

DATA ANALYST INTERNSHIP

TASK-3

DIABETES PREDICTION DATASET

BY :- SWARNAV KUMAR



Retrieve the Patient_id and ages of all patients.

Classroom x diabetesprediction_dataset

Limit to 1000 rows

```
1 • SELECT Patient_id, age
2 FROM employee
3 GROUP BY Patient_id, age;
4
5
```

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

	Patient_id	age
▶	PT101	80
	PT102	54
	PT103	28
	PT104	36
	PT105	76
	PT74389	20
	PT74390	36
	PT74391	80
	PT74392	32
	PT74393	59
	PT74394	11
	PT74395	25
	PT74396	0

employee 1 x Read Only

Result Grid

Form Editor

Field Types

Select all female patients who are older than 40.

Classroom x diabetesprediction_dataset

Limit to 1000 rows

```

1 • SELECT *
2   FROM employee
3   WHERE gender='Female' AND age>40;

```

Result Grid

	EmployeeName	Patient_id	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
▶	NATHANIEL FORD	PT101	Female	80	0	1	never	25.19	6.6	126	0
	Linette Martinez	PT101	Female	80	0	1	never	25.19	6.6	140	0
	GARY JIMENEZ	PT102	Female	54	0	0	No Info	27.32	6.6	80	0
	Martin S Bandvik	PT102	Female	54	0	0	No Info	27.32	6.6	80	0
	Terence G White	PT74393	Female	59	0	0	No Info	29.63	4	126	0
	Terence G White	PT74393	Female	59	0	0	No Info	29.63	4	126	0
	JOHN MARTIN	PT114	Female	67	0	0	never	25.69	5.8	200	0
	JOHN MARTIN	PT114	Female	67	0	0	never	25.69	5.8	200	0
	DAVID FRANKLIN	PT115	Female	76	0	0	No Info	27.32	5	160	0
	DAVID FRANKLIN	PT115	Female	76	0	0	No Info	27.32	5	160	0
	Andre L Brown	PT74401	Female	56	0	0	former	37	5.8	90	0
	Andre L Brown	PT74401	Female	56	0	0	former	37	5.8	90	0
	John M Marian	PT74405	Female	67	0	0	No Info	27.32	3.5	130	0

employee 7 x

Read Only

Calculate the average BMI of patients.

Classroom* x diabetesprediction_dataset

Limit to 1000 rows

```
1 • SELECT avg(bmi) FROM employee;  
2
```

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

	avg(bmi)
▶	27.313497019505462

Result Grid
Form Editor

List patients in descending order of blood glucose levels.

Classroom* x diabetesprediction_dataset

Limit to 1000 rows

```

1 • SELECT *
2   FROM employee
3   ORDER BY blood_glucose_level DESC;

```

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
▶	GLORIA CARLOS	PT19277	Male	42	0	0	never	29.26	6	300	1
	MARVIN MOUTON	PT19284	Male	41	1	0	former	27.32	6.1	300	1
	Maurice R Munsell	PT92530	Male	64	0	0	not current	30.2	6.6	300	1
	Edson Marquez	PT92581	Male	68	0	0	never	31.65	7.5	300	0
	DIANE CHEW	PT92581	Male	68	0	0	never	31.65	7.5	300	1
	John J Delfin	PT92677	Female	13	0	0	No Info	21.73	9	300	1
	Patricia Barragan	PT92758	Female	41	0	1	never	23.5	5.7	300	1
	Susan A Salvador	PT92859	Male	80	0	0	never	34.61	7	300	1
	ANDREA PITTMAN	PT19642	Female	12	0	0	never	53.4	5.8	300	1
	Karima C Baptiste	PT92895	Female	80	0	0	never	24.12	5.8	300	1
	LAURA KIDD	PT13300	Female	40	0	0	current	33.25	5.7	300	1
	LAURA KIDD	PT13300	Female	40	0	0	current	33.25	5.7	300	1
	Haroon A Razzak	PT86611	Female	32	0	1	former	25.59	9	300	1

employee 6 x

Result Grid
Form Editor
Field Types
Read Only

Find patients who have hypertension and diabetes.

