2004287 : Mert ACAR
2004099 : Celal Berkay ALTUNEL
2101196 : Mohammed Yazan KURDİ He made no contribution
to the PROJECT

Mert ACAR: https://github.com/AaCcRr/AIN
Celal Berkay ALTUNEL: https://github.com/AaCcRr/AIN

2023

1 Analysis

We analyzed our data sequentially with the purpose of obtaining our data in the most accurate form and ensuring that our model, which we will create, produces reliable results. We selected various methods of analysis based on the available data. We calculated the results using three different approaches: t-test, correlation, and percentage analysis.

1.1 Calculate the correlation between avg glucose level and stroke

- The correlation between "avg$_{g}$*lucose*$_{l}$*evel”and”stroke”iscalculatedas*0.15.*This valueindicatesapositiverelationshipbetweenthetwovariables.However,sincethe correlationcoefficientis*0.15,*thisrelationshipisconsideredweak.Inotherwords, thecorrelationbetween”avg*$_{g}$*lucose*$_{l}$*evel”and”stroke”isnotverystrong.*

- The correlation between "bmi" and "stroke" is calculated as 0.07. This value indicates a very weak relationship between the two variables. With a correlation coefficient of 0.07, the relationship between "bmi" and "stroke" is almost negligible.

- The correlation between "age" and "stroke" is calculated as 0.26. This value indicates a moderate positive relationship between the two variables. With a correlation coefficient of 0.26, the relationship between "age" and "stroke" is moderately positive.

These correlation values quantify the relationship between the variables. However, it's important to note that correlation measures only linear relationships and may disregard the influence of other factors. Therefore, it is important to consider other analysis methods and specific characteristics of the data as well.

1.2 T test

- The p-value of 0.9673092270304077 suggests that there is no statistically significant difference between the Rural and Urban areas. This means that the difference observed in the data is likely due to chance, and there is no strong evidence to conclude that there is a meaningful difference between the two regions.

- The p-value of 0.06459348879850953 indicates that there is no statistically significant difference between the male and female areas. This suggests that any observed differences in the data may be attributed to random variation, and there is not enough evidence to support the presence of a significant difference between the two genders.

- The extremely small p-value of 1.8922577018821537e-45 indicates that there is a statistically significant difference between the Yes and No areas. This means that the observed difference in the data is highly unlikely to occur by chance alone. The results provide strong evidence to support the presence of a meaningful difference between the two categories.

  In statistical analysis, the p-value helps determine the significance of the observed differences or relationships. A p-value less than a predetermined significance level (commonly 0.05) is typically considered statistically significant, suggesting that the observed difference is unlikely due to random chance. On the other hand, a p-value greater than the significance level indicates that there is not enough evidence to support the presence of a significant difference or relationship.

1.3 Analysis by Percentage

- "Residence type'e göre stroke yüzdeleri" means "Stroke percentages by Residence$_t$*ype”. Theoutputshowsthestrokepercentagesfordifferentresidencetypes(RuralandUrban). Accordingtotheresults, thestrokepercentageisapproximately*%4.14*forRuralareasand* %4.12*forUrbanareas.Thissuggeststhatthereisaslightdifferenceinstrokeprevalence basedontheresidencetype, butthedifferenceisnotsubstantial.*

- "smoking status'e göre stroke yüzdeleri" means "Stroke percentages by smoking$_s$*tatus”. Theoutputdisplaysthestrokepercentagesfordifferentsmokingstatuses(Unknown, formerly smoked, neversmoked, andsmokes).Theresultsshowthatthestrokepercentageis approximately* %2.38*forUnknownsmokingstatus,* %6.80*forformerlysmoked,* %4.09*fornever smoked, and*%5.04*forsmokes.Thisindicatesthattherearevariationsinstroke prevalencebasedonsmokingstatus, withhigherpercentagesobservedforformerlysmokedand smokescategoriescomparedtoothercategories.*

- Overall, these analyses provide insights into the relationship between
  and stroke prevalence. They help identify potential associations between variables and stroke occurrence, which can be useful for further investigation or predictive modeling.

Overall, these analyses provide insights into the relationship between different variables and stroke prevalence. They help identify potential associations between variables and stroke occurrence, which can be useful for further investigation or predictive modeling.

2 Model

2.1 One-Hot Encoding

One-Hot Encoding is a method for converting categorical variables to numerical representations. In this method, binary columns (dummy variables) are created for each category. The One-Hot Encoding processes used in the related data set are as follows:

One-Hot Encoding for the

"gender" column: [language=Python] pd.get$_{d}ummies(data[”gender”])$ This operation represents the "Male" and "Female" categories in the "gender" column with binary columns.

One-Hot Encoding is applied similarly for other categorical variables.

2.2 Deleting Original Columns

After One-Hot Encoding is complete, the encoded categorical columns are extracted from the original dataset. This is done because the coded columns already contain the information of the original categorical variables.

2.3 Creating the Updated DataFrame

An empty DataFrame is created to store the updated data and the corresponding data rows for each ID are pulled from the original dataset and added to this DataFrame.

2.4 Removing the ID Column

After the rows for each ID are combined, the ID column is removed. This means that the ID will not be used as an input to the model.

2.5 Preparing Input and Output Data

Input and output data are parsed from the updated DataFrame. Input data is all columns except the "stroke" column. The output data is the "stroke" column.

2.6 Creating the Model

A sequential model is defined in the section where the model is created. This model consists of dense layers and is configured with different activation functions.

2.7 Compilation of Model

The model is compiled by specifying the optimizer, loss function, and metrics. In this step, it is determined how the model will be learned and how its performance will be measured.

2.8 Training the Model

The model is trained in a specified number of epochs and mini-groups with the specified input and output data. In this step, the model improves its performance by learning the patterns on the data.

2.9 Result

The One-Hot Encoding process converts categorical variables to numerical representations, enabling machine learning models to process data more effectively. In this document, I've covered how One-Hot Encoding is implemented and other steps in the data preparation process.

3 Division of Labor

3.1 Celal Berkay Altunel:

Data analysis and simplification

3.2 Mert Acar:

Artificial intelligence modelings

References

References follow the acknowledgments in the camera-ready paper. Use unnumbered first-level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to `small` (9 point) when listing the references. Note that the Reference section does not count towards the page limit.

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.