DIABETES PREDICTION

Using MySQL

PRESENTATION 2024

By Rajashree Gavali

Table

EmployeeName	Patient_id	Gender	DOB	Age	Hypertension	Heart_disease	Smoking_history	BMI	HbA1c_level	Blood_glucose_level	diabetes
NATHANIEL FORD	PT101	Female	05-11-1992	31	0	1	never	25.19	6.6	140	0
GARY JIMENEZ	PT102	Female	11-11-1992	31	0	0	No Info	27.32	6.6	80	0
ALBERT PARDINI	PT103	Male	13-11-1992	31	0	0	never	27.32	5.7	158	0
CHRISTOPHER CHONG	PT104	Female	05-12-1992	31	0	0	current	23.45	5	155	0
PATRICK GARDNER	PT105	Male	03-01-1989	35	1	1	current	20.14	4.8	155	0
DAVID SULLIVAN	PT106	Female	05-01-1989	35	0	0	never	27.32	6.6	85	0
ALSON LEE	PT107	Female	23-01-1989	35	0	0	never	19.31	6.5	200	1
DAVID KUSHNER	PT108	Female	05-02-1989	35	0	0	No Info	23.86	5.7	85	0
MICHAEL MORRIS	PT109	Male	21-02-1989	35	0	0	never	33.64	4.8	145	0
JOANNE HAYES-WHITE	PT110	Female	09-03-1989	35	0	0	never	27.32	5	100	0
ARTHUR KENNEY	PT111	Female	19-03-1989	35	0	0	never	27.32	6.1	85	0
PATRICIA JACKSON	PT112	Female	01-04-1989	35	0	0	former	54.7	6	100	0
EDWARD HARRINGTON	PT113	Female	14-04-1989	35	0	0	former	36.05	5	130	0
JOHN MARTIN	PT114	Female	21-04-1989	35	0	0	never	25.69	5.8	200	0
DAVID FRANKLIN	PT115	Female	26-04-1989	35	0	0	No Info	27.32	5	160	0
RICHARD CORRIEA	PT116	Male	27-04-1989	35	0	0	No Info	27.32	6.6	126	0
AMY HART	PT117	Male	29-04-1989	35	0	0	never	30.36	6.1	200	0
SEBASTIAN WONG	PT118	Female	30-04-1989	35	0	0	never	24.48	5.7	158	0
MARTY ROSS	PT119	Female	10-05-1989	35	0	0	No Info	27.32	5.7	80	0
ELLEN MOFFATT	PT120	Male	10-05-1989	35	0	0	ever	25.72	3.5	159	0
VENUS AZAR	PT121	Male	12-05-1989	35	0	0	current	36.38	6	90	0
JUDY MELINEK	PT122	Male	11-06-1989	35	0	0	No Info	18.8	6.2	85	0
GEORGE GARCIA	PT123	Female	14-06-1989	35	0	0	never	21.24	4.8	85	0
VICTOR WYRSCH	PT124	Female	17-06-1989	35	0	1	former	27.94	6.5	130	0
JOSEPH DRISCOLL	PT125	Female	24-06-1989	35	0	0	No Info	13.99	4	140	0

```
4      -- 1. Retrieve the Patient_id and ages of all patients.
5
6  •   select Patient_id, Age
7      from diabetes_prediction;
8
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

Patient_id	Age
PT101	31
PT102	31
PT103	31
PT104	31
PT105	35
PT106	35
PT107	35
PT108	35
PT109	35
PT110	35
PT111	35
PT112	35
PT113	35
PT114	35

```sql
10
11      -- 2. Select all female patients who are olderthan 30.
12
13      select * from diabetes_prediction
14      where Gender='Female' and Age > 30;
15
16
17
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

