U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

André Sousa (up202005277@fe.up.pt)
Gonçalo Pinto (up202004907@fe.up.pt)
Pedro Fonseca (up202008307@fe.up.pt)

Supervised Learning

AI Second Project

Specification of the work

- **Goal:** Analyze and study the process of classification (data preprocessing, feature selection, model selection, model training and model evaluation). In this case, the goal is to use data collected from past records of an insurance company to predict whether a given client will file a claim in the next 6 months.

- **Project Objectives:**
  - Given a data set, predict the target feature
  - Compare multiple machine learning models and feature selection to achieve the best results

Software used and related work

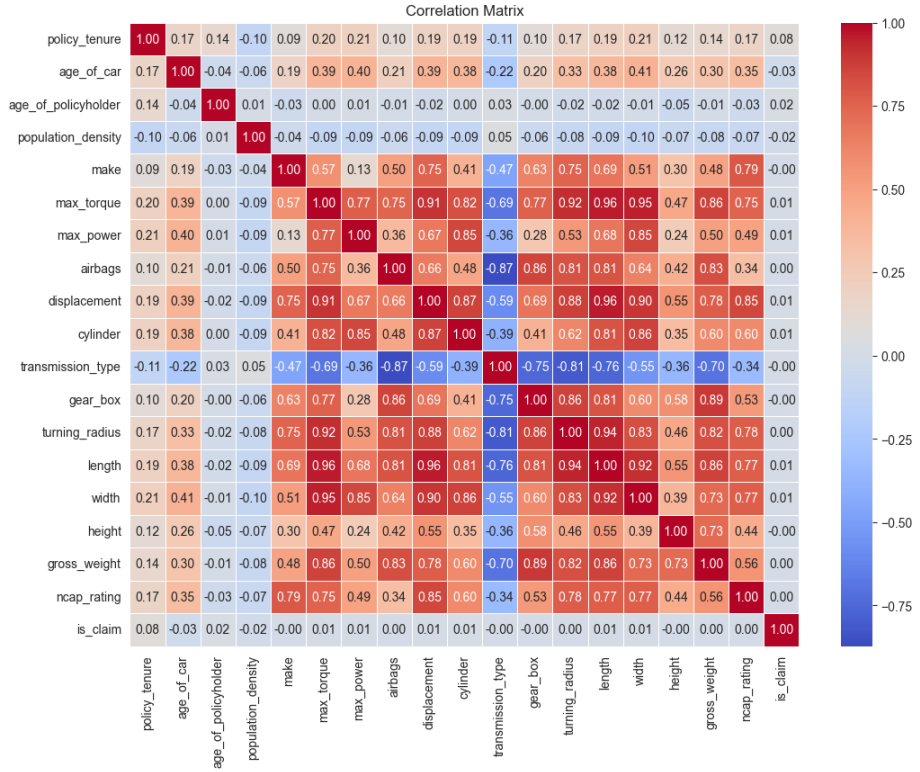
- Stuart Russell, Peter Norvig; Artificial intelligence. ISBN: 978-0-13-207148-2
- Richard S. Sutton; Reinforcement learning. ISBN: 978-0-262-03924-6
- Stuart Russel, Peter Norvig; Artificial Intelligence: A modern Approach.
- Exercise5 IART SupervisedLearning
- Other approaches to the problem

