TASK 3 :- PROVIDING INSIGHTS ON DIABETES PREDICTION DATA USING SQL

ABOUT DATA

- **Dataset consists of 1,00,000 patients diabetes records with 10 columns.**
- **Where, 3 categorical columns such as Employee name, patient id, smoking history and**
- **7 Numerical columns such as age,hypertension,heart disease,bmi,HbAlc level,blood glucose level and diabetes**

1. **Retrieve the Patient_id and ages of all patients.**

MYSQL CODE

```
SELECT Patient_Id,age FROM diabetes_pred_data;
```

OUTPUT

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

2. **Select all female patients who are older than 40**

MYSQL CODE

```
SELECT * FROM diabetes_pred_data WHERE gender="Female" and age>40;
```

OUTPUT

Employee Name	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

3. Calculate the average BMI of patients.

MYSQL CODE

```
SELECT ROUND(avg(bmi),2) as Diabetes_Patients from diabetes_pred_data WHERE diabetes=1;
```

OUTPUT

Average for only Diabetes patients

Diabetes_Patients
32.15

Average For ALL patients

Diabetes_Patients
27.24

4. List patients in descending order of blood glucose levels.

MYSQL CODE

```
SELECT * FROM diabetes_pred_data  ORDER BY blood_glucose_level DESC;
```

OUTPUT

Employee_Name	Patient_Ic	gender	age	hypertensior	heart_disease	smoking_history	bmi	HbA1c_le	blood_glucose_level	diabetes
Adrian G Mendez	PT98419	Male	57	0	0	Ex-smoker	27.32	6.5	300	1
Lenora G Banks	PT98454	Female	76	1	0	Ex-smoker	38.59	6.6	300	1
Dante Rogayan	PT98461	Male	72	0	0	Ex-smoker	27.72	6.6	300	1
Tinisha C Bishop	PT98500	Male	73	0	0	Ex-smoker	27.32	8.8	300	1
Tualatai Auimat...	PT98538	Female	17	0	0	never	26.52	8.2	300	1
Michelle D McGee	PT98852	Male	79	0	0	Ex-smoker	27.32	7.5	300	1
Lawrence Shum	PT98855	Male	43	0	0	former	48.56	6.8	300	1
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

5. Find patients who have hypertension and diabetes

MYSQL CODE

```
SELECT * FROM diabetes_pred_data WHERE hypertension =1 and diabetes =1;
```

OUTPUT

Employee Name	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

6. Determine the number of patients with heart disease.

MYSQL CODE

```
SELECT COUNT(*) AS HEART_DISEASES_PATIENTS FROM diabetes_pred_data WHERE
heart_disease =1;
```

OUTPUT

HEART_DISEASES_PATIENTS
3942

EmployeeName	Patient_id	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
NATHANIEL FORD	PT101	Female	80.0	0	1	never	25.19	6.6	140	0
PATRICK GARDNER	PT105	Male	76.0	1	1	current	20.14	4.8	155	0
VICTOR WYRSCH	PT124	Female	72.0	0	1	former	27.94	6.5	130	0
JOHN HANLEY	PT127	Male	67.0	0	1	not current	27.32	6.5	200	1
THOMAS SIRAGUSA	PT143	Female	77.0	1	1	never	32.02	5.0	159	0
...	...	...	...	...	...	...	...	...	...	...
Clyde L Woods	PT99927	Male	63.0	0	1	No Info	27.32	6.6	300	1
Erlinda Andres	PT99949	Male	80.0	1	1	former	28.79	5.8	90	0
Estelle Yancey	PT100013	Male	80.0	0	1	former	27.32	5.0	140	0
Stephanie Chang	PT100036	Female	65.0	1	1	never	33.55	8.2	140	1
Marquis D Walker	PT100039	Male	55.0	0	1	former	30.42	6.2	300	1

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

MYSQL CODE 1

```
SELECT smoking_history,COUNT(Patient_id) AS Total_Count FROM diabetes_pred_data
GROUP BY smoking_history;
```

MYSQL CODE 2

```
SELECT COUNT(Patient_Id) AS NON_SMOKERS,(SELECT COUNT(Patient_Id) as SMOKER
FROM diabetes_pred_data
WHERE smoking_history IN("ever","Ex-smoker","former","current","not current")) AS
SMOKERS,
(SELECT count(Patient_Id) AS NO_NFO FROM diabetes_pred_data WHERE
smoking_history IN ("No Info")) AS NO_INFO
FROM diabetes_pred_data
WHERE smoking_history IN ("never");
```

