Stroke Prediction Analysis

In this project, my objective is to utilize linear regression and multiple linear regression techniques to predict the probability of a stroke. The dataset I'm working with consists of 12 columns and 4935 rows. I will be examining the relationship between various independent variables such as age, hypertension, heart disease, marital status, work type, residence type, average glucose levels, BMI levels, and smoking status with the dependent variable, which is stroke.

In the initial phase of the analysis, I will extract valuable insights from the dataset. This includes determining the number of individuals who have experienced a stroke and categorizing them based on factors like work type and smoking status.

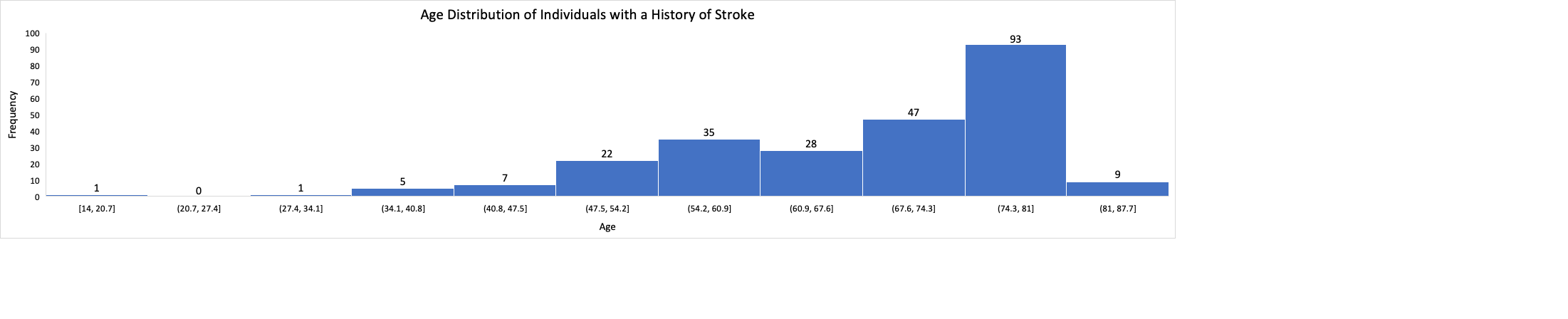
Moving on to the next phase, I will construct simple linear regression models by considering individual independent variables. The goal is to assess whether these models effectively capture a significant portion of the variation in the dependent variable.

Lastly, in the third phase, I will develop multiple regression models to explore the potential for stronger predictive abilities. By incorporating multiple independent variables, I aim to identify any variables that exhibit a substantial relationship with the dependent variable, allowing us to predict the probability of a stroke more accurately.

The columns:

1. ID: A unique number assigned to each individual.
2. Gender: Indicates the gender of the individuals.
3. Age: Represents the age of each individual.
4. Hypertension: Indicates whether the person has a history of hypertension. It takes a value of 1 for yes and 0 for no.
5. Ever\_married: Indicates whether the individual has been married. It can be either "Yes" or "No".
6. Work\_type: Categorizes the individual's employment status into four categories: Private (working for a privately owned company), Self-employed, Government employee, has children,or Never worked.
7. Smoking\_status: Classifies the individual's smoking status into four categories: Formerly smoked, Never smoked, Smokes, and Unknown.
8. BMI: Represents the Body Mass Index of each individual.
9. Residence\_type: Categorizes the individual's residence into two categories: Urban or Rural.
10. Avg\_glucose\_level: Indicates the average glucose level of the individual.
11. Stroke: Indicates whether the individual has a history of stroke. It takes a value of 1 for yes and 0 for no.

Part I



The histogram provides an overview of the age distribution among individuals who have a history of stroke. The distribution is skewed to the left, suggesting that a portion of the individuals who experienced a stroke did so at a relatively young age. The majority of individuals with a stroke history fall within the age range of 74.3 to 81 years. On average, the age of these individuals is approximately 68 years, while the median age is 71 years. Consequently, it can be inferred that the majority of individuals in this category are older adults.

