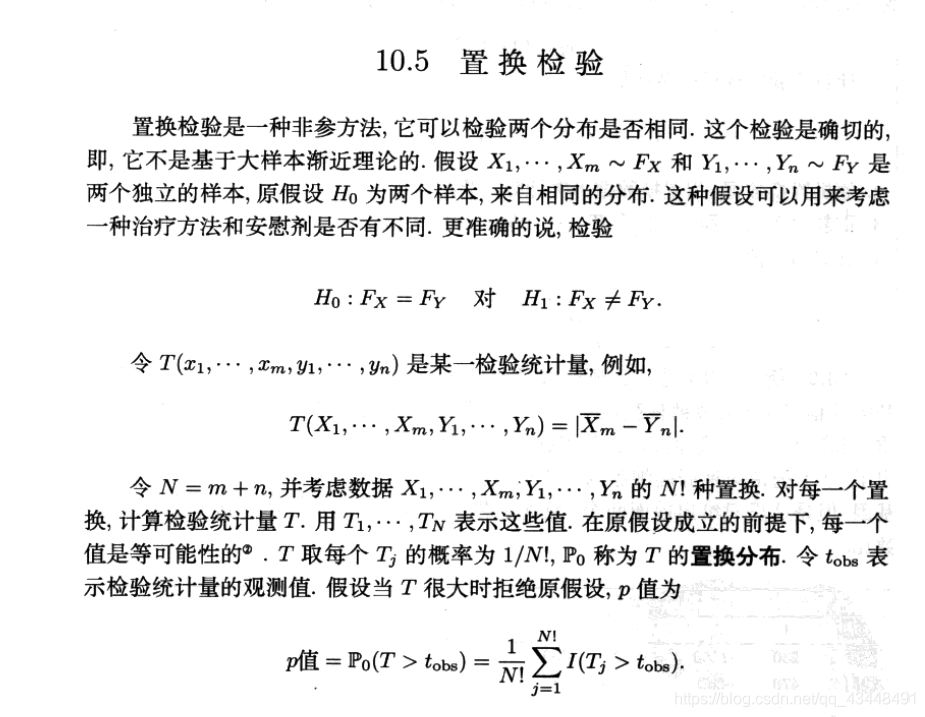
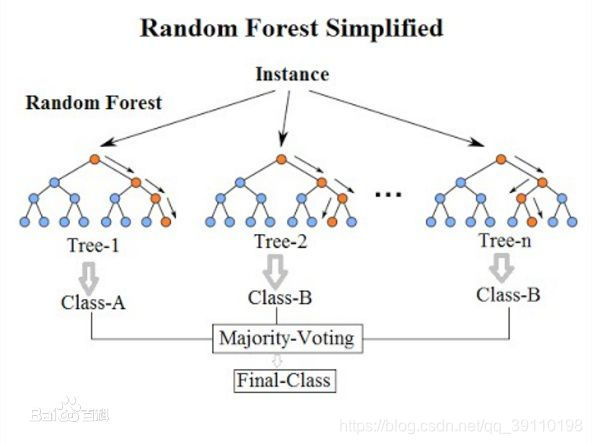
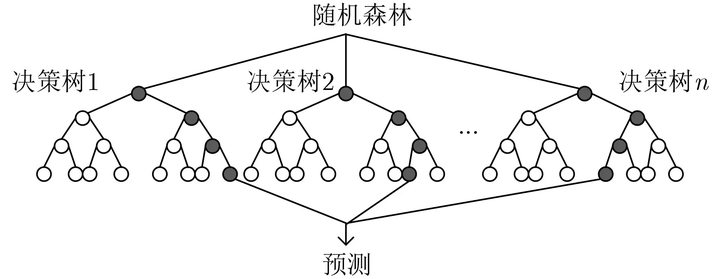
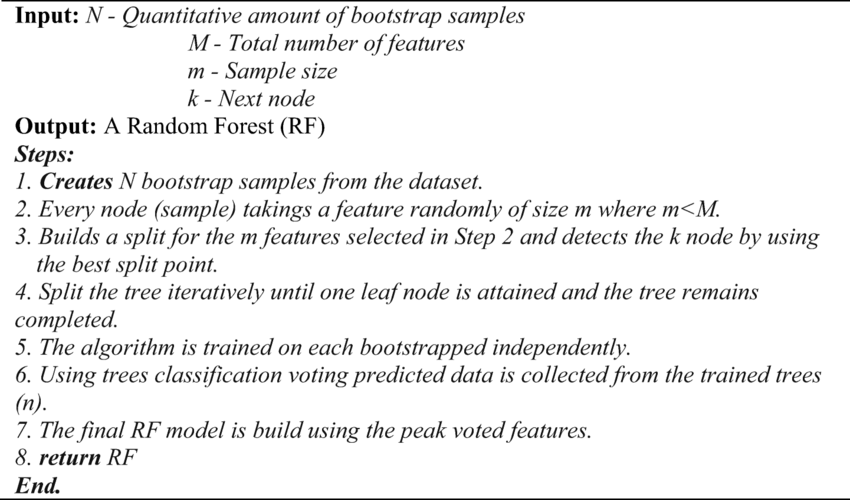
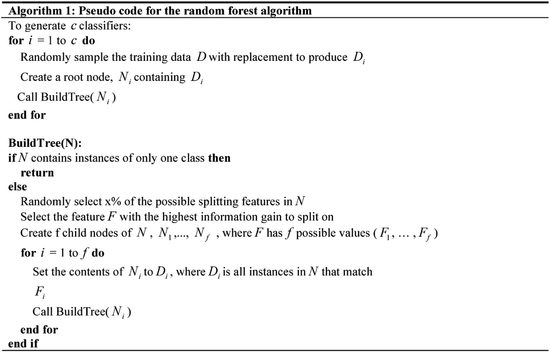
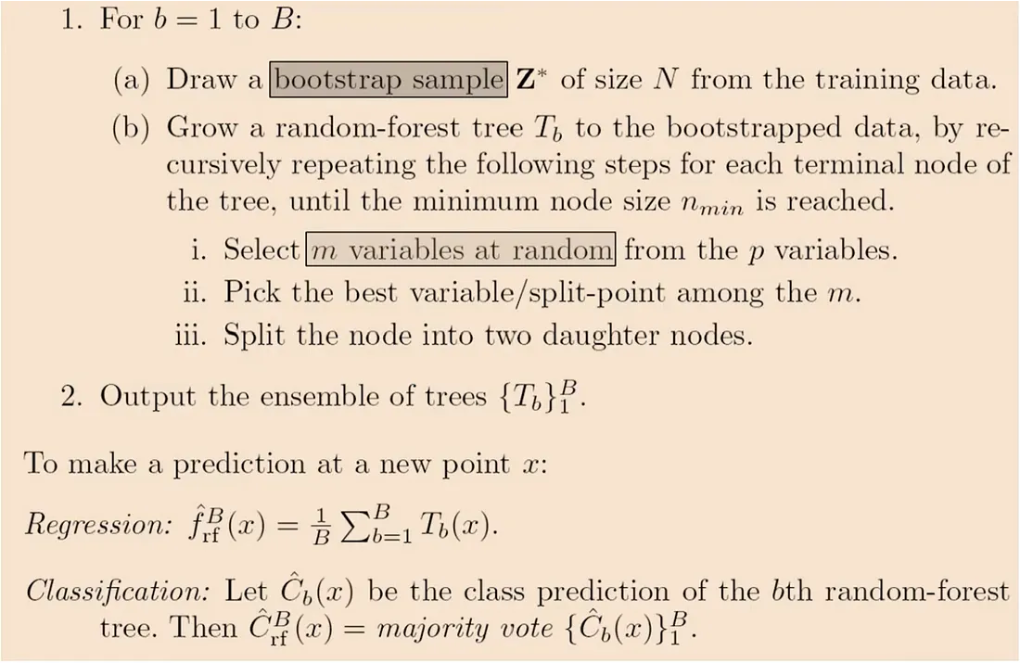
**第二问：采用置换检验来反驳教练的说法。**