OUTPUT

1

smoking_history	Total_Count
Ex-smoker	34822
never	21405
current	6005
No Info	25382
ever	2081
former	2995
not current	3370

2

NON_SMOKERS	SMOKERS	NO_INFO
22502	51194	26305

8. **Retrieve the Patient_ids of patients who have a BMI greater than the average BMI.**

MYSQL CODE

```
SELECT Patient_Id,Employee_Name,bmi
FROM diabetes_pred_data
 WHERE bmi> (SELECT avg(bmi) as Average_BMI FROM diabetes_pred_data);
```

OUTPUT

Patient_Id	Employee_Name	bmi
PT102	GARY JIMENEZ	27.32
PT103	ALBERT PARDINI	27.32
PT106	DAVID SULLIVAN	27.32
PT109	MICHAEL MORRIS	33.64
PT110	JOANNE HAYES-WHITE	27.32
PT111	ARTHUR KENNEY	27.32
PT112	PATRICIA JACKSON	54.7
PT113	EDWARD HARRINGTON	36.05
PT115	DAVID FRANKLIN	27.32
PT116	RICHARD CORRIEA	27.32
PT117	AMY HART	30.36

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

MYSQL CODE

```
SELECT
*  FROM diabetes_pred_data
WHERE HbA1c_level = (SELECT Min(HbA1c_level) as Min_Max_level
FROM   diabetes_pred_data)
OR
HbA1c_level = (SELECT Max(HbA1c_level) as Min_Max_level FROM
diabetes_pred_data);
```

OUTPUT

Employee_Name	Patient_Id	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
ELLEN MOFFATT	PT120	Male	37	0	0	ever	25.72	3.5	159	0
JOHN TURSI	PT134	Female	20	0	0	never	22.19	3.5	100	0
MICHAEL THOMPSON	PT141	Male	73	0	0	Ex-smoker	25.91	9	160	1
SHARON MCCOLE WICHER	PT145	Female	67	0	0	Ex-smoker	27.32	3.5	160	0
KEVIN CASHMAN	PT156	Male	50	0	0	former	37.16	9	159	1
MARK KEARNEY	PT158	Female	19	0	0	never	23.35	3.5	155	0
MONIQUE MOYER	PT174	Male	43	0	0	not current	27.32	3.5	126	0
JOHN HALEY JR	PT213	Male	37	0	0	No Info	27.14	3.5	90	0
KHAIRUL ALI	PT219	Female	12	0	0	No Info	20.9	3.5	158	0
MICHAEL CASTAGNOLA	PT221	Female	36	0	0	No Info	27.32	3.5	160	0
JOHN RAHAIM	PT233	Female	79	0	0	Ex-smoker	27.32	3.5	200	0

25 04:30:09 SELECT * FROM diabetes_pred_data WHERE HbA1c_level =... 8007 row(s) returned 0.140 sec / 0.141 sec

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

MYSQL CODE

```
SELECT Employee_Name,Patient_Id,age,YEAR(current_date())-age AS DOB
from diabetes_pred_data;
```

OUTPUT

Employee_Name	Patient_Id	age	DOB
GARY JIMENEZ	PT102	54	1969
ALBERT PARDINI	PT103	28	1995
CHRISTOPHER CHONG	PT104	36	1987
DAVID SULLIVAN	PT106	20	2003
ALSON LEE	PT107	44	1979
DAVID KUSHNER	PT108	79	1944
MICHAEL MORRIS	PT109	42	1981
JOANNE HAYES-WHITE	PT110	32	1991
ARTHUR KENNEY	PT111	53	1970
PATRICIA JACKSON	PT112	54	1969
EDWARD HARRINGTON	PT113	78	1945

11.Rank patients by blood glucose level within each gender group.

MYSQL CODE

```
SELECT *,
 row_number() over(partition by gender order by blood_glucose_level) AS
 Row_Number_Ranking,
dense_rank() over(partition by gender order by blood_glucose_level) AS Dense_Ranking
FROM diabetes_pred_data;
```