The bar graph illustrates the comparison of average age between individuals with a history of stroke and those without. The average age of individuals who have experienced a stroke is 68 years, while the average age of individuals without a stroke history is 43.48 years. This data suggests that older individuals tend to have a higher likelihood of experiencing a stroke compared to younger individuals.

The bar graph compares the number of individuals with and without a history of stroke in our dataset. Among the individuals in our dataset, 4,687 individuals have never experienced a stroke, while only 248 individuals have a history of stroke. This indicates that the majority of individuals in our dataset do not have a history of stroke, with a much smaller proportion having experienced a stroke.

The chart displays the average BMI levels for individuals with and without a history of stroke. The average BMI level for individuals with no history of stroke is 29.20, while the average BMI for individuals who have had a stroke is slightly higher at 30.47. There is not a significant difference in BMI levels between the two groups.

The chart compares the average glucose levels between the two groups. The average glucose level for individuals with no history of stroke is 105.14, while the average glucose level for individuals with a history of stroke is 132.80. It is important to note that a normal blood sugar level is typically considered to be less than 140 mg/dL (7.8 mmol/L). Based on this threshold, both groups fall within the normal range, and there is no significant difference in glucose levels between the two groups.

Out of the 248 individuals with a history of stroke, 28 individuals were never married, while 220 were married. This difference in marital status between the two groups is statistically significant. However, it is important to note that married individuals in our dataset tend to be older compared to individuals who have never been married. There is a moderate, positive correlation of 0.65 between age and the married category.Based on this analysis, we can conclude that being married does not increase the likelihood of a stroke. Other factors, such as age, may play a more significant role in determining the risk of a stroke.

The pie chart provides an overview of the individuals who have a history of stroke, categorized into four groups based on their smoking status. Among the individuals, 36.29% are classified as never smokers, indicating that they have never smoked. 16.94% are categorized as current smokers, indicating that they continue to smoke. Additionally, 28.23% are classified as former smokers, suggesting that they used to smoke but have since quit. If we combine the percentage of current smokers and former smokers, we get a total of 45.17%, indicating that almost half of the individuals in our dataset have a history of smoking. It is important to note that we do not have any information on the smoking status of 18.55% of the individuals.

This pie chart provides insights into the distribution of smoking habits among individuals with a history of stroke, highlighting the prevalence of smoking among this group.

Among the individuals with a history of stroke, 88.71% of them were married at some point in their lives, while 11.29% were never married. This indicates that a majority of individuals who experienced a stroke had a marital status of being married, while a smaller proportion were never married.

Among the individuals who experienced a stroke, 56.45% were female and 43.55% were male. This distribution suggests that there is no significant difference in the occurrence of stroke between genders.

Out of the individuals with a history of stroke, 54.04% live in an urban area and 45.97% live in a rural area. This distribution indicates that there is no significant difference in the prevalence of stroke between urban and rural areas.

The chart categorizes individuals who have experienced a stroke into four work type categories. Among these individuals, 60.1% work in the private sector, 26.2% are self-employed, 13.3% work in the government sector, 0.4% work with children, and 0% have never worked. It is notable that a higher percentage of individuals working in the private sector and those who are self-employed have a history of stroke compared to the other work types. However, it is important to note that these observations are based on a relatively small sample size of 248 individuals, so caution should be exercised in making assumptions about the general population.

According to the data, 81.05% of the individuals in our dataset have no history of heart disease, while only 18.95% have been diagnosed with heart disease. It is quite surprising to see a relatively low percentage of individuals with a history of heart disease in our dataset.

Part 2

In this next section, we will look at different simple linear regression models using single independent variables.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.248369332 |  |  |  |  |  |  |  |
| R Square | 0.061687325 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.061497114 |  |  |  |  |  |  |  |
| Standard Error | 0.211664474 |  |  |  |  |  |  |  |
| Observations | 4935 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 1 | 14.52965885 | 14.52965885 | 324.3093521 | 2.84609E-70 |  |  |  |
| Residual | 4933 | 221.0075245 | 0.04480185 |  |  |  |  |  |
| Total | 4934 | 235.5371834 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | -0.062235493 | 0.006935117 | -8.973964356 | 3.9913E-19 | -0.075831409 | -0.0486396 | -0.0758314 | -0.0486396 |
| age | 0.002515974 | 0.00013971 | 18.00859106 | 2.84609E-70 | 0.002242081 | 0.00278987 | 0.00224208 | 0.00278987 |

In this regression model, the stroke is the dependant variable and age is the independent variable.

