Exercise 4- Report

GROUP 2:

Amber Sethi (101328584)
Joshua Dmello (101346654)
Mohamed Bennis (101276333)
Muhammad Attique (101292217)
Tanvi Sharma (101265693)

Identifying Source, Derived and Target Variables along with Limitations and Advantages of the Data Set

Table of Contents

Source variables:

1. **Diabetes Pedigree function**

It provides some data on diabetes mellitus history in relatives and the genetic relationship of those relatives to the patient. This measure of genetic influence gave us an idea of the hereditary risk one might have with the onset of diabetes mellitus.

2. **BMI**

At birth, the lifetime risk of developing diabetes is one in three, but lifetime risks across BMI categories are still not fully ascertained. Being overweight and especially obesity, particularly at younger ages, substantially increases the lifetime risk of diagnosed diabetes. Lifetime risk estimates for diabetes according to BMI would be valuable for:

communicating an individual's risk of diabetes given his/her BMI and

Identifying groups of individuals who would benefit most from primary prevention.

3. **Age**

With an increase in age, the chances of developing diabetes are higher due to factors related to hereditary traits (diabetes pedigree) and lifestyle traits like overweight and obesity at younger ages. During periods of life such as teenage and middle-age years, factors such as pregnancy and an increase in BMI (age-related) as well as a shift to senior age category symptoms (decrease in metabolic rate, age decline, etc.) can be attributed to a source variable as Age.

4. **Serum Insulin**

Insulin is an anabolic hormone that promotes glucose uptake, glycogenesis, lipogenesis, and protein synthesis of skeletal muscle and fat tissue through the tyrosine kinase receptor pathway. Insulin is a hormone that helps move blood sugar, known as glucose. Insulin plays a key role in keeping glucose at the right levels. Insulin resistance may eventually lead to the development of type 2 diabetes. This happens when your pancreas is no longer able to compensate by secreting the large amounts of insulin required to keep the blood sugar normal.

5. **Glucose (Blood Glucose)**

Blood sugar, or glucose, is the main sugar found in your blood. It comes from the food you eat and is your body's main source of energy. Your blood carries glucose to all of your body's cells to use for energy. Insulin is responsible for allowing glucose into your body's cells. When glucose enters your cells, the amount of glucose in your bloodstream falls. Diabetes is characterized by a high blood sugar level over a prolonged period.

Derived variables:

Here are examples of 15 derived variables that might be useful in our analysis and why they might be useful:

1. Age category
2. BMI category
3. Diastolic Blood pressure categories
4. Serum insulin levels
5. Blood Glucose categories
6. Blood sugar levels (with diabetes)
7. Pregnancy/ conception frequency range
8. Diabetes pedigree function range
9. Plasma glucose levels
10. Triceps skinfold thickness range
11. Glucose level by BMI
12. Glucose level by class variable
13. BMI category by class variable
14. Average glucose levels by BMI
15. Blood pressure by class variable

These derived variables above are going be so useful for our analysis for many reasons such as the following:

Categories for the different variables (age, BMI, blood pressure, etc..) that are labelled with plain language make the data easier to read or digest.

Categories make the analysis simpler and can convey clearer messages or interpretations.

- Measuring different variables such as glucose levels, BMI or blood pressure against other variables such as class variables or BMI help determine if they have relationships amongst them.

Describe the logic we will use in creating these derived variables.