Classroom* x diabetesprediction_dataset

Limit to 1000 rows

```

1 • SELECT *
2 FROM employee
3 WHERE hypertension=1 AND diabetes=1;
4

```

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

	EmployeeName	Patient_id	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
▶	Garreth Miller	PT74437	Female	46	1	0	former	33.78	4.5	155	1
	Sheryl Calixta P Ronquillo	PT74478	Female	80	1	0	never	27.32	6.8	280	1
	Sheryl Calixta P Ronquillo	PT74478	Female	80	1	0	never	27.32	6.8	280	1
	ARTHUR STELLINI	PT343	Male	57	1	1	not current	27.77	6.6	160	1
	DEBBIE TAM	PT392	Female	65	1	0	never	38.56	8.8	280	1
	JOANNE HOEPER	PT400	Female	63	1	1	not current	42	6.6	160	1
	Joseph A Leonardini	PT74658	Female	39	1	0	never	31.11	4.8	159	1
	JOHNSON YOU	PT74665	Female	56	1	1	never	32.24	6.8	126	1
	DANIEL ARMENTA	PT476	Female	55	1	0	never	29.18	6.8	220	1
	Sidney K Sakurai	PT74741	Female	76	1	0	not current	30.43	6.8	280	1
	Kevin M Omalley	PT74792	Male	67	1	0	never	26.53	8.8	159	1
	JOHN HART	PT565	Male	48	1	0	never	27.32	6	300	1
	Triscila F Cael	PT74811	Female	80	1	0	former	27.32	6.5	159	1

employee 9 x

Result Grid
Form Editor
Field Types
Read Only

Determine the number of patients with heart disease.

Classroom* x diabetesprediction_dataset

Limit to 1000 rows

```
1 SELECT COUNT(EmployeeName)
2 FROM employee
3 WHERE heart_disease='1';
4
```

Result Grid

	COUNT(EmployeeName)
▶	1544

Export: | Wrap Cell Contents:

Result Grid

Form Editor

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

Classroom* x diabetesprediction_dataset

Limit to 1000 rows

```
1 • USE employee;
2
3 • SELECT smoking_history, COUNT(*) AS patient_count
4 FROM employee
5 WHERE smoking_history IN ('current', 'never')
6 GROUP BY smoking_history;
7
```

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

	smoking_history	patient_count
▶	never	13455
	current	3567

Result Grid
Form Editor



PSYLIQ

Retrieve the Patient ids of patients who have a BMI greater than the average BMI.

Classroom* x diabetesprediction_dataset

Limit to 1000 rows

```
1 • SELECT Patient_id,bmi
2 FROM employee
3 WHERE bmi> (SELECT AVG(bmi) FROM employee);
4
```

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

	Patient_id	bmi
▶	PT102	27.32
	PT102	27.32
	PT103	27.32
	PT103	27.32
	PT74389	27.32
	PT74389	27.32
	PT74390	27.32
	PT74390	27.32
	PT74391	27.32
	PT74391	27.32
	PT74392	32.41
	PT74392	32.41
	PT74393	29.63
	PT74393	29.63
	PT74395	32.56
	PT74395	32.56

employee 2 x

Result Grid
Form Editor
Field Types
Query Stats
Read Only



PSYUQ

Find the patient with the highest HbA1c level and the patient with the lowest HbA1clevel.

