Project_Big Data Analytics_Pyspark_GCP

1. Data set description: Provide a detailed description of the public data set you have selected, including its source, format, and any relevant details about the data.

Human Stress Detection: Human stress level detection using physiological data

About Dataset

"Humidity – Temperature – Step count – Stress levels" represents the titles for Stress-Lysis.csv file.

Based on the human's physical activity, the stress levels of the human being are detected and analyzed here. A dataset of 2001 samples is provided for human body humidity, body temperature and the number of steps taken by the user. Three different classifications of stress are performed, low stress, normal stress, and high stress. More information on how this data is analyzed can be found at "L. Rachakonda, S. P. Mohanty, E. Kougianos, and P. Sundaravadivel, "Stress-Lysis: A DNN-Integrated Edge Device for Stress Level Detection in the IoMT," IEEE Trans. Conum. Electron., vol. 65, no. 4, pp. 474–483, 2019."

The advantage of this research lies in its potential to provide a non-invasive, real-time method for detecting and monitoring human stress levels using easily measurable physiological data like body humidity, temperature, and step count. Early detection of stress can lead to timely interventions, helping prevent negative health outcomes related to chronic stress, such as cardiovascular issues and mental health disorders. This research can also contribute to the development of wearable health devices, improving personal health management and workplace productivity by reducing stress-related problems.

Human Stress Detection (kaggle.com)

2. Research question: Clearly define your research question and explain why studying is important. What do you want to learn from the data?

Importance of Studying This Topic:

Understanding and detecting stress is crucial because stress can have significant impacts on both physical and mental health. Chronic stress is linked to a variety of health problems, including cardiovascular disease, immune system dysfunction, and mental health disorders like anxiety and depression. Early detection of stress can lead to timely interventions that improve overall well-being, productivity, and quality of life.

By using physiological indicators such as humidity, temperature, and step count, we can analyze how these factors correlate with stress levels, offering a non-invasive and objective way to monitor stress in real time. This research is important for developing technologies, such as wearable health devices, that can monitor stress and provide real-time feedback to users.

Additionally, it could have applications in fields like occupational health, where understanding stress levels could improve workplace environments and reduce burnout.

What We Want to Learn from the Data:

From this dataset, we aim to learn how well body humidity, temperature, and step count can serve as indicators of stress. Specifically, we seek to:

1. Determine the relationship between these physiological metrics and stress levels.
2. Develop a model that can classify stress levels (low, normal, and high) based on the input features.
3. Evaluate the accuracy of our classification model and understand which physiological factors contribute most to predicting stress levels.

Ultimately, this research will help us understand whether these indicators are sufficient for accurate stress detection and what improvements can be made for future stress-monitoring technologies.

3. Machine Learning model: Specify the type of machine learning model you plan to use, such as classification or clustering, and explain why you have chosen this model.

For our above-mentioned dataset, where the goal is to classify stress levels (low, normal, and high) based on numerical physiological data (humidity, temperature, and step count), **one of the following three classification models in PySpark are recommended and will be done using one or more models:**

1. Logistic Regression (Multinomial / Softmax Regression):
   Logistic regression (multinomial logistic regression (also known as softmax regression) is a simple yet powerful model for classification tasks, especially when there are multiple categories, as in this case with three stress levels (low, normal, high). It is interpretable and can be trained efficiently on numerical datasets like the one you have. Logistic regression can handle multiclass classification using the one-vs-rest (OvR) or multinomial methods, which makes it suitable for this task.
2. Random Forest Classifier:
   Random Forest is a robust and flexible model that can handle complex patterns in the data and performs well even when there are non-linear relationships between the input features (humidity, temperature, step count) and the target (stress level). It is an ensemble method that combines multiple decision trees to improve classification accuracy and reduce overfitting. Random forests work well with tabular data and can handle imbalanced classes or noisy data effectively.
3. Decision Tree Classifier:

The Decision Tree Classifier is a simple yet powerful model that can handle complex patterns in the data and performs well even when there are non-linear relationships between the input features (humidity, temperature, step count) and the target (stress level). It works by recursively partitioning the data based on the most significant features to create a tree-like model of decisions. Decision Trees work well with tabular data and are easy to interpret, making them suitable for understanding the relationships between features and the target variable. However, they can be prone to overfitting if not properly pruned, so techniques like pruning or setting a maximum depth are often used to improve generalization.

4. Gradient-Boosted Trees (GBT):
   Gradient-Boosted Trees are another ensemble-based model that builds trees sequentially, where each tree tries to correct the errors of the previous one. This method can produce very accurate results for classification tasks. GBTs are well-suited for medium-sized datasets like yours (2001 samples) and can handle both numerical and categorical data efficiently, offering high performance for classification problems like stress level prediction.

Note: one of the above three classification models in PySpark will be done using one or more models.

Why These Models?
These models were selected because they are well-suited for classification tasks involving numerical tabular data. Logistic regression offers a simple, interpretable baseline, while Random Forest and Gradient-Boosted Trees provide more powerful and flexible models that can capture complex relationships in the data. Additionally, all three models are well-supported in PySpark, making them suitable for scalable analysis on large datasets if needed in the future.

By comparing the performance of these models (e.g., f- score, accuracy, precision, recall), we can select the best approach for detecting human stress levels based on the physiological data provided.

4. Expected outcomes: What do you expect to achieve after implementing your learning model? What do you hope to learn or discover from your data analysis?

After implementing the learning model, I expect to achieve an accurate classification of human stress levels (low, normal, high) based on physiological data (humidity, temperature, step count). I hope to discover how well these features predict stress levels and which factors contribute most to the model's accuracy. Additionally, I aim to identify patterns and correlations between physical activity and stress, providing insights that could improve real-time stress monitoring and management. Moreover, extracting important features, predicting for new data, creating a mobile app to detect quickly. Finally, we can change the parameters and features and provide the physician patient some information that what factors are more effective for reducing stress.

5. Evaluation plan: Explain how you plan to evaluate your project and assess the correctness of your model. What metrics or methods will you use to evaluate the effectiveness of your learning model? How well do you expect the model to work, and how will you measure its performance?

To evaluate the project and assess the correctness of the model, I plan to use the following metrics and methods:

1. Accuracy: This will measure the proportion of correctly classified stress levels out of all predictions, giving a general sense of the model's performance.
2. Precision, Recall, and F1-Score: These metrics will provide a deeper understanding of the model's performance for each stress level (low, normal, high). Precision will show how many of the predicted stress levels are correct, while recall will indicate how many actual stress levels were correctly identified. The F1-score balances both precision and recall.
3. Confusion Matrix: I will use this to visualize the performance of the model by showing where the model is making correct predictions and where it is misclassifying different stress levels.
4. Cross-Validation: To ensure the model's robustness and avoid overfitting, I will use k-fold cross-validation to evaluate the model on multiple splits of the data and assess its generalization capability.

I expect the model to work reasonably well, with a good balance between precision and recall, especially for the dominant or most distinguishable stress classes. Performance will be measured by comparing these metrics across different models (e.g., Logistic Regression, Random Forest, Gradient-Boosted Trees) to select the most effective one.