基于机器学习的刑事案件分类技术

1. **项目背景**

文本分类用计算机对文本集(或其他实体或物件)按照一定的分类体系或标准进行自动分类标记。 它根据一个已经被标注的训练文档集合, 找到文档特征和文档类别之间的关系模型, 然后利用这种学习得到的关系模型对新的文档进行类别判断[1]。文本分类是自然语言处理的核心问题之一，相关研究成果已广泛应用于各行各业的文本处理中。

目前，文本分类的主流方式是采用机器学习方法（包括深度学习方法）。对于中文文本来说，其主要思想，先对文本进行分词，去停用词，在形成词典的基础上，对文本进行向量化，从而生成文本向量。然后，采用机器学习方法进行文本分类。

本项目旨在采用机器学习方法对大量刑事案件判决书进行学习的基础上，实现给定案情的案件可能的判决进行预测。

**2． 数集简介**

给定的数据集有4098份刑事案件判决书，涵盖了刑法中规定的大多数罪名。请以这些判决书为学习样本，采用至少两种机器学习算法，实现给定案情的预测判决。预测结果至少包括可能的罪名（注意一个案件，有可以有一个以上的罪名成立）， 能给出刑期或罚金则更好。

1. **完成要求**
2. 描述在进行机器学习前如何处理数据的？
3. 描述项目完成过程中用到的技术？
4. 结合项目本身，描述所使用的机器学习算法？
5. 描述你是如何进行算法验证的？
6. 算法性能比较。至少比较所选的两种机器学习算法在学习时间、预测准确率、召回率、ROC，并给出分类结果的混淆矩阵。
7. 项目完成心得。
8. 列出参考文献。

参考文献

[1] 文本分类， 百度百科，https://baike.baidu.com/item/文本分类/7267115?fr=aladdin