## Assignment 1 specification and instructions

This assignment is worth 30% of your overall grade for this module.

## Aim

This short project tests your ability to construct and evaluate machine learning models. You are asked to produce two machine learning models and perform a comparative analysis of measures of performance. It is not necessary to obtain the best possible performance by searching far and wide for the most state-of-the-art algorithms. You should instead use reasonable choices for model classes and performance measures - similar to those discussed in our lectures and readings. Which models you use, and how you evaluate performance, is up to you. If your models perform poorly by your selected metric, do not worry. Your goal is to find a sensible approach, present any evidence that one model outperforms another, and to produce clear, concise, understandable code and text documenting your effort. Do not attempt to code everything from scratch.  You are expected to use packages discussed in the lectures and readings, or similar. However, you should understand, and be capable of explaining, the packages you use.

这个简短的项目测试您构建和评估机器学习模型的能力。您需要制作两个机器学习模型并对性能指标进行比较分析。您不必通过广泛搜索最先进的算法来获得最佳性能。您应该使用合理的模型类别和性能指标选择 - 类似于我们在讲座和阅读材料中讨论的那些。您使用哪种模型以及如何评估性能取决于您自己。如果您的模型在您选择的指标下表现不佳，请不要担心。您的目标是找到一种合理的方法，提出任何证据表明一个模型优于另一个模型，并编写清晰、简洁、易懂的代码和文本来记录您的努力。不要试图从头开始编写所有内容。您需要使用讲座和阅读材料中讨论的软件包或类似软件包。但是，您应该理解并能够解释您使用的软件包。

## Data

You will use the data in P1data5117.csv. This is obtained by emailing the module lecturer(s) on  They will sample an existing dataset (both instances and attributes) and email you with your specific data file. The instances not sampled for your file will be used as holdout data for your model(s).

The task is to predict Y from the X variables. This is a classification problem, so your model should return a value between zero and one (that we interpret as a predicted percentage of having the actual value one).

数据

您将使用 P1data5117.csv 中的数据。这可以通过向 模块讲师发送电子邮件获得。他们将对现有数据集（实例和属性）进行采样，并通过电子邮件向您发送您的特定数据文件。未为您的文件采样的实例将用作模型的保留数据。

任务是从 X 变量预测 Y。这是一个分类问题，因此您的模型应该返回一个介于零和一之间的值（我们将其解释为实际值为一的预测百分比）。

## Instructions

After cleaning the data and converting to numeric, examine the data and perform cross-correlation, recording any insights or observations found.

Define an evaluation function that will return metrics and validation information for any classifier model applied to any test/train split of the original data.

Choose two classifier model classes from the following site:

[Scikit-learn: Supervised learning](https://scikit-learn.org/stable/supervised_learning.html)

For each class, find a choice of hyperparameters that gives superior model performance as set out in your evaluation function.

Compare your best model from each class in terms of estimated predictive performance for unseen data, and the overfit/underfit tradeoff. If you use any randomisation (in, say, shuffling and/or splitting data), use 5117 as the random key where 5117 denotes the last four digits in your matriculation number.

Write a short report summarising your methods and findings. In particular: Can you identify any X variables that are important (or unimportant) for predicting new instances of this data class? Which specific model would you deploy based on these analyses, and why?

说明

清理数据并转换为数字后，检查数据并执行互相关，记录发现的任何见解或观察结果。

定义一个评估函数，该函数将返回应用于原始数据的任何测试/训练分割的任何分类器模型的指标和验证信息。

从以下站点选择两个分类器模型类：

Scikit-learn：监督学习

对于每个类，找到一个超参数选择，该超参数可提供评估函数中规定的卓越模型性能。

比较每个类别中的最佳模型，比较它们对未见数据的预测性能估计值以及过度拟合/欠拟合权衡。如果您使用任何随机化（例如，改组和/或拆分数据），请使用 5117 作为随机密钥，

撰写一份简短的报告，总结您的方法和发现。特别是：您能否确定任何对预测此数据类的新实例重要（或不重要）的 X 变量？您将根据这些分析部署哪种特定模型，以及原因是什么？

## Key points

Here are some key points that will help you in completing this assignment:

* We are not looking for a final model that performs well: we are looking to see if you can build a sensible model and a sensible evaluation of its performance, and also if you can clearly document your effort.
* Presentation counts. A clear and concise Jupyter notebook will earn more marks than an unstructured notebook full of opaque and poorly commented code.
* If you do not understand something or have questions, you are encouraged to discuss it with your peers (say, via the Teams channel) or the module lecturer(s). However, the deliverables that you submit must comply with the policy on good academic practice.
* Note that any discussions involving other students taking x will be affected by the inherent differences in the underlying data, the model classes chosen, and any randomization performed as part of your workflow. Hence discussions will involve options for techniques and reporting, rather than comparing actual results.

要点

以下是一些有助于您完成此作业的要点：

• 我们并不要求最终模型表现良好：我们要求您构建合理的模型并对其性能进行合理的评估，以及您是否可以清楚地记录您的努力。

• 演示很重要。简洁明了的 Jupyter 笔记本比充满晦涩难懂且注释不清的代码的非结构化笔记本能获得更多分数。

• 如果您不理解某些内容或有疑问，我们鼓励您与同学（例如，通过 Teams 频道）或模块讲师讨论。但是，您提交的可交付成果必须符合良好学术实践政策。

• 请注意，涉及其他参加 x 的学生的任何讨论都将受到基础数据、所选模型类别以及作为工作流程一部分执行的任何随机化的固有差异的影响。因此，讨论将涉及技术和报告选项，而不是比较实际结果。

## Submission

1. The code of your solution in a Jupyter notebook.
2. A brief, clear and concise summary describing your models, your comparative measures of performance, and your final result, in a maximum three page PDF file.

提交：

通过作业 1 提交点上传以下两项内容：

1. Jupyter 笔记本中您的解决方案的代码。

2. 简短、清晰、简洁的摘要，描述您的模型、性能比较指标和最终结果，最多三页 PDF 文件。