Machine Learning Assignment

PROJECT REPORT

TEAM ID: 4

PROJECT TITLE: COVID-19 Vaccine Adverse Event Risk Prediction

Name	SRN
C VISHWA	PES2UG23CS139
GAUTAM MENON	PES2UG23CS196

Problem Statement

The widespread administration of mRNA-based COVID-19 vaccines has been crucial in mitigating the pandemic's impact. However, certain individuals experience adverse events following vaccination, varying in type and severity. Identifying individuals at higher risk of such reactions can enhance post-vaccination safety monitoring and personalized medical guidance.

Objective / Aim

Predicting the Adverse Events Following Receipt of mRNA Based COVID 19 Vaccines.

Dataset Details

- **Source:** VAERS – Vaccine Adverse Event Reporting System
- **Size:** 8175 Samples, 17 Features
- **Key Features:** Dosage, Symptoms, Vaccination Site, etc.
- **Target Variable:** IS_SEVERE

Methodology

2 Models used – Logistic Regression and Random Forest

A. Data Cleaning and Target Creation

1. **Target Definition:** We created the **IS_SEVERE** target variable by searching the symptom columns (SYMPTOM1 to SYMPTOM5) for key severe terms: 'DEATH', 'SEIZURE', 'ISCHAEMIC STROKE', etc. Any report containing these terms was labelled **1** (Severe). (Based on Model 1)
2. **Missing Data:** Missing values in selected features (VAX_DOSE_SERIES, VAX_ROUTE, VAX_SITE) were filled using the **mode** (most frequent value). (Based on Model 2)

B. Modelling Pipeline

1. **Splitting:** The cleaned data was split into **70% for training** and **30% for testing** (Based on Model 1) or **80% training** and **20% testing** (Based on Model 2). Crucially, **stratification** was used to maintain the rare event proportion in both sets.

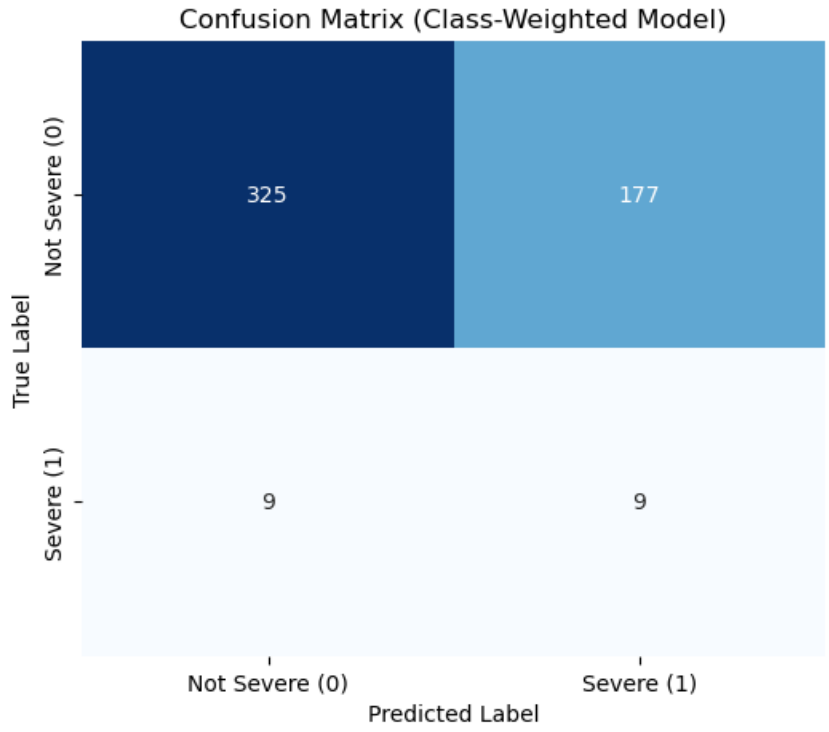
2. **Imbalance Handling:** Both models used the **class_weight='balanced'** parameter to tell the algorithm to pay more attention to the small 'Severe' class, preventing it from just predicting 'Non-Severe' all the time.

