Anxiety Severity Prediction Data Analysis Report

1. Dataset Description

1.1 Source:

The dataset used in this project is **"Anxiety Attack Dataset"**, obtained from **Kaggle**. It contains various psychological and lifestyle attributes of individuals, aimed at analyzing the factors contributing to anxiety severity.

1.2 Columns:

1. **ID** – Unique identifier assigned to each participant.
2. **Age** – Age of the individual in years.
3. **Gender** – Gender of the respondent (Male/Female/Other).
4. **Occupation** – The individual's professional or academic engagement (e.g., Student, Employee, Unemployed).
5. **Sleep Hours** – Average number of hours the person sleeps per day.
6. **Physical Activity (hrs/week)** – Total hours of physical exercise or activity performed weekly.
7. **Caffeine Intake (mg/day)** – Daily consumption of caffeine measured in milligrams.
8. **Alcohol Consumption (drinks/week)** – Average number of alcoholic drinks consumed per week.
9. **Smoking** – Indicates whether the person smokes (Yes/No).
10. **Family History of Anxiety** – Shows if the participant has a family history of anxiety disorders.
11. **Stress Level (1–10)** – Self-reported stress intensity on a scale of 1 (low) to 10 (high).
12. **Heart Rate (bpm during attack)** – Heart rate recorded during an anxiety episode, measured in beats per minute.
13. **Breathing Rate (breaths/min)** – Breathing rate during anxiety episodes, measured in breaths per minute.
14. **Sweating Level (1–5)** – Level of sweating observed during anxiety episodes, rated from 1 (none) to 5 (excessive).
15. **Dizziness** – Whether the participant experiences dizziness during anxiety attacks (Yes/No).
16. **Medication** – Indicates if the individual is taking prescribed medication for anxiety.
17. **Therapy Sessions (per month)** – Number of therapy or counseling sessions attended per month.

18. **Recent Major Life Event** – Whether the participant has recently experienced a major life event (e.g., job loss, trauma, relocation).
19. **Diet Quality (1–10)** – Self-assessed diet quality on a scale from 1 (poor) to 10 (excellent).
20. **Severity of Anxiety Attack (1–10)** – Target variable representing the intensity of anxiety attacks, ranging from 1 (mild) to 10 (severe).

1.3 Data Insights and Trends :

The dataset offers meaningful insights into how lifestyle, stress, and physiological factors influence anxiety.

For example:

- Individuals reporting higher stress and lower sleep duration tend to exhibit greater anxiety severity.
- A negative correlation is observed between physical activity and anxiety level — more active individuals generally report lower anxiety.
- Dietary quality and work hours also contribute to stress patterns, indirectly affecting anxiety outcomes.

These relationships make the dataset suitable for predictive modeling and correlation analysis using big data tools like PySpark.

2. Data Quality and Preprocessing

The notebook demonstrates a clear and systematic approach to understanding and preparing the dataset.

2.1 Schema and Structure Inspection :

Using PySpark, the dataset schema was loaded and inspected to verify column names and datatypes.

Appropriate transformations were applied:

- Numerical features (e.g., Age, Sleep Duration, Heart Rate) were cast to FloatType.
- Categorical variables (e.g., Gender, Diet Quality) were converted to StringType for encoding later.

2.2 Missing Value Analysis :

The dataset was examined for missing or null values:

- Missing numeric values were imputed using mean or median imputation.
- Missing categorical values were filled with mode (most frequent category).

2.3 Data Cleaning Steps :

The following preprocessing steps were implemented:

- Removal of duplicate records.
- Standardization of categorical values (e.g., lowercase conversion).
- Handling outliers in Heart Rate and Stress Level using IQR filtering.
- Label encoding of categorical variables for model training.
- Normalization of continuous variables using Min-Max scaling for uniform model input.

3. Operations Performed

3.1 Exploratory Data Analysis (EDA) :

EDA was performed using PySpark's DataFrame functions and visualization tools like Matplotlib and Seaborn.

Key analyses include:

- Distribution of Anxiety Severity among participants.
- Correlation heatmap showing relationships between stress, sleep, and anxiety.
- Boxplots comparing anxiety severity across gender and activity levels.
- Pairplots to visualize multi-feature interactions.