EmployeeName	Patient_id	Gender	DOB	Age	Hypertension	Heart_disease	Smoking_history	BMI	HbA1c_level	Blood_glucose_level	diabetes
GARY JIMENEZ	PT102	Female	11-11-1992	31	0	0	No Info	27.32	6.6	80	0
CHRISTOPHER CHONG	PT104	Female	05-12-1992	31	0	0	current	23.45	5	155	0
DAVID SULLIVAN	PT106	Female	05-01-1989	35	0	0	never	27.32	6.6	85	0
ALSON LEE	PT107	Female	23-01-1989	35	0	0	never	19.31	6.5	200	1
DAVID KUSHNER	PT108	Female	05-02-1989	35	0	0	No Info	23.86	5.7	85	0
JOANNE HAYES-WHITE	PT110	Female	09-03-1989	35	0	0	never	27.32	5	100	0
ARTHUR KENNEY	PT111	Female	19-03-1989	35	0	0	never	27.32	6.1	85	0
PATRICIA JACKSON	PT112	Female	01-04-1989	35	0	0	former	54.7	6	100	0
EDWARD HARRINGTON	PT113	Female	14-04-1989	35	0	0	former	36.05	5	130	0
JOHN MARTIN	PT114	Female	21-04-1989	35	0	0	never	25.69	5.8	200	0
DAVID FRANKLIN	PT115	Female	26-04-1989	35	0	0	No Info	27.32	5	160	0
SEBASTIAN WONG	PT118	Female	30-04-1989	35	0	0	never	24.48	5.7	158	0
MARTY ROSS	PT119	Female	10-05-1989	35	0	0	No Info	27.32	5.7	80	0
GEORGE GARCIA	PT123	Female	14-06-1989	35	0	0	never	21.24	4.8	85	0
JOSEPH DRISCOLL	PT125	Female	24-06-1989	35	0	0	No Info	13.99	4	140	0
HARLAN KELLY-JR	PT131	Female	06-07-1989	35	0	0	No Info	31.75	4	200	0
GARY AMELIO	PT133	Female	22-07-1989	35	0	0	current	22.01	6.2	126	0
JOHN TURSI	PT134	Female	24-07-1989	35	0	0	never	22.19	3.5	100	0
JOSE VELO	PT135	Female	30-07-1989	35	0	0	never	23.55	5	85	0
SUSAN CURRIN	PT137	Female	04-08-1989	34	0	0	No Info	21.76	4.5	130	0

```
10
11      -- 3. Calculate the average BMI of patients.
12
13 •    Select avg(BMI) as Avg_BMI from diabetes_prediction;
14
15
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Avg_BMI
27.327955333227067

```
14
15      -- 4. List patients in descending order of blood glucose levels.
16
17 •    select Patient_id,Blood_glucose_level from diabetes_prediction order by Blood_glucose_level desc;
18
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

