Project Journal

Team Member: Rahul Muggalla

Contribution: House Project (Real Estate Pricing Predictions)

- **Tasks Performed**:
  - **Data Acquisition**:
    - Scraped property data from Makaan.com using BeautifulSoup (10 hours).
    - Challenges: Encountered dynamic website structures; resolved by updating XPath selectors.
  - **Data Storage**:
    - Designed and stored data in a MySQL database for efficient querying (5 hours).
    - Challenges: Optimized schema to handle large datasets and ensure scalability.
  - **Data Preprocessing**:
    - Cleaned missing values, standardized price formats, and encoded categorical variables (8 hours).
    - Challenges: Addressed inconsistent pricing formats like "Lakhs" and "Crores."
  - **Model Development**:
    - Trained regression models (Linear Regression, Random Forest, Decision Tree, XGBoost) to predict property prices (12 hours).
    - Challenges: Hyperparameter tuning for XGBoost required significant computational effort; resolved using GridSearchCV.
  - **Visualization**:
    - Created heatmaps and scatter plots using Plotly to identify feature correlations (6 hours).
    - Challenges: Ensured visualizations were clear and stakeholder-friendly.

- **Total Time Spent**: 41 hours.
- **Challenges Faced**:
  - Handling large datasets efficiently during preprocessing.
  - Balancing model complexity with interpretability for stakeholders.
- **Solutions Implemented**:
  - Used MySQL for structured data storage and seamless integration with Python.
  - Leveraged ensemble models like Random Forest and XGBoost for better predictive accuracy.

Team Member: Ayyappa Gorantla

Contribution: Crime Case Project (Binary Classification Task)

- **Tasks Performed**:
  - **Dataset Selection**:
    - Identified publicly available datasets relevant to binary classification tasks (4 hours).
    - Challenges: Ensured balanced class distribution in the dataset.
  - **Data Preprocessing**:
    - Cleaned missing values, encoded categorical variables, and scaled numerical features using StandardScaler (6 hours).
    - Challenges: Required stratified sampling to maintain class balance during train-test splits.
  - **Model Development**:
    - Implemented Logistic Regression as a baseline model (4 hours).
    - Explored advanced models like Random Forest and XGBoost for improved accuracy (10 hours).
    - Challenges: Hyperparameter tuning using GridSearchCV was resource-intensive.
  - **Evaluation**:

        - Assessed model performance using precision, recall, F1-score, and confusion matrices (5 hours).
        - Visualized ROC curves to analyze classification thresholds (3 hours).
    - **Optimization**:
        - Fine-tuned hyperparameters to balance precision and recall for critical decision-making scenarios (4 hours).
- **Total Time Spent**: 32 hours.
- **Challenges Faced**:
    - Balancing false positives and false negatives due to equal importance in classification tasks.
    - Computational intensity of ensemble models like Random Forest.
- **Solutions Implemented**:
    - Used stratified sampling to ensure balanced class representation.
    - Focused on explainability by analyzing feature importance in Random Forest models.

Team Member: Shubham Pandurang Kawade

Contribution: BigBasket Project (E-commerce Demand Forecasting)

- **Tasks Performed**:
    - **Dataset Creation**:
        - Simulated a dataset representing real-world e-commerce scenarios with over 10,000 records (6 hours).
        - Challenges: Ensured that the dataset captured seasonal trends and regional variations.
    - **Feature Engineering**:
        - Derived new features such as "average sales per region" and "demand variation rate" (5 hours).

        - Challenges: Required domain expertise to create meaningful features.
    - **Model Development**:
        - Trained regression models like Linear Regression and Gradient Boosting for demand forecasting (10 hours).
        - Challenges: Gradient Boosting required careful tuning of learning rates and tree depths.
    - **Evaluation**:
        - Assessed model performance using Mean Squared Error (MSE) and $$R^2$$ scores (6 hours).
        - Challenges: Moderate accuracy metrics highlighted the need for richer datasets or external factors.
    - **Visualization**:
        - Developed interactive dashboards using Plotly to present demand trends across regions and categories (5 hours).
- **Total Time Spent**: 32 hours.
- **Challenges Faced**:
    - Capturing seasonal trends effectively in a simulated dataset.
    - Balancing overfitting in Gradient Boosting models.
- **Solutions Implemented**:
    - Incorporated temporal variables to improve model accuracy.
    - Used interactive visualizations for better stakeholder communication.

Summary of Individual Contributions

The project journal satisfies all requirements by providing comprehensive documentation of each team member's contributions:

1. Rahul Muggalla focused on real estate pricing predictions using advanced regression techniques.
2. Ayyappa Gorantla worked on binary classification tasks with an emphasis on balanced evaluation metrics.

3. Shubham Pandurang Kawade concentrated on demand forecasting in e-commerce with innovative feature engineering.