Big Dataset 1
(diabetes _ binary _ health _ indicators _ BRFSS2015.csv/ Big Dataset_1.csv)
https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

diabetes _ binary _ health _ indicators _ BRFSS2015.csv is a clean dataset of 253,680 survey responses to the CDC's BRFSS2015. The target variable Diabetes_binary has 2 classes. 0 is for no diabetes, and 1 is for prediabetes or diabetes. This dataset has 21 feature variables and is not balanced.

Big Dataset 2 (diabetes _ binary _ 5050split _ health _ indicators _ BRFSS2015.csv/ Big Dataset_2.csv)
https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

diabetes _ binary _ 5050split _ health _ indicators _ BRFSS2015.csv is a clean dataset of 70,692 survey responses to the CDC's BRFSS2015. It has an equal 50-50 split of respondents with no diabetes and with either prediabetes or diabetes. The target variable Diabetes_binary has 2 classes. 0 is for no diabetes, and 1 is for prediabetes or diabetes. This dataset has 21 feature variables and is balanced.

Original Source:
https://www.kaggle.com/code/alexteboul/diabetes-health-indicators-dataset-notebook/output

Data description of the original source:
https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf

Variables:

Response Variable / Dependent Variable:

1) **Diabetes_binary** (binary categorical)
    - Question: (Ever told) you have diabetes

      Responses:
    - 0: No (no diabetes)
    - 1: Yes (prediabetes or borderline diabetes or diabetes)

Predictor Variables / Independent Variables:

2) **HighBP** (binary categorical)
    - Question: Adults who have been told they have high blood pressure by a doctor, nurse, or other health professional

      Responses:
    - 0: No
    - 1: Yes

3) **HighChol** (binary categorical)
    - Question:
      Have you EVER been told by a doctor, nurse or other health professional that your blood cholesterol is high?

      Responses:
    - 0: No
    - 1: Yes

4) **CholCheck** (binary categorical)
    - Question: Cholesterol check within past five years

      Responses:
    - 0: No
    - 1: Yes

5) **BMI** (float numerical)
    - Body Mass Index: BMI = weight (in kilograms) / [height (in metres)]^2
      (Has 2 implied decimal places)

6) **Smoker** (binary categorical):
    - Question: Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]

      Responses:
    - 0: No
    - 1: Yes

7) **Stroke** (binary categorical):
    - Question: (Ever told) you had a stroke.

      Responses:
    - 0: No
    - 1: Yes

8) **HeartDiseaseorAttack** (binary categorical):
    - Question: Adults who reported having coronary heart disease (CHD) or myocardial infarction (MI)

      Responses:
    - 0: No
    - 1: Yes

9) **PhysActivity** (binary categorical):
    - Question: Adults who reported doing physical activity or exercise during the past 30 days other than their regular job

      Responses:
    - 0: No
    - 1: Yes

10) **Fruits** (binary categorical):
    - Question: Consume Fruit 1 or more times per day

      Responses:
    - 0: No
    - 1: Yes

11) **Veggies** (binary categorical):
    - Question: Consume Vegetables 1 or more times per day

      Responses:
    - 0: No
    - 1: Yes

12) **HvyAlcoholConsump** (binary categorical):
    - Question: Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)

      Responses:
    - 0: No
    - 1: Yes

13) **AnyHealthcare** (binary categorical):

- Question: Do you have any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare, or Indian Health Service?

  Responses:
- 0: No
- 1: Yes

14) **NoDocbcCost** (binary categorical):
- Question: Was there a time in the past 12 months when you needed to see a doctor but could not because of cost?

  Responses:
- 0: No
- 1: Yes

15) **GenHlth** (categorical):
- Question: Would you say that in general your health is: Scale 1-5

  Responses:
- 1 = excellent
- 2 = very good
- 3 = good
- 4 = fair
- 5 = poor

16) **MentHlth** (int numerical):
- Question: Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?

  Responses:
- Scale 1-30 days

17) **PhysHlth** (int numerical):
- Question: Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?

  Responses:
- Scale 1-30 days

18) **DiffWalk** (binary categorical):
- Question: Do you have serious difficulty walking or climbing stairs?

  Responses:
- 0 is no
- 1 is yes

19) **Sex** (binary categorical)
- Question: Indicate sex of respondent.

  Responses:
- 0 is female
- 1 is male

20) **Age** (categorical): 13-level age category
- Responses:
  1: Age 18 to 24
  2: Age 25 to 29
  3: Age 30 to 34
  4: Age 35 to 39
  5: Age 40 to 44
  6: Age 45 to 49
  7: Age 50 to 54
  8: Age 55 to 59
  9: Age 60 to 64
  10: Age 65 to 69
  11: Age 70 to 74
  12: Age 75 to 79
  13: Age 80 or older

21) **Education** (categorical):
- Questions: What is the highest grade or year of school you completed?

  Responses:
- 1: Never attended school or only kindergarten
- 2: Grades 1 through 8 (Elementary)
- 3: Grades 9 through 11 (Some high school)
- 4: Grade 12 or GED (High school graduate)
- 5: College 1 year to 3 years (Some college or technical school)
- 6: College 4 years or more (College graduate)

22) **Income** (categorical):
- Question: Is your annual household income from all sources:

  Responses:
- 1 or 2: Less than $15,000
- 3 or 4: $15,000 to less than $25,000
- 5: $25,000 to less than $35,000
- 6: $35,000 to less than $50,000
- 7 or 8: $50,000 or more

Small Dataset ***(Small Dataset.csv)***
https://www.kaggle.com/datasets/vikasukani/diabetes-data-set

Variables:

Response Variable / Dependent Variable:

1) **Outcome** (binary categorical): target feature

    - 0: No (no diabetes)
    - 1: Yes (prediabetes or borderline diabetes or diabetes)

Predictor Variables / Independent Variables:

2) **Pregnancies** (int numerical): Number of pregnancies.
    - Range of values: 0 to 17

3) **Glucose** (int numerical): Glucose (a simple monosaccharide) level
    - Range of values: 0 to 199

4) **BloodPressure** (int numerical): Blood pressure level. Blood pressure is the force of your blood pushing against the walls of your arteries
    - Range of values: 0 to 122

5) **SkinThickness** (int numerical): Triceps skin fold thickness in mm
    - Range of values: 0 to 110

6) **Insulin** (int numerical): Insulin level. Insulin is a polypeptide hormone that regulates carbohydrate metabolism.
    - Range of values: 0 to 744

7) **BMI** (float numerical) : weight in kg / (height in m)^2
    - Range of values: 0 to 80.6

8) **DiabetesPedigreeFunction** (float numerical): Diabetes pedigree function (a function which scores likelihood of diabetes based on family history)
    - Range of values: 0.08 to 2.42

9) **Age** (int numerical): Patient age
    - Range of values: 21 to 81