Analysis of the Dataset

Statistically Significant Features

The given dataset has 33 features. Among them there are two type of variables: categorical and continuous. For the categorical variables chi-square has been used to find out significant ones.

First of all, household_id, user_id, profile_name, father_name & mother_name are not important because almost all of them are unique. However, categorical variables like union_name, gender, had_stroke and diabetic are mostly statistically significant because important health conditions like cardiovascular disease, hypertension, stroke are dependent on them. Also total_income, is_freedom_fighter feature are significant but less than the formers.

Algorithm for chi-square:

1. Data_contingency_table ← make contingency table from two features
2. (dof,expected) ← calculate degree of freedom and expected values
3. (prob, critical, stat) ← calculate probability value, critical value and stat value
4. based on prob, critical, stat values accept or reject H0

Here, H0 means two features are independent.

Mathematical Explanation:

Let's say for the features "had_stroke" and "has_cardiovascular_disease" we will calculate chi-square. The contingency table will be like:

has_cardiovascular_disease → had_stroke	0	1
0	29948 (29942.02)	28 (33.97)
1	17 (22.97)	6 (0.0261)

Table 1: Contingency Table for chi-square calculation

The values in brackets are expected frequency which are calculated using the formula (row total * column total) / total. Now chi-square table is formed using (observed frequency-expected frequency)/expected frequency and the summation is taken for finding out critical value along with degree of freedom. Here the value is 3.841.

if (critical value >= calculated value)   H0 is accepted

else   H0 is rejected

The continuous variables were tested with Kolmogorov-Smirnov test and QQ plot. None of them are from gaussian distribution. So, Mann Whitney test was used to find out which features were from same distribution to test the statistical significance. From Mann Whitney test it was found that only the HEIGHT, WEIGHT, BMI features were from same distribution. So, continuous features except these three are statistically significant.

Algorithm for Mann Whitney:

1. Set rank value for all the data of two samples
2. Calculate rank sum for both samples
3. Find U*stat* individually using the formula RankSUM – n(n+1)/2
4. Take the lowest U*stat* as final and find U*critical* for the two samples
5. if (U*stat*<U*critical* at alpha=0.05)   reject H0
   else   accept H0

Here, H0 means two datasets are from same distribution.

Possible Analytical Outcomes

The Pearson's correlation-coefficients among the continuous variables are as below:

	Age	**SYSTOLIC**	**DIASTOLIC**	**HEIGHT**	**WEIGHT**	**BMI**	**SUGAR**	**PULSE_RATE**
Age		.363	.274	-0.062	-0.062	-0.062	0.106	0.017
SYSTOLIC	.363		.754	-0.125	-0.124	-0.125	-0.039	0.148
DIASTOLILC	.274	.754		-0.071	-0.071	-0.071	-0.046	0.321
HEIGHT	-0.062	-0.125	-0.071		1	0.999	-0.0024	-0.004
WEIGHT	-.062	-0.124	-0.071	1		1	-0.024	-0.004
BMI	-0.062	-0.125	-0.071	0.999	1		-0.024	-0.004
SUGAR	0.106	-0.039	-0.046	-0.0024	-0.024	-0.024		-0.093
PULSE_RATE	0.017	0.148	0.321	-0.004	-0.004	-0.004	-0.093	

Table 2: Pearson's correlation-coefficients result

From the table it is clear that Age & SYSTOLIC; DIASTOLIC & SYSTOLIC; HEIGHT & WEIGHT & BMI are strongly correlated.

Applying logistic regression, prediction on "has_cardiovascular_disease" feature based on age, had_stroke and diabetic is possible. The confusion matrix is:

7489 (TP)	2 (FN)
7 (FP)	2 (TN)

Table 3: Confusion Matrix

From the table precision is .999066, recall is .9997330 and accuracy is 0.998.