OUTPUT

Ranking order of Female

Employee_Name	Patient_Ic	gender	age	hyper	hear	smoking_h	bmi	HbA1c_le	blood_glucose_level	diabet	Row_Number_Rankin	Dense_Rank
Eve Bekker	PT100...	Female	30	0	0	never	22.88	4.5	80	0	1	1
Jacqueline C Ro...	PT99894	Female	72	1	0	Ex-smo...	50.85	3.5	80	0	2	1
Aisha M Malone	PT100...	Female	0	0	0	No Info	18.37	3.5	80	0	3	1
James E Nelson	PT100...	Female	31	0	0	current	20.23	6	80	0	4	1
Joshua R Mcdo...	PT100...	Female	47	1	0	current	25.48	4.5	80	0	5	1
Edward A Ang	PT100...	Female	21	0	0	No Info	23.04	6.5	80	0	6	1
Jensa Woo	PT100...	Female	54	0	0	Ex-smo...	27.32	6.5	80	0	7	1
Sharon S Young	PT100...	Female	57	0	0	Ex-smo...	33.64	6.2	80	0	8	1
Judi Soto	PT99994	Female	20	0	0	current	28.06	6	80	0	9	1
Benny M Choi	PT99774	Female	57	0	0	Ex-smo...	29.56	4	80	0	10	1
John E Long	PT99842	Female	19	0	0	never	27.32	6.1	80	0	11	1
Michael T Feist	PT99910	Female	34	0	0	No Info	27.32	4.5	80	0	12	1

Ranking order of Male

Employee_Name	Patient_Ic	gender	age	hyper	hear	smoking_h	bmi	HbA1c_le	blood_glucose_level	diabet	Row_Number_Rankin	Dense_Rank
Erica Cline	PT53627	Female	51	0	0	Ex-smo...	31.1	6.2	300	1	56990	18
Pierangelo Espi...	PT54533	Female	49	1	0	ever	34.74	6	300	1	56991	18
Carlos Chavez	PT52821	Male	21	0	0	never	21.48	4	80	0	1	1
Wade Wietgrefe	PT54456	Male	26	0	0	never	25.06	6.5	80	0	2	1
Laura Mae Alpert	PT53426	Male	34	0	0	never	36.28	4	80	0	3	1
Craig Allen	PT53142	Male	56	0	0	Ex-smo...	27.32	5.7	80	0	4	1
Annabelle Arenas	PT53118	Male	62	1	0	Ex-smo...	28.09	6.5	80	0	5	1
Ara Balian	PT54597	Male	8	0	0	No Info	24.18	6.1	80	0	6	1
Cynthia Nicholson	PT54599	Male	31	0	0	never	32.4	5.7	80	0	7	1
Alice Wong	PT53168	Male	6	0	0	No Info	27.32	6.5	80	0	8	1
Alan Thoburn	PT52811	Male	58	0	0	Ex-smo...	21.87	6.1	80	0	9	1
April High	PT53115	Male	16	0	0	current	32.41	4.8	80	0	10	1
Fay Hui	PT54258	Male	41	0	0	current	27.32	5.7	80	0	11	1

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

MYSQL CODE

```
UPDATE diabetes_pred_data
SET smoking_history = 'Ex-smoker'
WHERE age > 50;
```

OUTPUT

Employee_Name	Patient_Ic	gender	age	hyper	heart_d	smoking_history	bmi	HbA1c_le	blood_glucose_level	diabet	Row_Number_Rankin	Dense_Rank
NATHANIEL FORD	PT101	Female	80	0	1	Ex-smoker	25.19	6.6	140	0	26604	7
Eve Bekker	PT100...	Female	30	0	0	never	22.88	4.5	80	0	1	1
Jacqueline C Ro...	PT99894	Female	72	1	0	Ex-smoker	50.85	3.5	80	0	2	1
Aisha M Malone	PT100...	Female	0	0	0	No Info	18.37	3.5	80	0	3	1
James E Nelson	PT100...	Female	31	0	0	current	20.23	6	80	0	4	1
Joshua R Mcdo...	PT100...	Female	47	1	0	current	25.48	4.5	80	0	5	1
Edward A Ang	PT100...	Female	21	0	0	No Info	23.04	6.5	80	0	6	1
Jensa Woo	PT100...	Female	54	0	0	Ex-smoker	27.32	6.5	80	0	7	1
Sharon S Young	PT100...	Female	57	0	0	Ex-smoker	33.64	6.2	80	0	8	1
Judi Soto	PT99994	Female	20	0	0	current	28.06	6	80	0	9	1
Benny M Choi	PT99774	Female	57	0	0	Ex-smoker	29.56	4	80	0	10	1
John E Long	PT99842	Female	19	0	0	never	27.32	6.1	80	0	11	1
Michael T Feist	PT99910	Female	34	0	0	No Info	27.32	4.5	80	0	12	1