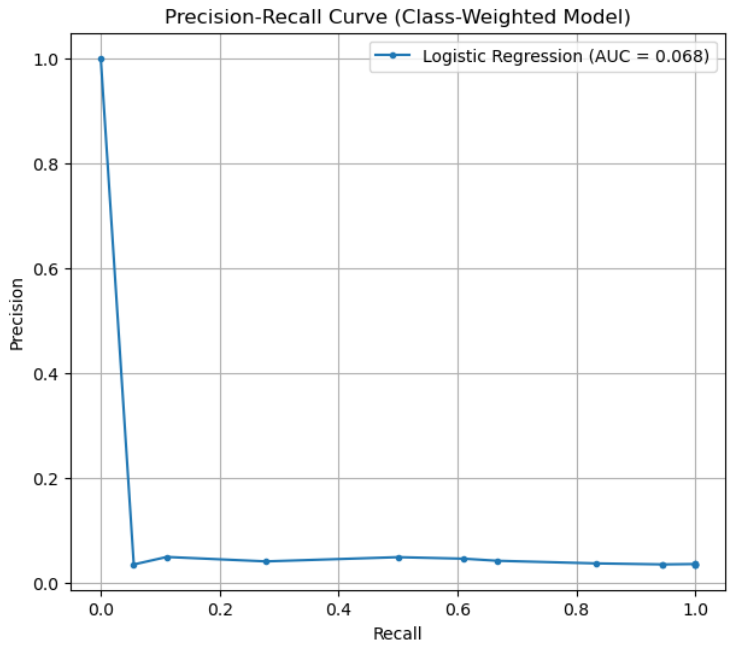
Results & Evaluation

Due to the imbalance, **Recall** (correctly identifying severe cases) and **Precision-Recall AUC (AP)** are the most critical metrics, as simple Accuracy is misleading.

Model 1: Logistic Regression

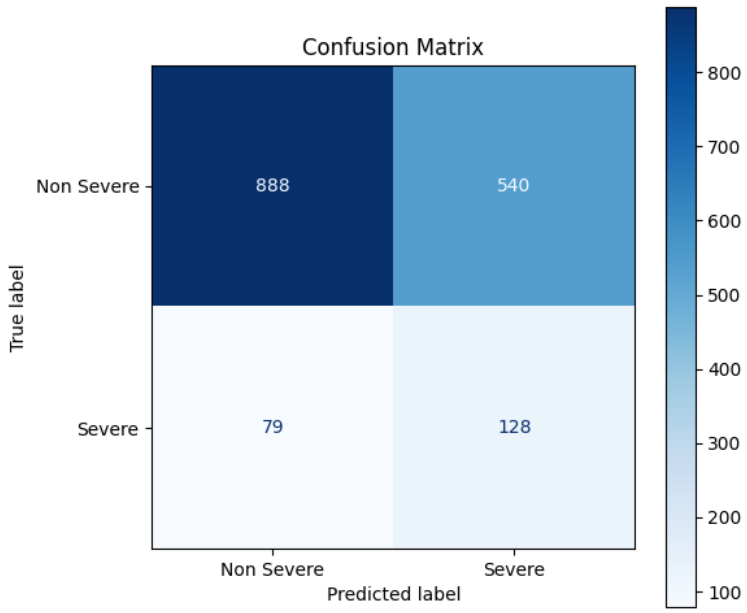
Metric	Score
Accuracy	0.6423
Recall (Severe)	**0.50**
Precision (Severe)	0.05
F1-score (Severe)	0.09
PR AUC	**0.134**

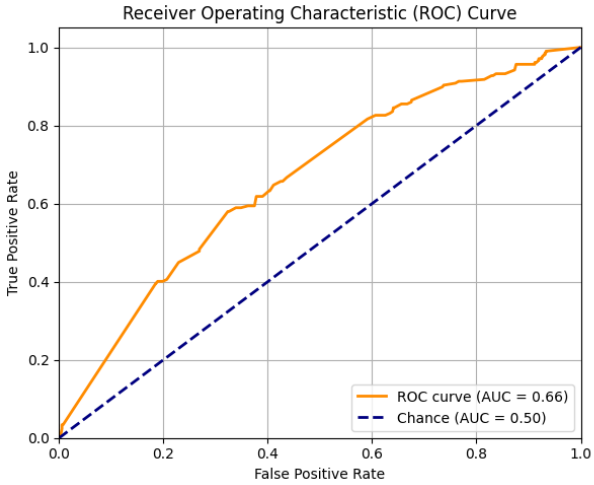


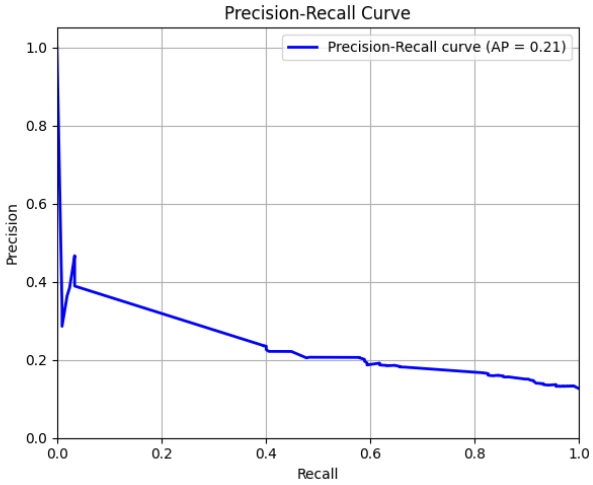


Model 2: Random Forest Classifier

Metric	Score
Accuracy	0.6214
Recall (Severe)	**0.62**
Precision (Severe)	0.19
F1-score (Severe)	0.29
PR AUC	**0.29**







Conclusion

The **Random Forest Classifier** is the better model for this task. It offers a superior balance of identifying the rare, important severe events (higher Recall) while being less noisy with its predictions (higher Precision and PR AUC) compared to the simpler Logistic Regression model.