🔍 Post-Session Notes: Revision - Building Models & Selecting the Right Model

---

1. Understanding Data and Feature Selection

- **Importance: Do not rush to use complex models. Start with a thorough understanding of your data.**
- **Exploratory Data Analysis (EDA) helps identify:**
  - **Missing values (e.g., 1% missingness handled with mean/median imputation)**
  - **Feature relevance and potential improvements (e.g., adding BMI as a useful feature)**
- **Feature selection ensures you focus on meaningful predictors, reducing noise and improving model interpretability.**

---

2. Correlation Analysis and Hypothesis Testing

- **Correlation coefficients quantify the strength and direction of linear relationships:**
  - **Example: Strong negative correlation (-0.93) between stress and sleep quality.**
- **Interpretation:**

    - Negative correlation: One variable increases as the other decreases.

- Hypothesis Testing:
    - Null hypothesis H0H_0: No relationship.
    - Alternative hypothesis H1H_1: Significant relationship exists.
- Use two-tailed tests to detect any kind of relationship.
- P-values < 0.05 suggest rejecting H0H_0, confirming significant correlation.
- R-squared (R^2) value explains variance explained by model; higher values indicate better fit.

---

3. Building Linear Regression Models

- Use linear regression when relationship appears linear (e.g., stress level predicting sleep quality).
- Model purpose: Predict future outcomes accurately, not just fit existing data.
- Prefer simpler models if performance is comparable.

---

4. Polynomial Regression

- **Tested quadratic terms to capture potential nonlinearities.**
- **No significant improvement over linear regression observed.**
- **Polynomial regression increases risk of overfitting—use cautiously.**

---

5. Categorical Variables and Logistic Regression

- **Categorical variables (e.g., mood, activity type) need appropriate encoding.**
- **Continuous variables can be converted to categories, e.g., sleep hours → binary "well-rested" or not.**
- **Logistic regression models probability of class membership (e.g., well-rested vs not).**
- **Poor accuracy (~51%) indicated logistic regression was insufficient for the problem and features.**
- **Thresholds for categorization impact results; extreme cutoffs may skew data and bias models.**

---

6. Support Vector Machines (SVM) for Classification

- **Used as an alternative to logistic regression for classification tasks.**
- **Aim: Find optimal decision boundaries maximizing margin between classes.**

- **Observed poor performance (~50%), likely due to complex, noisy data.**
- **Visualization with PCA (dimensionality reduction to 2 components) showed scattered data points, many as support vectors — indicating data overlap and noise.**

---

7. Dimensionality Reduction with Principal Component Analysis (PCA)

- **PCA reduces feature space while retaining maximum variance.**
- **Useful to visualize complex, high-dimensional data.**
- **In this case, showed no clear separability between classes, explaining poor classifier performance.**

---

8. Model Selection and Evaluation Strategy

- **Start with EDA: Understand your data, distributions, missing values, and feature relationships.**
- **Use statistical tools (correlation, hypothesis testing) to guide feature choice.**
- **Begin with simple models (linear regression, logistic regression).**
- **Consider complex models only if simpler ones fail or data complexity demands it.**

- Beware of overfitting, especially with polynomial terms or high-dimensional feature spaces.
- Use visualization tools like PCA to assess data separability and model limitations.
- Interpret accuracy and other metrics with caution, especially on imbalanced or noisy data.
- Continuous feature engineering and improved data collection may be more valuable than chasing model complexity.

🔑 Key Takeaways

- Never blindly apply advanced ML models without understanding the data.
- Data understanding drives good model choice.
- Simpler models often suffice and are easier to interpret.
- Classification thresholds must be chosen carefully considering data distribution.
- Visualization and dimensionality reduction are critical to understanding data and model behavior.
- Good models come from a cycle of data exploration → statistical analysis → modeling → evaluation → refinement.

- Real-world datasets may require creative solutions beyond standard models.

📈 Practical Advice

- Always start your project with detailed EDA.
- Document your data cleaning and imputation strategy.
- Compare models systematically with consistent evaluation metrics.
- Use domain knowledge to engineer and select meaningful features.
- Validate assumptions behind models (e.g., linearity for linear regression).
- Leverage visualization techniques early and often.
- Understand limitations of your data before drawing conclusions from model outputs.

This session reinforced the importance of thoughtful, stepwise model building and selection grounded in solid data understanding, especially in healthcare or other critical domains.
