**头脑风暴会议记录**

**1）议题：（所选题目）**

如何制定搭建【软件缺陷分析预测系统】的具体路线

**2）参会人员：**

第 8 组：（名单）苏琪、卫可昕、吴芳玲、马佳卉、周昱辰

主持人：马佳卉

记录员：卫可昕

**3）主持人发言：（解释题目）**

我们要搭建一个软件缺陷分析预测系统，要实现的两个主要功能就是“分析”和“预测”，一方面是要分析现有的数据集，根据数据集的特征和标签训练一个模型；一方面是要通过我们训练出来的这个模型，输入一组特征值，来预测这条数据的标签，也就是这行代码是否有bug。在这个过程中需要使用机器学习的方法做分类预测，我们之前也接触过很多典型的分类算法并尝试做过相关的项目，所以这次我们可以从中选择几个算法，进行试验和性能比较，最终讨论得出几种方案，然后去选择最优的那个。现在让我们开始此次的头脑风暴吧。

**4）成员发表意见：（每位成员独立发表意见，每人可以发表多条意见）**

马佳卉：我觉得我们可以采用决策树算法，这是一个机器学习中很常用的分类算法。我们可以把特征值向量作为输入，返回一个最终的决策，也就是标签。我们先将数据划分为训练集和测试集，然后用决策树模型进行训练和预测。决策树模型可以生成可理解的规则，可以发现特征的重要程度，而且它的计算复杂度比较低。所以我觉得我们可以试着采用决策树模型来进行代码的正确性分类预测。

周昱辰：我认为我们可以采用SVM算法，它的大致思想是：假设样本空间上有两类点，即一行代码是否有bug，我们可以通过找到一个划分超平面，将这两类样本分开，而划分超平面应该选择泛化能力最好的，也就是能使得两类样本中距离它最近的样本点距离最大。SVM算法在样本量中等偏小的情况也有较好的效果，但是在样本量巨大且使用核函数时计算量很大。

吴芳玲：这是一个二分类问题，而且有61个特征值（不计标签），均为数字型且连续，我觉得可以用朴素贝叶斯分类器，假设各特征相互独立，利用高斯分布，计算出各个特征的条件概率，所有特征相乘，得到综合的概率，判断代码是否有缺陷。这种方法的优势是简单易行，缺点是可能存在较大误差，因为实际各特征之间会存在些联系。在这基础上，可以考虑用贝叶斯网络。相比朴素贝叶斯分类器，贝叶斯网络考虑了各特征之间存在关联，更加普适。另外，我们应该需要进行特征提取。可以尝试朴素贝叶斯分类器，各个特征分别用来预测标签，计算准确率，从准确率上可以看出特征与标签之间的相关性。

卫可昕：我认为我们可以采用KNN算法，KNN是一种非参的，惰性的算法模型,这个模型不会对数据做出任何的假设，KNN建立的模型结构是根据数据来决定的，这也比较符合现实的情况， KNN算法的核心思想是如果样本在特征空间的 k个最相邻的样本中大部分属于某一类，那么该样本也属于这一类，并且具有这类样本的特征。在分类决策中，该方法仅根据样本的一个或少数几个样本的类别确定待分样本所属的类别。KNN方法仅适用于极少数相邻样本的分类决策。在 kNN方法中，由于 KNN方法主要依赖于周围有限的邻域样本，而不是依赖域判别法来确定类别，因此， kNN方法比其他方法更适合于交叉或重叠较多的待分样本集。

苏琪：我也认为这是一个二分类问题，因此我们可以考虑使用机器学习课上正在讲授的逻辑回归来解决这个问题。假设各个样本之间独立同分布且各个特征相互独立，寻找一个合适的预测函数，在此基础上构造损失函数，并用梯度下降法来优化我们的模型。最后通过正则化解决可能出现的过拟合问题以及利用合适的模型来进行模型评估指标。

马佳卉：我觉得我们应该使用数据库来存储和管理数据，这样数据的读写速度会快一些，也不容易丢失或者混乱，可以提高我们应用开发的效率。

1. **方案讨论排序决策过程：**

贝叶斯算法具有较大的缺陷，因为该算法假设各特征相互独立的特性，可能会存在较大误差，因为实际各特征之间会存在些联系。

决策树更适用于离散值数据的处理，而给定数据集特征中，存在较多连续型数据，并且数据集的取值范围跨度较大，较稀疏，不便进行连续值的离散化。一般经过我们的研究探讨，认为所给的训练集、测试集中存在数据不满足离散值的特性，因此，使用不太适用于连续值的决策树并不合适。

SVM算法可用于线性/非线性分类，比较适用于样本量中等偏小的情况，有较好的效果，可以解决高维问题，也可以处理大型特征空间问题。但SVM算在样本量巨大、处理线性不可分问题进行高维映射、使用核函数等情况下，计算量很大，且对缺失数据和噪声数据较为敏感，因此优先级稍后。

KNN算法相比其他算法，算是比较简洁明了的算法。即使没有很高的数学基础也能搞清楚它的原理。KNN算法是惰性的，且模型训练时间快，此外，KNN算法预测效果好，对异常值不敏感。当需要使用分类算法，且数据比较大时可以尝试使用KNN算法进行分类，适用于本课题。

Logistic 回归实际上是使用线性回归模型的预测值逼近分类任务真实标记的对数几率，其优点有：直接对分类的概率建模，无需实现假设数据分布，从而避免了假设分布不准确带来的问题（区别于生成式模型）；不仅可预测出类别，还能得到该预测的概率，这对一些利用概率辅助决策的任务很有用；对数几率函数是任意阶可导的凸函数，有许多数值优化算法都可以求出最优解，logistic回归能因此提高准确率。

因此，形成排序：

1. Logistic 回归
2. KNN算法
3. SVM算法
4. 决策树
5. 贝叶斯算法

**6）结论：（讨论确定至少2个最终方案）**

方案1：采用Logistic 回归算法，使用数据库

方案2：采用KNN算法，使用数据库