**SRTP项目前期准备**

**标题拟定：**基于LLM和卫星图像的社会学统计数据发掘与建模

**背景**

近年来，亚非拉地区的一些发展中国家由于长期战乱和医疗资源不足等因素，面临着严重的公共卫生挑战。同时，这些地区的公共卫生数据收集和监测体系相对薄弱，导致了政府和国际组织对于卫生资源的分配和干预措施的难以制定。目前对这些国家的公共卫生数据调研形式多以实地考察为主，一些研究团队，如斯坦福大学，每年花费大量的经费，冒着安全风险进入一些亚非拉地区国家调研这些地区的公共卫生数据收集。

**主要内容：**

1. **前期准备：**

与导师及斯坦福大学的教授进行讨论，选择1~3个典型社会学统计指标作为验证集，并选择合适的国家及地区作为模型验证的样本。

1. **数据收集与整理：**
   1. 利用LLM收集近期与该统计学数据相关的权威报告。
   2. 利用多模态的LLM对收集到的报告进行分析，提取其中的具体数据并进行整理。
2. **数据标注**

从GoogleEarth等卫星地图平台上下载对应地区的卫星图片，对卫星图片的尺度进行统一调整，并根据划分的区域进行分组，完成图片的数据标注。

1. **模型设计与训练**
   1. 阅读与学习神经网络相关论文，选择合适的神经网络的模型，完成该项目的模型设计
   2. 在服务器上进行模型的训练，并根据训练结果及时调整模型及参数，达到更好的训练结果。

**5. 模型验证**

斯坦福大学方面提供其收集的相关统计数据，与模型的预测结果进行比较，评估该模型的预测准确度。

**项目预期成果**

1. 验证评估：较好地契合斯坦福大学的真实数据，表现出较高地准确率，展示项目成果在公共卫生和人口统计领域的应用潜力和价值。

2. 论文成果：总结项目中涉及的卫星图像处理技术、大语言模型应用技术和统计数据挖掘方法，为相关领域的研究和应用提供参考和借鉴。

3. 推广成果：将这个模型推广到一些社会学研究领域，运用到对于社会学数据的统计之中，可以为政府和国际组织制定针对性的卫生干预措施提供科学依据。

4. 社会效益：达到对于亚非拉地区社会学数据的预测，促进对于亚非拉地区第三世界国家的人文关怀

**前期尝试：**

**效果较好的指标：**

IMR (U5MR)

Maternal Health

infection rate(COVID-19)

