KING'S OWN INSTITUTE*

Success in Higher Education

INDIVIDUAL ASSIGNMENT COVERSHEET

Family Name: TASNIM Given Name: SABABA

Student Number: 20028086 Lecturer's/ Tutor's Name: Dr. Firoz Anwar

Subject Code & Name: ICT370 Data Analytics

Assignment Title: Week 4 Individual Progress Report

Declaration

(This declaration must be completed by the student or the assignment will not be marked.)

I certify the following:

- I have read and understood the *Student Academic Misconduct Policy*.
- This assignment is my own work based on my personal study and or research.
- I have acknowledged all material and sources used in the preparation of this assignment including any material generated in the course of my employment.
- The assignment has not previously been submitted for assessment.
- I have not copied in part or in whole or otherwise plagiarised the work of other students.
- I have read and I understand the criteria used for assessment.
- The assignment is within the word and page limits specified in the unit outline.
- The use of any material in this assignment does not infringe the intellectual property / copyright of a third party.
- I understand that this assignment may undergo electronic detection for plagiarism, and an anonymous copy of the assignment may be retained on the database and used to make comparisons with other assignments in future.
- By completing this coversheet in full and submitting this assignment electronically, I am bound by the conditions of the KOI's *Student Academic Misconduct Policy* and the declaration on this coversheet.

sababa
Signature

26 / 07 / 2025
Date

Assignment Receipt

Family Name: TASNIM Given Name: SABABA

Student Number: 20028086 Lecturer's/ Tutor's Name: Dr. Firoz Anwar

Subject Code & Name: ICT370 Data Analytics

Assignment Title: Week 4 Individual Progress Report

sababa
Signature

31 / 07 / 2025
Date

Name/Title of the data:

Name: Diabetes Prediction Dataset

Data URL: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

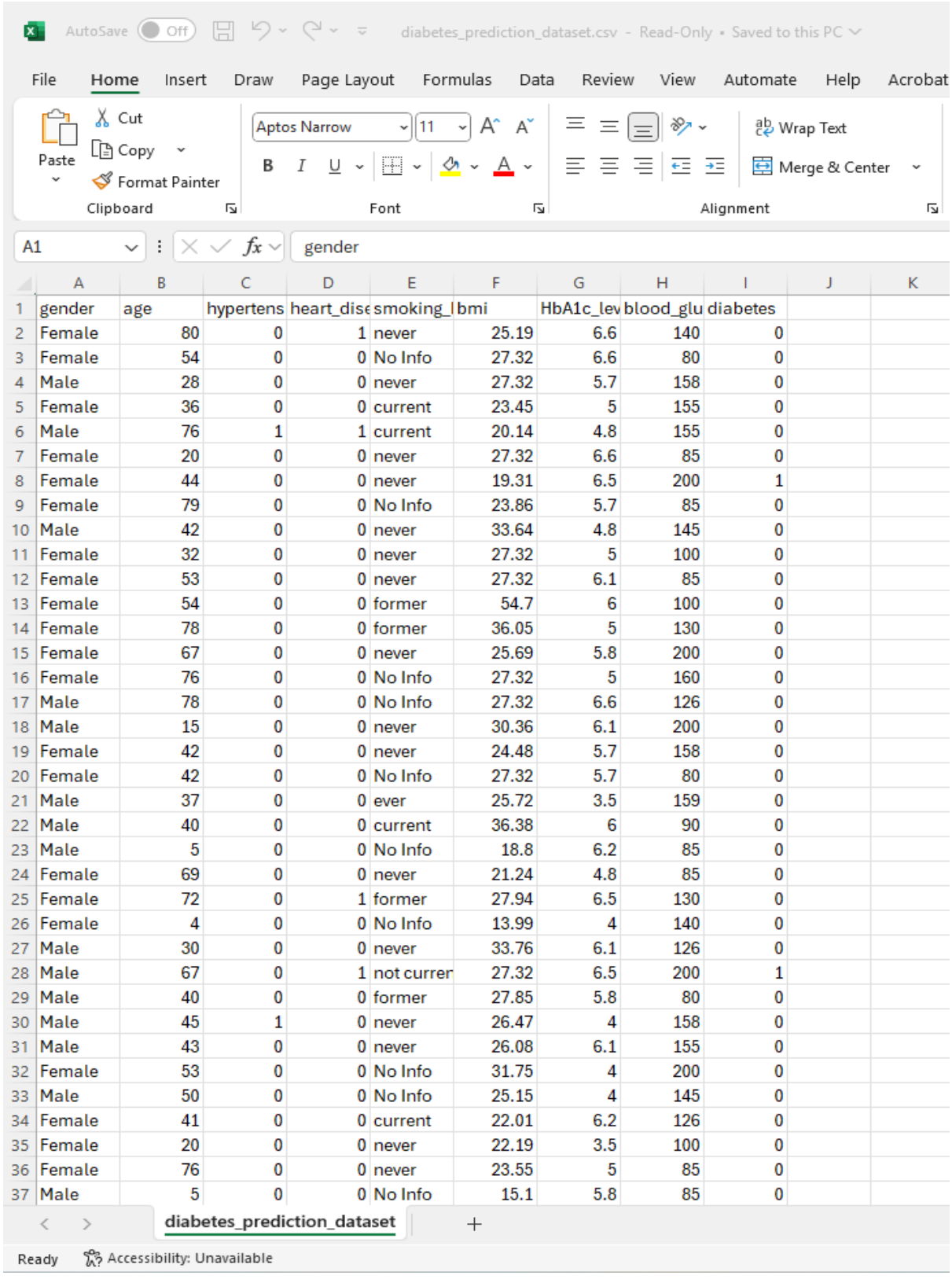
Description of dataset:

This dataset includes synthetic information of patients' records pertaining to Diabetes and includes patients' Age, BMI, their Diabetes risk assessment through HbA1c levels, and blood glucose levels. The dataset aims to help in training machine learning models or conducting an evaluation on healthcare analytics pertaining to Diabetes. (Mustafa, 2023).