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

MYSQL CODE

```
INSERT INTO diabetes_pred_data
VALUES("HEROSAKI","PT001","Male",25,0,0,"never",33.55,6.3,145,0);
```

OUTPUT

Employee_Name	Patient_I	gender	age	hype	heart_d	smoking_history	bmi	HbA1c_le	blood_glucose_level	diabet	Row_Number_Rankin	Dense_Rank
HEROSAKI	PT001	Male	25	0	0	never	33.55	6.3	145	0	20427	8
SHERYL BREGMAN	PT1000	Female	34	0	0	ever	26.15	6	126	0	16704	5
JOHN HOFFMAN	PT10000	Female	46	0	0	No Info	36.9	6.1	200	0	53981	13
Meredith H Red...	PT100...	Male	45	0	0	never	28.61	3.5	80	0	1784	1
Minouche Kandel	PT100...	Male	57	0	0	Ex-smoker	32.73	5.7	126	0	11660	5

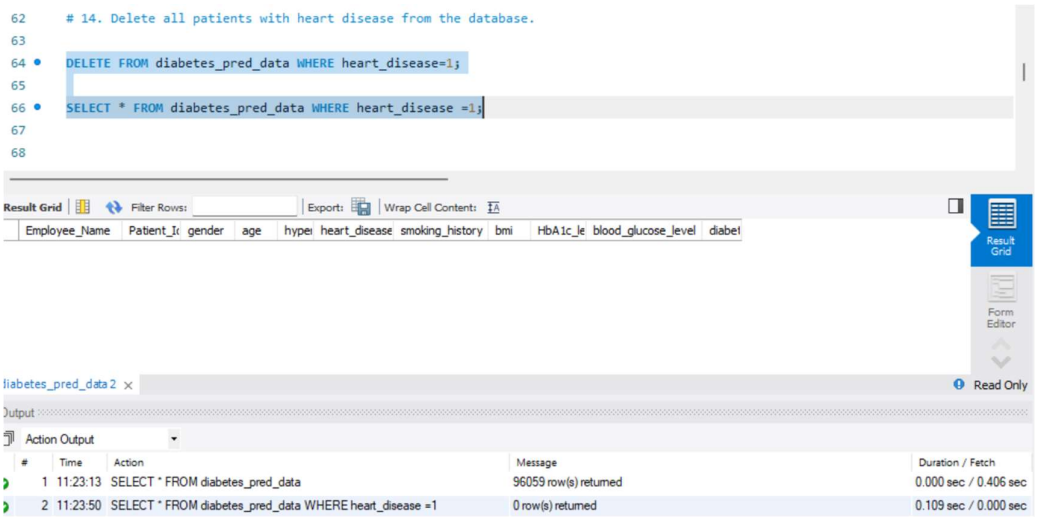
14.Delete all patients with heart disease from the database.

MYSQL CODE

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

OUTPUT

After Deleting the Record



15.Find patients who have hypertension but not diabetes

MYSQL CODE

```
SELECT * FROM diabetes_pred_data where hypertension =1 AND
diabetes =0;
```

OUTPUT

Employee_Name	Patient_Id	gender	age	hypertensior	heart_disease	smoking_history	bmi	HbA1c_le	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 SAK...	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 TACC...	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

abetes_pred_data31 ×

34 17:57:31 SELECT * FROM diabetes_pred_data where hypertension =1 AND diabet... 5397 row(s) returned 0.015 sec / 0.110 sec

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

MYSQL CODE

```
Alter table diabetes_pred_data
ADD CONSTRAINT uniq_patient_id UNIQUE(Patient_id);
```

