Group Task-3

ML Ethics & Bias Case Study

1. Objective

- To analyze real-world bias in deployed machine learning systems.
- To identify technical causes of algorithmic unfairness.

2. Case Study Selection

Chosen Case: Bias in Automated Hiring Systems

- AI-based resume screening tools used for recruitment automation.
- Models trained on historical hiring data.
- Deployed in large-scale enterprise recruitment pipelines.

3. Problem Definition

Task: Candidate ranking and shortlisting using supervised learning.

- Input: Resumes (textual, semi-structured data).
- Output: Candidate suitability score.
- Model Type: NLP classification or ranking model.

4. Observed Bias

- System favored candidates from historically dominant demographics.
- Penalized resumes containing gender-indicative keywords.

5. Root Causes of Bias

5.1 Data-Level Bias

- Historical hiring data reflects societal and organizational bias.
- Class imbalance across gender and educational backgrounds.

5.2 Feature-Level Bias

- Text embeddings capturing gender-correlated semantics.
- Keyword-based features reinforcing stereotypes.
- Loss functions optimized only for accuracy.

6. Ethical Implications

- Violation of fairness and equal opportunity principles.
- Discriminatory hiring outcomes.
- Legal risks (employment discrimination laws).

7. Bias Detection Techniques

- **Fairness Metrics**
  - Demographic parity difference.
  - Equal opportunity gap.
  - Stratified performance evaluation across groups.
- **Explainability Tools**
  - SHAP/LIME for feature attribution.
  - Bias surfacing via feature importance analysis.

---

8. Bias Mitigation Strategies

8.1 Pre-processing Methods

- Dataset rebalancing (oversampling minority groups).
- Removal of sensitive attributes and proxies.
- Fair representation learning.

8.2 In-processing Methods

- Fairness-aware loss functions.
- Adversarial debiasing networks.
- Regularization for demographic parity.

9. Governance and Ethical Guidelines

- Implement fairness audits before deployment.
- Maintain model documentation (Model Cards).
- Ensure human-in-the-loop decision making.

10. Technical Best Practices

- Use diverse and representative datasets.
- Apply fairness-aware evaluation pipelines.
- Maintain transparency in feature engineering.

11. Trade-offs and Challenges

- Fairness vs accuracy trade-off.
- Difficulty defining universal fairness metrics.

12. Future Scope

- Development of standardized fairness benchmarks.
- Interpretable and inherently fair ML architectures.
- Integration of ethics into ML lifecycle (MLOps + Responsible AI).

---

13. Conclusion

- Bias in machine learning often originates from data and evaluation practices.
- Ethical AI requires both technical and organizational interventions.
- Fairness-aware modeling improves trust and societal acceptance.
- Responsible AI development must integrate fairness, transparency, and accountability as core design principles.