Stroke = -0.0622 + 0.00251 \* Age

The R-squared value of 0.062 suggests that only 6.2% of the variation in the dependent variable (stroke) can be explained by the predictor variable (age). This means that age alone does not have a strong influence on predicting the probability of stroke.

The adjusted R-squared value being identical to the R-squared value indicates that adding additional predictor variables would not improve the model's performance in explaining the variation in stroke.

The standard error of 0.211 represents the average deviation between the observed and predicted values of the dependent variable. A smaller standard error indicates a better fit of the model.

The intercept of -0.0062 indicates the predicted value of stroke when all predictors are zero. However, since age cannot be zero in reality, the interpretation of the intercept in this context might not be meaningful.

The estimated coefficient for age of 0.0025 suggests that for every one unit increase in age, the estimated probability of stroke increases by 0.25%. This implies that age has a small positive effect on the likelihood of experiencing a stroke.

Although the model shows statistical significance with a low p-value of 2.85E-70, indicating that age is a useful predictor in determining the probability of stroke, the low R-squared value suggests that age alone is not sufficient to accurately predict stroke.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.102354522 |  |  |  |  |  |  |  |
| R Square | 0.010476448 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.010275855 |  |  |  |  |  |  |  |
| Standard Error | 0.217363815 |  |  |  |  |  |  |  |
| Observations | 4935 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 1 | 2.467593082 | 2.467593082 | 52.22747703 | 5.69995E-13 |  |  |  |
| Residual | 4933 | 233.0695903 | 0.047247028 |  |  |  |  |  |
| Total | 4934 | 235.5371834 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 0.017699115 | 0.005464923 | 3.238676299 | 0.001208812 | 0.006985435 | 0.0284128 | 0.00698544 | 0.0284128 |
| Married | 0.047913769 | 0.006629959 | 7.226858033 | 5.69995E-13 | 0.0349161 | 0.06091144 | 0.0349161 | 0.06091144 |

In this model, the independent variable is whether the person was ever married or not.

The R-squared value of 0.0104 indicates that only 1.04% of the variation in stroke can be explained by whether the person was ever married or not. This suggests that the variable of being married has a very weak relationship with the likelihood of experiencing a stroke.

The adjusted R-squared value being identical to the R-squared value implies that adding additional predictor variables will not improve the model's performance in explaining the variation in stroke.

The intercept of 0.017699 indicates that there is a 1.77% chance of a person experiencing a stroke when the person has never been married.

The coefficient for the married variable is 0.0479, meaning that if a person was ever married, their chances of experiencing a stroke would increase by 4.79%. However, it's important to note that the coefficient is relatively small, indicating a weak association between being married and the probability of stroke.

The p-value being extremely low suggests that the married variable is statistically significant and has some utility in predicting stroke. However, the overall model indicates a poor fit due to the low R-squared value, indicating that being married alone is not a strong predictor of stroke.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.12617152 |  |  |  |  |  |  |  |
| R Square | 0.015919252 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.015719763 |  |  |  |  |  |  |  |
| Standard Error | 0.216765194 |  |  |  |  |  |  |  |
| Observations | 4935 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 1 | 3.749575889 | 3.74957589 | 79.8000293 | 5.74686E-19 |  |  |  |
| Residual | 4933 | 231.7876075 | 0.04698715 |  |  |  |  |  |
| Total | 4934 | 235.5371834 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 0.041018706 | 0.003254204 | 12.6048338 | 7.0595E-36 | 0.034639017 | 0.0473984 | 0.03463902 | 0.0473984 |
| hypertension | 0.091511414 | 0.010244098 | 8.93308621 | 5.7469E-19 | 0.071428423 | 0.11159441 | 0.07142842 | 0.11159441 |

The intercept of 0.041 represents the predicted value of the dependent variable (probability of getting a stroke) when the person has no prior history with hypertension.

