

# Diabetes Prediction Analysis

PYSLIQ TASK III

# ❖ Retrieve the Patient\_id and ages of all patients.

➤ **Select Pateint\_ID,  
Age FROM  
DiabetesPredict ;**

patient_id	age
PT102	54
PT103	28
PT104	36
PT106	20
PT107	44
PT108	79
PT109	42
PT110	32
PT111	53
PT112	54
PT113	78
PT114	67
PT115	76
PT116	78
PT117	15
PT118	42
PT119	42
PT120	37
PT121	40
PT122	5
PT123	69

# ❖ Select all female patients who are older than 40.

Select \* from DiabetesPredict  
Where gender = 'Female' AND  
age > 40;

Result Grid

Filter Rows

Export

Wrap Cell Content:

Fetch rows

	EmployeeName	patient_id	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
▶	GARY JIMENEZ	PT102	Female	54	0	0	Ex-smoker	27.32	6.6	80	0
	ALSON LEE	PT107	Female	44	0	0	never	19.31	6.5	200	1
	DAVID KUSHNER	PT108	Female	79	0	0	Ex-smoker	23.86	5.7	85	0
	ARTHUR KENNEY	PT111	Female	53	0	0	Ex-smoker	27.32	6.1	85	0
	PATRICIA JACKSON	PT112	Female	54	0	0	Ex-smoker	54.7	6	100	0
	EDWARD HARRINGTON	PT113	Female	78	0	0	Ex-smoker	36.05	5	130	0
	JOHN MARTIN	PT114	Female	67	0	0	Ex-smoker	25.69	5.8	200	0
	DAVID FRANKLIN	PT115	Female	76	0	0	Ex-smoker	27.32	5	160	0
	SEBASTIAN WONG	PT118	Female	42	0	0	never	24.48	5.7	158	0
	MARTY ROSS	PT119	Female	42	0	0	No Info	27.32	5.7	80	0
	GEORGE GARCIA	PT123	Female	69	0	0	Ex-smoker	21.24	4.8	85	0
	HARLAN KELLY-JR	PT131	Female	53	0	0	Ex-smoker	31.75	4	200	0
	GARY AMELIO	PT133	Female	41	0	0	current	22.01	6.2	126	0
	JOSE VELO	PT135	Female	76	0	0	Ex-smoker	23.55	5	85	0
	MICHAEL THOMPSON	PT144	Female	66	0	0	Ex-smoker	29.3	4.8	159	0
	SHARON MCCOLE WIC...	PT145	Female	67	0	0	Ex-smoker	27.32	3.5	160	0
	EDWIN LEE	PT146	Female	44	0	0	never	24.93	6.1	100	0
	TRENT RHORER	PT148	Female	60	0	0	Ex-smoker	18.03	4	159	0

## ❖ Calculate the average BMI of patients

Select AVG(bmi) FROM DiabetesPredict;

	AVG(bmi)
▶	27.238479476151184

## ❖ List patients in descending order of blood glucose levels

Select \* FROM DiabetesPredict ORDER BY blood\_glucose\_level Desc

EmployeeName	patient_id	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
Seth I Rubenstein	PT98911	Female	60	0	0	Ex-smoker	40.18	9	300	1
Philip Tran	PT99008	Male	69	0	0	Ex-smoker	31.56	7	300	1
Gilbert J Fragoso	PT99638	Female	67	1	0	Ex-smoker	34.3	5.7	300	1
Amado A Lumas Jr	PT99663	Male	56	1	0	Ex-smoker	28.47	6.1	300	1
Shanice M Guidry	PT99672	Male	57	1	0	Ex-smoker	41.93	5.7	300	1
Angelica J Young	PT99764	Male	80	0	0	Ex-smoker	34	9	300	1
Flor D Roman	PT99809	Male	62	0	0	Ex-smoker	27.32	6	300	1
Josephine C Cabrera	PT99968	Male	64	1	0	Ex-smoker	33.12	5.7	300	1
Lubov Ovtchinikova	PT96144	Male	43	1	0	former	35.39	9	300	1
John G Alexander	PT96269	Male	62	1	0	Ex-smoker	26.32	7	300	1
Erin C Joakimson	PT96328	Female	65	1	0	Ex-smoker	30.66	6.1	300	1
Janice Lee	PT96346	Female	80	0	0	Ex-smoker	21.63	6.1	300	1
Benjamin K Wong	PT96351	Male	80	1	0	Ex-smoker	30.68	5.7	300	1
Nadine R Gordon	PT96371	Female	55	0	0	Ex-smoker	26.95	5.8	300	1