Classroom* x diabetesprediction_dataset

Limit to 1000 rows

```
1 • SELECT *
2 FROM employee
3 WHERE HbA1c_level = (SELECT MAX(HbA1c_level) FROM employee)
4 OR HbA1c_level = (SELECT MIN(HbA1c_level) FROM employee);
5
-
```

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
▶	ALSON LEE	PT74390	Female	36	0	0	never	27.32	3.5	100	0
	ALSON LEE	PT74390	Female	36	0	0	never	27.32	3.5	100	0
	John M Marian	PT74405	Female	67	0	0	No Info	27.32	3.5	130	0
	John M Marian	PT74405	Female	67	0	0	No Info	27.32	3.5	130	0
	Amen Y Chow	PT74413	Male	44	0	0	never	31.91	3.5	100	0
	HARLAN KELLY-JR	PT74413	Male	44	0	0	never	31.91	3.5	100	0
	Heather A Piper	PT74433	Female	8	0	0	No Info	21.01	3.5	130	0
	KIRK RICHARDSON	PT74433	Female	8	0	0	No Info	21.01	3.5	130	0
	Kevin K Chin	PT74435	Male	39	0	0	No Info	27.32	3.5	90	0
	MICHAEL ROLOVICH	PT74435	Male	39	0	0	No Info	27.32	3.5	90	0
	Tara M Steeley	PT74441	Male	48	1	0	No Info	31.53	3.5	158	0
	DOUGLAS RIBA	PT74441	Male	48	1	0	No Info	31.53	3.5	158	0
	Kathryn L Miller	PT74442	Female	37	0	0	current	35.93	3.5	159	0
	AI-KYUNG CHUNG	PT74442	Female	37	0	0	current	35.93	3.5	159	0
	Matthew J Lee	PT74454	Male	43	0	0	not current	27.32	3.5	126	0
	Louis W Wong	PT74469	Female	30	0	0	No Info	27.32	3.5	160	0

employee 3 x

Result Grid
Form Editor
Field Types
Query Stats
Read Only



PSYUQ

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

Classroom* x

Limit to 1000 rows

```
1 • SELECT EmployeeName, Patient_id,  
2     YEAR(NOW()) - age AS Birth_year,  
3     YEAR(NOW()) - YEAR(NOW()) + age AS Current_age  
4 FROM employee;  
5
```

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

	EmployeeName	Patient_id	Birth_year	Current_age
▶	GARY JIMENEZ	PT102	1969	54
	Martin S Bandvik	PT102	1969	54
	ALBERT PARDINI	PT103	1995	28
	ALBERT PARDINI	PT103	1995	28
	CHRISTOPHER CHONG	PT104	1987	36
	Michele L Grindstaff	PT104	1987	36
	Mary A Angel	PT74389	2003	20
	Mary A Angel	PT74389	2003	20
	ALSON LEE	PT74390	1987	36
	ALSON LEE	PT74390	1987	36
	Kimberly K Hiroshima	PT74392	1991	32
	Kimberly K Hiroshima	PT74392	1991	32
	Terence G White	PT74393	1964	59
	Terence G White	PT74393	1964	59
	Vicente Mayor	PT74394	2012	11
	Vicente Mayor	PT74394	2012	11
	Kevin G Labanowski	PT74395	1998	25
	Kevin G Labanowski	PT74395	1998	25
	John M Robertson	PT74396	2015	8
	John M Robertson	PT74396	2015	8

Result 1 x

Read Only

Result Grid
Form Editor
Field Types
Query Stats
Execution Plan



PSYUQ

Rank patients by blood glucose level within each gender group.

