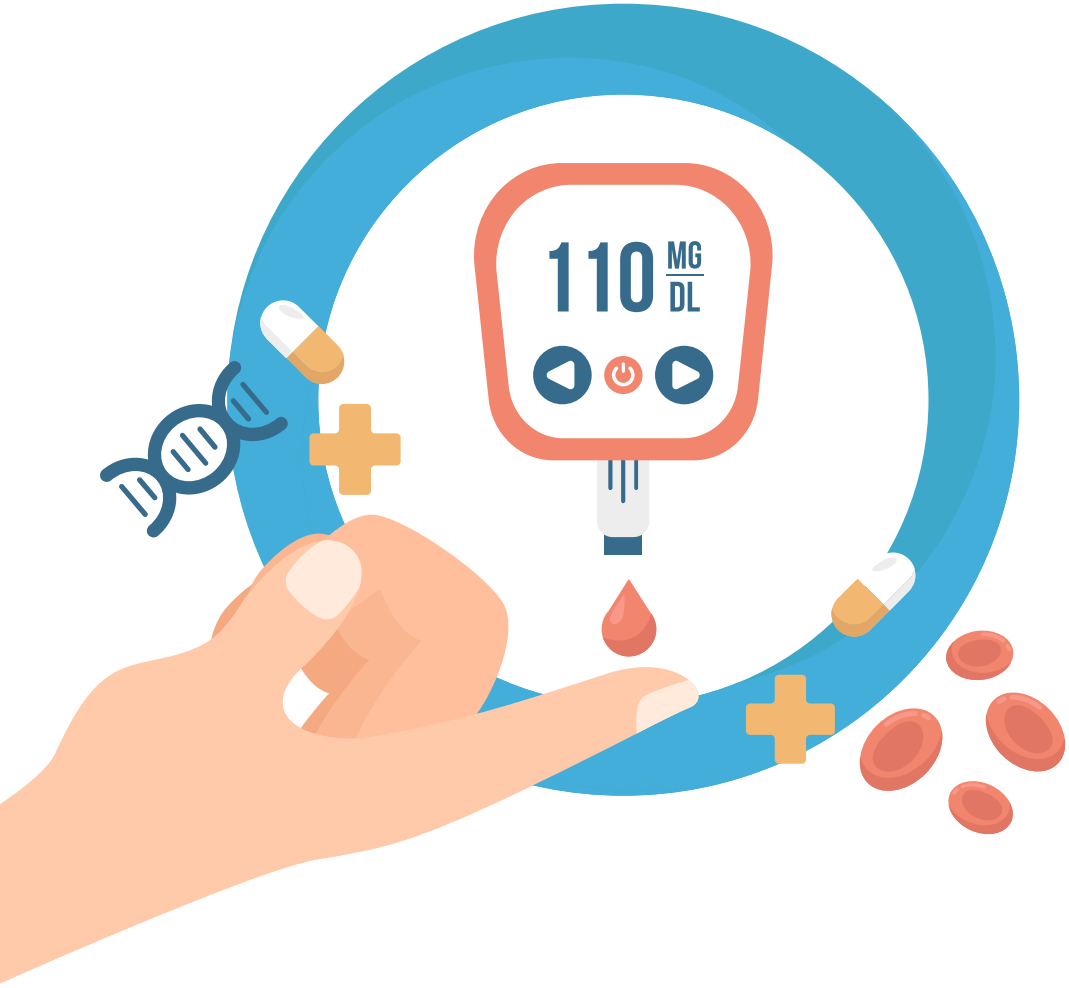


PSYLIQ

Diabetes Prediction

SHOUVIK SAHA
DATA ANALYST INTERN

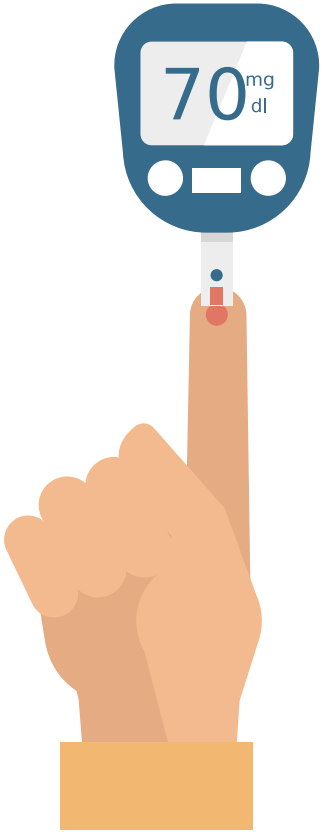
Acknowledgment

I want to express my sincere gratitude to Psyliq for providing me with the opportunity to intern as a data analyst. I am thankful for the support and guidance I received from the team at Psyliq throughout my internship. Additionally, I appreciate the collaborative environment that allowed me to learn and grow professionally. I am grateful for the contributions of my colleagues and mentors, which have been instrumental in my development as a data analyst. Lastly, I want to thank my family and friends for their unwavering support. This internship has been a valuable experience that has enhanced my skills and knowledge in data analytics. Thank you to Psyliq for this opportunity.

