**1. Understanding of the Problem**

**🔹 Objective**

To build predictive models that estimate the likelihood of a policyholder filing a car insurance claim within the next 6 months, using available policyholder and vehicle-related features.

**🔹 Assumptions Made**

* The training data is representative of future policyholders.
* PCA-transformed features retain relevant predictive information.
* Class imbalance exists and requires adjustment (handled via SMOTE).
* Missing values and outliers have been appropriately addressed.

### 🔹 ****Handling Missing Data (5 pts)****

No missing values were found in the dataset. This ensured all variables were complete and suitable for analysis without the need for imputation or removal of rows/columns.

### 🔹 ****Feature Understanding (5 pts)****

We examined the dataset’s structure, data types, and summary statistics. A new variable, volume, was created by multiplying length, width, and height to represent the size of a vehicle. The original length, width, and height columns were dropped to reduce redundancy and multicollinearity. policy\_id was also dropped, as it served only as a unique identifier and had no predictive value.

### 🔹 ****Exploratory Analysis (5 pts)****

Univariate analysis was used to understand the distribution of individual features. Bivariate analysis helped identify relationships between predictors and the target variable is\_claim. This confirmed class imbalance and provided early insight into which variables may influence claim likelihood.

### 🔹 ****Encoding/Transformation (5 pts)****

Categorical variables were one-hot encoded to convert them into numeric format suitable for modeling. Continuous variables were scaled using the min\_maax method to ensure uniform feature contribution, especially before applying PCA and distance-based models.

**Dimensionality Reduction Using PCA**

**PCA** was applied using the correlation matrix to reduce dimensionality and multicollinearity. The first 15 components, explaining about **79.62%** of total variance, were retained for modeling.

### Model Performance Comparison

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1 Score** | **AUC** |
| --- | --- | --- | --- | --- | --- |
| Random Forest (500 trees) | 0.8067 | 0.0866 | 0.2166 | 0.1209 | 0.5743 |
| Logistic Regression | 0.5850 | 0.0775 | 0.5117 | 0.1341 | 0.5697 |
| Decision Tree | 0.5726 | 0.0704 | 0.4738 | 0.1222 | 0.5317 |

| **Data Type** | **Description** | **Count** |  |
| --- | --- | --- | --- |
| Chr | Character (categorical variables) | 24 |  |
| Int | Integer (discrete numerical values) | 11 |  |
| Num | Numeric (continuous numerical values) | 5 |  |

| **Variable Type** | **Count** | **Percentage** |  |
| --- | --- | --- | --- |
| Character (Categorical) | 24 | 53.3% |  |
| Numeric (Integer/Continuous) | 21 | 46.7% |  |
| **Total** | **45** | **100%** |  |

Great follow-up!

The **0.05** value (or **5%**) is a **common threshold** used in statistics to test **significance**. Here’s why:

**🎯 Why p < 0.05?**

* The **p-value** tells us how likely it is that a result happened **just by chance**.
* A **p-value < 0.05** means there’s **less than a 5% chance** the result is random.
* So, we say the result is **statistically significant** — it likely reflects a real effect.

📌 **Why 5%?**  
It’s a widely accepted rule in statistics — a balance between being **too strict** and **too lenient**. It’s not magic, just a standard.

Great question!

The decision tree uses only **PC11** and **PC12** because of how decision trees **select features**:

**🌲 Why Only PC11 and PC12?**

1. **Tree chooses the best split** at each step — the feature that best separates the data into the target classes (claims or no claims).
2. In your case, **PC11 and PC12** had the **strongest splitting power** (i.e., they reduced uncertainty the most) among all 15 principal components.
3. Even though there are 15 PCs, the tree **only goes as deep as needed** to make decisions — and these two were enough for the splits it found useful.

**📌 Bonus Insight:**

* This doesn’t mean other PCs aren’t important — just that in this tree structure, **PC11 and PC12 gave the best immediate splits** for the training data.
* If you grow a deeper tree, or use Random Forest, you might see **more PCs** being used.