# ❖ Find patients who have hypertension and diabetes.

Select \* FROM DiabetesPredict WHERE hypertension = 1 AND diabetes =1;

EmployeeName	patient_id	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
JONES WONG	PT139	Male	50	1	0	current	27.32	5.7	260	1
PATRIC STEELE	PT205	Female	80	1	0	Ex-smoker	27.32	6.8	280	1
CHAD LAW	PT355	Male	63	1	0	Ex-smoker	35.06	5.8	200	1
CATHERINE JAMES	PT451	Female	52	1	0	Ex-smoker	50.3	6.6	155	1
JOHN HART	PT565	Male	48	1	0	current	36.12	6.8	140	1
JOHN BARKER	PT567	Female	79	1	0	Ex-smoker	27.32	6.5	159	1
ROBERT BONNET	PT632	Female	49	1	0	not current	36.93	8.8	155	1
VITANI BENJAMIN	PT727	Male	43	1	0	not current	40.86	6.6	159	1
LANNIE ADELMAN	PT828	Female	38	1	0	not current	27.32	6.1	160	1
JOEL DELIZONNA	PT852	Female	28	1	0	never	20.09	6.6	200	1
KAREN KUBICK	PT861	Male	59	1	0	Ex-smoker	25.94	9	140	1
ANA GONZALEZ	PT983	Female	75	1	0	Ex-smoker	27.32	6.6	240	1
LARRY CAMILLERI	PT1075	Female	44	1	0	former	36.8	6.5	126	1
THOMAS CULLINAN	PT1183	Female	53	1	0	Ex-smoker	41.76	6.8	300	1
CURTIS CHAN	PT1222	Male	59	1	0	Ex-smoker	23.55	5.7	300	1
JAMES CUNNINGHAM	PT1232	Female	78	1	0	Ex-smoker	32.92	7.5	126	1
VICTOR WONG	PT1242	Female	54	1	0	Ex-smoker	22.48	9	126	1
DAVID DELBON	PT1271	Male	50	1	0	not current	25.49	6.1	260	1

## ❖ Determine the number of patients with heart disease

```
Select count(*) FROM heart_disease FROM DiabetesPredict WHERE  
heart_disease = 1
```

	A	B
1	COUNT(*)	
2	3942	
3		
.		

❖ Group patients by smoking history and count how many smokers and nonsmokers there are

► **Select smoking\_history,count(\*) AS number\_of\_patients FROM DiabetesPredict GROUP BY Smoking\_history**

smoking_history	number_of_patients
Ex-smoker	34821
never	21404
current	6006
No Info	25382
ever	2081
former	2995
not current	3370



- ❖ Retrieve the Patient\_ids of patients who have a BMI greater than the average BMI.

- **Select \* FROM DiabetesPredict WHERE bmi > (Selct AVG(bmi) FROM DiabetesPredict) ORDER BY bmi;**

A	B	C	D	E	F	G	H	I	J	K
EmployeeName	patient_id	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
Julan L Cheung	PT98610	Male	45	0	0	current	27.2	5.7	80	0
Jeremiah D Lehane	PT84399	Male	30	0	0	current	27.2	6.1	158	0
Victor Raquinan	PT87215	Female	80	0	0	Ex-smoker	27.2	4.5	85	0
Cassie Naughton	PT70939	Male	24	0	0	not current	27.2	5	126	0
Anthony Joslin	PT72343	Female	80	0	0	Ex-smoker	27.2	4.8	155	0
Keena Middleton	PT72648	Female	80	0	0	Ex-smoker	27.2	4.8	130	0
Wan Hong Kuang	PT66590	Female	66	0	0	Ex-smoker	27.2	7.5	160	1
Megan Alferness	PT68923	Female	69	0	0	Ex-smoker	27.2	5.8	85	0
Cristian Vargas	PT69268	Female	50	0	0	former	27.2	5	130	0
MARICHU GLOVER	PT23721	Female	41	0	0	No Info	27.2	6.5	130	0
WILSON PHAM	PT8549	Male	44	0	0	former	27.2	3.5	160	0
NOIME VENTENILLA	PT3863	Male	17	0	0	No Info	27.2	6.2	160	0
CECILIA ADIAZ	PT14514	Female	57	0	0	Ex-smoker	27.2	5.8	100	0
MARTIN SMITH	PT15688	Female	54	0	0	Ex-smoker	27.2	5.8	126	1

- ❖ Find the patient with the highest HbA1c level and the patient with the lowest HbA1c level.

**Select \* FROM DiabetesPredict ORDER BY HbA1c\_level LIMIT 1;**

	EmployeeName	patient_id	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
▶	Meredith H Reddoch-Ho	PT100000	Male	45	0	0	never	28.61	3.5	80	0

- ❖ Calculate the age of patients in years (assuming the current date as of now).

