# 参考文献

一、赢比赛重要指标：

1. 赢下第一盘或第二盘的前八场比赛中的任何一场

2. 第一盘的第8、10和11局以及第二盘的第4场

Richardson, P. A., Adler, W., & Hankes, D. (1988). Game, Set, Match: Psychological Momentum in Tennis. The Sport Psychologist, 2(1), 69–76. doi:10.1123/tsp.2.1.69

二、

转换破发后赢一局概率变大



三、

建立特定模型分析势能的必要性



四、

二元逻辑回归介绍

[1] Guo Wenfeng, Fan Chao, Guo Xindong. MOOC retreat lesson based on binary logistic regression model to predict [J]. Journal of computer age, 2017, (12) : 50-53. DOI: 10.16644 / j.carol carroll nki cn33-1094 / tp. 2017.12.014.

五 随机森林介绍

[1] Dong Shishi, Huang Shihishi. Analysis on random forest theory [J]. Integrated Technology,2013,2(01):1-7.

# 符号说明

# 假设和理由

1.**假设运动员自身的体能参数，例如运动耐力，身体机能没有显著差异。**因为我们所得到的数据只有针对比赛每一次发球的参数，并以此来建立模型。我们的模型是建立在发球数据的基础上的，如果需要根据运动员自身体能参数差异来进行加成，则标准会过于主观，且身体参数本身就难以量化。

**2.假设match之间是相互独立的。**即运动员不会因为先前的match的表现而影响到本次match的发挥，继而影响预测。其原因一方面是因为match之间间隔时间非常长，不太容易短期内产生心理变化，另一方面是我们选用的模型训练数据最大是以match为单位的，match之间由于比赛参数不同（发生在什么时间，运动员水平如何）会对模型造成偏差，因此不适宜对多个match进行一起训练。

**3.假设运动员的比赛发挥会受到比赛具体的发球行为影响。**例如运动员达成了ace, double\_fault, unforced\_error会有相应的正面、负面效果。把这些因素纳入考量，可以考虑人自身的心理状态对比赛发挥的影响，使模型更为完善。

**4.假设旁边没有教练指导，或者教练指导不会对运动员发挥造成影响。**所给的数据集中没有提供教练的行为，此外教练对于选手的影响难以量化，不适合纳入模型分析考量。

**5.排除观众反应、裁判造成的有争议判罚对运动员发挥造成的影响。**所给的数据集中没有提供相关的数据，此外这些因素的随机性较强，对于运动员心理影响具有黑箱性，不适合纳入模型分析考量。

# Assumptions and Justifications

1. **Athletes' physical parameters, such as endurance and bodily functions, show no significant differences.** The data obtained is specific to parameters for each serve in a match, forming the basis for our model. Our model relies solely on serve data, and introducing adjustments based on individual athletes' physical parameters would be overly subjective and challenging to quantify due to the inherent difficulty in measuring physical parameters.

2.**Matches are independent of each other.** Athletes' performance in previous matches does not impact their performance in the current match, thereby not influencing predictions. This is attributed to the substantial time gap between matches, making it unlikely for short-term psychological changes to occur. Additionally, the model's training data is match-centric, and combining data from different matches for training would introduce bias due to varying match parameters (timing, athlete skill levels).

3.**Athletes' match performance is influenced by specific serving behaviors, such as achieving aces, double faults, or unforced errors, resulting in corresponding positive or negative effects.** Incorporating these factors allows for the consideration of an athlete's psychological state in enhancing the model's comprehensiveness.

4.**Absence of coach guidance or negligible impact of coach guidance on athletes' performance.** The provided dataset lacks information on coaches' actions, and quantifying the influence of coaches on athletes is impractical, making it unsuitable for inclusion in model analysis.

5. **Exclusion of the impact of audience reactions and controversial referee decisions on athletes' performance.** The dataset does not provide relevant information on these factors, and their inherent randomness, coupled with their opaque psychological impact on athletes, makes them unsuitable for consideration in model analysis.