Patient_id	Blood_glucose_level
PT91144	300
PT91135	300
PT90561	300
PT88440	300
PT90590	300
PT91562	300
PT91095	300
PT90006	300
PT89505	300
PT89546	300
PT89934	300
PT90086	300
PT89191	300
PT91250	300
PT89960	300
PT90569	300
PT89459	300

```sql
15
16      -- 5. Find patients who have hypertension and diabetes.
17
18      select * from diabetes_prediction
19      where hypertension = 1 and diabetes = 1;
20
21
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

EmployeeName	Patient_id	Gender	DOB	Age	Hypertension	Heart_disease	Smoking_history	BMI	HbA1c_level	Blood_glucose_level	diabetes
JONES WONG	PT139	Male	09-08-1989	34	1	0	current	27.32	5.7	260	1
PATRIC STEELE	PT205	Female	04-06-1997	27	1	0	never	27.32	6.8	280	1
ARTHUR STELLINI	PT343	Male	07-09-1997	26	1	1	not current	27.77	6.6	160	1
CHAD LAW	PT355	Male	12-09-1997	26	1	0	ever	35.06	5.8	200	1
CATHERINE JAMES	PT451	Female	21-10-1997	26	1	0	never	50.3	6.6	155	1
JOHN HART	PT565	Male	10-11-1997	26	1	0	current	36.12	6.8	140	1
JOHN BARKER	PT567	Female	11-11-1997	26	1	0	former	27.32	6.5	159	1
ROBERT BONNET	PT632	Female	01-12-1997	26	1	0	not current	36.93	8.8	155	1
VITANI BENJAMIN	PT727	Male	24-12-1997	26	1	0	not current	40.86	6.6	159	1
LANNIE ADELMAN	PT828	Female	11-01-1999	25	1	0	not current	27.32	6.1	160	1
JOEL DELIZONNA	PT852	Female	14-01-1999	25	1	0	never	20.09	6.6	200	1
KAREN KUBICK	PT861	Male	16-01-1999	25	1	0	ever	25.94	9	140	1
ANA GONZALEZ	PT983	Female	02-02-1999	25	1	0	No Info	27.32	6.6	240	1
LARRY CAMILLERI	PT1075	Female	11-02-1999	25	1	0	former	36.8	6.5	126	1
EDWARD LEE	PT1123	Female	17-02-1999	25	1	1	former	44.23	8.2	145	1
THOMAS CULLINAN	PT1183	Female	24-02-1999	25	1	0	never	41.76	6.8	300	1
CURTIS CHAN	PT1222	Male	01-03-1999	25	1	0	never	23.55	5.7	300	1
JAMES CUNNINGH...	PT1232	Female	02-03-1999	25	1	0	ever	32.92	7.5	126	1
ELLEN BRIN	PT1236	Female	02-03-1999	25	1	1	never	43.16	8.8	280	1
VICTOR WONG	PT1242	Female	03-03-1999	25	1	0	never	22.48	9	126	1
DAVID DELBON	PT1271	Male	05-03-1999	25	1	0	not current	25.49	6.1	260	1

```sql
23
24      -- 6. Determine the number of patients with heart disease.
25
26      select count(*) as Total_heart_disease_patients from diabetes_prediction where Heart_disease=1;
27
28
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Total_heart_disease_patients
3937

```
25
26      -- 7. Group patients by smoking history and count how many smokers and non-smokers there are.
27
28 •    select smoking_history, count(*) as Total_patients from diabetes_prediction group by smoking_history;
29
30
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

smoking_history	Total_patients
never	35045
No Info	35753
current	9265
former	9324
ever	3997
not current	6434

```sql
-- 8. Retrieve the Patient_id of patients who have a BMI greater than the average BMI.

select Patient_id
from diabetes_prediction
where bmi > (select avg(BMI) from diabetes_prediction);
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

Patient_id
PT109
PT112
PT113
PT117
PT121
PT124
PT126
PT128
PT131
PT140
PT143
PT144
PT149
PT153
PT156
PT160
PT161
PT165
PT168
PT176
PT179
PT181

```sql
-- 9. Find the patient with the highest HbA1c level and the patient with the lowest HbA1clevel.
select Patient_id, MAX(HbA1c_level) as highest_HbA1c, MIN(HbA1c_level) as lowest_HbA1c
from diabetes_prediction group by Patient_id;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

Patient_id	highest_HbA1c	lowest_HbA1c
PT101	6.6	6.6
PT102	6.6	6.6
PT103	5.7	5.7
PT104	5	5
PT105	4.8	4.8
PT106	6.6	6.6
PT107	6.5	6.5
PT108	5.7	5.7
PT109	4.8	4.8
PT110	5	5
PT111	6.1	6.1
PT112	6	6
PT113	5	5
PT114	5.8	5.8
PT115	5	5
PT116	6.6	6.6
PT117	6.1	6.1
PT118	5.7	5.7
PT119	5.7	5.7
PT120	3.5	3.5

```
38      -- 10. Calculate the age of patients in years (assuming the current date as of now).
39
40  •   select Patient_id, timestampdiff(year, STR_TO_DATE(DOB, '%d-%m-%Y'), curdate()) AS Age
41      FROM diabetes_prediction;
42
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

