Part 3: Ethical Reflection

Prompt:

Your predictive model from Task 3 is deployed in a company. Discuss:

- Potential biases in the dataset (e.g., underrepresented teams).
- How fairness tools like IBM AI Fairness 360 could address these biases.

1. Potential Biases in the Dataset

Although this model uses a medical dataset (breast cancer data), in a **real-world company deployment context**, similar predictive models might suffer from biases due to:

- **Demographic Skew**: If the dataset is not balanced across gender, ethnicity, age groups, or geographical regions, the model might be more accurate for overrepresented groups and inaccurate for others.
- **Sampling Bias**: The dataset might predominantly contain data from patients in specific hospitals or regions, reducing generalizability.
- **Label Bias**: Human decisions (e.g., labeling a case as “high priority”) might reflect subjective biases, which get encoded into the training data.
- **Feature Bias**: Some features might correlate with protected attributes (like race or gender), even unintentionally, leading to discriminatory predictions.

2. Using IBM AI Fairness 360 to Address Bias

IBM AI Fairness 360 (AIF360) is an open-source Python toolkit designed to detect and mitigate bias in machine learning models.

Key Uses in This Context:

- **Bias Detection**:
  - It can identify if the model performs differently for protected groups (e.g., one gender or age group).

  - Metrics like **disparate impact**, **equal opportunity difference**, and **statistical parity** help measure fairness.

- **Bias Mitigation**:
  - Offers preprocessing algorithms like **Reweighing** and **Disparate Impact Remover** to balance the training data.
  - In-processing algorithms (like **Adversarial Debiasing**) can be applied during training.
  - Post-processing techniques (e.g., **Equalized Odds**) adjust model outputs after training.

- **Audit and Explainability**:
  - It provides dashboards and tools to **visualize fairness metrics** and explain which features contribute most to biased outcomes.