本项目的目标是利用机器学习方法对数据集进行处理，并对相应的数据集进行预测。

该项目至少应包含以下4个部分

1.数据集的分析

2.机器学习和预测

3.预测结果的可视化

4.理论形式主义

数据集的分析

为了分析数据集，你必须从给定的数据集中提取一些统计信息，例如数据的类型、缺失值、离群值、变量之间的相关性等。这一部分还应该包含对领域应用的分析，并解释预测的目标。

预测

1.使用回归线来预测一些变量的值（可以选择任意的

变量作为目标。计算误差。

2.使用机器学习方法，以预测一组新对象的类别。你可以使用的方法是K-最近邻（K-NN），支持向量机（SVM），决策树，神经网络......所得到的结果应该使用一些外部指标来验证，如预测误差或其他。解释超参数是如何固定的，并使用Cros Valdiation。

在报告中应对所获得的结果进行分析，并提供改进结果的解决方案，例如将多种方法结合起来。

视觉化

您可以将从分类（预测）中提取的知识可视化，以便呈现结果，例如，使用预测颜色的散点图，决策树，......

理论细节 从算法上（数学上）给出该方法的形式论，能得到最好的结果。

解释所用方法的所有参数及其对结果的影响。

我应该做一些比较来结束这个项目。

报告应该包含以上4个部分，你应该解释所用方法的动机，并讨论所获得的结果。

请注意，你可以使用AWS、Python、R和或Matlab和或Knime进行实验。

每组(2名学生)将使用以下网络链接中的10个数据集中的一个。

httplipn.univ-paris13.fr~grozavuPredAdataProject。

每个文件夹都包含一个包含数据集的文件.data和一个包含相应数据集信息的文件.names。这些数据集是从UCI存储库中提取的。

您也可以使用UCI网站或其他平台的不同数据集。