Patient_id	Age
PT101	31
PT102	31
PT103	31
PT104	31
PT105	35
PT106	35
PT107	35
PT108	35
PT109	35
PT110	35
PT111	35
PT112	35
PT113	35
PT114	35
PT115	35
PT116	35

```
44    -- 11. Rank patients by blood glucose level within each gender group.
45
46 •  select *, rank() over (partition by Gender order by Blood_glucose_level desc) as glucose_rank
47    from diabetes_prediction;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

EmployeeName	Patient_id	Gender	DOB	Age	Hypertension	Heart_disease	Smoking_history	BMI	HbA1c_level	Blood_glucose_level	diabetes	glucose_rank
Tualatai Auimatagi	PT98538	Female	30-09-1995	28	0	0	never	26.52	8.2	300	1	1
Seth I Rubenstein	PT98911	Female	30-09-1995	28	0	0	current	40.18	9	300	1	1
Lenora G Banks	PT98454	Female	29-09-1995	28	1	0	never	38.59	6.6	300	1	1
Gilbert J Fragoso	PT99638	Female	23-09-1995	28	1	0	ever	34.3	5.7	300	1	1
Warren Wong	PT97955	Female	28-09-1995	28	0	0	former	37.42	6.1	300	1	1
Clair Wildman	PT92189	Female	21-09-1995	28	0	0	No Info	27.32	6.2	300	1	1
Mary Ann Moran	PT92871	Female	23-09-1995	28	0	0	never	53.4	5.8	300	1	1
Michele A Flowers	PT93343	Female	24-09-1995	28	1	0	current	47.23	6.5	300	1	1
Marc S Slavin	PT93637	Female	25-09-1995	28	0	0	current	45	8.2	300	1	1
Francis W Morris	PT86986	Female	03-09-1995	28	1	0	never	38.58	8.8	300	1	1
James W Vaughn	PT86048	Female	31-08-1995	28	0	0	never	27.32	6.1	300	1	1
Barry K Davis	PT86328	Female	01-09-1995	28	0	0	current	28.12	9	300	1	1
Stephen M Samuel...	PT86684	Female	02-09-1995	28	1	0	never	56.88	5.8	300	1	1
Emina H Abrams	PT87333	Female	04-09-1995	28	0	0	current	27.7	6.5	300	1	1
Jemal J Bailey	PT87598	Female	06-09-1995	28	0	0	never	27.32	6.8	300	1	1
Richard D Vargas	PT87322	Female	04-09-1995	28	1	0	never	32.01	8.2	300	1	1
Lauren A Lester	PT86621	Female	02-09-1995	28	0	0	never	37.85	9	300	1	1
Adoracion Ozaraga	PT90590	Female	16-09-1995	28	0	0	No Info	21.03	6.5	300	1	1
Haroon Ahmad	PT89934	Female	14-09-1995	28	0	0	not current	33.52	5.8	300	1	1
Zandra L Thompson	PT91135	Female	18-09-1995	28	1	0	never	34.51	8.2	300	1	1
Esther E Velonza	PT91743	Female	20-09-1995	28	0	1	not current	22.66	6.8	300	1	1
Sandra R Scott	PT90561	Female	16-09-1995	28	0	0	former	22.81	8.8	300	1	1
Ligia Afu-Li	PT89960	Female	14-09-1995	28	0	0	current	43.25	5.7	300	1	1
Sharanjit K Grewal	PT90569	Female	16-09-1995	28	0	0	ever	38.41	6.2	300	1	1

```sql
-- 12. Update the smoking history of patients who are olderthan 40 to "Ex-smoker."

update diabetes_prediction
set smoking_history = 'Ex-smoker'
where Age > 40;
```

```sql
-- 13. Insert a new patient into the database with sample data.

insert into diabetes_prediction
(EmployeeName, Patient_id, Gender, DOB, Age, Hypertension, Heart_disease,
Smoking_history, BMI, HbA1c_level, Blood_glucose_level, diabetes)
values ('DEVID LEE', 'PT1101', 'Male', '01-01-1990', 34, 0, 0,
'Non-smoker', 25.0, 5.5, 100, 0);

-- 14. Delete all patients with heart disease from the database.

