**Insurance data report- Hagay Ringel (May 16, 2023)**

In this project I used a dataset related to vehicle insurance from Kaggle.

This project aimed to assess whether we can predict vehicle damage based on the factors of previous insured and the age of the vehicle.

The methodology is preparing the dataset includes checking for nulls which we didn’t have, then, converting categorical values to numerical values: The factor “previous insured” is a binary variable which means there are only two answers: yes or no. The factor “vehicle age” is originally included the values of <1, 1-2, >2, but I later converted it to binary values: less than 2 years or more than 2 years.

Later, I added two visuals which are not necessarily needed to the aim of this project but give insights on our data.

The next step was preparing the dataset to our machine learning model using the algorithm Logistic Regression, for the following reasons:

1. LR is an algorithm that fits binary classification problems like this.
2. LR is simple to understand and works well with small datasets.
3. LR can provide insight into the importance of different features in determining the outcome.

First, we split the data to training set and testing set on the ratio of 70% training and 30% testing.

Second, we made predictions based on the testing set.

Third, producing the confusion matrix.

The last step was creating a classification report which is basically the summary of our model.

According to our model these are the results:

A screenshot of a computer

Description automatically generated

1. Precision- For class 0 (damage), the precision is 0.90, meaning that out of all instances predicted as class 0, 90% were actually correct. For class 1 (No damage), the precision is 0.95, indicating that out of all instances predicted as class 1, 95% were correct.
2. Sensitivity (Recall)- Measures the model's ability to correctly identify positive instances. For class 0 (damage), the recall is 0.95, suggesting that the model correctly identified 95% of the actual class 0 instances. For class 1 (No damage), the recall is 0.90, indicating that the model identified 90% of the actual class 1 instances.
3. F1-score is the mean of precision and recall, providing a balanced measure of the model's performance. In this case, both classes (damage and no damage) have an F1-score of 0.93.
4. Accuracy- The overall accuracy of the model is 0.93, which means that it correctly predicted the class labels for 93% of the instances in the dataset.

Note: This Project’s goal is to assess whether we can predict vehicle damage based on factors of vehicle age and previous insured. We still don’t know in what rate these factors predict damage to the vehicle, and still have other factors that may influence the model for example, age, gender etc.