Brief description of every column:

Outlined below is a brief description of each column in the dataset**.**

1. **Gender:** Patients categories are divided in 'Male' 'Female' and 'Other'.
2. **Age:** Age of the patient when diagnosed.
3. **Hypertension:** Does patient have hypertension, coded as 0 (no) and 1 (yes).
4. **Heart-disease:** Does patient suffer from a heart disease, 0 (no) and 1 (yes).
5. **Smoking-history:** Categorized into 'never', 'current', or 'former' etc.
6. **BMI:** Body mass index (float value).
7. **HbA1c_level:** The blood sugar levels for the past 2 to 3 months.
8. **Blood glucose level:** The blood glucose level measurement at the time of the assessment.
9. **Diabetes:** The target variable, indicates if the individual is a diabetic (0 = No, 1 = Yes).



	A	B	C	D	E	F	G	H	I	J	K
1	gender	age	hypertens	heart_dise	smoking_l	bmi	HbA1c_lev	blood_glu	diabetes		
2	Female	80	0	1	never	25.19	6.6	140	0		
3	Female	54	0	0	No Info	27.32	6.6	80	0		
4	Male	28	0	0	never	27.32	5.7	158	0		
5	Female	36	0	0	current	23.45	5	155	0		
6	Male	76	1	1	current	20.14	4.8	155	0		
7	Female	20	0	0	never	27.32	6.6	85	0		
8	Female	44	0	0	never	19.31	6.5	200	1		
9	Female	79	0	0	No Info	23.86	5.7	85	0		
10	Male	42	0	0	never	33.64	4.8	145	0		
11	Female	32	0	0	never	27.32	5	100	0		
12	Female	53	0	0	never	27.32	6.1	85	0		
13	Female	54	0	0	former	54.7	6	100	0		
14	Female	78	0	0	former	36.05	5	130	0		
15	Female	67	0	0	never	25.69	5.8	200	0		
16	Female	76	0	0	No Info	27.32	5	160	0		
17	Male	78	0	0	No Info	27.32	6.6	126	0		
18	Male	15	0	0	never	30.36	6.1	200	0		
19	Female	42	0	0	never	24.48	5.7	158	0		
20	Female	42	0	0	No Info	27.32	5.7	80	0		
21	Male	37	0	0	ever	25.72	3.5	159	0		
22	Male	40	0	0	current	36.38	6	90	0		
23	Male	5	0	0	No Info	18.8	6.2	85	0		
24	Female	69	0	0	never	21.24	4.8	85	0		
25	Female	72	0	1	former	27.94	6.5	130	0		
26	Female	4	0	0	No Info	13.99	4	140	0		
27	Male	30	0	0	never	33.76	6.1	126	0		
28	Male	67	0	1	not currer	27.32	6.5	200	1		
29	Male	40	0	0	former	27.85	5.8	80	0		
30	Male	45	1	0	never	26.47	4	158	0		
31	Male	43	0	0	never	26.08	6.1	155	0		
32	Female	53	0	0	No Info	31.75	4	200	0		
33	Male	50	0	0	No Info	25.15	4	145	0		
34	Female	41	0	0	current	22.01	6.2	126	0		
35	Female	20	0	0	never	22.19	3.5	100	0		
36	Female	76	0	0	never	23.55	5	85	0		
37	Male	5	0	0	No Info	15.1	5.8	85	0		

Figure: Screenshot of dataset

Business Questions:

These questions seek to gain valuable answers from the data:

- How are diabetes cases distributed in the population concerning different age cohorts?
- Are there any significant differences in BMI of diabetic patients when compared to non-diabetic patients?
- To what extent does a smoker's history impact the likelihood of a diabetes diagnosis?
- Are patients suffering from hypertension more prone to diabetes?
- Among individuals with elevated HbA1c levels, what proportion are likely to be diabetic?

Missing Values:

The dataset has some missing values in the smoking_history and bmi columns.

- **smoking_history:** Certain patients lack corresponding smoking data.
- **bmi:** some records contain null, or zero values for BMI.

Completeness of this dataset is not an issue, but the researchers identified a lack of data regarding a family history of diabetes. Moreover, the data gap does not accommodate questions regarding age, sex, family history or alcohol consumption, which are additional risk factors. To improve the analysis, the following data would be helpful:

• Patient's dietary habits
• Frequency of physical exercise
• Genetic history or family history of diabetes
• Duration of chronic diabetes risk factors such as hypertension.

Categorical Data and Grouping:

The dataset has the following grouped or categorical columns:

•**gender** → Classified as Male/Female/Other

•**smoking_history** → Several options such as 'never', 'current', 'former'

•**diabetes, hypertension, and heart_disease** → Binary categories as well (0/1)

Consolidation, cleaning or transformation:

The following actions are necessary:

- Fill or delete null entries for bmi and smoking_history fields
- Convert categorical variables to numerical values for easier analysis (e.g., employing One-Hot Encoding)
- Normalize continuous variables where necessary, such as bmi and blood_glucose_level.

Bibliography:

Mustafa Tariq. (2023). *Diabetes Prediction Dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset [Accessed 26 June 2025]