AIMS Lab Research Engineer Selection Test

Name : Md Al Amin Tokder

Designation : Machine Learning Engineer , Devolved AI

Institution : Rajshahi University Of Engineering and Technology (RUET)

Contact Info : alamintokdercse@gmail.com

Mobile number : 01750206042

Problem-01 Solution:

Motivation:

- Increasing prevalence of overweight and obesity among children globally and its potential to lead to adult obesity and related cardiovascular risks.
- Specifically, the study aimed to explore how childhood anthropometric measurements (like BMI and skinfold thickness) could predict cardiovascular risk factors in adulthood.

Problems Addressed:

- The paper focuses on determining the long-term impact of childhood body measurements on the likelihood of developing cardiovascular diseases and other associated risk factors in adulthood.
- It addresses whether early life BMI and skinfold thickness are reliable predictors for adult diseases like metabolic syndrome, hyperglycemia, type 2 diabetes, and other related conditions.

Challenges:

- One of the main challenges is linking data across a long timeframe (35 years in this study), which involves keeping track of participants over decades.
- The study must account for various confounding factors that could influence the outcomes, such as sex, physical activity, alcohol consumption, smoking, and family history of obesity.

Existing Ways to Solve:

- The use of longitudinal cohort studies to track health metrics from childhood into adulthood. This method is already established for studying long-term health outcomes.
- Applying logistic regression models adjusted for multiple variables to assess the strength and nature of the associations between childhood measurements and adult health outcomes.

Future Scope:

- Further research might explore interventions that could be implemented during childhood to mitigate the identified risks.
- Studies could also examine the biological mechanisms underlying the observed associations, potentially leading to more targeted prevention strategies.
- Expanding the research to diverse populations or different age cohorts could provide broader insights into the generalizability of the findings

Problem-02 Solution:

Dataset description:

The dataset has 29,999 entries and 34 features. The columns include various demographic, health and socioeconomic attributes.

- **Demographic Information**: This includes household_id, user_id, profile_name, father_name, mother_name, birthday, age, and gender.
- **Socioeconomic Status**: Attributes like total_income which indicates the income category.
- **Health Information**: Includes whether the individual is classified as poor (is_poor), whether they are a freedom fighter (is_freedom_fighter), whether they had a stroke (had_stroke), whether they have cardiovascular diseases (has_cardiovascular_disease) and diabetes status (diabetic).
- **Health Measurements**: Such as SYSTOLIC, DIASTOLIC, PULSE_RATE, HEIGHT, WEIGHT, BMI, SUGAR and SPO2.

Algorithm :

- **Calculate the Chi-square Statistic**:

  Observed Frequencies (O): These are the values in the contingency table.
  Expected Frequencies (E): These are calculated under the assumption that there is no association between the variables. For each cell in the table, the expected frequency is calculated as:

  $$E_{i,j} = \frac{(Row\ Total\ i) \times (Column\ Total\ j)}{Total\ N}$$

  The chi-square statistic is calculated as:

  $$\chi^2 = \sum \left(\frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \right)$$

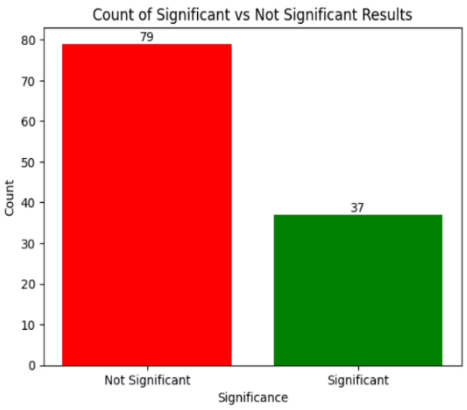
- **Degrees of Freedom (df):**
  The degrees of freedom for the chi-square test are calculated as:
  df=(number of rows−1)×(number of columns−1)
  This reflects the number of independent ways the cell counts can vary.
  Degrees of freedom vary across the tests which reflects the number of levels each variable has.

- **P-value:** The p-value is calculated from the chi-square distribution using the chi-square value and the degrees of freedom. It tells the probability of observing a chi-square statistic at least as extreme as the one computed if the null hypothesis (no association between the variables) were true.
  The very low p-value indicates a strong, statistically significant association between two features whereas extremely low p-value indicates a very strong and statistically significant association between features.
  Typically, if the p-value is less than a threshold (commonly 0.05), we reject the null hypothesis that indicate a statistically significant association between the variables.

Result Analysis:

In this dataset we have found total 34 features . There are 561 unique pairs of features that can be created from the 34 features in the dataset.among these pairs,all pairs are not valid for Chi-square test as there were null valued columns,some feature had categorical value.

	Feature 1	Feature 2	Chi-square Statistic	P-value	Degrees of Freedom	Interpretation
0	total_income	gender	1.189601	7.554994e-01	3	Not Significant
1	total_income	is_poor	0.000000	1.000000e+00	0	Not Significant
2	total_income	is_freedom_fighter	14.214156	2.627679e-03	3	Significant
3	total_income	had_stroke	0.900317	8.253513e-01	3	Not Significant
4	total_income	has_cardiovascular_disease	4.929155	1.770589e-01	3	Not Significant
5	total_income	disabilities_name	8.481028	9.030546e-01	15	Not Significant
6	total_income	diabetic	38.433059	2.288488e-08	3	Significant
7	total_income	profile_hypertensive	9.958818	1.891946e-02	3	Significant
8	total_income	RESULT_STAT_BP	159.239608	1.173648e-24	18	Significant
9	total_income	RESULT_STAT_BMI	45.668069	6.003162e-05	15	Significant
10	total_income	TAG_NAME	14.035484	2.924229e-02	6	Significant



After applying Chi-square test ,we found there were 37 feature that was significant and about 79 features were not significant.

Here is my analysis code link for chi-square test:

https://github.com/Shoukhin1803078/UIU-test-myself/blob/main/UIU_test.ipynb