# 优缺点

### 主成分分析（PCA）

**优点：**

1. **降维：** 主成分分析可以用于降低数据的维度，消除冗余信息，提高计算效率。

**缺点：**

1. **解释性差：** 主成分往往难以解释，因为它们是原始特征的线性组合。

### 二元逻辑回归

**优点：**

1. **简单而有效：** 逻辑回归是一种简单而有效的分类算法，易于实现和理解。
2. **可解释性：** 模型的输出是概率，易于解释每个特征对预测的影响。
3. **不容易过拟合：** 在数据量较小的情况下，逻辑回归通常不容易过拟合。
4. **在线学习：** 逻辑回归支持在线学习，能够适应动态数据。

**缺点：**

1. **线性决策边界：** 逻辑回归假设数据是线性可分的，对于非线性问题可能表现不佳。
2. **对异常值敏感：** 对于异常值比较敏感，可能影响模型性能。

### 随机森林

**优点：**

1. **高准确性：** 随机森林通常能够提供较高的准确性，对于复杂的数据集表现良好。
2. **处理大量特征：** 能够处理大量的输入特征，并能够评估它们的相对重要性。
3. **抗过拟合：** 由于采用了多个树的投票机制，随机森林对过拟合有较强的抵抗力。
4. **对缺失值不敏感：** 随机森林对于数据中的缺失值和异常值相对不敏感。
5. **易于调参：** 随机森林的超参数相对较少，调参相对容易。

**缺点：**

1. **黑盒模型：** 随机森林很难解释，它是一个黑盒模型，不如逻辑回归那样易于解释。
2. **可能过度拟合：** 在一些情况下，随机森林可能对一些噪声敏感，导致过度拟合。

**Principal Component Analysis (PCA):**

**Advantages:**

1. *Dimensionality Reduction:* PCA effectively reduces data dimensionality, eliminating redundancy and enhancing computational efficiency.
2. *Data Visualization:* By retaining principal components, data can be visualized in lower dimensions, aiding in understanding its structure.
3. *Decorrelation:* PCA eliminates correlations between features, contributing to improved model stability.

**Disadvantages:**

1. *Linear Assumption:* PCA relies on a linear assumption and may not be suitable for data with nonlinear relationships.
2. *Interpretability Challenge:* Principal components are often challenging to interpret as they are linear combinations of original features.

**Binary Logistic Regression:**

**Advantages:**

1. *Simplicity and Effectiveness:* Logistic regression is a straightforward and effective classification algorithm, easy to implement and comprehend.
2. *Interpretability:* Model output is in probabilities, facilitating the interpretation of each feature's impact on predictions.
3. *Resilience to Overfitting:* Logistic regression tends to be less prone to overfitting in scenarios with limited data.
4. *Online Learning:* Logistic regression supports online learning, adapting to dynamic data.

**Disadvantages:**

1. *Linear Decision Boundary:* Logistic regression assumes data is linearly separable, performing suboptimally for nonlinear problems.
2. *Sensitivity to Outliers:* Logistic regression is sensitive to outliers, potentially affecting model performance.

**Random Forest:**

**Advantages:**

1. *High Accuracy:* Random forests typically provide high accuracy and excel on complex datasets.
2. *Feature Handling:* Capable of managing a large number of input features and assessing their relative importance.
3. *Overfitting Resistance:* Due to the voting mechanism of multiple trees, random forests exhibit strong resistance to overfitting.
4. *Insensitivity to Missing Values:* Random forests are relatively insensitive to missing values and outliers.
5. *Easy Parameter Tuning:* Random forests have fewer hyperparameters, making tuning relatively straightforward.

**Disadvantages:**

1. *Black Box Model:* Random forests are challenging to interpret, being a black box model unlike logistic regression.
2. *Potential Overfitting:* In some cases, random forests may be sensitive to noise, leading to overfitting.