Collect the dataset or create the dataset

TEAM ID	PNT2022TMID04216
PROJECT NAME	STATISTICAL MACHINE LEARNING APPROACHES TO LIVER DISEASE PREDICTION

Collect the dataset or create the dataset

There's a good story about bad data from Columbia University. A healthcare project was aimed to cut costs in the treatment of patients with pneumonia. It employed machine learning (ML) to automatically sort through patient records to decide who has the lowest death risk and should take antibiotics at home and who's at a high risk of death from pneumonia and should be in the hospital. The team used historic data from clinics, and the algorithm was accurate.

But there was with an important exception. One of the most dangerous conditions that may accompany pneumonia is asthma, and doctors always send asthmatics to intensive care resulting in minimal death rates for these patients. So, the absence of asthmatic death cases in the data made the algorithm assume that asthma isn't that dangerous during pneumonia, and in all cases the machine recommended sending asthmatics home, while they had the highest risk of pneumonia complications.

ML depends heavily on data. It's the most crucial aspect that makes algorithm training possible and explains why machine learning became so popular in recent years. But regardless of your actual terabytes of information and data science expertise, if you can't make sense of data records, a machine will be nearly useless or perhaps even harmful.

The thing is, all datasets are flawed. That's why data preparation is such an important step in the machine learning process. In a nutshell, data preparation is a set of procedures that helps make your dataset more suitable for machine learning. In broader terms, the data prep also includes establishing the right data collection mechanism. And these procedures consume most of the time spent on machine learning. Sometimes it takes months before the first algorithm is built!

The dataset that you use to train your machine learning models can make or break the performance of your applications. For example, using a text dataset that contains loads of biased information can significantly decrease the accuracy of your machine learning model.

This was what happened to Amazon's initial tests. They trained a machine learning model for their automated hiring system. This tool was designed to pinpoint the best candidates across a batch of applicants for the job vacancies at their engineering departments. However, because they used the wrong dataset, their algorithm produced results that were considerably biased towards male candidates.

DATASET:

https://www.kaggle.com/datasets/uciml/indian-liver-patient-records