Select EmployeeName,Pateint\_id,ABS

(age-Year(now()))

AS year\_of\_birth

FROM DiabetesPredict


	A	B	C
1	EmployeeName	patient_id	year_of_birth
2	GARY JIMENEZ	PT102	1970
3	ALBERT PARDINI	PT103	1996
4	CHRISTOPHER CHONG	PT104	1988
5	DAVID SULLIVAN	PT106	2004
6	ALSON LEE	PT107	1980
7	DAVID KUSHNER	PT108	1945
8	MICHAEL MORRIS	PT109	1982
9	JOANNE HAYES-WHITE	PT110	1992
10	ARTHUR KENNEY	PT111	1971
11	PATRICIA JACKSON	PT112	1970
12	EDWARD HARRINGTON	PT113	1946

## ❖ Rank patients by blood glucose level within each gender group

Select employeename,patient\_id,gender,blood\_glucose\_level,Row\_Number()  
Over(ORDER BY blood\_glucose\_level DESC)

AS patient\_rank FROM DiabetesPredict

	A	B	C	D	E
1	employeename	patient_id	gender	blood_glucose_level	patient_rank
2	Grace Gancayco	PT97671	Female	300	1
3	Idalia R Farina	PT97708	Female	300	2
4	Warren Wong	PT97955	Female	300	3
5	Adrian G Mendez	PT98419	Male	300	4
6	Lenora G Banks	PT98454	Female	300	5
7	Dante Rogayan	PT98461	Male	300	6
8	Tinisha C Bishop	PT98500	Male	300	7
9	Tualatai Auimatagi	PT98538	Female	300	8
10	Michelle D McGee	PT98852	Male	300	9
11	Lawrence Shum	PT98855	Male	300	10
12	Seth I Rubenstein	PT98911	Female	300	11
13	Philip Tran	PT99008	Male	300	12
14	Gilbert J Fragoso	PT99638	Female	300	13
15	Amado A Lumas Jr	PT99663	Male	300	14
16	Shanice M Guidry	PT99672	Male	300	15
17	Angelica J Young	PT99764	Male	300	16
18	Flor D Roman	PT99809	Male	300	17
19	Josephine C Cabrera	PT99968	Male	300	18
20	Mark S Lui	PT83573	Female	300	19



❖ Update the smoking history of patients who are older than 50 to "Ex-smoker"

► `UPDATE DiabetersPredict SET smoking_history = "Ex-smoker" WHERE age > 50;`

## ❖ Insert a new patient into the database with sample data

```
INSERT INTO DiabetesPredict VALUES ("ABC", "PT12345", "Female", 20, 0, 0,  
"Current", 21.1, 5.2, 77, 0);
```

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

```
DELETE FROM DiabetesPredict WHERE heart_disease = 1;
```

# ❖ Find patients who have hypertension but not diabetes using the EXCEPT operator.

Select \* FROM DiabetesPredict WHERE hypertension = 1 EXCEPT Select \* FROM DiabetesPredict WHERE diabetes =1

A	B	C	D	E	F	G	H	I	J	K
EmployeeName	patient_id	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
DENISE SCHMITT	PT129	Male	45	1	0	never	26.47	4	158	0
RAY CRAWFORD	PT155	Female	45	1	0	never	23.05	4.8	130	0
KENNETH SMITH	PT161	Male	44	1	0	current	27.86	6.6	145	0
CHARLES SCOTT	PT215	Female	55	1	0	Ex-smoker	34.2	5.7	140	0
SHANNON SAKOWSKI	PT227	Male	79	1	0	Ex-smoker	28.73	6.6	160	0
MARISA MORET	PT241	Female	80	1	0	Ex-smoker	44.06	6.5	160	0
STEPHEN TACCHINI	PT326	Female	48	1	0	never	36.73	6.6	126	0
ANDREW LOGAN	PT339	Male	59	1	0	Ex-smoker	25.31	6	130	0
HAGOP HAJIAN	PT357	Female	52	1	0	Ex-smoker	21.46	4	80	0
PERRY LEONG	PT377	Female	48	1	0	No Info	24.29	3.5	90	0
MELISSA LERMA	PT379	Female	59	1	0	Ex-smoker	27.4	5.7	140	0
JOHN KOSTA	PT446	Female	52	1	0	Ex-smoker	22.48	5	158	0
MOHAMMED NURU	PT474	Female	42	1	0	never	39.29	6.6	159	0
JASON PAW	PT475	Female	60	1	0	Ex-smoker	48.02	6	85	0
DANIEL ARMENTA	PT476	Female	55	1	0	Ex-smoker	32.2	6	140	0
JALAL AINEB	PT506	Male	66	1	0	Ex-smoker	31.28	3.5	80	0
MICHAEL DALY	PT507	Male	77	1	0	Ex-smoker	26.32	5.8	90	0



- ❖ Define a unique constraint on the "patient\_id" column to ensure its values are unique

```
ALTER TABLE DiabetesPredict MODIFY patient_id VARCHAR(255) UNIQUE;
```

❖ Create a view that displays the Patient\_ids, ages, and BMI of patients.