delete from diabetes_prediction where Heart_disease = 1;
```

```sql
-- 15. Find patients who have hypertension but not diabetes using the EXCEPT operator

select * from diabetes_prediction
where Hypertension = 1
Except
select * from diabetes_prediction
where diabetes = 1;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

EmployeeName	Patient_id	Gender	DOB	Age	Hypertension	Heart_disease	Smoking_history	BMI	HbA1c_level	Blood_glucose_level	diabetes
DENISE SCHMITT	PT129	Male	29-06-1989	35	1	0	never	26.47	4	158	0
RAY CRAWFORD	PT155	Female	02-01-1997	27	1	0	never	23.05	4.8	130	0
KENNETH SMITH	PT161	Male	09-03-1997	27	1	0	current	27.86	6.6	145	0
CHARLES SCOTT	PT215	Female	08-06-1997	27	1	0	never	34.2	5.7	140	0
SHANNON SAKOWSKI	PT227	Male	02-07-1997	27	1	0	No Info	28.73	6.6	160	0
MARISA MORET	PT241	Female	13-07-1997	27	1	0	never	44.06	6.5	160	0
STEPHEN TACCHINI	PT326	Female	28-08-1997	26	1	0	never	36.73	6.6	126	0
ANDREW LOGAN	PT339	Male	05-09-1997	26	1	0	No Info	25.31	6	130	0
HAGOP HAJIAN	PT357	Female	13-09-1997	26	1	0	never	21.46	4	80	0
PERRY LEONG	PT377	Female	25-09-1997	26	1	0	No Info	24.29	3.5	90	0
MELISSA LERMA	PT379	Female	26-09-1997	26	1	0	never	27.4	5.7	140	0
JOHN KOSTA	PT446	Female	20-10-1997	26	1	0	not current	22.48	5	158	0

```sql
-- 16. Define a unique constraint on the "patient_id" column to ensure its values are unique.

alter table diabetes_prediction
add constraint unique_patient_id UNIQUE (Patient_id);

-- 17. Create a view that displays the Patient_ids, ages, and BMI of patients.

create view patient_info as
select Patient_id,Age,BMI
FROM diabetes_prediction;

select * from patient_info;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

Patient_id	Age	BMI
PT102	31	27.32
PT103	31	27.32
PT104	31	23.45
PT106	35	27.32
PT107	35	19.31
PT108	35	23.86
PT109	35	33.64
PT110	35	27.32
PT111	35	27.32

18. Suggest improvements in the database schema to reduce data redundancy and improve data integrity.

Here are some suggestions to improve the schema:

- **Normalization:** Consider normalizing the data by breaking it into separate tables. For example, create a separate table for patient demographics (including gender, DOB, etc.) and another for health-related information (hypertension, heart disease, etc.). This reduces redundancy and ensures data integrity.
- **Foreign Keys:** Use foreign keys to establish relationships between related tables. For instance, link patient demographics to health data using a common identifier (e.g., Patient_id).
- **Data Types:** Ensure appropriate column data types (e.g., use DATE for D.O.B, INT for hypertension, etc.).
- **Constraints:** Besides the unique constraint, consider other constraints like NOT NULL, CHECK, and DEFAULT to enforce data rules.

19. Explain how you can optimize the performance of SQL queries on this dataset.

To enhance query performance:

- **Indexes**: Create indexes on frequently queried columns (e.g., Patient_id, gender, etc.). Indexes speed up data retrieval.
- **Avoid SELECT**: Specify only the necessary columns in SELECT statements instead of using SELECT *.
- **JOIN Optimization**: Optimize JOIN operations by choosing the right JOIN type (INNER, LEFT, etc.) and ensuring indexed columns are used.
- **Query Tuning**: Analyze query execution plans, identify bottlenecks, and optimize accordingly.
- **Caching**: Use caching mechanisms (e.g., database-level caching or application-level caching) to reduce redundant queries.
- **Partitioning**: If the dataset is large, consider partitioning tables based on specific criteria (e.g., date ranges).

Thank You