機器學習概論——KNN練習

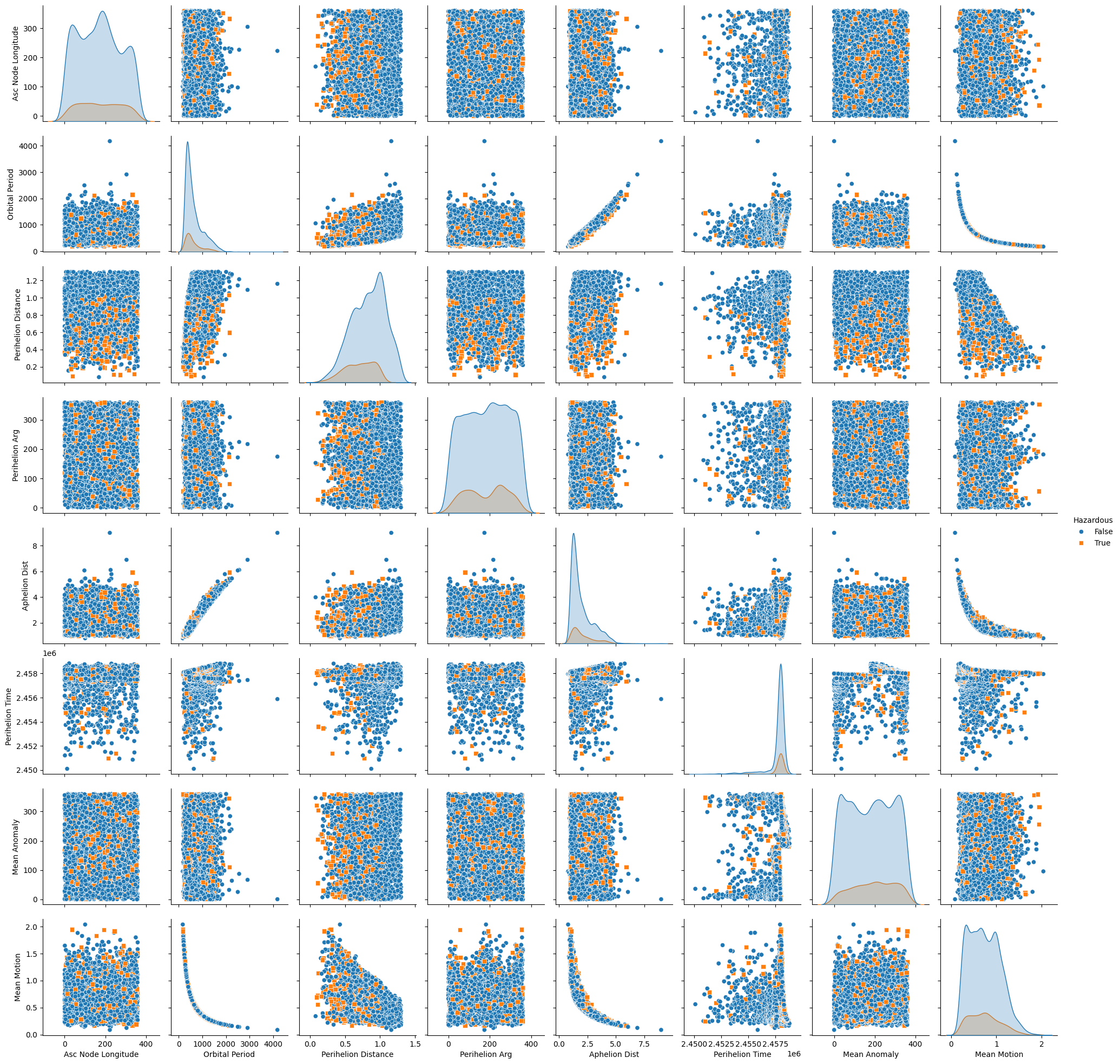
# 主旨：

# 利用訓練集資料來預測小行星撞地球的機率

# 資料集來源：

# <https://www.kaggle.com/datasets/shrutimehta/nasa-asteroids-classification>

# PairPlot



* 这张图是一个配对图（pair plot），显示了多个变數之间的散点图、密度分布以及各自的核密度估计曲线。主要变數依旧包括：

1. Asc Node Longitude（升交点经度）
2. Orbital Period（轨道周期）
3. Perihelion Distance（近日点距离）
4. Perihelion Arg（近日点角距）
5. Aphelion Dist（远日点距离）
6. Perihelion Time（近日点时间）
7. Mean Anomaly（平近点角）
8. Mean Motion（平均运动）

* 在配对图的每个对角线上是各个变量的密度分布图（波形图），展示了它们的分布形态。非对角线的图则是变量之间的散点图，展示变量之间的关系。
* 例子：

**Mean Anomaly（平近点角）的分布图**

在配对图的对角线上，可以看到 **Mean Anomaly** 的波形图（或密度分布图），其表现为一个波浪形的曲线，展示了平近点角（0° 到 360°）的数据分布。蓝色表示非危险天体（"False"），橙色表示危险天体（"True"）。

1. **单峰与多峰**

* 如果密度曲线是单峰的（只有一个主要波峰），说明 **Mean Anomaly** 在某个角度附近数据点较多，表示大部分天体在该位置附近聚集。