Create Table bmi\_of\_patient AS (Select patient\_id,age,bmi FROM DiabetesPredict);

Select \* From bmi\_of\_patient;

	A	B	C
1	patient_id	age	bmi
2	PT102	54	27.32
3	PT103	28	27.32
4	PT104	36	23.45
5	PT106	20	27.32
6	PT107	44	19.31
7	PT108	79	23.86
8	PT109	42	33.64
9	PT110	32	27.32
10	PT111	53	27.32
11	PT112	54	54.7
12	PT113	78	36.05
13	PT114	67	25.69
14	PT115	76	27.32
15	PT116	78	27.32
16	PT117	15	30.36
17	PT118	42	24.48
18	PT119	42	27.32
19	PT120	37	25.72

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

## 1. Normalization :

- Decompose tables into smaller, related entities to eliminate redundant data.
- Apply 1st, 2nd, or 3rd normal form to organize data efficiently.

## 2. Use of Foreign Keys :

- Establish relationships between tables using foreign keys.
- Ensure referential integrity by enforcing constraints on foreign key relationships.

## 3. Elimination of Repeated Data :

- Identify repeating groups of data and create separate tables for them.
- Use lookup tables for frequently repeated values to reduce storage and improve consistency.

## 4. Normalization Forms :

- Evaluate database schema against normalization forms and normalize further if needed.
- Avoid data duplication by organizing data into atomic, smallest logical units.

## 5. Unique Constraints :

- Apply unique constraints to ensure each value in a column is unique.
- Prevent duplicate entries and enforce data integrity.



## **6. Use of Composite Keys :**

- Combine multiple columns to create composite keys for uniquely identifying records.
- Enhance data integrity by ensuring unique combinations of values.

## **7. Data Redundancy Analysis :**

- Analyze existing data for redundant information.
- Identify patterns or duplicated data points and refactor the schema accordingly.

## **8. Normalization Guidelines :**

- Follow normalization guidelines such as minimizing data redundancy and dependency.
- Split large tables into smaller ones to manage data more effectively.

## **9. Normalization Benefits :**

- Improved data consistency and accuracy.
- Reduced storage requirements and improved query performance.

## **10. Regular Schema Review :**

- Schedule periodic reviews of the database schema to identify areas for optimization.
- Continuously refine the schema based on changing business requirements and usage patterns.

Implementing these improvements will lead to a more efficient database schema with reduced data redundancy and enhanced data integrity.

# ❖ Explain how you can optimize the performance of SQL queries on this dataset

## 1. Understanding Indexing :

- Use indexing on frequently queried columns.
- Analyze the data distribution for effective index creation.

## 2. Query Optimization Techniques :

- Utilize proper join strategies (nested loops, hash joins, etc.).
- Rewrite complex queries to simpler forms.
- Employ WHERE clause efficiently for filtering.

## 3. Database Statistics Maintenance :

- Regularly update statistics to ensure query optimizer's accuracy.
- Monitor and analyze execution plans for query optimization opportunities.

## 4. Table Partitioning :

- Partition large tables to enhance query performance.
- Implement partition pruning for minimizing data access.

## 5. Optimizing Joins :

- Choose appropriate join methods based on data size and indexing.
- Avoid unnecessary joins and use join hints if necessary.



## **6. Data Normalization and Denormalization :**

- Normalize database schema for efficient storage.
- Consider denormalization for read-heavy applications to reduce join overhead.

## **7. Query Caching :**

- Implement caching mechanisms for frequently accessed data.
- Utilize database cache or application-level caching where applicable.

## **8. Resource Utilization :**

- Optimize server resources like memory, CPU, and disk I/O.
- Allocate sufficient resources for the database server.

## **9. Monitoring and Profiling :**

- Monitor query performance using tools like SQL Profiler.
- Profile queries to identify bottlenecks and areas for improvement.

## **10. Regular Maintenance :**

- Schedule routine database maintenance tasks like index reorganization and database cleanup.
- Regularly review and optimize SQL code for efficiency.

Implementing these strategies can significantly enhance the performance of SQL queries on the dataset.