Derived Variables:	The logic behind each variable:
Age Category	• 18 to 35: young adult • 36 to 55: middle age • 56 to 99: older adult
BMI category	• Below 18.5: Underweight • 18.5-24.9: Normal or Healthy Weight • 25.0-29.9: Overweight • 30.0 and Above: Obese
Diastolic Blood pressure categories	• Less than 80: Normal • 80-89: High Blood Pressure Stage 1(Hypertension Stage 1) • 90 or Higher: High Blood Pressure Stage (Hypertension Stage 2) • Higher than 120: Hypertensive Crisis
Serum insulin levels (Within Normal or Medically Supervised range in the case of testing)	• < 25 mIU/L: Fasting • 30-230 mIU/L: 30 minutes after glucose administration • 18-276 mIU/L: 1 hour after glucose administration • 16-166 mIU/L: 2 hours after glucose administration
Blood Glucose categories	• < 53 mg/dL: Severe Hypoglycemia • < 70 mg/dL: Hypoglycemia • < 125 mg/dL: Normal • < 200 mg/dL: High

	• < 200 – 500+ mg/dL: Metabolic
Blood sugar levels (With diabetes)	• Fasting: 80-130 mg/dL • 1-2 hours after meals: 180 mg/dL • A1C test: <7%
Pregnancy/ Conception frequency range	• 0: None • 0-3: low frequency • 4-6: medium frequency • 7+: high frequency
Diabetes pedigree function range	• More than 0.5: from a parent, full sibling • 0.25-0.5: from half-sibling, grandparent, aunt, or uncle • Less than 0.25: from a half aunt, half-uncle, or first cousin
Plasma glucose levels	• Below 11.1 mmol/l Below 200 mg/dl: Random • Below 5.5 mmol/l Below 100 mg/dl: Fasting • Below 7.8 mmol/l Below 140 mg/dl: 2 hour post-prandial
Triceps skinfold thickness range	• 6-12mm: lean individuals and between. • 40-50mm: obese individuals.

Target variables

Outcome

This is a class Variable where 0 is assigned if non-diabetic and 1 if diabetic. This is the target variable used as a comparison standard for the source and derived variables, i.e. through analysis, patterns might be formed on the population sample assigned 0(False) vs the population assigned 1(True).

This might also be useful in applying the same trend to predict based on the variables, a shift from non-diabetic (0) to diabetic (1).

Limitations

- Diabetes mellitus is classified into six categories: type 1 diabetes, type 2 diabetes, hybrid forms of diabetes, hyperglycemia first detected during pregnancy, "unclassified diabetes", and "other specific types". The "hybrid forms of diabetes" contains slowly evolving, immune-mediated diabetes of adults and ketosis-prone type 2 diabetes. The "hyperglycemia first detected during pregnancy" contains gestational diabetes mellitus and diabetes mellitus in pregnancy (type 1 or type 2 diabetes first diagnosed during pregnancy). The "other specific types" are a collection of a few dozen individual causes. Diabetes is a more variable disease than once thought and people may have combinations of forms.
  This can skew the analysis of the data environment due to the outcome not having any descriptors to the type of diabetes.
- The variables measure is at a single instance of time rather than multiple instances like a time series which would more likely be a more accurate dataset on which to base a model.

Advantages

- Since this is a closed population group (Females- 21 to 81), the variable value can be closely compared to medical standardized levels. (Male physiology vs female physiology) This makes it easier to identify outliers and replace null values to mean values.
- Since the Target variable (Outcome) is binary, we can see more patterns between a single variable (BMI, Diabetes Pedigree function, Age, Pregnancies, Insulin, Glucose, Blood Pressure) and the predicted effect on the target variable or collective effect, i.e., BMI and age is a predictor of diabetes likelihood when compared objectively and independently.

Random Sample:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
103	1	81	72	18	40	26.6	0.283	24	0
436	12	140	85	33	0	37.4	0.244	41	0
760	2	88	58	26	16	28.4	0.766	22	0
238	9	164	84	21	0	30.8	0.831	32	1
365	5	99	54	28	83	34.0	0.499	30	0
...	...	...	...	...	...	...	...	...	...
27	1	97	66	15	140	23.2	0.487	22	0
51	1	101	50	15	36	24.2	0.526	26	0
30	5	109	75	26	0	36.0	0.546	60	0
346	1	139	46	19	83	28.7	0.654	22	0
249	1	111	86	19	0	30.1	0.143	23	0

100 rows × 9 columns