The p-value being extremely small indicates that the hypertension variable is statistically significant and provides useful information for estimating the probability of getting a stroke.

However, the overall model's R-squared value of 1.59% indicates that the model explains only a small portion of the variation in the dependent variable (stroke). This suggests that hypertension alone has a weak association with the probability of getting a stroke, and there may be other factors that are not captured in the model that contribute to the variation in stroke occurrence.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.134163823 |  |  |  |  |  |  |  |
| R Square | 0.017999931 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.017800864 |  |  |  |  |  |  |  |
| Standard Error | 0.216535915 |  |  |  |  |  |  |  |
| Observations | 4935 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 1 | 4.23965315 | 4.23965315 | 90.4212377 | 2.9249E-21 |  |  |  |
| Residual | 4933 | 231.29753 | 0.0468878 |  |  |  |  |  |
| Total | 4934 | 235.537183 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 0.043133047 | 0.00317203 | 13.5979412 | 2.2747E-41 | 0.03691446 | 0.04935163 | 0.03691446 | 0.04935163 |
| heart\_disease | 0.127776044 | 0.01343737 | 9.50900824 | 2.9249E-21 | 0.10143282 | 0.15411926 | 0.10143282 | 0.15411926 |

In this model, heart disease is the independent variable.

The intercept of 0.043 represents the predicted value of the dependent variable (probability of getting a stroke) when the individual doesn't have a heart disease.

The coefficient of heart disease is 0.1227, indicating that if the individual has a heart disease, their chances of getting a stroke increase by 12.27%.

The small p-value suggests that heart disease is statistically significant and provides valuable information for predicting the occurrence of a stroke.

However, the overall R-square of 1.79% indicates that only a small portion of the variation in stroke occurrence is explained by heart disease. This implies that there are other factors not captured in the model that contribute to the occurrence of stroke, and heart disease alone may not be a strong predictor of stroke.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |  |
| Multiple R | 0.003792787 |  |  |  |  |  |  |  |  |
| R Square | 1.43852E-05 |  |  |  |  |  |  |  |  |
| Adjusted R Square | -0.000188328 |  |  |  |  |  |  |  |  |
| Standard Error | 0.218509869 |  |  |  |  |  |  |  |  |
| Observations | 4935 |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |  |
| Regression | 1 | 0.003388257 | 0.00338826 | 0.07096337 | 0.78995038 |  |  |  |  |
| Residual | 4933 | 235.5337951 | 0.04774656 |  |  |  |  |  |  |
| Total | 4934 | 235.5371834 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |  |
| Intercept | 0.049253731 | 0.00487386 | 10.1056932 | 8.8697E-24 | 0.0396988 | 0.05880867 | 0.0396988 | 0.05880867 |  |
| Work\_type | 0.00168644 | 0.006330728 | 0.26638951 | 0.78995038 | -0.0107246 | 0.01409748 | -0.0107246 | 0.01409748 |  |

In this model, the predictor variable is work type, with a value of 1 assigned to workers in the private sector and 0 assigned to workers in all other types.

The high p-value suggests that the work type variable is not statistically significant and does not provide useful information for predicting the probability of getting a stroke.

Both the R-square and adjusted R-square values are extremely low, indicating that the model does not explain any of the variation in the probability of getting a stroke. This means that work type alone is not a strong predictor of stroke.

Overall, the model is weak and lacks predictive power in determining the likelihood of a stroke based on work type. It suggests that other factors not included in the model may play a more significant role in predicting the occurrence of a stroke.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.132115487 |  |  |  |  |  |  |  |
| R Square | 0.017454502 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.017255324 |  |  |  |  |  |  |  |
| Standard Error | 0.216596042 |  |  |  |  |  |  |  |
| Observations | 4935 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 1 | 4.111184218 | 4.11118422 | 87.632642 | 1.168E-20 |  |  |  |
| Residual | 4933 | 231.4259992 | 0.04691385 |  |  |  |  |  |
| Total | 4934 | 235.5371834 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | -0.016971818 | 0.007815136 | -2.17166 | 0.02992882 | -0.032293 | -0.0016507 | -0.032293 | -0.0016507 |
| avg\_glucose\_level | 0.000631064 | 6.74125E-05 | 9.3612308 | 1.168E-20 | 0.00049891 | 0.00076322 | 0.00049891 | 0.00076322 |
|  |  |  |  |  |  |  |  |  |

In this model, the independent variable is the average glucose level, which is found to be statistically significant in predicting the occurrence of a stroke, as indicated by the low p-value.