Root Node (Start)

* The tree first splits based on PC11:
* If PC11 ≥ –0.67, the model predicts class 0 (no claim) with 49% confidence.
* If PC11 < –0.67, it moves to the next decision.

Second Split (PC12)

* Now it checks PC12 ≥ 1.6:
  + If true → predicts class 1 (claim) with 55% confidence
  + If false → predicts class 0 with 35% confidence

| **Rank** | **Principal Component** | **Importance Score (MeanDecreaseGini)** | **Interpretation** |
| --- | --- | --- | --- |
| 1 | PC12 | 3217.80 | **Most influential** - Captures the strongest predictive patterns |
| 2 | PC13 | 3216.72 | Nearly equal importance to PC12 - Likely complementary patterns |
| 3 | PC11 | 3075.06 | Strong contributor, but ~5% less important than top PCs |
| 4 | PC9 | 2877.94 | Moderate importance - May represent secondary trends |
| 5 | PC10 | 2874.26 | Similar to PC9 - Consider analyzing together |

### 🔧 ****Feature Engineering and Preprocessing****

After exploratory data analysis, several preprocessing steps were applied to enhance data quality and prepare it for modeling.  
A new feature, **volume**, was created by multiplying **length × width × height** to represent the car's size as a single numeric metric. The original length, width, and height columns were dropped to reduce redundancy.  
Non-informative identifiers like **policy\_id** were also removed.  
Categorical variables were **one-hot encoded** to make them machine-readable, and **binary columns** were converted into 1 and 0 for consistency.  
All numeric features were **scaled using Min-Max normalization** to ensure equal weighting, especially important for PCA and distance-based algorithms.  
The **max\_torque** and **max\_power** text columns were parsed to extract their numeric values (e.g., torque in Nm and power in bhp) to make them usable in modeling.

### ****3. Policyholder Age vs Claim (Rplot31.jpeg)****

* **Pattern**: U-shaped relationship – highest claims for youngest (<0.4) and oldest (>0.8) policyholders.
* **Why It Matters**: Target risk education programs to young/senior drivers.

### ****4. Car Age vs Claim (Rplot30.jpeg)****

* **Pattern**: Strong inverse relationship – newer cars (0.0–0.25) file significantly more claims.
* **Business Impact**: Adjust depreciation curves or offer new-car diagnostics.

### ****1. Claim Rate by Steering Type (Rplot36.jpeg)****

* **Key Pattern**:
  + **Power steering** shows the highest claim rate (~0.06)
  + **Manual** and **Electric** are significantly lower (~0.02-0.03)
* **Business Implication**:  
  Power steering systems may correlate with riskier driving behavior or vehicle types. Investigate whether this reflects specific car models.

### ****2. Claim Rate by Segment (Rplot35.jpeg)****

* **Key Pattern**:
  + Clear variation across segments (specific categories not labeled)
  + One segment stands out with ~2x higher claim rate than others

### ****1. Policyholder Age Distribution (Rplot28.jpeg)****

![Policyholder Age Histogram]

* **Key Pattern**:
  + Right-skewed distribution with peak at ~0.6 (normalized age)
  + Few very young (<0.4) or senior (>0.8) policyholders
* **Risk Insight**:
  + Younger drivers (left tail) are typically higher risk but underrepresented

### ****2. Car Age Distribution (Rplot27.jpeg)****

![Car Age Histogram]

* **Key Pattern**:
  + Sharp peak at newest cars (~0.0-0.25)
  + Rapid drop-off for older vehicles
* **Risk Insight**:
  + New cars have higher claim probability but dominate your portfolio

### ****3. Policy Tenure Distribution (Rplot26.jpeg)****

![Policy Tenure Histogram]

* **Key Pattern**:
  + Uniform distribution across tenure lengths
  + Slight dip at mid-tenure (~0.5)
* **Risk Insight**:
  + No natural customer retention point visible