Objective

In this project, we'll utilize SQL querying and data analysis skills to examine a detailed dataset comprising demographic, clinical, and lifestyle data of individuals. The dataset will encompass variables like EmployeeName, Patient_id, gender, D.O.B, hypertension, heart_disease, smoking_history, BMI, HbA1c_level, blood_glucose_level, and diabetes. It will provide insights into age, gender, body mass index (BMI), blood pressure, family history of diabetes, dietary habits, physical activity levels, and laboratory test results.

1. Retrieve the Patient_id and ages of all patients



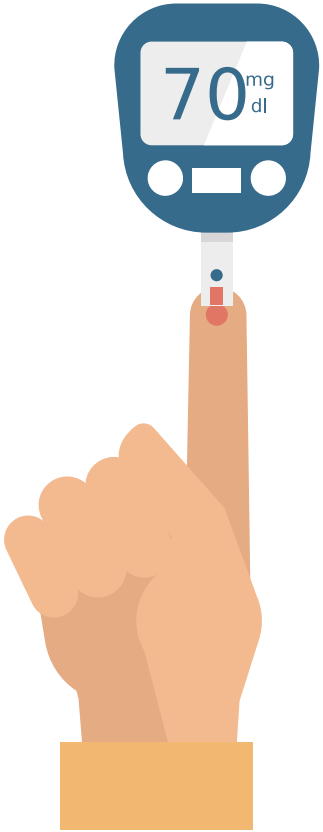
Limit to 1000 rows

```
1
2     # check all columns
3 •   select * from diabetes_prediction;
4
5     /*1. Retrieve the Patient_id and ages of all patients.*/
6 •   SELECT patient_id, age FROM diabetes_prediction;
7
8
9
10
```

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

patient_id	age
PT101	31
PT102	31
PT103	31
PT104	31
PT105	35
PT106	34
PT107	35
PT108	34
PT109	35
PT110	34
PT111	35
PT112	35
PT113	34
PT114	34
PT115	34
PT116	34

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



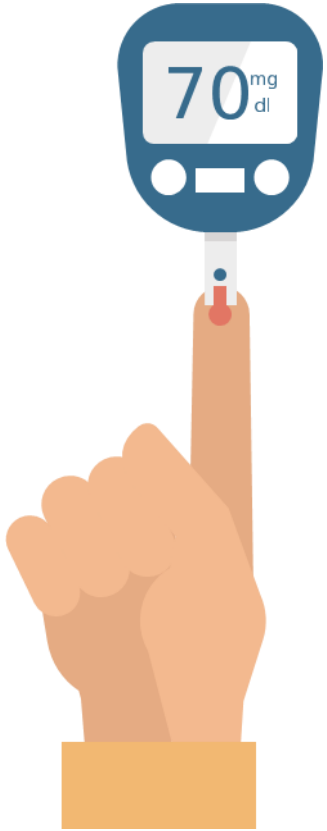

```
8   # 2. Select all female patients who are older than 40.
9   select * from diabetes_prediction where gender = 'Female' and age > 40;
10  # There are no female patients older than 40
11
12
13
```

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

EmployeeName	Patient_id	gender	D.O.B	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes

In the given dataset there are no female patient older than 40 years.