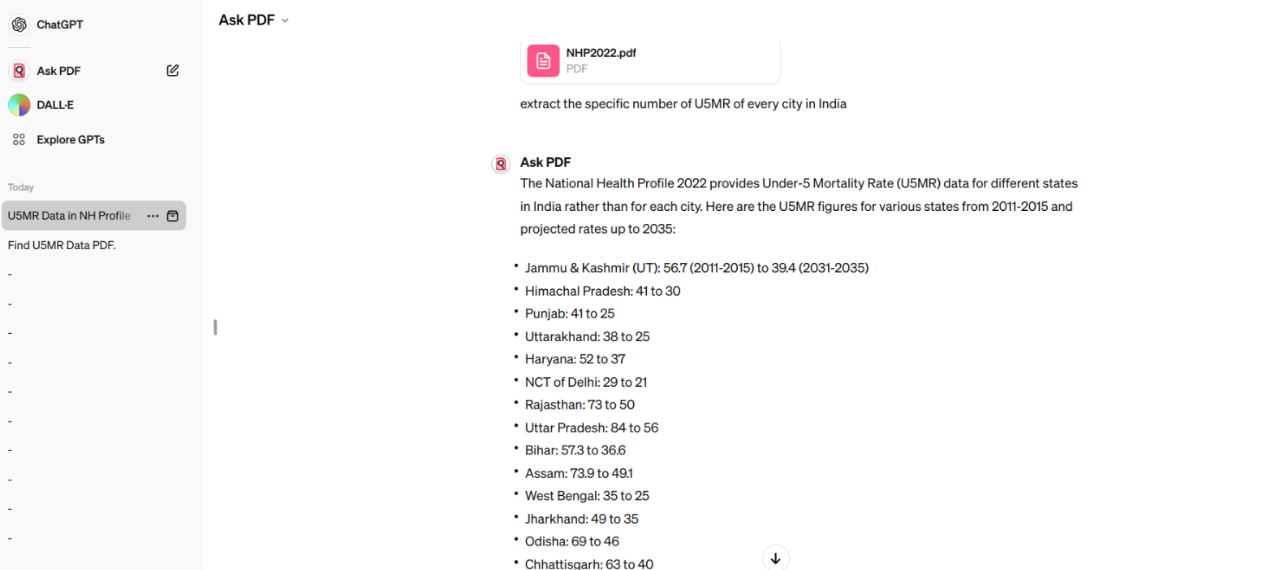
**report website**

IMR (U5MR)

Maternal Health

infection rate(COVID-19)

多模态LLM数据提取：

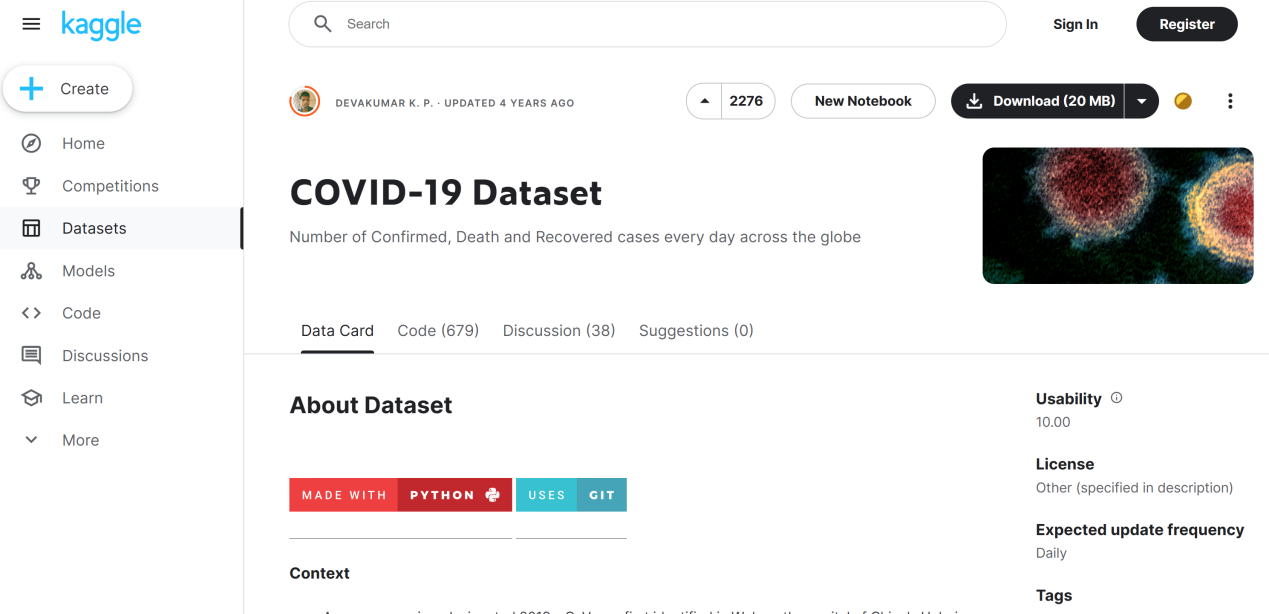
****

**公共卫生数据查找**

[Child Mortality - UNICEF DATA](https://data.unicef.org/topic/child-survival/under-five-mortality/)

[世界卫生数据平台 (who.int)](https://www.who.int/zh/data)

[COVID-19 Dataset | Kaggle](https://www.kaggle.com/datasets/imdevskp/corona-virus-report)

这些指标大多以国家为单位，同时大多数是时序数据。

一些印度的测试可用报告（NHFS，NHP）