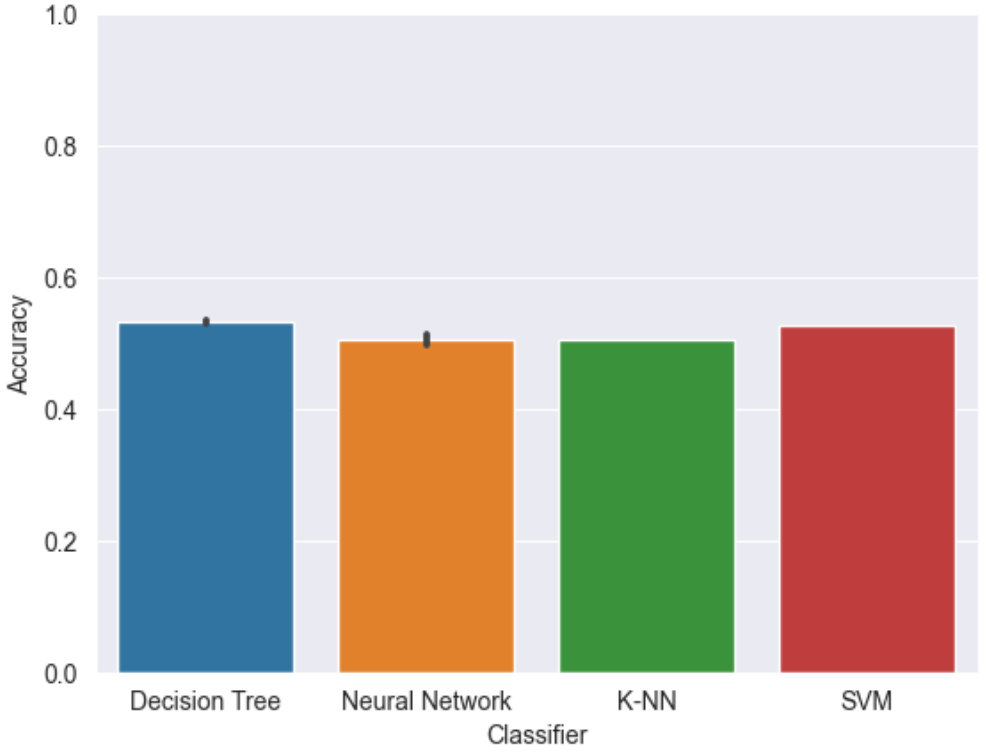
Tools and algorithms

- PyCharm Professional
- Python 3.9
- Git
- Python packages:
  - scikit-learn
  - Pandas
  - Seaborn
  - Matplotlib
- Machine Learning models:
  - Decision Tree
  - Neural Network (MLP)
  - k-nearest neighbors (KNN)
  - Support Vector Machine (SVM)

Implementation work already carried out

- Data preprocess
- Correlation analysis
- Feature selection (based on correlation results)
- Initial algorithm comparison





Data Preprocessing

1. Balancing the Dataset
   - Ensure equal representation of each target column value by creating a balanced dataset.

2. Transforming Nonnumeric Columns
   - Convert nonnumeric columns into numeric ones for compatibility with machine learning models.
   - Use techniques such as one-hot encoding, label encoding and string slicing.

3. Removing Redundancy from the Dataset
   - Identification and removal of columns with high correlation values to eliminate redundancy.
   - Test different correlation thresholds (0.70, 0.80, 0.90) to determine the optimal level of removal.
   - Creation of a dataset without columns that show no correlation with the target column.

4. Creating Combinations of Features
   - Introduction of a new column representing a numerical combination of three other columns.
   - Selection of columns with higher correlation to the target column for creating the combined feature.
   - Enhancement of the dataset with additional information for improved model performance.

5. Scaling and Adding Features
   - Application of feature scaling using the **StandardScaler** module to ensure consistent ranges for numeric features.
   - Usage of the **PolynomialFeatures** module to generate polynomial combinations of existing features.
   - Enhancement of the dataset by adding new features to capture non-linear relationships.

Developed models and comparison

We used five distinct machine learning models from scikit-learn for this task:

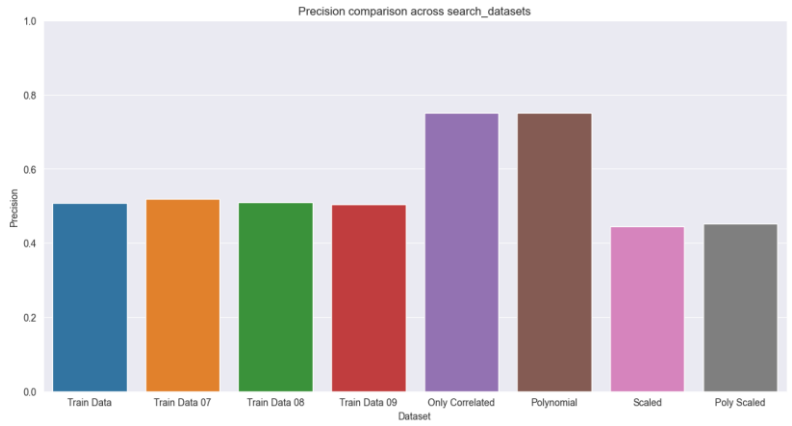
- Tree: DecisionTreeClassifier
- Neural Network: MLPClassifier
- Neighbours: KNeighborsClassifier
- Support Vector Machine: SVC
- Stochastic Gradient Descent : SGD

