**Final Assignment**

Monish Kamal Bathini

2250471

**Database**: This dataset is extracted from Kaggle. The dataset is diabetes dataset. It contains the features such as Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, Outcome.

DIABETES.CSV

**Factor Analysis: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, Outcome**

Maximum Likelihood Factor Analysis of the Correlation Matrix

**Unrotated Factor Loadings and Communalities**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Factor1** | **Factor2** | **Factor3** | **Communality** |
| Pregnancies | 0.003 | -0.386 | -0.561 | 0.464 |
| Glucose | 0.454 | -0.732 | 0.269 | 0.814 |
| BloodPressure | 0.280 | -0.115 | -0.241 | 0.150 |
| SkinThickness | 0.826 | 0.393 | -0.113 | 0.849 |
| Insulin | 0.558 | -0.033 | 0.163 | 0.339 |
| BMI | 0.488 | -0.032 | -0.045 | 0.241 |
| DiabetesPedigreeFunction | 0.253 | -0.042 | 0.031 | 0.067 |
| Age | 0.042 | -0.547 | -0.589 | 0.648 |
| Outcome | 0.309 | -0.449 | -0.007 | 0.297 |
|  |  |  |  |  |
| Variance | 1.6769 | 1.3570 | 0.8347 | 3.8686 |
| % Var | 0.186 | 0.151 | 0.093 | 0.430 |

**Rotated Factor Loadings and Communalities**

Varimax Rotation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Factor1** | **Factor2** | **Factor3** | **Communality** |
| Pregnancies | -0.049 | -0.677 | -0.059 | 0.464 |
| Glucose | 0.173 | -0.132 | -0.875 | 0.814 |
| BloodPressure | 0.258 | -0.279 | -0.075 | 0.150 |
| SkinThickness | 0.914 | 0.052 | 0.108 | 0.849 |
| Insulin | 0.497 | 0.103 | -0.284 | 0.339 |
| BMI | 0.456 | -0.076 | -0.164 | 0.241 |
| DiabetesPedigreeFunction | 0.223 | -0.004 | -0.131 | 0.067 |
| Age | -0.058 | -0.780 | -0.192 | 0.648 |
| Outcome | 0.156 | -0.233 | -0.468 | 0.297 |
|  |  |  |  |  |
| Variance | 1.4666 | 1.2347 | 1.1673 | 3.8686 |
| % Var | 0.163 | 0.137 | 0.130 | 0.430 |

**Factor Score Coefficients**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Factor1** | **Factor2** | **Factor3** |
| Pregnancies | 0.005 | -0.346 | 0.060 |
| Glucose | 0.055 | 0.092 | -0.826 |
| BloodPressure | 0.045 | -0.091 | 0.010 |
| SkinThickness | 0.823 | -0.043 | 0.218 |
| Insulin | 0.092 | 0.060 | -0.078 |
| BMI | 0.079 | -0.023 | -0.024 |
| DiabetesPedigreeFunction | 0.030 | 0.004 | -0.022 |
| Age | 0.002 | -0.583 | 0.038 |
| Outcome | 0.023 | -0.053 | -0.100 |

From the above three tables, we can only come to a conclusion that skin thickness is highly expressed in factor 1 where as the remaining factors doesn’t mean anything. This is because of the data selected.

2)**Database:** This database is extracted from Kaggle. The database is Social Network Ads. I considered the independent variable as age and the dependent variable as Purchased.

SOCIAL\_NETWORK\_ADS.CSV

**Binary Logistic Regression: Purchased versus Age**

**Method**

|  |  |
| --- | --- |
| Link function | Logit |
| Rows used | 400 |

**Response Information**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Value** | **Count** |  |
| Purchased | 1 | 143 | (Event) |
|  | 0 | 257 |  |
|  | Total | 400 |  |

**Regression Equation**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| P(1) | | = | | exp(Y')/(1 + exp(Y')) | |
| Y' | = | | -8.044 + 0.1889 Age | |

**Coefficients**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Term** | **Coef** | **SE Coef** | **Z-Value** | **P-Value** | **VIF** |
| Constant | -8.044 | 0.784 | -10.26 | 0.000 |  |
| Age | 0.1889 | 0.0192 | 9.87 | 0.000 | 1.00 |

**Odds Ratios for Continuous Predictors**

|  |  |  |
| --- | --- | --- |
|  | **Odds Ratio** | **95% CI** |
| Age | 1.2080 | (1.1635, 1.2542) |

**Model Summary**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Deviance R-Sq** | **Deviance R-Sq(adj)** | **AIC** | **AICc** | **BIC** | **Area Under ROC Curve** |
| 35.53% | 35.34% | 340.26 | 340.29 | 348.24 | 0.8686 |

**Analysis of Variance**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Wald Test** | |
| **Source** | **DF** | **Chi-Square** | **P-Value** |
| Regression | 1 | 97.34 | 0.000 |
| Age | 1 | 97.34 | 0.000 |

Verifying:

1. Suppose (age) = 26 then

YI = -8.044 + 0.1889\*Age = -8.044 + 0.1889\*26

Antilog(-3.1326) = 0.04178238774 ~ 0

P = s/(1+s) = 0.04 ~ 0

1. Suppose age = 45

Y'=-8.044+0.1889\*45 = 0.4565

Antilog(0.4565) = 1.5785394 ~ 1

1. Suppose age = 54

Y'=-8.044+0.1889\*54 = 2.1566

Antilog(2.1566) = 0.7685329

P = 0.43455957 ~ 1

So from the above three examples, we can conclude that the values are same.