# 多元统计期末考核

## **关于数据集**

冠状病毒病（COVID-19）是由一种新发现的冠状病毒引起的传染病。大多数感染 COVID-19 病毒的人会出现轻度至中度呼吸道疾病，无需特殊治疗即可康复。老年人以及患有心血管疾病、糖尿病、慢性呼吸道疾病和癌症等潜在健康问题的人更容易患上严重疾病。  
在整个大流行病过程中，医疗保健提供者面临的主要问题之一是医疗资源短缺以及有效分配这些资源的适当计划。在这些困难时期，能够预测一个人在检测呈阳性时或在此之前可能需要什么样的资源，这将对政府有巨大帮助，因为他们将能够采购和安排必要的资源挽救那个病人的生命。

该项目的主要目标是建立一个机器学习模型，根据 Covid-19 患者的当前症状、状态和病史，预测患者是否处于高危状态。

数据预处理，数据清洗，变量相关性，降维，回归模型，聚类，判别法，建议？

## 内容

该数据集由墨西哥政府提供。该数据集包含大量匿名的患者相关信息，包括先决条件。原始数据集包含 21 个独特的特征和 1,048,576 名独特的患者。**在布尔特征中，1 表示“是”，2 表示“否”。值 97 和 99 是缺失数据**。

* 性别：1代表女性，2代表男性。
* 年龄：患者。
* 分类：covid 测试结果。值 1-3 表示患者被诊断为不同  
  程度的 covid。4 或更高意味着患者不是 covid 的携带者或测试没有定论。
* 患者类型：患者在单位接受的护理类型。1 人回家，2 人住院。
* 肺炎：患者是否已经有气囊炎症。
* 怀孕：患者是否怀孕。
* 糖尿病：患者是否患有糖尿病。
* copd：表示患者是否患有慢性阻塞性肺疾病。
* 哮喘：患者是否患有哮喘。
* inmsupr：患者是否免疫抑制。
* 高血压：患者是否患有高血压。
* 心血管：患者是否患有心脏或血管相关疾病。
* 肾慢性：患者是否患有慢性肾病。
* 其他疾病：患者是否患有其他疾病。
* 肥胖：患者是否肥胖。
* 烟草：患者是否是烟草使用者。
* usmr：表示患者是否接受过一级、二级或三级医疗单位的治疗。
* 医疗单位：提供医疗服务的国家卫生系统机构类型。
* 插管：患者是否连接到呼吸机。
* icu：表示患者是否已入住重症监护病房。
* 死亡日期：如果患者死亡，请注明死亡日期，否则为 9999-99-99。

数据集：<https://www.kaggle.com/datasets/meirnizri/covid19-dataset>

参考代码：

逻辑回归：<https://www.kaggle.com/code/yasirakyzl/covid-19-ml-model-90-accuracy>

平时作业：20% 期中：5份上机报告，40% 期末：40%

截止日期：2023.1.13