Classroom* x diabetesprediction_dataset

Limit to 1000 rows

```
1 • SELECT Patient_id, gender, blood_glucose_level,
2     RANK() OVER (PARTITION BY gender ORDER BY blood_glucose_level) AS blood_glucose_rank
3     FROM employee;
4
```

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

	Patient_id	gender	blood_glucose_level	blood_glucose_rank
▶	PT19032	Female	80	1
	PT19274	Female	80	1
	PT19171	Female	80	1
	PT19387	Female	80	1
	PT19176	Female	80	1
	PT18621	Female	80	1
	PT91790	Female	80	1
	PT91855	Female	80	1
	PT19621	Female	80	1
	PT19021	Female	80	1
	PT92381	Female	80	1
	PT92084	Female	80	1
	PT91857	Female	80	1
	PT19139	Female	80	1
	PT19139	Female	80	1
	PT92603	Female	80	1

Result 4 x

Result Grid
Form Editor
Field Types
Query Stats
Read Only

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

Classroom* x diabetesprediction_dataset

Limit to 1000 rows

```

1 • SET SQL_SAFE_UPDATES=0;
2 • UPDATE employee
3   SET smoking_history= "Ex-smoke"
4   WHERE age> "50";
5
6 • SELECT * FROM employee;

```

Result Grid

	EmployeeName	Patient_id	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
▶	NATHANIEL FORD	PT101	Female	80	0	1	Ex-smoke	25.19	6.6	126	0
	Linette Martinez	PT101	Female	80	0	1	Ex-smoke	25.19	6.6	140	0
	GARY JIMENEZ	PT102	Female	54	0	0	Ex-smoke	27.32	6.6	80	0
	Martin S Bandvik	PT102	Female	54	0	0	Ex-smoke	27.32	6.6	80	0
	ALBERT PARDINI	PT103	Male	28	0	0	never	27.32	5.7	158	0
	ALBERT PARDINI	PT103	Male	28	0	0	never	27.32	5.7	158	0
	CHRISTOPHER CHONG	PT104	Female	36	0	0	current	23.45	5	155	1
	Michele L Grindstaff	PT104	Female	36	0	0	current	23.45	5	155	0
	PATRICK GARDNER	PT105	Male	76	1	1	Ex-smoke	20.14	4	155	0
	PATRICK GARDNER	PT105	Male	76	1	1	Ex-smoke	20.14	4	155	0
	Mary A Angel	PT74389	Female	20	0	0	never	27.32	6.6	85	0
	Mary A Angel	PT74389	Female	20	0	0	never	27.32	6.6	85	0
	ALSON LEE	PT74390	Female	36	0	0	never	27.32	3.5	100	0
	ALSON LEE	PT74390	Female	36	0	0	never	27.32	3.5	100	0
	John R Torrise	PT74391	Male	80	0	1	Ex-smoke	27.32	5	160	0
	John R Torrise	PT74391	Male	80	0	1	Ex-smoke	27.32	5	160	0
	Kimberly K Hiroshima	PT74392	Male	32	0	0	not current	32.41	6.5	80	0
	Kimberly K Hiroshima	PT74392	Male	32	0	0	not current	32.41	6.5	80	0

employee 1 x

Read Only



Insert a new patient into the database with sample data.

Classroom* x diabetesprediction_dataset

Limit to 1000 rows

```
1  INSERT INTO employee
2  VALUES("JACK JEN", "PT100101", "Male", "48", "0", "0", "never", "18.15", "6", "168", "0");
3
4  • SELECT * FROM employee
5  WHERE EmployeeName= "JACK JEN";
6  |
```

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

	EmployeeName	Patient_id	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
▶	JACK JEN	PT100101	Male	48	0	0	never	18.15	6	168	0

Result Grid
Form Editor
Field Types



PSYUQ

Delete all patients with heart disease from the database.

Classroom x diabetesprediction_dataset

Limit to 1000 rows

```
1 • DELETE FROM employee
2   WHERE heart_disease='1';
3
4 • SELECT * FROM employee;
5
```