The intercept of -0.01697 represents the predicted value of stroke when the average glucose level is zero, which doesn't make practical sense since glucose is an essential component for bodily functions.

The coefficient of the average glucose level suggests that for every unit increase in the average glucose level, the probability of having a stroke increases by 11%. This means that higher glucose levels are associated with a higher risk of stroke.

However, it's important to note that the overall model has a low R-square value, indicating that only a small portion of the variation in stroke can be explained by the average glucose level alone. Therefore, while average glucose level may have some influence on the probability of stroke, there are likely other factors not included in the model that play a more significant role.

In conclusion, while the average glucose level is a useful variable in predicting stroke, this model is not considered strong or reliable due to its limited explanatory power.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.02764613 |  |  |  |  |  |  |  |
| R Square | 0.000764309 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.000561747 |  |  |  |  |  |  |  |
| Standard Error | 0.21842792 |  |  |  |  |  |  |  |
| Observations | 4935 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 1 | 0.180023077 | 0.18002308 | 3.773217862 | 0.05213619 |  |  |  |
| Residual | 4933 | 235.3571603 | 0.04771076 |  |  |  |  |  |
| Total | 4934 | 235.5371834 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 0.026949705 | 0.012393223 | 2.17455173 | 0.02971109 | 0.00265347 | 0.05124594 | 0.00265347 | 0.05124594 |
| bmi - no missing | 0.000797531 | 0.000410574 | 1.94247725 | 0.052136188 | -7.3771E-06 | 0.00160244 | -7.377E-06 | 0.00160244 |

In this model, the independent variable is the BMI (Body Mass Index) levels. However, the p-value associated with the BMI variable is larger than 5%, indicating that it is not statistically significant and not useful for predicting the occurrence of a stroke.

Furthermore, the R-square value is extremely small, suggesting that only a tiny portion of the variation in stroke can be explained by BMI levels alone. This indicates that BMI levels are not a strong predictor of stroke in this model.

Based on these results, we can conclude that BMI levels do not have a significant relationship with the likelihood of having a stroke in the given dataset. Other variables or factors not included in the model may have a more substantial impact on stroke occurrence.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.00968947 |  |  |  |  |  |  |  |
| R Square | 9.3886E-05 |  |  |  |  |  |  |  |
| Adjusted R Square | -0.0001088 |  |  |  |  |  |  |  |
| Standard Error | 0.21850118 |  |  |  |  |  |  |  |
| Observations | 4935 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 1 | 0.022113619 | 0.02211362 | 0.46318261 | 0.49617226 |  |  |  |
| Residual | 4933 | 235.5150698 | 0.04774277 |  |  |  |  |  |
| Total | 4934 | 235.5371834 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 0.05192244 | 0.003960982 | 13.1084795 | 1.2929E-38 | 0.04415716 | 0.05968773 | 0.04415716 | 0.05968773 |
| Smokes | -0.0043537 | 0.00639714 | -0.6805752 | 0.49617226 | -0.016895 | 0.00818751 | -0.016895 | 0.00818751 |

In this model, the independent variable is "never smoked" in the smoking status column. You assigned the number one to individuals who never smoked since the largest percentage of individuals with a history of stroke falls into this category.

However, the results indicate that the independent variable "never smoked" is not statistically significant in predicting strokes, as indicated by the large p-value. The R-square value is also very low, suggesting that this variable explains only a small portion of the variation in stroke occurrence.

Overall, the models built using the individual independent variables have not been strong in predicting strokes, as seen by their low R-square values. It is worth exploring whether combining multiple variables in a multiple regression model could lead to better predictions.

