1) Choose a Business Problem Business/System description (10-12 lines) Problem description (5 to 10 lines)

Business/System Description:

The insurance industry is critical for providing financial security against unforeseen losses. Insurance providers handle vast amounts of claims across various domains, including automotive, property, and personal injury. A significant challenge faced by this industry is fraudulent claims, which cost billions of dollars annually, impacting profitability and increasing premiums for honest customers. Efficient claim processing systems combine manual review and automated tools, but scalability and accuracy remain limitations. Leveraging advanced technologies like Machine Learning (ML), insurers aim to streamline the claim review process, reduce operational costs, and improve fraud detection accuracy. This not only safeguards insurers' revenues but also ensures fair premium pricing for policyholders.

Problem Description:

Insurance fraud is a persistent issue, often characterized by false information or exaggerated claims. Fraudulent activities disrupt operational efficiency and customer trust. Identifying fraudulent claims is challenging, as they often mimic legitimate cases. Relying solely on manual reviews is resource-intensive and prone to errors. With increasing claim volumes, there is a need for scalable, data-driven solutions to distinguish between fraudulent and legitimate claims effectively.

2) How Machine Learning Can Help Address the Problem:

Machine Learning (ML) offers robust and scalable solutions for addressing insurance fraud by automating fraud detection and enhancing accuracy.

1. Pattern Recognition:

ML algorithms can analyze vast amounts of historical claim data to identify subtle patterns and anomalies associated with fraudulent claims. These insights help flag suspicious claims for further investigation.

2.Real-Time Detection:

By integrating ML models into claim processing systems, insurance providers can evaluate claims in real time, drastically reducing the time required for manual reviews.

3. Reducing False Positives:

ML models can optimize fraud detection by distinguishing legitimate claims from fraudulent ones with high precision, minimizing disruptions for honest customers.

4. Feature Importance Analysis:

Algorithms like Random Forest and Gradient Boosting can identify the most significant factors contributing to fraud, enabling insurers to focus on critical data points for fraud prevention strategies.

5.Scalability:

ML solutions can handle the increasing volume of claims efficiently, providing scalability that manual methods cannot match.

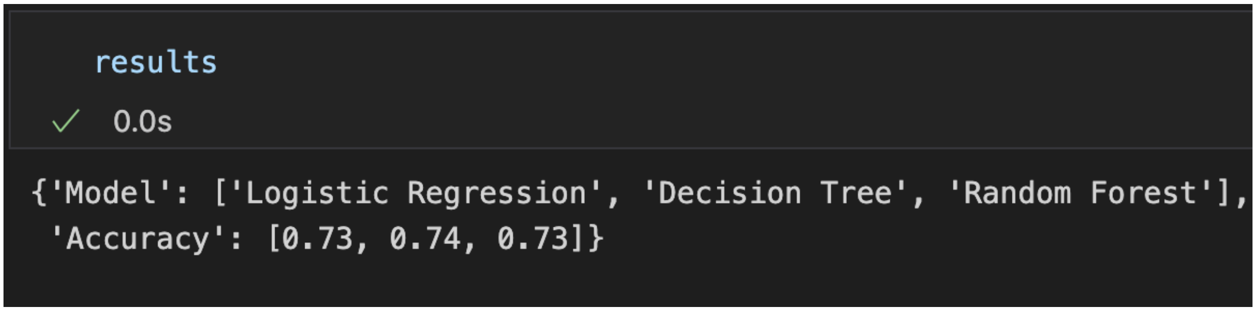
6. Cost Reduction:

By automating fraud detection, ML reduces the reliance on resource-intensive manual processes, cutting operational costs while maintaining accuracy.