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-smoke	27.32	6.6	80	0
	Martin S Bandvik	PT102	Female	54	0	0	Ex-smoke	27.32	6.6	80	0
	ALBERT PARDINI	PT103	Male	28	0	0	never	27.32	5.7	158	0
	ALBERT PARDINI	PT103	Male	28	0	0	never	27.32	5.7	158	0
	CHRISTOPHER CHONG	PT104	Female	36	0	0	current	23.45	5	155	1
	Michele L Grindstaff	PT104	Female	36	0	0	current	23.45	5	155	0
	Mary A Angel	PT74389	Female	20	0	0	never	27.32	6.6	85	0
	Mary A Angel	PT74389	Female	20	0	0	never	27.32	6.6	85	0
	ALSON LEE	PT74390	Female	36	0	0	never	27.32	3.5	100	0
	ALSON LEE	PT74390	Female	36	0	0	never	27.32	3.5	100	0
	Kimberly K Hiroshima	PT74392	Male	32	0	0	not current	32.41	6.5	80	0
	Kimberly K Hiroshima	PT74392	Male	32	0	0	not current	32.41	6.5	80	0
	Terence G White	PT74393	Female	59	0	0	Ex-smoke	29.63	4	126	0
	Terence G White	PT74393	Female	59	0	0	Ex-smoke	29.63	4	126	0
	Vicente Mayor	PT74394	Male	11	0	0	No Info	21.26	6.2	130	0
	Vicente Mayor	PT74394	Male	11	0	0	No Info	21.26	6.2	130	0
	Kevin G Labanowski	PT74395	Male	25	0	0	No Info	32.56	6	85	0
	Kevin G Labanowski	PT74395	Male	25	0	0	No Info	32.56	6	85	0
	John M Robertson	PT74396	Male	8	0	0	former	36.05	5	130	0
	John M Robertson	PT74396	Male	8	0	0	former	36.05	5	130	0

employee 3 x

Result Grid
Form Editor
Field Types
Query Stats
Execution Plan
Read Only

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

Classroom x diabetesprediction_dataset

Limit to 1000 rows

```

1 SELECT * FROM employee
2 WHERE hypertension='1'
3 EXCEPT
4 SELECT * FROM employee
5 WHERE diabetes='1';
6

```

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

	EmployeeName	Patient_id	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
▶	Charles F Schuler	PT74418	Male	69	1	0	Ex-smoke	32.86	6.6	155	0
	KHOA TRINH	PT74418	Male	69	1	0	Ex-smoke	32.86	6.6	155	0
	RAY CRAWFORD	PT74437	Female	46	1	0	former	33.78	4.5	155	0
	Tara M Steeley	PT74441	Male	48	1	0	No Info	31.53	3.5	158	0
	DOUGLAS RIBA	PT74441	Male	48	1	0	No Info	31.53	3.5	158	0
	CHARLES SCOTT	PT215	Female	55	1	0	Ex-smoke	34.2	5.7	140	0
	SHANNON SAKOWSKI	PT227	Male	79	1	0	Ex-smoke	28.73	6.6	160	0
	Karen J Heald	PT227	Male	79	1	0	Ex-smoke	28.73	6.6	160	0
	MARISA MORET	PT241	Female	80	1	0	Ex-smoke	44.06	6.5	160	0
	William B Griffin	PT241	Female	80	1	0	Ex-smoke	44.06	6.5	160	0
	Mark E Mahoney	PT74560	Female	71	1	0	Ex-smoke	24.65	4.5	145	0
	Jimmy D Bui	PT74638	Female	40	1	0	current	27.4	5.7	140	0
	RAYMOND CHAVEZ	PT380	Male	14	1	0	current	27.32	6.5	126	0
	Kevin A Lee	PT74649	Female	65	1	0	Ex-smoke	38.56	8.8	155	0
	MORGAN PETITI	PT401	Female	39	1	0	never	31.11	4.8	160	0
	Daniel Gray	PT74685	Male	80	1	0	Ex-smoke	20.08	4.8	145	0
	MICHAEL GONZALES	PT74685	Male	80	1	0	Ex-smoke	20.08	4.8	145	0
	Dustin L Daza	PT74690	Male	62	1	0	Ex-smoke	47.18	5.8	159	0
	ROBERT MAERZ	PT74690	Male	62	1	0	Ex-smoke	47.18	5.8	159	0

Result 6 x

Read Only



Define a unique constraint on the "patient_id" column to ensure its values are unique.

