TGVD teamwork result sheet

|  |
| --- |
| **Team number:** *16* |
| **1. Describe the nodes at your cluster (i.e., one master, four workers, one client)** |
| One master, four workers. |
| **2. Have you used any tool or format from Unit 5: Other Big Data tools? If yes, describe which ones and how did you use them.** |
| No. |
| **3. Explain the solution for the analytical questions from section 4.1.** |
| Query 1: Average age of patients who have heart disease versus patients who do not have heart disease. In our given dataset, age is depicted as string categories, so we must convert these strings to numerical values. Then, we simply filter the patients which have had heart disease and those who have not into separate groups and get the average age of each of them. It’s important to add the empty groupBy() function call to later be able to get the sum of the ages.  Query 2: Percentage of Smoker with Heart Disease. We filtered the rows which’s SmokerStatus column indicates that the patient is a smoker and got overall how many smokers are in the dataset. Then we got the count of smokers with heart disease. Lastly, we divided the counts and multiplied by 100% to get the percentage.  Query 3: Percentage of alcohol consumers who have heart disease. We filtered the rows where the HadHeartAttack column is 'Yes' to get the total number of patients with heart disease. Next, we filtered the rows where both HadHeartAttack and AlcoholDrinkers columns are 'Yes' to get the number of alcohol consumers with heart disease. Finally, we calculated the percentage by dividing the number of alcohol consumers with heart disease by the total number of heart disease patients and multiplying by 100%.  Query 4: BMI Difference of Patients with and without Heart Disease. We calculated the average Body Mass Index (BMI) for two groups of patients: those who have had a heart attack and those who have not. By filtering the dataset based on the HadHeartAttack column, we computed the average BMI for each group separately. This helps to understand whether there is a significant difference in BMI between patients with and without heart disease.  Query 5: Average Sleeping Time Difference of Patients with and without heart disease. We calculated the average sleeping time for two groups of patients: those who have had a heart attack and those who have not. By filtering the dataset based on the HadHeartAttack column, we computed the average sleeping hours for each group separately. This helps to understand whether there is a significant difference in sleeping patterns between patients with and without heart disease.  Query 6: Percentage of Patients with Heart Disease Who Partake in Physical Activity. We calculated the percentage of patients with heart disease who engage in physical activity. First, we filtered the dataset to find the total number of patients with heart disease. Then, we filtered again to count the number of these patients who reported partaking in physical activity. Finally, we calculated the percentage by dividing the number of active heart disease patients by the total number of heart disease patients and multiplying by 100. |
| **4. Explain the algorithms used in section 4.2 and the different results you have obtained.** |
| We chose Logistic Regression as the model for our problem solution, utilizing methods from the PySpark MLlib libraries. Initially, we transformed all categorical variables into numerical ones. To achieve this, we employed the StringIndexer method, which assigned a unique numerical value to each category. Subsequently, we utilized the OneHotEncoder method to convert these numerical values into binary vectors. This transformation was particularly suitable for Logistic Regression, as binary vectors are often more effective than decimal numbers in this context. In our modeling process, we decided to include all available variables except for the predicted one. Initially, we experimented with a subset of variables, but found that this led to significant inaccuracies in the results. For the final model, we achieved an overall accuracy of 94.9% and a high specificity of 98.9%. However, our sensitivity was notably lower at 25%, indicating a substantial number of false negatives. This discrepancy highlights the model's struggle to correctly identify positive cases, suggesting potential areas for improvement in our approach. We filtered the top 10 variables that had the most impact on the model. Interestingly, height has the highest impact on the prediction, and it affects the prediction negatively. Age is the second most impactful variable and has a positive influence. Having had skin cancer is the third most important predictor and also has a positive impact. It's important to note that height is a numerical variable, while the others are categorical. This difference in scales can affect our interpretation of their relative impacts. |
| **5. Which tool have you used to present the results?** |
| Locally: We plotted the coefficients of the machine learning model with matplotlib.  The other results are displayed as data frames.  In the cluster: we are saving the results in .csv files. There are two folders, where one of them contains results from the analytical queries and the logistic regression model performance metrics, while the other contains logistic regression coefficient values for each variable. |