3. Calculate the average BMI of patients.




```
13    # 3. Calculate the average BMI of patients.
14 •  select avg(bmi) as average_bmi from diabetes_prediction;
15
16
17
18
19
20
21
22
```

Limit to 1000 rows

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

average_bmi
27.321033180382557

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

```
16    # 4. List patients in descending order of blood glucose levels.
17 •  select patient_id, blood_glucose_level
18    from diabetes_prediction
19    order by blood_glucose_level desc;
20
21
22
23
24
25
```

Limit to 1000 rows

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

patient_id	blood_glucose_level
PT97934	300
PT97570	300
PT97955	300
PT97622	300
PT98852	300
PT98855	300
PT97141	300
PT97671	300
PT95937	300
PT98911	300
PT99672	300
PT97708	300
PT99663	300
PT99008	300
PT96057	300
PT96062	300

5. Find patients who have hypertension and diabetes.

```
20
21      # 5. Find patients who have hypertension and diabetes.
22      select patient_id, hypertension, diabetes from diabetes_prediction
23      where hypertension = 1 and diabetes = 1;
24
```

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

patient_id	hypertension	diabetes
PT139	1	1
PT205	1	1
PT343	1	1
PT355	1	1
PT451	1	1
PT565	1	1
PT567	1	1
PT632	1	1
PT727	1	1
PT828	1	1
PT852	1	1
PT861	1	1
PT983	1	1
PT1075	1	1
PT1123	1	1
PT1183	1	1

6. Determine the number of patients with heart disease.

```
25
26      # 6. Determine the number of patients with heart disease.
27      select count(*) as heart_patient from diabetes_prediction
28      where heart_disease = 1;
29
30
```

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

heart_patient
3937

7. Group patients by smoking history and count how many smokers and non-smokers there are.

```
30
31    # 7. Group patients by smoking history and count how many smokers and non-smokers there are.
32
33    select smoking_history, count(smoking_history) as no_of_patient
34    from diabetes_prediction
35    where smoking_history in ('current', 'never')
36    group by smoking_history;
37
38
39
```

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

smoking_history	no_of_patient
never	35045
current	9265

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

```
40
41 •  select patient_id, bmi from diabetes_prediction where bmi >
42    (select avg(bmi) as avg_bmi from diabetes_prediction);
43
44
45
46
```

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

patient_id	bmi
PT109	33.64
PT112	54.7
PT113	36.05
PT117	30.36
PT121	36.38
PT124	27.94
PT126	33.76
PT128	27.85
PT131	31.75
PT140	56.43
PT143	32.02
PT144	29.3
PT149	28.27
PT153	28.12
PT156	37.16
PT160	63.48

Here we have shown the patient who have BMI geater than the average BMI

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

```
45    # 9. Find the patient with the highest HbA1c level and the patient with the lowest HbA1clevel.
46
47    # Highest HbA1c_level level:
48    select patient_id, HbA1c_level from diabetes_prediction
49    where HbA1c_level =
50    (select max(HbA1c_level) from diabetes_prediction);
51
52
53
```

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

patient_id	HbA1c_level
PT141	9
PT156	9
PT236	9
PT270	9
PT400	9
PT519	9
PT673	9
PT710	9
PT861	9
PT907	9
PT1242	9
PT1319	9
PT1332	9
PT1375	9
PT1470	9
PT1502	9

Here we have shown the patients with the highest HbA1c level

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

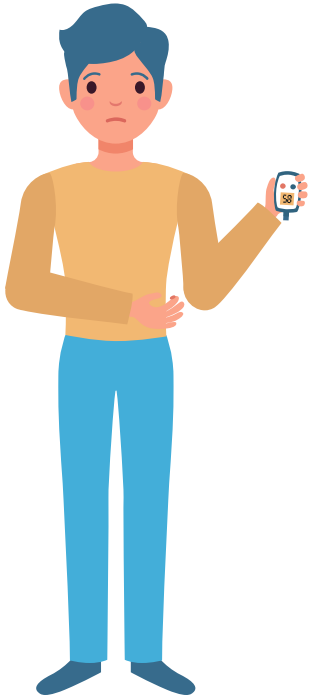
```
53
54      # Lowest HbA1c_level level:
55      select patient_id, HbA1c_level from diabetes_prediction
56      where HbA1c_level =
57      (select min(HbA1c_level) from diabetes_prediction);
58
59
60
```

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

patient_id	HbA1c_level
PT120	3.5
PT134	3.5
PT145	3.5
PT158	3.5
PT174	3.5
PT213	3.5
PT219	3.5
PT221	3.5
PT233	3.5
PT250	3.5
PT265	3.5
PT269	3.5
PT279	3.5
PT292	3.5
PT304	3.5
PT305	3.5

