Title:- Adversarial ML Attacks: How Hackers Manipulate ML Models

Project Idea:

Adversarial Machine Learning studies how maliciously crafted inputs (adversarial examples) can deceive AI and machine learning models, causing misclassification or system failures. The project explores attack types, real-world implications, and defense strategies to strengthen AI robustness in critical sectors like autonomous vehicles, finance, healthcare, and cybersecurity.

Project Summary:

The project covers:

- Understanding attacks: White-box, black-box, gray-box models; evasion, poisoning, transfer, inference, and model extraction attacks.
- Real-world impact: Examples include Tesla Autopilot misreading stop signs, Microsoft Tay chatbot manipulation, and credit card fraud detection failures.
- Defense strategies: Adversarial training, defensive distillation, input preprocessing, robust architectures, and enterprise-level ML security practices.
- Threat modeling: Classifying adversaries by access and knowledge to guide defensive measures.

What Has Been Done and Working:

- Literature review: Documented major research milestones and attack methods.
- Threat model taxonomy: Defined white-box, black-box, and gray-box scenarios.
- Defense strategies implemented: Highlighted proactive, detection, and enterprise-level measures.
- Examples analyzed: Real-world case studies to show attack impact and practical mitigation techniques.

Conclusion:

Adversarial Machine Learning poses serious risks to AI integrity and security. Effective defense requires combining proactive model robustness, continuous monitoring, and enterprise-level security practices. Strengthening models against adversarial attacks ensures safe, reliable, and trustworthy AI deployment

Name: Shezonia Idrees
Rollno: BITF22M044
Subject: Information Security