**Columns in the Dataset:**

1. **months\_as\_customer**:
   * **Description**: The number of months the customer has been with the insurance provider.
   * **Use for AI**: This is important as **longer-term customers** might be less likely to commit fraud compared to newly signed customers. **Short-term customers** with sudden claims may trigger fraud flags.
2. **age**:
   * **Description**: The age of the customer.
   * **Use for AI**: Age can be a factor in determining risk. Younger and older customers might have different risk profiles. For example, younger customers might be more likely to make high-value claims, and older customers may have more medical claims.
3. **policy\_bind\_date**:
   * **Description**: The date when the insurance policy was bound.
   * **Use for AI**: This date is useful for **time-based fraud detection**, particularly when combined with the **time\_difference** column. Claims made **soon after policy binding** might be flagged as suspicious.
4. **time\_difference**:
   * **Description**: The time difference (in days) between the claim submission and policy binding.
   * **Use for AI**: This can be used to detect **suspicious timing** of claims. Claims filed **very shortly after the policy is active** (e.g., within a few days or weeks) might be fraudulent.
5. **fraud\_reported**:
   * **Description**: Indicates whether fraud was reported for the claim (yes or no).
   * **Use for AI**: This is the **target variable** for the fraud detection AI model. The AI will learn from historical data with **fraud reported** and **no fraud** to predict future fraud.
6. **claim\_type**:
   * **Description**: The type of claim (e.g., vehicle claim, property claim, injury claim).
   * **Use for AI**: Different claim types might have different patterns of fraud. For instance, **vehicle claims** could be fraudulent if the damage is inconsistent with the policy terms, while **property claims** might show patterns in location or claim amount.
7. **total\_claim\_amount**:
   * **Description**: The total amount being claimed.
   * **Use for AI**: A **higher claim amount** can indicate potential fraud, especially if it’s disproportionate to the customer’s history. The AI can learn thresholds for what’s considered **normal** for each claim type.
8. **insured\_sex**:
   * **Description**: The sex of the insured.
   * **Use for AI**: While sex might not directly indicate fraud, it can be used in combination with other features (such as age and claim amount) to see if there are patterns in **fraudulent claims** by gender.
9. **insured\_education\_level**:
   * **Description**: The education level of the insured.
   * **Use for AI**: This could provide insights into the customer’s background, potentially helping identify any patterns in fraud by education level. For example, it could be used in conjunction with other factors to detect if more **educated** individuals are less likely to commit fraud.
10. **policy\_annual\_premium**:
    * **Description**: The annual premium paid for the policy.
    * **Use for AI**: **High premiums** could indicate high-value claims, which may be more likely to be fraudulent. The AI can use the premium value to determine if a claim is disproportionately high compared to the premium paid.
11. **umbrella\_limit**:
    * **Description**: The umbrella limit associated with the policy.
    * **Use for AI**: A higher **umbrella limit** may allow for larger claims, and the AI can learn patterns based on the size of claims compared to the umbrella limit. **Excessive claims** relative to the umbrella limit can trigger fraud detection.
12. **vehicle\_claim**:
    * **Description**: The amount claimed for vehicle-related claims (if applicable).
    * **Use for AI**: This is specific to **vehicle-related claims**. The AI can detect patterns in vehicle claims, such as whether certain **vehicle types** are more prone to fraudulent claims. It can also track the **frequency and amount** of vehicle-related claims to detect suspicious patterns.
13. **time\_since\_policy\_activation**:
    * **Description**: The number of months since the insurance policy was activated.
    * **Use for AI**: **Newer policies** could be more likely to have **fraudulent claims**. A claim filed **immediately after activation** could be flagged as suspicious. This column helps the model detect early claims, which may indicate fraud.
14. **claim\_frequency**:
    * **Description**: The number of claims made in the last 12 months.
    * **Use for AI**: High **claim frequency** is often correlated with fraud. Customers who file multiple claims in a short period are more likely to be flagged as fraudulent.
15. **claim\_size\_ratio**:
    * **Description**: The ratio of the **claim amount** to the **annual premium**.
    * **Use for AI**: **Large claims relative to premium** could indicate fraudulent activity, especially if the ratio is unusually high compared to typical claims for that policy type.
16. **claim\_type\_frequency**:
    * **Description**: The number of times a certain **claim type** has been filed.
    * **Use for AI**: Repetitive claims for the same **type of damage** might indicate fraud. The AI can learn which types of claims are **more frequent** and might flag claims that fall outside the normal range.
17. **average\_claim\_amount\_last\_12\_months**:
    * **Description**: The average claim amount over the last 12 months.
    * **Use for AI**: If a claim is much higher than the **average claim amount**, it could be flagged as suspicious.
18. **claim\_flag**:
    * **Description**: Indicates whether the claim is a **large claim** (1 = large, 0 = not large).
    * **Use for AI**: Large claims are often **more prone to fraud**, so the AI can use this flag to differentiate between **normal claims** and **potentially fraudulent large claims**.

**How These Columns Will Be Used for AI:**

* **Features for Predictive Modeling**:
  + Columns like **claim\_type**, **total\_claim\_amount**, **claim\_frequency**, and **claim\_size\_ratio** will help the AI model **detect patterns** of fraud by understanding **claim behaviors**, **fraudulent trends**, and **risk factors**.
* **Target Variable**:
  + **fraud\_reported** will be the **target variable** for the machine learning model (the label the AI will try to predict). It will be trained to predict whether a claim is **fraudulent (yes)** or **non-fraudulent (no)** based on the other columns.
* **Pattern Detection**:
  + The AI will look for patterns in **frequent claim submissions**, **large claims**, **short policy durations**, and **high claim amounts** to differentiate between **genuine claims** and **fraudulent claims**.
* **Feature Engineering**:
  + The **claim\_flag** column will provide useful information on claims that are **outliers** or unusually large. The model can use this flag to detect **disproportionate claims**.
  + The **time\_since\_policy\_activation** and **time\_difference** columns will help in detecting **suspicious claims**, particularly those filed soon after the policy starts.