ROYAL UNIVERSITY OF
PHNOM PENH

DIABETES PREDICTION USING MACHINE LEARNING

LECTURE: KHIM CHAMREOUN

GROUP TWO

Phork Norak, Phon Lihour, Vay Mithona, Sim Liheng

December 16 2025

Bussiness Understanding

The Challenge: A Chronic Metabolic Disorder

Diabetes mellitus is a chronic metabolic disorder characterized by the body's impaired ability to utilize blood sugar (glucose) effectively. It poses a significant challenge to global public health, necessitating better tools for early diagnosis and management.

Type 1 Diabetes

An autoimmune response, often manifesting in childhood, leading to an insulin deficiency. Caused by the destruction of insulin-producing beta cells.

Type 2 Diabetes

The most prevalent type, arising from insufficient insulin secretion or insulin resistance. Associated with family history, obesity, and inactivity.

Gestational Diabetes (GDM)

A temporary type developing during pregnancy due to hormonal changes. While it resolves post-childbirth, it increases the mother's future risk of Type 2 diabetes.

Predictive Modeling for Early Diagnosis

Many machine learning models have been built to assist doctors in the diagnosis of diabetes, leveraging patient data to identify patterns and predict outcomes. These models serve as powerful decision-support tools.

The Precedent

The PIMA Indian Diabetes Dataset

A foundational dataset that has been instrumental in developing and benchmarking diabetes prediction models. Your work will build upon this legacy.

The Advancement

Our Mission

We will use a more recent, larger dataset to build a new predictive model and deploy it as a functional web application.

Data Understanding

The Taipei Medical Center Dataset

Based on the 2023 study by Chou et al., J. Pers. Med., your work utilizes a contemporary and relevant dataset.

Name: 'TAIPEI diabetes. csv'

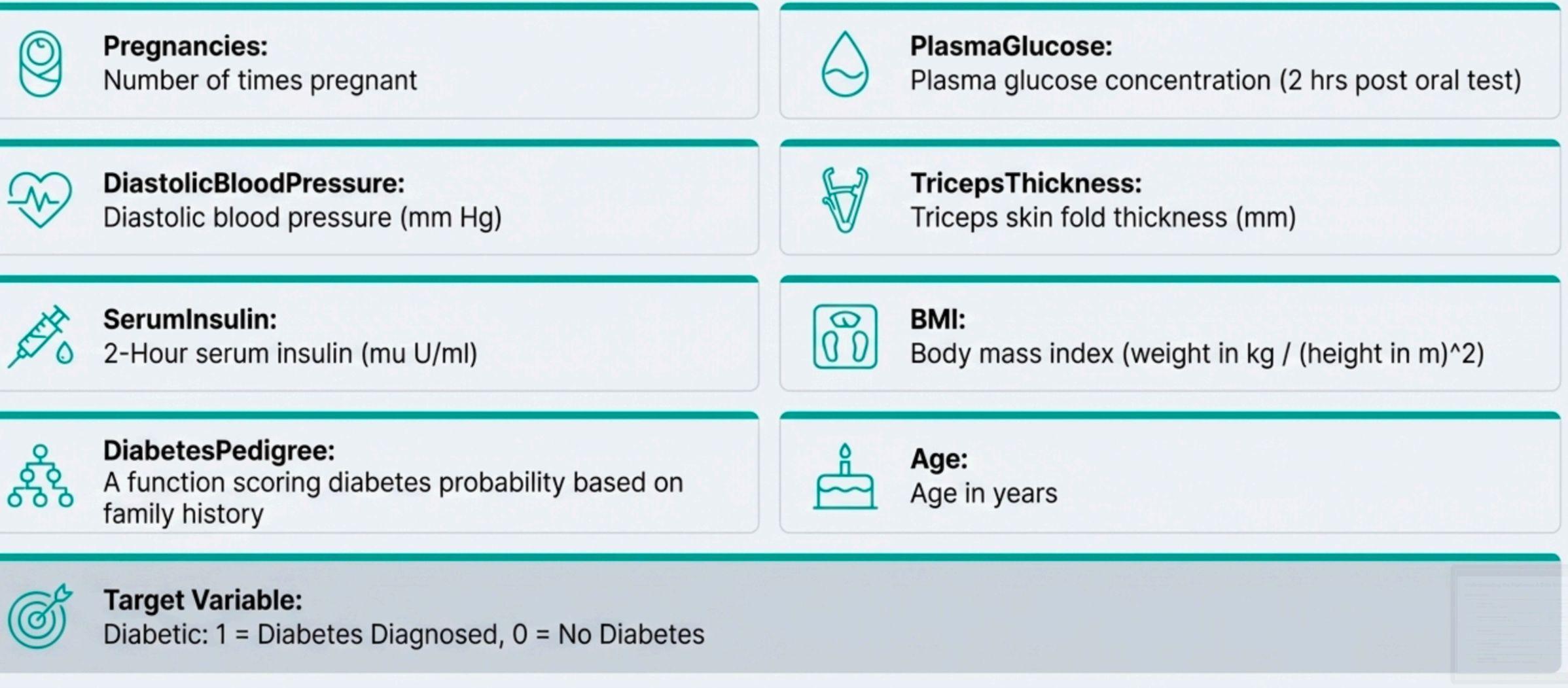
Subjects: 15,000 women

Age Range: 20 to 80 years

Source: Patients at Taipei Municipal Medical Center Collection Period: 2018-2020 and 2021-2022