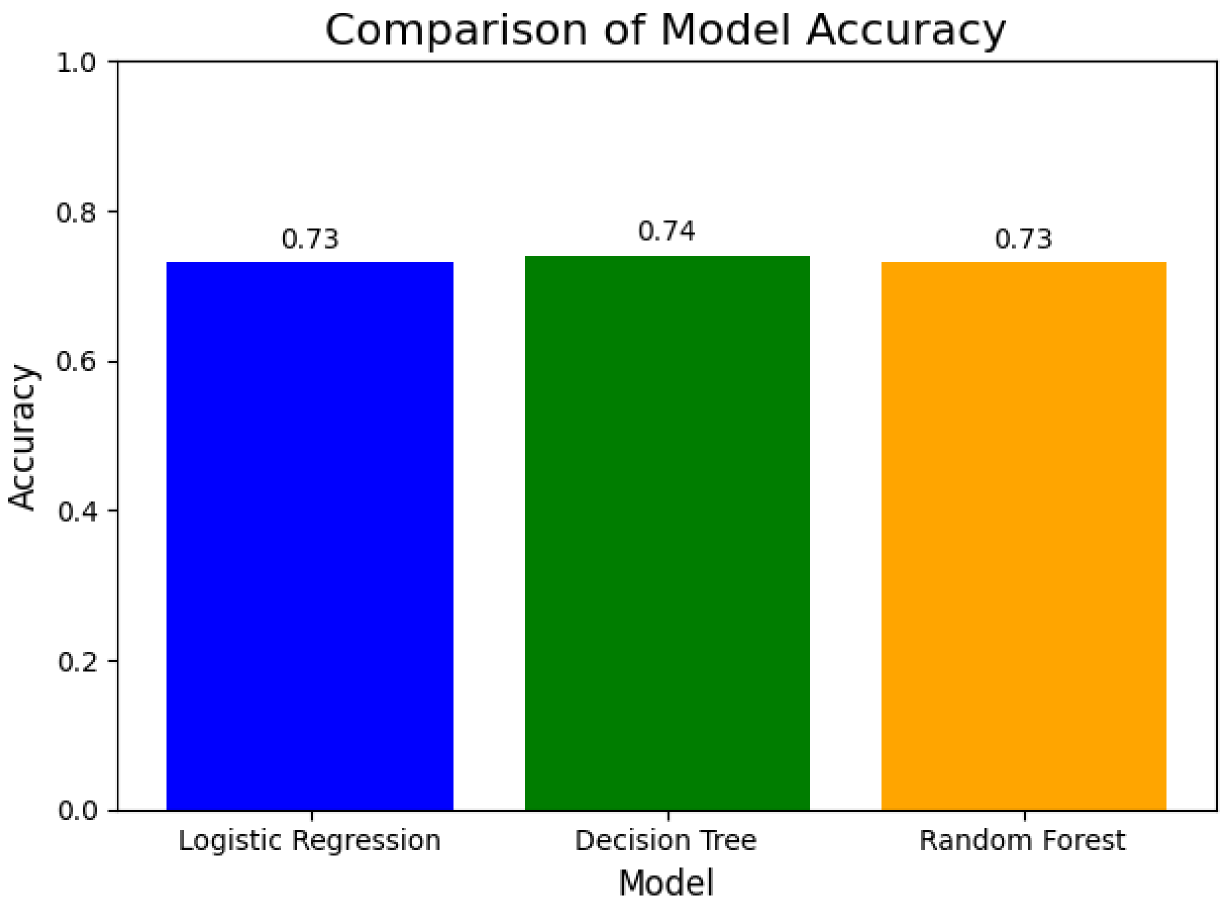
3) Data Description:

The dataset, "insurance_claims.csv ", is a detailed collection of insurance claim records, where each row represents an individual claim, and columns provide features describing the claimant, policy, and incident details. Key features include customer information ("age ", "insured_education_level"), policy specifics (e.g.,"policy_deductable ", "policy_annual_premium "), and incident attributes ("incident_type ", "incident_severity "). Financial data such as "total_claim_amount ", "injury_claim ", and "vehicle_claim " provide insights into the claim’s value. The target variable, "fraud_reported ", indicates whether a claim is fraudulent ("Y ") or legitimate ("N "). To ensure privacy, sensitive personal identifiers have been anonymized. This dataset, with its mix of numerical and categorical features, serves as a robust foundation for building machine learning models to detect fraudulent claims and improve operational efficiency in the insurance sector.

4) Compare Results of all Models Applied:



```
results
✓ 0.0s
{'Model': ['Logistic Regression', 'Decision Tree', 'Random Forest'],
 'Accuracy': [0.73, 0.74, 0.73]}
```



Analysis:

1. **Decision Tree** achieves the highest accuracy (0.74), though the improvement over the other models is marginal.
2. **Logistic Regression** and **Random Forest** both achieve 0.73 accuracy, performing similarly on this dataset.
3. The small difference in accuracy across models suggests that the dataset might require further feature engineering or hyperparameter tuning to exploit the strengths of more complex models like Random Forest.

5) How the Results Helped in Addressing the Problem:

The results from applying machine learning models like Logistic Regression, Decision Tree, and Random Forest provided valuable insights into addressing the issue of detecting fraudulent insurance claims. By achieving accuracies in the range of 73%–74%, the models demonstrated their ability to distinguish between fraudulent and legitimate claims with reasonable reliability.

1.Improved Fraud Detection:

The models identified patterns and relationships between features (`incident_severity`, `policy_deductable`, `property_damage`) and the likelihood of fraud. These insights can enhance decision-making processes in claim assessments.

2.Automation Potential:

With comparable accuracies, models such as Decision Tree offer explainability, making them practical for deployment in automated systems to flag suspicious claims for further review.

3.Operational Efficiency:

The results indicate that using machine learning can reduce reliance on manual reviews, saving time and resources. Claims can be processed more quickly, with fraudulent ones flagged for investigation, ensuring fairness for legitimate policyholders.