🔍 Post-Session Notes: Revision – Results and Visualization

🧠 1. Project Overview and Dataset Description

- The session continued work on a synthetic healthcare dataset containing 11 features inspired by real-world smart healthcare use cases.
- The dataset was shared via Google Colab, and students could either download it or use it directly.
- About 1% of the data had missing values, making data cleaning a crucial first step.

🧹 2. Data Cleaning and Handling Missing Values

- Categorical features: Missing values were imputed using the mode.
- Numerical features: Missing values handled using mean or median, depending on the distribution.
- This phase ensured that downstream machine learning models had a clean and consistent dataset to work with.

📊 3. Univariate Analysis and Visualization

- **Techniques used:**
  - **Count plots for categorical data (e.g., gender).**
  - **Box plots for numerical variables (e.g., daily calorie intake).**
- **Outlier detection and distribution shapes (e.g., normal, skewed, Poisson) were examined to understand data characteristics.**

🔗 4. Relationship Between Variables: Bivariate and Multivariate Analysis

- **Bivariate analysis:**
  - **Boxplots comparing calorie intake across physical activity levels.**
  - **Showed no significant variation in calorie intake by activity type.**
- **Multivariate analysis:**
  - **Scatter plots: Stress level vs sleep quality showed a negative correlation.**
  - **Added third variable using hue for mood (happy, neutral, sad) to extract richer insights.**
  - **Pair plots helped visualize interrelationships between multiple numerical variables such as stress, sleep hours, and mood.**

📈 5. Interpreting Statistical Tests and Model Fit

- Focus was on interpreting model results beyond accuracy.
- Emphasis on:
  - Class balance and the support values for each class.
  - Understanding misleading metrics when dealing with imbalanced data.
- Encouraged viewing metrics in context rather than relying on a single value.

🤖 6. Support Vector Machine (SVM) Kernels and Model Experimentation

- Models tested:
  - Linear Regression: For continuous target variables.
  - Logistic Regression: For categorical target variables.
  - SVMs with various kernels for classification.
- Kernels explored:
  - RBF (Radial Basis Function)
  - Linear
  - Polynomial (degree 3)

    - Sigmoid
- Key insights:
    - Accuracy was generally low (26–41%).
    - Surprisingly, Linear kernel sometimes outperformed RBF.
    - Sigmoid kernel struggled due to multi-class nature of the dataset.
    - Use of `class_weight='balanced'` had limited improvement, highlighting that kernel choice alone doesn't solve all problems.

⚖️ 7. Importance of Handling Imbalanced Data

- The dataset had class imbalance issues that led to misleadingly high accuracy for the majority class.
- Techniques used:
    - Adjusted class weights during model training.
    - Checked support values to verify if the model was learning minority classes.
- Conclusion: Handling imbalance is essential for building fair and generalizable models.

📉 8. Visualization Techniques for Data Interpretation

- **Visualization was positioned not just as a result-presentation tool but as a thinking tool for:**
    - **Hypothesis generation**
    - **Identifying trends**
    - **Spotting data issues (e.g., skewness, outliers)**
- **Tools used:**
    - **Matplotlib**
    - **Seaborn (sns): For box plots, KDEs, heatmaps, and pair plots.**
- **Introduced KDE (Kernel Density Estimation) as a smooth alternative to histograms for distribution visualization.**

🌡️ 9. Correlation Analysis Using Heatmaps

- **Used heatmaps to visualize relationships between numerical features.**
- **Strong correlation (~0.86) between daily steps and calorie intake was observed.**
- **Most other features showed weak or no significant correlations.**
- **Heatmaps provide a quick overview of feature interactions, valuable for feature selection and modeling strategies.**

📝 10. Key Takeaways on Model Selection and Data Visualization

- **Don't blindly trust accuracy – always assess context, class balance, and metric interpretation.**
- **Visualize early and often – start with univariate, move to bivariate, and then explore multivariate relationships.**
- **Simple models often perform better when the data is not complex or when well-understood.**
- **Use class balancing techniques when working with imbalanced datasets to prevent misleading outcomes.**
- **Visualization helps build intuition and supports more informed modeling choices.**

📌 Next Steps and Looking Ahead

- **The project will evolve into more advanced modeling, including neural networks and deep learning.**
- **Future sessions will:**
    - **Introduce new evaluation metrics.**
    - **Focus on model explainability and interpretability.**
    - **Emphasize real-world challenges in healthcare-related ML systems.**

✅ Closing Thoughts

This session highlighted the importance of data-driven thinking, model experimentation, and visual interpretation over simply chasing higher accuracy. As we move toward deeper models, this foundational approach will ensure robust and insightful data science workflows.