Here we have shown the patients with the lowest HbA1c level

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

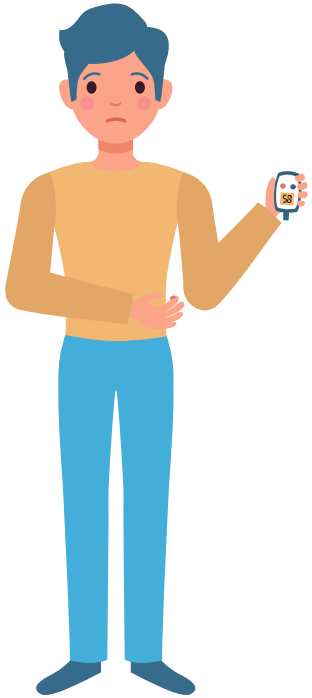


```
58
59      # 10. Calculate the age of patients in years (assuming the current date as of now).
60      select patient_id, age as age_of_patient
61      from diabetes_prediction;
62
63
```

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

patient_id	age_of_patient
PT101	31
PT102	31
PT103	31
PT104	31
PT105	35
PT106	34
PT107	35
PT108	34
PT109	35
PT110	34
PT111	35
PT112	35
PT113	34
PT114	34
PT115	34
PT116	34

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

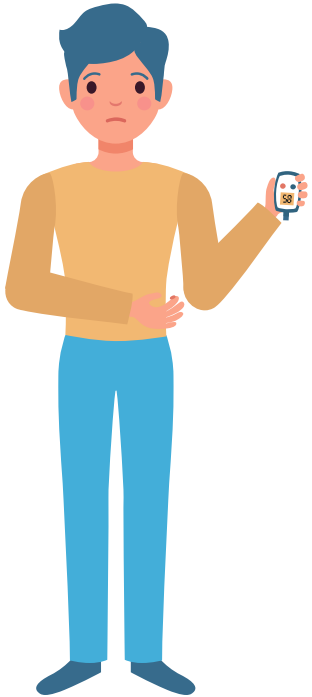


```
62
63      # 11. Rank patients by blood glucose level within each gender group.
64 •    select patient_id, blood_glucose_level, gender,
65      dense_rank() over(partition by gender order by blood_glucose_level) patient_rank
66      from diabetes_prediction;
67
68
69
```

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

patient_id	blood_glucose_level	gender	patient_rank
PT98584	80	Female	1
PT96552	80	Female	1
PT98149	80	Female	1
PT97519	80	Female	1
PT96556	80	Female	1
PT97682	80	Female	1
PT97685	80	Female	1
PT99316	80	Female	1
PT99179	80	Female	1
PT96389	80	Female	1
PT97756	80	Female	1
PT98295	80	Female	1
PT97868	80	Female	1
PT97210	80	Female	1
PT98457	80	Female	1
PT98289	80	Female	1

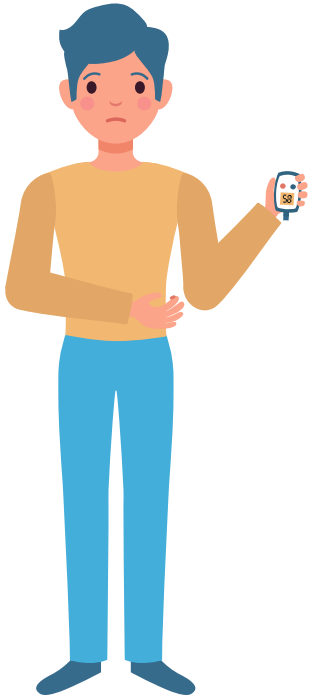
12. Update the smoking history of patients who are older than 50 to "Ex-smoker."



```
68    # 12. Update the smoking history of patients who are older than 50 to "Ex-smoker".
69 •  update diabetes_prediction set smoking_history = "Ex-smoker"
70    where age > 50;
71    # There are no patients in the database aged over 50.
72
```

There are no patient above 50 years in the dataset.

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



```
75    # 13. Insert a new patient into the database with sample data.
76
77    INSERT INTO diabetes_prediction VALUES ('Subham Saha', 'PT409991', 'Male', '11-05-1992', 34, 0, 1, 'never', 25.5, 7, 110, 1);
78
79    select * from diabetes_prediction where Patient_id = 'PT409991'
80
81
```

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

EmployeeName	Patient_id	gender	D.O.B	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
Subham Saha	PT409991	Male	11-05-1992	34	0	1	never	25.5	7	110	1

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

```
82
83    # 14. Delete all patients with heart disease from the database.
84    delete from diabetes_prediction where heart_disease = 1;
85
86    # check if records are deleted
87    select * from diabetes_prediction where heart_disease = 1;
```

