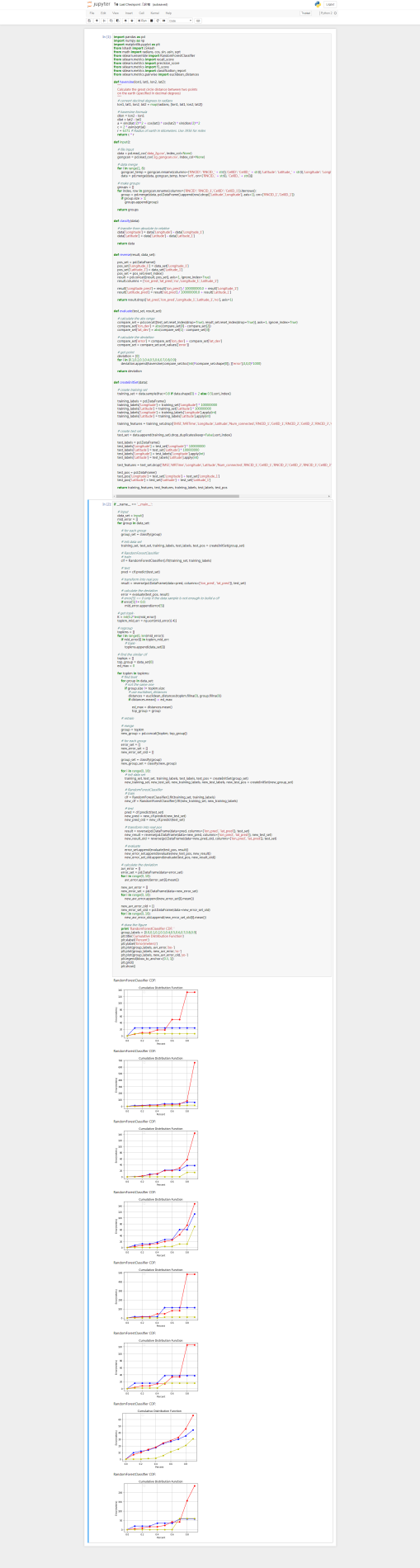
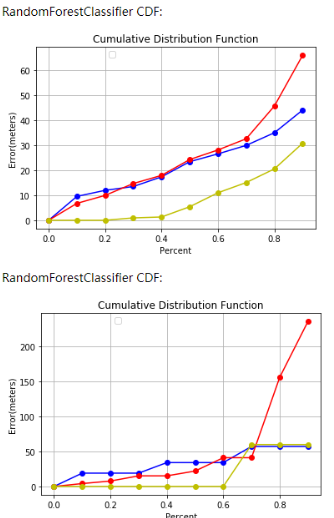
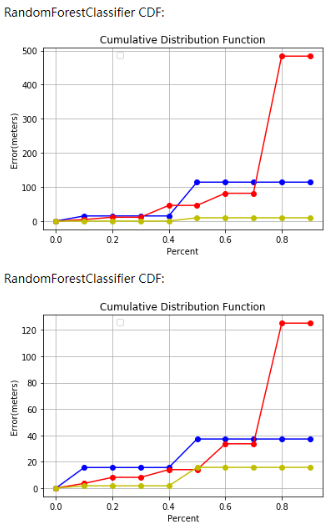
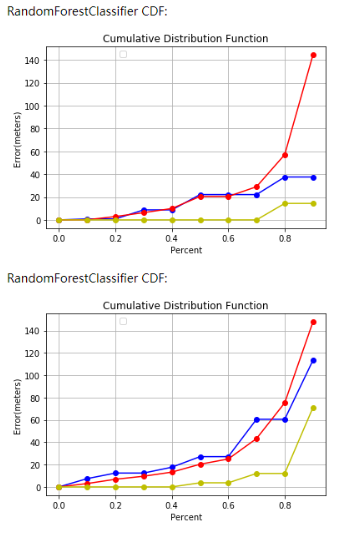
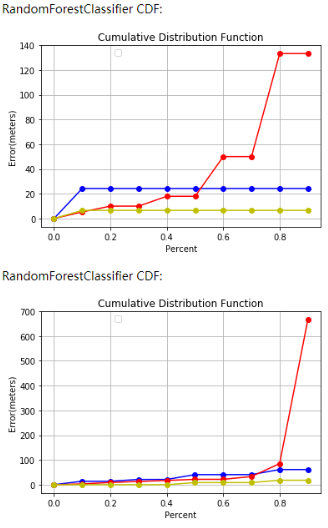
## 代码运行

[点击查看原图](e.png)



### 讨论分析

1. 本题改进前一小题中使用较好数据集的方式，转为使用相似数据集进行混合以期提高预测结果。
2. 本题获得相似数据集的方式为，遍历计算各数据集间的欧氏距离，获得的距离越小，相似度越大，原理类似KNN算法。
3. 图线绘制方法同d，蓝线为原曲线，黄线为旧测试集曲线，红线为新测试集曲线。
4. 以下为结果截图：



* 1. 本题内大量出现了原数据集数据量较少的情况，这一点的原因来自随机森林的随机性，导致恰好多个小数据集的误差最大，被划入topk-集合。
  2. 图中大部分红线均具有一定特点，即，中位误差均一定程度小于蓝线，大于中位的部分误差最终都会远大于蓝线，这一点来源于加入的数据集本身并不优秀，由于叠加效果，用全部新数据进行训练和验证的情况下，最大误差将大于两个原数据集自身的最大误差，表现为红线尾端大幅上升。
  3. 但各图中黄线误差均整体小于蓝线误差，表明这些数据集都由于新相似数据的加入提升了预测的结果，而非加入数据本身训练出来的模型优秀。
  4. 此种方法和d中使用优秀数据集的方法相比
     1. 是针对于原数据集的优化，通过增加相似数据量来提升模型准确度。
     2. 避免了由于两数据集数据量的差异导致的反优化作用。
     3. 但两种方法都可以一定程度的提高模型准确率，两题中的黄线误差均相对蓝线误差有所下降，而具体优化效果各有优劣。

### 性能比较

此处性能比较将比较和e训练和验证耗时，主要自变量为加入数据集的不同，除去获得topk的耗时，d、e两题耗时基本相近缺少比较意义。