Understanding the Features: A Data Dictionary

Input Variables



Pregnancies:
Number of times pregnant

PlasmaGlucose:
Plasma glucose concentration (2 hrs post oral test)

DiastolicBloodPressure:
Diastolic blood pressure (mm Hg)

TricepsThickness:
Triceps skin fold thickness (mm)

SerumInsulin:
2-Hour serum insulin (mu U/ml)

BMI:
Body mass index (weight in kg / (height in m)^2)

DiabetesPedigree:
A function scoring diabetes probability based on family history

Age:
Age in years

Target Variable:
Diabetic: 1 = Diabetes Diagnosed, 0 = No Diabetes

Result: Exploratory data analysis

```
df.duplicated().sum()
```

[4] Python

```
0
```

```
print("\nShowing the number of rows and columns of the dataset:")
df.shape
```

[5] Python

```
Showing the number of rows and columns of the dataset:

(15000, 10)
```

```
df.info()
```

[3] Python

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15000 entries, 0 to 14999
Data columns (total 10 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   PatientID              15000 non-null  int64
 1   Pregnancies            15000 non-null  int64
 2   PlasmaGlucose          15000 non-null  int64
 3   DiastolicBloodPressure 15000 non-null  int64
 4   TricepsThickness       15000 non-null  int64
 5   SerumInsulin           15000 non-null  int64
 6   BMI                    15000 non-null  float64
 7   DiabetesPedigree       15000 non-null  float64
 8   Age                    15000 non-null  int64
 9   Diabetic               15000 non-null  int64
dtypes: float64(2), int64(8)
memory usage: 1.1 MB
```

```
print("\nShowing first 5rows of the dataset:")
df.head()
```

[6] Python

Showing first 5rows of the dataset:

	PatientID	Pregnancies	PlasmaGlucose	DiastolicBloodPressure	TricepsThickness	SerumInsulin	BMI	DiabetesPedigree	Age	Diabetic
0	1354778	0	171	80	34	23	43.509726	1.213191	21	0
1	1147438	8	92	93	47	36	21.240576	0.158365	23	0
2	1640031	7	115	47	52	35	41.511523	0.079019	23	0
3	1883350	9	103	78	25	304	29.582192	1.282870	43	1
4	1424119	1	85	59	27	35	42.604536	0.549542	22	0

```
print("\nSummary Statistics of Dataset:")
df.describe()
```

[7] Python

Summary Statistics of Dataset:

	PatientID	Pregnancies	PlasmaGlucose	DiastolicBloodPressure	TricepsThickness	SerumInsulin	BMI	DiabetesPedigree	Age	Diabetic
count	1.500000e+04	15000.000000	15000.000000	15000.000000	15000.000000	15000.000000	15000.000000	15000.000000	15000.000000	15000.000000
mean	1.502922e+06	3.224533	107.856867	71.220667	28.814000	137.852133	31.509646	0.398968	30.137733	0.333333
std	2.892534e+05	3.391020	31.981975	16.758716	14.555716	133.068252	9.759000	0.377944	12.089703	0.471420
min	1.000038e+06	0.000000	44.000000	24.000000	7.000000	14.000000	18.200512	0.078044	21.000000	0.000000
25%	1.252866e+06	0.000000	84.000000	58.000000	15.000000	39.000000	21.259887	0.137743	22.000000	0.000000
50%	1.505508e+06	2.000000	104.000000	72.000000	31.000000	83.000000	31.767940	0.200297	24.000000	0.000000
75%	1.755205e+06	6.000000	129.000000	85.000000	41.000000	195.000000	39.259692	0.616285	35.000000	1.000000
max	1.999997e+06	14.000000	192.000000	117.000000	93.000000	799.000000	56.034628	2.301594	77.000000	1.000000

```
print("\nMissing Values in Dataset:\n")
print(df.isnull().sum())
```

[8] Python

```
Missing Values in Dataset:

PatientID                 0
Pregnancies               0
PlasmaGlucose             0
DiastolicBloodPressure    0
TricepsThickness          0
SerumInsulin              0
BMI                       0
DiabetesPedigree          0
Age                       0
Diabetic                  0
dtype: int64
```