By including multiple variables together in a multiple regression model, we can account for the collective influence of these variables and potentially improve the predictive power of the model. It would be interesting to assess the performance of such a model and see if it can better explain the variation in stroke occurrence.

Third Part

In this part, you plan to use a combination of age, hypertension, heart disease, marital status, and average glucose levels as predictor variables to predict the probability of stroke. It seems you have excluded some variables from your analysis due to their poor ability to predict stroke.

By incorporating these selected variables into a multiple regression model, you aim to capture their collective impact on stroke occurrence and potentially improve the model's predictive ability. This approach allows you to assess the significance and contribution of each variable while considering their interrelationships.

By focusing on variables that demonstrate stronger relationships with the dependent variable and excluding those with weaker predictive abilities, you can enhance the model's overall performance and increase its ability to accurately predict the probability of stroke.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | age | hypertension | heart\_disease | Married | avg\_glucose\_level |
| age | 1 | 0.272396547 | 0.267175533 | 0.65022485 | 0.239339405 |
| hypertension | 0.272396547 | 1 | 0.106308349 | 0.1537228 | 0.162321208 |
| heart\_disease | 0.267175533 | 0.106308349 | 1 | 0.1081857 | 0.162321208 |
| Married | 0.650224851 | 0.153722801 | 0.108185705 | 1 | 0.149724058 |
| avg\_glucose\_level | 0.239339405 | 0.173325779 | 0.162321208 | 0.14972406 | 1 |

To ensure the accuracy of my models and avoid multicollinearity, I conducted a correlation analysis among the variables. From the correlation matrix, I observed that age and marital status exhibit a strong relationship. To prevent redundancy and multicollinearity issues, I made the decision not to include both variables in the models I constructed. This approach ensures that each variable in my models contributes unique information and enhances the reliability and interpretability of the results. By selecting variables with distinct relationships to the dependent variable and minimizing collinearity, I aim to develop robust models that offer valuable insights into the predictors of stroke.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.175118989 |  |  |  |  |  |  |  |
| R Square | 0.03066666 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.030273581 |  |  |  |  |  |  |  |
| Standard Error | 0.215156653 |  |  |  |  |  |  |  |
| Observations | 4935 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 2 | 7.223138788 | 3.611569394 | 78.0164894 | 0 |  |  |  |
| Residual | 4932 | 228.3140446 | 0.046292385 |  |  |  |  |  |
| Total | 4934 | 235.5371834 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 0.035487333 | 0.00329257 | 10.77800368 | 8.675E-27 | 0.02903243 | 0.04194224 | 0.02903243 | 0.04194224 |
| hypertension | 0.082094531 | 0.010226029 | 8.027996901 | 1.2286E-15 | 0.062046962 | 0.1021421 | 0.06204696 | 0.1021421 |
| heart\_disease | 0.116316121 | 0.013427869 | 8.66229169 | 6.1795E-18 | 0.089991521 | 0.14264072 | 0.08999152 | 0.14264072 |
|  |  |  |  |  |  |  |  |  |

In this multiple linear regression model, I have included hypertension and heart disease as predictor variables to estimate the probability of experiencing a stroke. The intercept value is 0.036, indicating that an individual without hypertension or heart disease has a baseline likelihood of 3.5% of getting a stroke.

If the person has heart disease, the probability of stroke increases by 11.63%. Similarly, if the person has hypertension, the probability of stroke increases by 8.2%. Both variables show statistical significance in predicting stroke, as evidenced by their low p-values.