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

EmployeeName	Patient_id	gender	D.O.B	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes

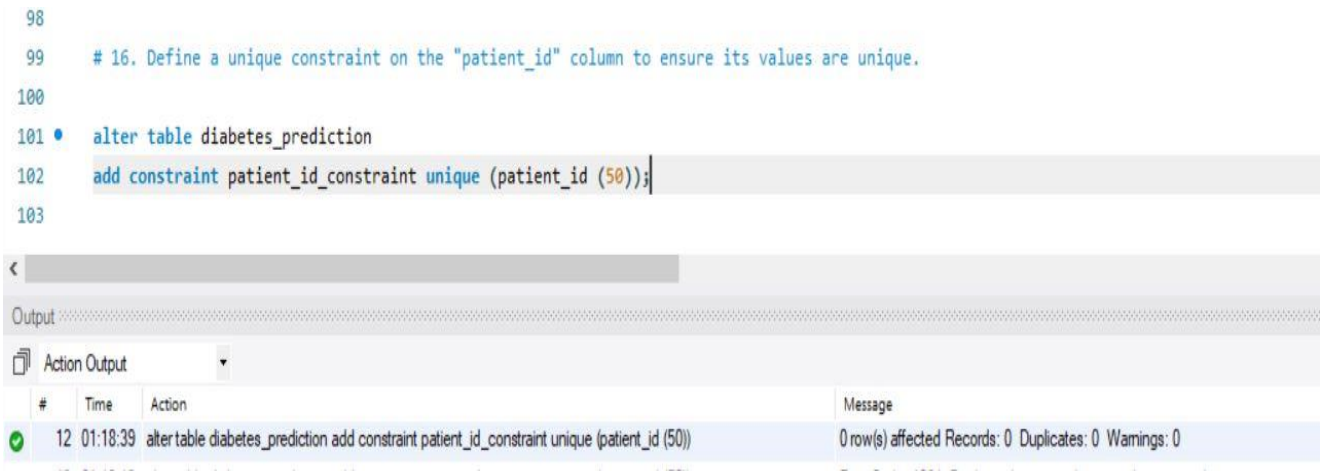
15. Find patients who have hypertension but not diabetes using the EXCEPT operator.

```
91    -- 15. Find patients who have hypertension but not diabetes using the EXCEPT operator.
92    select patient_id, hypertension, diabetes from diabetes_prediction
93    where hypertension = 1
94    except
95    select patient_id, hypertension, diabetes from diabetes_prediction
96    where diabetes = 1;
97
98
```

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

patient_id	hypertension	diabetes
PT129	1	0
PT155	1	0
PT161	1	0
PT215	1	0
PT227	1	0
PT241	1	0
PT326	1	0
PT339	1	0
PT357	1	0
PT377	1	0
PT379	1	0
PT446	1	0
PT474	1	0
PT475	1	0
PT476	1	0

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



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

```
105     # 17. Create a view that displays the Patient_ids, ages, and BMI of patients.
106
107 •   create view patient_bmi_details as
108     select patient_id, age, bmi
109     from diabetes_prediction;
110
111 •   select * from patient_bmi_details;
112
```

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	34	27.32
PT107	35	19.31
PT108	34	23.86
PT109	35	33.64
PT110	34	27.32
PT111	35	27.32
PT112	35	54.7
PT113	34	36.05
PT114	34	25.69
PT115	34	27.32
PT116	34	27.32
PT117	34	30.36

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

To reduce data redundancy and improve data integrity in the database schema, we can consider the following improvements:

- **Normalization:** Break down the dataset into separate tables to store related information, reducing redundancy.
- **Primary and Foreign Keys**: Utilize primary keys to uniquely identify records in each table and foreign keys to establish relationships between tables.
- **Data Types and Constraints:** Choose appropriate data types for columns and apply constraints to ensure data integrity.
- **Indexing:** Implement indexing on frequently queried columns to enhance query performance.

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

To optimize the performance of SQL queries we can consider the following improvements:

- **Indexing:** Create indexes on columns frequently used in WHERE, JOIN, and ORDER BY clauses to improve data retrieval and minimize full table scans.
- **Query Tuning:** Analyze query execution plans, optimize join techniques, minimize subqueries, and use specific filtering and aggregation to reduce the amount of processed data.
- **Normalization and Joins:** Ensure proper normalization to minimize redundancy, optimize table relationships, and use efficient join techniques to streamline data retrieval without unnecessary complexity.

Thank You

PSYLIQ