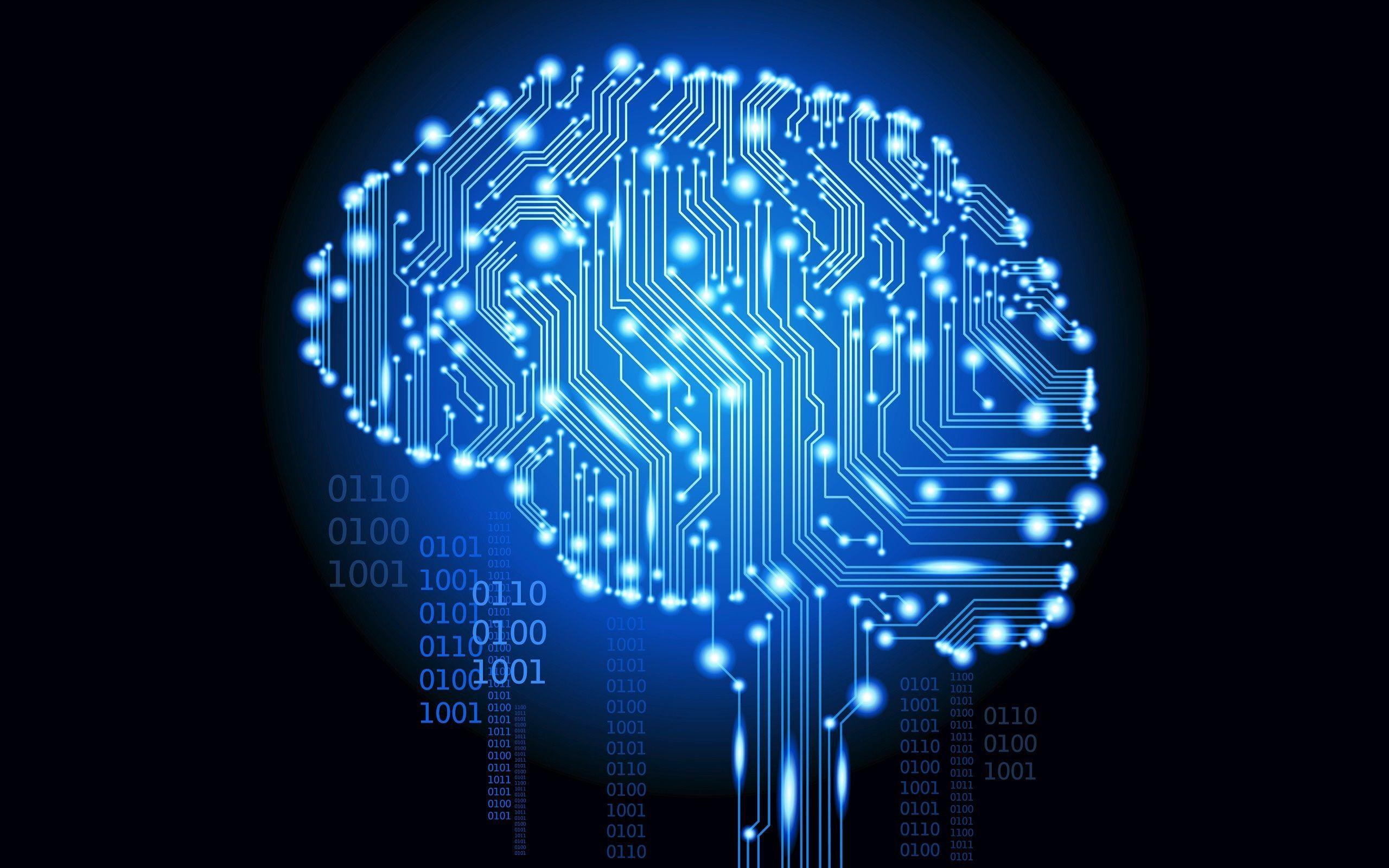
|  |
| --- |
| Project  Machine Learning DS675  Milestone 1:  Pick Datasets |
|  |



**By: Aditya Patel**

**Ap2935@njit.edu**

## Milestone 1: Pick Datasets

### Task:

Browse through the datasets on [kaggle.com](https://www.kaggle.com/). Pick two datasets that you find interesting. Please pick datasets that have a high usability rating (>9). Then, for each dataset, give a link to it, and write 1-2 paragraphs about what it is, and what learning task you want to use it for (e.g., classification, regression). Some of the datasets come with an explicit learning task, but you can always define your own learning task

### Dataset 1:

### Smoking and Drinking Dataset with body signal

**Link:** <https://www.kaggle.com/datasets/sooyoungher/smoking-drinking-dataset>

**Usability:** 10

**Size:** ~ 991k Entries with 24 Features (Columns). Dataset also contains feature Descriptions, Units of Measurement, Rounding Information and Data Format

**Description**

This dataset presents a comprehensive collection of health-related information, focusing primarily on smoking and drinking habits in conjunction with vital body signals and demographic details.

The dataset comprises data sourced from the National Health Insurance Service in Korea, ensuring the exclusion of personal and sensitive information. Its primary objectives are twofold:

1. Analysis of Body Signals
2. Classification of Smokers and Drinkers

**Use-case Planned**

The dataset contains multiple bio signals, which are key elements in a pathology and toxicology report. If a regression machine learning model is trained using these features, a model can be developed that can help in predicting the disease or underlying medical conditions, such as the individual's likelihood of experiencing heart disease, stroke, or diabetes.

This prediction is based on various factors such as blood pressure, blood sugar levels, and cholesterol levels. Additionally, the dataset allows for estimating the risk of kidney disease by analyzing serum creatinine levels and identifying anemia through hemoglobin levels.

Moreover, this valuable information can be leveraged to spread awareness and self-checkup platforms to identify the future consequences based on one's vitals.

Though the prediction will not provide absolute certainty or definitive results regarding one's health status, which can only be done by a doctor, it will certainly help in developing medical reports and making informed decisions about lifestyle changes.

### Dataset 2:

**Lung Cancer Prediction**

Air Pollution, Alcohol, Smoking & Risk of Lung Cancer

**Link:** <https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link>

**Usability:** 10

**Size:** ~ Usability: 10

Size: ~ 1k Entries with 26 Features (Columns). The dataset contains feature Descriptions and categorical distribution information

**Description**

This dataset contains categorical information on patients with lung cancer, including their age, gender, air pollution exposure, alcohol use, dust allergy, occupational hazards, genetic risk, chronic lung disease, balanced diet, obesity, smoking, passive smoking, chest pain, coughing of blood, fatigue, weight loss, shortness of breath, wheezing, swallowing difficulty, clubbing of fingernails and snoring. The data is categorized under different categories such as high-pollution group, low-pollution group, the level of occupational hazards of the patient. (Categorical) and many more.

**Use-case Planned**

The dataset is similar to dataset 1 and contains multiple bio signals, which are important factors in predicting the lifestyle of the person and their chances of getting lung cancer. If a classification-based machine learning model is trained using multiple features, a model can be developed that can categorize person into low, medium, and high-risk classification of getting cancer or it can be used to identify the current mental state of subject i.e., depression, anxiety or stress

Though the prediction will not provide absolute certainty or definitive results regarding one's health status, which can only be done by a doctor, it will certainly help in developing medical reports and making informed decisions about lifestyle changes.