My script

Logistic regression

“We trained a Logistic Regression model on the PCA-transformed dataset to predict claim occurrence.  
The model achieved an **AUC of 0.5697**, slightly better than random guessing. While overall accuracy was moderate

In terms of interpretability, the model offered clear insights:

* **PC12** had the strongest **negative effect** on claim likelihood (Estimate = –0.171).
* **PC13** had a strong **positive effect** (+0.127), and
* **PC11** and **PC14** also showed significant negative influence.

Components like **PC3**, **PC5**, and **PC15** were **not statistically significant**, meaning they didn’t meaningfully contribute to predictions.”

Decision tree

Next we trained a Decision Tree model and tested it to get the confusion matrix.  
Although the Decision Tree achieved **moderate recall**, its **AUC was 0.5317**, suggesting that it struggled to distinguish between the classes better than random chance.

The model primarily relied on **PC11 as the root node** and **PC12 as the decision**  to split the data. Most predictions of the minority class (claims) were made in a specific branch of the tree, which highlights the persistent **class imbalance**, even after applying SMOTE.

Random forest

“We trained Random Forest models with 10, 100, and 500 trees. As the number of trees increased, **recall and F1 score slightly improved**, helping detect more claim cases. and the **ROC AUC** rose from **0.5598** to **0.5747**, indicating slightly better class separation.

However, beyond 100 trees, AUC **plateaued**, showing **diminishing returns** with more trees.

However, the overall **accuracy remained high and stable**, around 80%. The model with **500 trees performed best** in terms of balancing precision and recall.”

Overall, Random Forest achieved the highest accuracy and precision but had the lowest recal which means it failes to identify a large amount of clamants,

Logistic Regression outperformed in recall and F1 score however its precision and accuracy were relatively low

Decision Tree performed slightly below logistic regression, with recall of 47.38% and F1 score of 12.22%, showing that it detected some claims but also produced many false positives.

the relatively low AUC values across all models (0.5317 to 0.5743) show that the models have difficulty distinguishing between claimants and non-claimants. This could be due to the class imbalance and/or possible information loss from PCA.

**🌲 How the Tree Makes Decisions:**

1. **First Decision (Root Node – PC11):**
   * The first rule checks: **Is PC11 ≥ -0.67?**
     + **Yes** → The model predicts **no claim (class 0)**, with **49% confidence**.
     + **No** → The model moves to a **second rule** based on PC12.
2. **Second Decision (PC12):**
   * The next check is: **Is PC12 ≥ 1.6?**
     + **Yes** → The model predicts **a claim (class 1)**, with **51% confidence**.
     + **No** → The model again predicts **no claim (class 0)**, with **49% confidence**.

**🔍 What This Means:**

* The tree is using **PC11 as the main split**, suggesting it's a key factor in separating claim vs. no-claim cases.
* The **claim prediction (class 1)** happens **only if**:
  + PC11 is quite **low** (less than -0.67), **and**
  + PC12 is **high** (greater than or equal to 1.6).
* This means **claims are mostly associated with a specific range of component values**, which might represent combinations of risk-related vehicle or policyholder features (though PCs are not directly interpretable).

**⚠️ Takeaway:**

* The decision tree’s logic is **simple and easy to follow**, but the **low confidence scores (~50%)** and **class imbalance** limit its performance.
* It's still useful for understanding **how certain feature combinations contribute to claims**.