关于数据集

冠状病毒病（新冠肺炎）是一种由新发现的冠状病毒引起的传染病。大多数感染新冠肺炎病毒的人将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疾病的人更有可能患上严重疾病。

在整个疫情期间，医疗保健提供者面临的主要问题之一是医疗资源短缺和有效分配的适当计划。在这些困难时期，能够预测一个人在检测呈阳性时甚至在此之前可能需要什么样的资源，将对当局有巨大帮助，因为他们将能够获得和安排必要的资源来挽救患者的生命。

本项目的主要目标是建立一个机器学习模型，该模型根据新冠肺炎患者的当前症状、状态和病史，预测患者是否处于高危状态。

1. 定义“高危状态“。例如基于插管，ICU，死亡等状态
2. 随机选取一部分数据作为训练集，一部分作为验证集，一部分作为测试集
3. 设计机器学习分类算法并编程实现
4. 攥写完整的报告

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，包括预先条件。原始数据集由21个独特特征和1048576名独特患者组成。数据集特征中，1表示“是”，2表示“否”。值97，98，99是其他含义。

* sex：1：女性，2：男性
* age：患者的年龄（岁）
* classification：新冠肺炎试验结果。值1-3表示患者在不同的度。4或更高意味着患者不是新冠肺炎病毒携带者或检测结果不确定。
* patient type：患者接受的护理类型。1：回家治疗，2：住院治疗。
* pneumonia：患者是否已经有气囊炎症。1：是，2：否，99：未知
* pregnancy：患者是否怀孕
  + 1：是
  + 2：否
  + 97：表示男性不能怀孕，此项无统计意义。（sex = 2）
  + 99：因为当医生没有足够的关于怀孕的信息，或者他们有一些误导性的结果时，这个记录没有意义
* diabetes：患者是否患有糖尿病。1：是，2：否，98：未知
* copd：表示患者是否患有慢性阻塞性肺病。1：是，2：否，98：未知
* asthma：患者是否患有哮喘。1：是，2：否，98：未知
* inmspr：患者是否患有免疫抑制。1：是，2：否，98：未知
* hipertension：患者是否患有高血压。1：是，2：否，98：未知
* cardiovascular：患者是否患有心脏或血管相关疾病。1：是，2：否，98：未知
* renal chronic：患者是否患有慢性肾脏疾病。1：是，2：否，98：未知
* other disease：患者是否患有其他疾病。1：是，2：否，98：未知
* obesity：患者是否肥胖。1：是，2：否，98：未知
* tobacco：患者是否吸烟。1：是，2：否，98：未知
* usmr：表示患者是否接受了第1第2或第3级医疗单位的治疗。1：是，2：否
* medical unit：国家卫生系统提供护理的机构类型，共13种类，分别为1~13
* intubed：患者是否连接气管插管呼吸机。
  + 1：是
  + 2：否
  + 97：患者回家治疗，不具备气管插管呼吸机条件，此项无统计意义（patient type = 1）
  + 99：异常值
* icu：表示患者是否已入住重症监护室。
  + 1：是
  + 2：否
  + 97：患者回家治疗，不具备ICU条件，此项无统计意义（patient type = 1）
  + 99：异常值
* date died：如果患者死亡，为死亡日期，否则为9999-99-99。