

Diabetes Health Indicators Dataset Project

Context

- This dataset includes patient profiles with features based on demographics, lifestyle habits, family history, and clinical measurements that are well-established indicators of diabetes risk. All data is generated using statistical distributions inspired by real-world medical research, ensuring privacy preservation while reflecting realistic health patterns.

Attribute Information:

- 1- patient_id: Unique patient identifier
- 2- age: Age of patient in years
- 3- gender : Patient gender
- 4- ethnicity : Ethnic background
- 5- education_level: Highest completed education
- 6- income_level: Income category
- 7- employment_status: Employment type
- 8- smoking_status: Smoking behavior
- 9- alcohol_consumption_per_week: Drinks consumed per week
- 10- physical_activity_minutes_per_week: Physical activity (weekly minutes)
- 11- diet_score: Diet quality (higher = healthier)
- 12- sleep_hours_per_day: Average daily sleep hours
- 13- screen_time_hours_per_day: Average daily screen time hours
- 14- family_history_diabetes: Family history of diabetes
- 15- hypertension_history: Hypertension history
- 16- cardiovascular_history: Cardiovascular history
- 17- bmi: Body Mass Index (kg/m²)
- 18- waist_to_hip_ratio: Waist-to-hip ratio
- 19- systolic_bp: Systolic blood pressure (mmHg)
- 20- diastolic_bp: Diastolic blood pressure (mmHg)
- 21- heart_rate: Resting heart rate (bpm)
- 22- cholesterol_total: Total cholesterol (mg/dL)
- 23- hdl_cholesterol: HDL cholesterol (mg/dL)
- 24- ldl_cholesterol: LDL cholesterol (mg/dL)
- 25- triglycerides: Triglycerides (mg/dL)

- 26- glucose_fasting: Fasting glucose (mg/dL)
- 27- glucose_postprandial: Post-meal glucose (mg/dL)
- 28- insulin_level: Blood insulin level (μ U/mL)
- 29- hba1c: HbA1c (%)
- 30- diabetes_risk_score: Risk score (calculated, 0–100)
- 31- diabetes_stage: Stage of diabetes
- 32- diagnosed_diabetes: Target: Diabetes diagnosis

Data Information:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 31 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   age                                       100000 non-null  int64
1   gender                                   100000 non-null  object
2   ethnicity                               100000 non-null  object
3   education_level                         100000 non-null  object
4   income_level                           100000 non-null  object
5   employment_status                      100000 non-null  object
6   smoking_status                         100000 non-null  object
7   alcohol_consumption_per_week           100000 non-null  int64
8   physical_activity_minutes_per_week     100000 non-null  int64
9   diet_score                             100000 non-null  float64
10  sleep_hours_per_day                    100000 non-null  float64
11  screen_time_hours_per_day              100000 non-null  float64
12  family_history_diabetes                 100000 non-null  int64
13  hypertension_history                   100000 non-null  int64
14  cardiovascular_history                  100000 non-null  int64
15  bmi                                    100000 non-null  float64
16  waist_to_hip_ratio                     100000 non-null  float64
17  systolic_bp                            100000 non-null  int64
18  diastolic_bp                           100000 non-null  int64
19  heart_rate                             100000 non-null  int64
...
29  diabetes_stage                          100000 non-null  object
30  diagnosed_diabetes                      100000 non-null  int64
dtypes: float64(8), int64(16), object(7)
memory usage: 23.7+ MB
```

Data overview:

	age	gender	ethnicity	education level	income level	employment status	smoking status	alcohol consumption per week	physical activity minutes per week	diat score	...	hdl cholesterol	ldl cholesterol	triglycerides	glucose fasting	glucose postprandial	insulin level	hba1c	diabetes risk score	diabetes stage	diagnosed diabetes
0	58	Male	Asian	Highschool	Lower-Middle	Employed	Never	0	215	5.7	...	47	160	145	116	126	6.56	8.19	296	Type 1	1
1	48	Female	White	Highschool	Middle	Employed	Former	1	140	6.7	...	35	90	33	95	170	100	5.62	22.8	No Diabetes	0
2	58	Male	Hispanic	Highschool	Middle	Unemployed	Never	1	57	6.4	...	46	89	36	115	181	5.57	2.51	44.7	Type 2	1
3	74	Female	Black	Highschool	Low	Retired	Never	0	48	3.4	...	30	76	140	159	253	5.19	9.83	38.2	Type 2	1
4	46	Male	White	Graduate	Middle	Retired	Never	1	189	7.2	...	52	121	68	157	184	12.74	2.20	22.0	Type 1	1
...
9995	46	Male	Other	Graduate	Upper-Middle	Unemployed	Former	1	156	8.3	...	45	100	116	113	181	14.58	5.55	26.1	Pre-Diabetes	0
9996	48	Female	White	Graduate	Middle	Employed	Never	1	16	8.8	...	55	121	146	96	146	9.02	5.97	24.4	Pre-Diabetes	0
9997	57	Female	Black	No formal	Upper-Middle	Employed	Former	4	129	9.9	...	38	111	184	85	137	2.57	5.21	27.8	No Diabetes	0
9998	42	Female	Black	Highschool	Lower-Middle	Retired	Never	3	52	5.9	...	48	81	116	96	197	9.85	5.53	26.4	Pre-Diabetes	0
9999	52	Female	White	Postgraduate	Middle	Employed	Never	1	248	1.7	...	31	81	87	85	142	15.85	5.73	27.9	Pre-Diabetes	0

10000 rows x 21 columns

Data Information (Cont.):

- data with (100000) record & (32) features (columns).
- Data with (13) float columns & (12) int & (7) strings.

