AI-Powered Bug Predictor — Final Report

Project Summary

This project applies machine learning techniques to simulate a bug prediction system. Due to the unavailability of the `pc1.csv` defect dataset, we use the breast cancer classification dataset to represent a binary classification problem and explore a structured AI development workflow.

Objectives

- Build a supervised ML model using a known dataset.
- Analyze the feature importance in classification.
- Evaluate model accuracy and visualize insights.

Methodology

1. **Data Preparation**:
   - Used `sklearn.datasets.load_breast_cancer`.
   - Extracted feature matrix and labels.
   - Performed train-test split (80:20 ratio).

2. **Model Training**:
   - Used `RandomForestClassifier` for robustness and feature importance capability.
   - Trained model and calculated accuracy.

3. **Feature Importance Analysis**:

- Extracted top 10 most important features.

- Visualized using a horizontal bar chart.

4. **Output**:

- Achieved over 95% accuracy (sample output).

- Displayed feature importance of predictors like `mean concave points`, `worst concave points`, etc.

Results

The visualization shows features most influential in predicting outcomes. This is useful in software defect analysis, where understanding the root cause is as important as prediction.

![Feature Importance Output](./0ce83521-5229-43e5-ab87-c2792360b156.png)

Challenges

- The `pc1.csv` dataset could not be accessed (404 error).

- Used a substitute dataset for demonstration purposes.

- Visualization and logic still follow best practices for bug prediction systems.

Recommendations

- Replace mock data with a real defect dataset (e.g., NASA PROMISE).

- Incorporate explainability tools (e.g., SHAP).

- Add automation and deployment capability.

Conclusion

The project successfully simulates a bug prediction pipeline. It demonstrates how ML models can analyze structured data to make accurate predictions and explain why specific features are important.

Tools Used

- Google Colab
- Scikit-learn
- Pandas & Matplotlib
- Breast Cancer Dataset

By: Caren Rayon