The screenshot shows the SQL Classroom interface. The main editor displays the following SQL code:

```
1 • ALTER TABLE employee
2   ADD CONSTRAINT Patient_id UNIQUE (Patient_id);
3
4 • INSERT INTO employee
5   VALUES ("JACK PERRY", "PT100101", "Male", "38", "0", "0", "never", "30.50", "5.7", "210", "0");
6
```

The right sidebar shows a message: "Automatic context help is disabled. Use the toolbar to manually get help for the current caret position or to toggle automatic help." Below this are tabs for "Context Help" and "Snippets".

The bottom section is the "Output" window, showing a table of actions:

#	Time	Action	Message	Duration / Fetch
1	19:14:31	ALTER TABLE employee ADD CONSTRAINT Patient_id UNIQUE (Patient_id)	Error Code: 1170. BLOB/TEXT column 'Patient_id' used in key specification without a key length	0.000 sec
2	19:14:43	INSERT INTO employee VALUES ("JACK PERRY", "PT100101", "Male", "38", "0", "0", "never", "30.50", "5.7...	1 row(s) affected	0.000 sec

Since, we added a UNIQUE constraint to 'Patient_id', when we insert a duplicate 'Patient_id' it should show an error.



Classroom

```

1 • CREATE VIEW Patient_BMI_View AS
2   SELECT Patient_id, age, bmi
3   FROM employee;
4
5 • SELECT * FROM Patient_BMI_View;
6

```

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

Patient_id	age	bmi
PT102	54	27.32
PT102	54	27.32
PT103	28	27.32
PT103	28	27.32
PT104	36	23.45
PT104	36	23.45
PT74389	20	27.32
PT74389	20	27.32
PT74390	36	27.32
PT74390	36	27.32
PT74392	32	32.41
PT74392	32	32.41
PT74393	59	29.63
PT74393	59	29.63
PT74394	11	21.26
PT74394	11	21.26
PT74395	25	32.56
PT74395	25	32.56

Patient_BMI_View 3 x

Read Only



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

1. Normalization: Break data into smaller tables to reduce redundancy and maintain consistency.
2. Use Foreign Keys: Link tables via keys to avoid duplicate data and enforce referential integrity.
3. Avoid Multi-valued Attributes: Create separate tables for complex attributes to streamline data.
4. Implement Unique Constraints: Ensure uniqueness of critical fields to prevent duplicates.
5. Utilize Indexing: Improve query performance and enforce uniqueness for faster data retrieval.
6. Regular Maintenance: Conduct periodic checks and updates to maintain data consistency.
7. Use Triggers and Constraints: Enforce business rules and maintain data integrity at the database level.
8. Data Validation: Implement checks to ensure accurate data entry and prevent inconsistencies.
9. Utilize Views: Abstract complex queries to enhance data retrieval efficiency and minimize errors.
10. Proper Data Types: Choose appropriate data types to optimize storage and enforce data consistency.



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

Indexing:

Identify frequently used columns in WHERE clauses and apply indexing to these columns. This helps accelerate query execution by facilitating faster data retrieval.

Query Optimization:

Use proper JOINS, avoid unnecessary SELECT * queries, and employ WHERE clauses efficiently to limit the data processed, thereby reducing query execution time.

Database Statistics Update:

Frequent statistic updates aid query optimizer for accurate query plans based on current data distribution, enhancing performance.

Normalization and Denormalization:

Normalize tables to remove redundancy; denormalize for read-heavy tasks, cutting JOINS to enhance query performance and data retrieval.

Bonus: Consider utilizing stored procedures or views to encapsulate complex queries, optimizing repetitive tasks, and simplifying query execution for better performance.