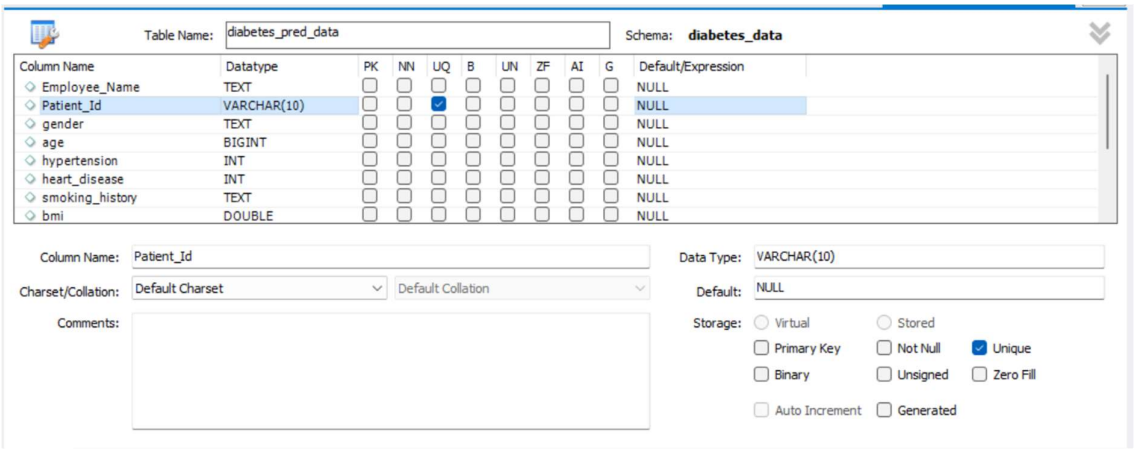
OUTPUT

```
76
77    # 16. Define a unique constraint on the "patient_id" column to ensure its values are unique.
78
79 •  Alter table diabetes_pred_data
80    ADD CONSTRAINT uniq_patient_id UNIQUE(Patient_id);
81
82 •  INSERT INTO diabetes_pred_data VALUES("HELLO","PT101","Female",25,1,0,"never",33.55,6.3,145,0);
83    # value check insertion
84 •  INSERT INTO diabetes_pred_data VALUES("HELLO","PT101","Female",25,1,0,"never",33.55,6.3,145,0);
85
86
```

Output

Action Output

#	Time	Action	Message	Duration / Fetch
6	11:35:06	NSERT INTO diabetes_pred_data VALUES("HELLO","PT101","Female",25,...	Error Code: 1064. You have an error in your SQL syntax; check the man...	0.000 sec
7	11:35:14	INSERT INTO diabetes_pred_data VALUES("HELLO","PT101","Female",25...	1 row(s) affected	0.078 sec
8	11:35:46	INSERT INTO diabetes_pred_data VALUES("HELLO","PT101","Female",25...	Error Code: 1062. Duplicate entry 'PT101' for key 'diabetes_pred_data.u...	0.079 sec

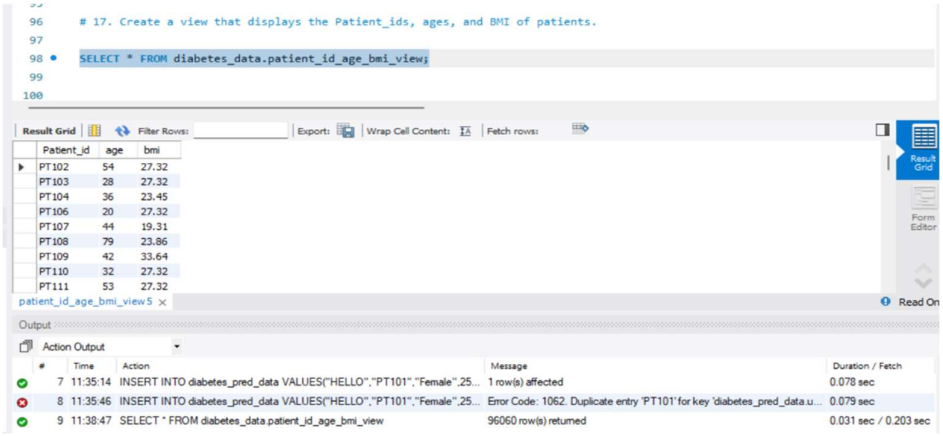


17. Create a view that displays the Patient_ids, ages, and BMI of patients

MYSQL CODE

```
SELECT * FROM diabetes_data.patient_id_age_bmi_view;
```

OUTPUT



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

1. **Normalization:** Break down large tables into smaller ones to eliminate duplicate data.This involves organizing tables to minimize redundancy and dependency.

2. **Primary Keys and Foreign Keys:** Ensure each table has a primary key (unique identifier) and use foreign keys to create relationships between tables. This maintains data integrity and prevents orphaned records.

3. **Constraint Enforcement:** Implement constraints such as NOT NULL, UNIQUE, and CHECK constraints to maintain data integrity.

4. **Regular Maintenance:** Perform regular data cleaning, remove duplicate records, and update outdated information.

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

Limit SELECT : Instead of retrieving all columns, specify only the required columns to reduce data transfer and processing time

Review Query Execution Plan : Analyze the query execution plan to identify bottlenecks and optimize accordingly.