However, it is important to note that the overall model only explains a small portion of the variation in stroke, suggesting that there may be other factors not accounted for in this model. Therefore, while hypertension and heart disease provide some predictive power, additional variables may need to be considered to develop a more accurate and comprehensive model for predicting stroke.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.271733155 |  |  |  |  |  |  |  |
| R Square | 0.073838907 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.073087458 |  |  |  |  |  |  |  |
| Standard Error | 0.210353405 |  |  |  |  |  |  |  |
| Observations | 4935 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 4 | 17.39180827 | 4.347952067 | 98.262013 | 1.39568E-80 |  |  |  |
| Residual | 4930 | 218.1453751 | 0.044248555 |  |  |  |  |  |
| Total | 4934 | 235.5371834 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | -0.080920808 | 0.009149645 | -8.84414693 | 1.2636E-18 | -0.098858187 | -0.0629834 | -0.0988582 | -0.0629834 |
| age | 0.002038926 | 0.000151286 | 13.4772569 | 1.1096E-40 | 0.001742337 | 0.00233551 | 0.00174234 | 0.00233551 |
| hypertension | 0.038711323 | 0.010404758 | 3.720540251 | 0.00020101 | 0.018313363 | 0.05910928 | 0.01831336 | 0.05910928 |
| heart\_disease | 0.061222778 | 0.013625659 | 4.493197442 | 7.1767E-06 | 0.034510418 | 0.08793514 | 0.03451042 | 0.08793514 |
| avg\_glucose\_level | 0.000306929 | 6.82395E-05 | 4.49781425 | 7.0233E-06 | 0.000173149 | 0.00044071 | 0.00017315 | 0.00044071 |

In this model, I have expanded the predictor variables to include age and average glucose levels alongside hypertension and heart disease. The intercept is -0.08, representing the predicted value when all variables are zero.

For each one-unit increase in age, the likelihood of experiencing a stroke increases by 0.2%, while a one-unit increase in average glucose levels results in a 0.03% increase in the likelihood of stroke.

Having heart disease increases the likelihood of stroke by 6.11%, and hypertension increases it by 3.87%.

Overall, this model shows a slightly higher R-square value compared to the previous model, indicating a slightly better fit. However, it is important to note that the model still does not provide a good fit. An interesting observation is that the sum of squared residuals (SSE) is larger than the sum of squared regression, suggesting that the model does not effectively explain the variation in the dependent variable.

Therefore, while the inclusion of age and average glucose levels improves the model to some extent, there may be other important variables or factors not included in the model that could further enhance its predictive capability for stroke.

Summary Output

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.208843518 |  |  |  |  |  |  |  |
| R Square | 0.043615615 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.042839644 |  |  |  |  |  |  |  |
| Standard Error | 0.213758069 |  |  |  |  |  |  |  |
| Observations | 4935 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 4 | 10.27309915 | 2.568274787 | 56.2077827 | 1.97031E-46 |  |  |  |
| Residual | 4930 | 225.2640842 | 0.045692512 |  |  |  |  |  |
| Total | 4934 | 235.5371834 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | -0.028036503 | 0.008484269 | -3.3045279 | 0.00095816 | -0.044669448 | -0.0114036 | -0.0446694 | -0.0114036 |
| hypertension | 0.065093941 | 0.01037215 | 6.275838982 | 3.7754E-10 | 0.044759909 | 0.08542797 | 0.04475991 | 0.08542797 |
| heart\_disease | 0.098151733 | 0.013526439 | 7.256287636 | 4.5988E-13 | 0.071633889 | 0.12466958 | 0.07163389 | 0.12466958 |
| avg\_glucose\_level | 0.00043116 | 6.87449E-05 | 6.271876559 | 3.872E-10 | 0.000296389 | 0.00056593 | 0.00029639 | 0.00056593 |
| Married | 0.029909737 | 0.006671283 | 4.483355954 | 7.5145E-06 | 0.016831052 | 0.04298842 | 0.01683105 | 0.04298842 |

In this model, we have replaced the age variable with the ever\_married (or married) variable. The R-square value is slightly lower compared to the previous model, indicating that this model explains a smaller portion of the variation in stroke.

The p-values for all variables in this model are small, indicating that all variables are statistically significant and useful for predicting stroke. The coefficient of the married variable is 0.02999, suggesting that being married increases the likelihood of getting a stroke by 2.999%.

However, it is worth noting that the intercept is negative, which may not make intuitive sense in the context of stroke likelihood.

Overall, our models have demonstrated that the independent variables we used were generally useful but not strong predictors of stroke. The models, in general, have performed poorly in accurately predicting the likelihood of stroke. This suggests that there may be other important variables or factors that were not included in our models, and further research is needed to improve the predictive capability for stroke.