**Data Leakage Prevention (DLP) : Develop AI models to detect and prevent unauthorized access and transfer of sensitive data**

**Employee Salary Analysis**

**SHANTHRAJ K 1, 1RVU23BSC132, B.Sc (H)**

**HARSHINI R 2, 1RVU23BSC053, B.Sc (H)**

**Problem Statement:**

To analyze the factors influencing employee salaries and provide insights for

better workforce management.

**Introduction:**

This analysis aims to explore the key factors that impact employee salaries.

By understanding these factors, organizations can make informed decisions to manage

their workforce more effectively. The study involves statistical analysis and data visualization

techniques to uncover trends and relationships within the dataset.

**Dataset:**

The dataset `insurance.csv` comprises information about insurance policies,

including their respective premiums. It consists of 1000 rows and 10 columns.

Each row represents an individual policyholder, and the columns include features such as

'age,' 'sex,' 'bmi,' 'children,' 'smoker,' 'region,' and 'charges.'

**Observations and Analysis:**

Statistical Analysis:

We performed statistical analysis on the dataset to understand the basic

characteristics of the data.

**Table 1. Statistical Analysis**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Feature | Mean | Median | Std Dev | Min | Max |
| Age | 39.2 | 39 | 14.1 | 18 | 64 |
| BMI | 30.66 | 30.4 | 6.39 | 15.96 | 53.13 |
| Children | 1.09 | 1 | 1.21 | 0 | 5 |
| Charges | 13270 | 9382 | 12110 | 1122 | 63770 |

**Key Findings:**

- The average age of policyholders is approximately 39 years.

- The average BMI of policyholders is around 30.66, indicating a tendency towards overweight.

- The median number of children covered by the insurance is 1.

- The average insurance charges are $13,270, with a wide range from $1,122 to $63,770.

**Conclusion:**

-Smoker Status: Policyholders who smoke have significantly higher insurance charges.

- Age: Older policyholders tend to incur higher charges.

- BMI and Children: These factors show some influence on charges

but are less significant compared to smoking status.

- Region: The region has minimal impact on insurance charges.

**Recommendations:**

For further analysis, it would be beneficial to explore the impact of other variables,

such as employment status or income level, on insurance charges. Additionally,

implementing machine learning models to predict charges based on these

features could provide deeper insights and more accurate predictions.

**References:**

1. Kaggel

2.Next DLP (https://www.nextdlp.com/resources/blog/how-machine-learning-improves-data-loss-prevention)

3.netskope (https://www.netskope.com/products/data-loss-prevention)