**在我们的模型中，就是将双方势头作为随机森林机器学习模型的输入，来预测比赛的输赢。然后随机排列势头重新使用机器学习来预测输赢，计算p值输出的图像分别如下图所示（原图群里有取最后一次的图）。P=0拒绝零假设，说明势头在比赛中起作用。**

****

//Random Forest model is utilized for training, prediction, and assessing feature importance.

Algorithm 1: Random Forest Classifier Training

Data:

训练数据集 (X\_train, y\_train)

测试数据集 (X\_test, y\_test)

Results:

训练好的模型 (model)

预测准确率 (*actual\_acc*)

特征重要性 (feature\_importance\_df)

中间流程:

1 训练随机森林分类器

2 调用 fit 方法使模型学习如何将特征映射到标签。

3 使用训练好的模型对测试数据 (X\_test) 进行预测，获取预测标签 (y\_pred)。

4 使用 accuracy\_score 方法计算模型在测试数据上的准确率 (*actual\_acc*)

5 使用模型的 feature\_importances\_ 属性获取每个特征对模型预测结果影响的重要程度。

6 将特征重要性与对应的特征名称存入DataFrame (feature\_importance\_df) 并按重要性排序。

//Permutation test is employed to assess the statistical significance of the model's accuracy under the null hypothesis.

Algorithm 2: Permutation Test

Data:

Real results of the test set *real*

Predicted results of the model on the test set *pred*

Results:

List of accuracy values generated during iterations *acc\_list*

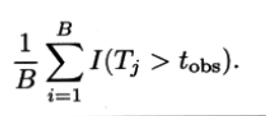
Probability *p*

1 *n* <- 1000

2 for *i* <- 1 to *n* do

3 *perm\_y* <- random\_permutation(*real*)

4 *acc\_list[i]* <- accuracy\_score(*perm\_y*, *pred*)

6 *p* <-  (*acc\_list* >= *actual\_acc*)图片中的B换成n

**7 return** *p*