* 如果密度曲线呈多峰分布，说明 **Mean Anomaly** 值在多个位置上都有数据聚集，可能意味着天体群体内存在多个不同的子群或轨道模式。

1. **分类特征**

* 如果蓝色和橙色曲线的分布形态差异明显，例如蓝色波峰在角度较低的区域，而橙色波峰在角度较高的区域，则可以推测 **Mean Anomaly** 对区分危险和非危险天体有帮助。

**Perihelion Distance 和 Aphelion Dist 的散點圖**

在配对图的对角线上，可以看到 **Mean Anomaly** 的波形图（或密度分布图），其表现为一个波浪形的曲线，展示了平近点角（0° 到 360°）的数据分布。蓝色表示非危险天体（"False"），橙色表示危险天体（"True"）。

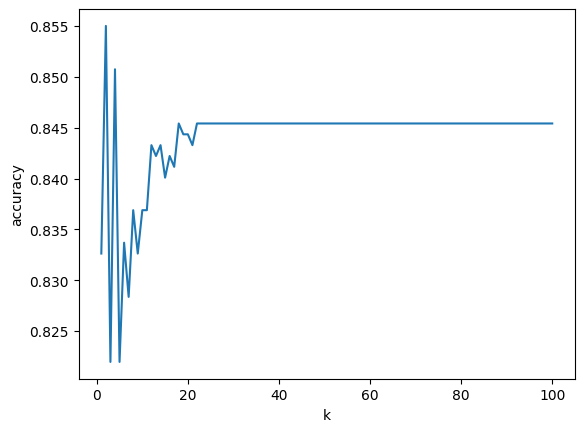
1. **正相關性**

* 如果這個散點圖中的點形成一條向上的趨勢線，這表示近日點距離和遠日點距離之間存在正相關。這意味著如果天體的近日點距離較大，它的遠日點距離也會較大

1. **分類差異**

* 通過觀察散點的顏色分佈（藍色代表"False"，橙色代表"True"），我們可以看到不同類別在這兩個變量上的分佈是否有明顯差異。
* 假設橙色點（"True"）集中在散點圖的某一特定區域（如更高的近日點和遠日點距離範圍），而藍色點（"False"）則分佈在另一個範圍，這意味著近日點距離和遠日點距離可能對於分類具有一定區分力。

最佳K值

-

* 從這張圖（迭代100次）可以看出最好的K值為**2**，準確率落在**86%**；隨著迭代次數增加則準確率趨近85%