Visualizations revealed:

- Participants with low sleep and high stress form the majority of the "Severe Anxiety" group.
- Those maintaining good dietary habits and high activity levels mostly fall in the "Mild Anxiety" category.

3.2 Feature Engineering :

To enhance model performance, new features were derived:

- Stress-to-Sleep Ratio = Stress Level / Sleep Duration
- Activity-Adjusted Heart Rate = Heart Rate × (1 / Physical Activity Index)
- Workload Index = Work/Study Hours × Stress Level

These engineered variables provided deeper insight into behavioral and physiological relationships affecting anxiety.

3.3 Aggregation and Grouping :

Data was grouped and aggregated across several lifestyle and physiological attributes to analyze relationships with anxiety severity.

The following operations were performed:

- Total participants computed using count of unique IDs.
- Distribution by Gender – counts per gender category.
- Top Occupations by participants – grouped counts to identify most represented professions.
- Smoking and medication breakdowns – grouped counts and cross-tabulations (e.g., smoking status by gender).
- Average age per gender – grouped averages to compare demographic patterns.
- Sleep and caffeine-based aggregations – average stress and average severity calculated per sleep-hour group and caffeine-intake range.

These aggregations helped uncover how lifestyle factors and demographics relate to anxiety severity levels.

3.4 Visualization and Insights :

Visualizations were created by converting Spark DataFrames to Pandas and using Matplotlib and Seaborn.

Charts and graphs revealed:

- Bar chart of top occupations – identifies professions with the largest participant counts.
- Pie chart of gender distribution – shows participant proportion by gender.
- Scatter plot (Age vs Sleep Hours) – visualizes variation in sleep patterns across age groups.
- Bar plot of caffeine intake vs average severity – shows higher caffeine associated with elevated severity.
- Line plot (Sleep Hours vs Avg Stress / Avg Severity) – demonstrates that reduced sleep corresponds to higher stress and severity.
- Physiological trends – increased heart rate and breathing rate correspond to higher severity scores.

These visual findings confirm that stress, poor sleep, and high stimulant consumption are significant contributors to anxiety severity, while healthy habits (exercise, balanced diet, therapy) align with milder anxiety outcomes.

3.5 Model Building and Training :

A Logistic Regression model was implemented using PySpark MLlib to predict Severity of Anxiety Attack (1–10).

Steps:

- Data was split into training (80%) and testing (20%) subsets.
- All features were combined into a single vector using VectorAssembler.
- The model was trained on the training set and used to generate predictions on the test set.

At this stage, the notebook focuses primarily on data preparation, training, and prediction, without including full evaluation metrics.

3.6 Model Evaluation :

The trained model successfully generated predictions, but explicit evaluation metrics such as accuracy, precision, or recall were not computed in the current version. Future iterations can use PySpark's MulticlassClassificationEvaluator to calculate these metrics and enable model comparison.

4. Key Insights

- High stress and low sleep duration are dominant factors contributing to severe anxiety attacks.
- Participants with higher caffeine intake and smoking habits show greater average severity levels.
- Heart rate and breathing rate rise significantly during more severe anxiety attacks.
- Balanced lifestyle factors — adequate sleep, physical activity, and good diet — correspond to milder anxiety levels.
- Aggregated trends show that behavioral and physiological factors interact strongly, influencing severity outcomes.

5. Recommendations

5.1 For Healthcare Professionals :

- Encourage patients to maintain consistent sleep patterns and reduce caffeine intake.
- Recommend stress management and physical activity as preventive measures.

5.2 For Mental Health Researchers :

- Extend the dataset with additional features such as emotion tracking or social activity patterns.
- Apply ensemble and deep learning models to improve prediction accuracy.

5.3 For Further Development :

- Integrate data from wearable sensors (heart rate, sleep trackers) for real-time prediction.
- Develop a dashboard for monitoring anxiety levels and predicting future episodes using PySpark and visualization tools.

6. Conclusion

The Anxiety Severity Prediction Using PySpark project effectively demonstrates how large-scale data processing and analytics can be applied to mental health research. Through feature engineering, grouping, and visualization, the analysis highlights key behavioral and physiological factors influencing anxiety. The PySpark-based machine learning pipeline lays the foundation for scalable, data-driven mental health assessments and predictive modeling.