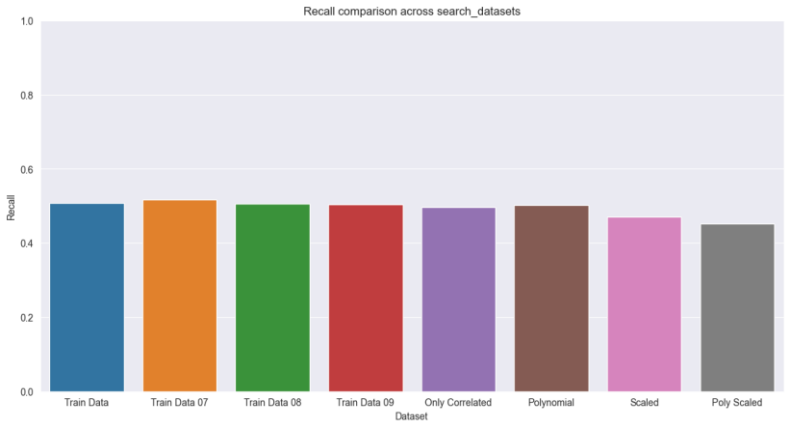
Regarding the division of the dataset into training and testing sets, we employed two different methods:

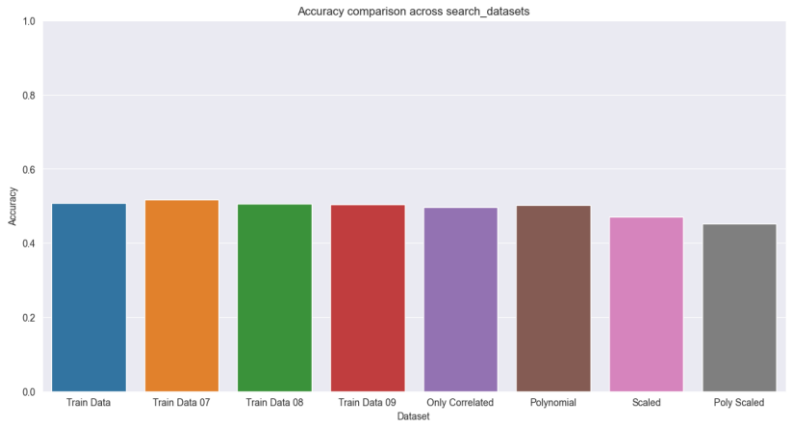
- Stratified K Folding
- Random Division

We applied all the models to each dataset and assessed their performance by employing various metrics such as confusion matrices, accuracy, recall, F1 measure, and precision. Additionally, we used the **GridSearchCV** module to determine the optimal parameters for each model and dataset.

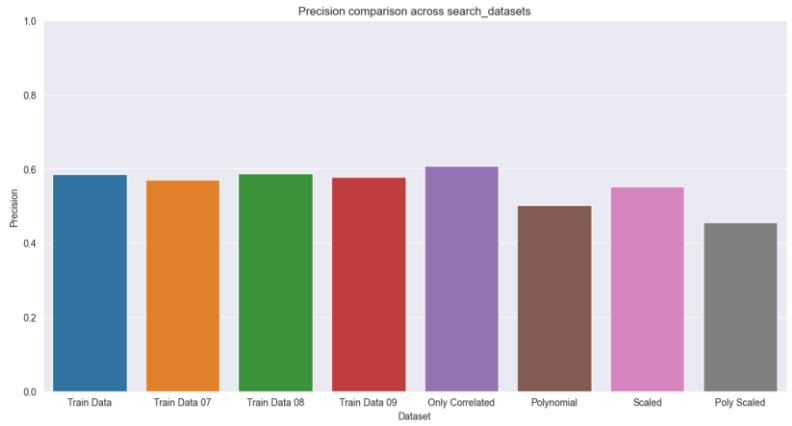
Neural Network

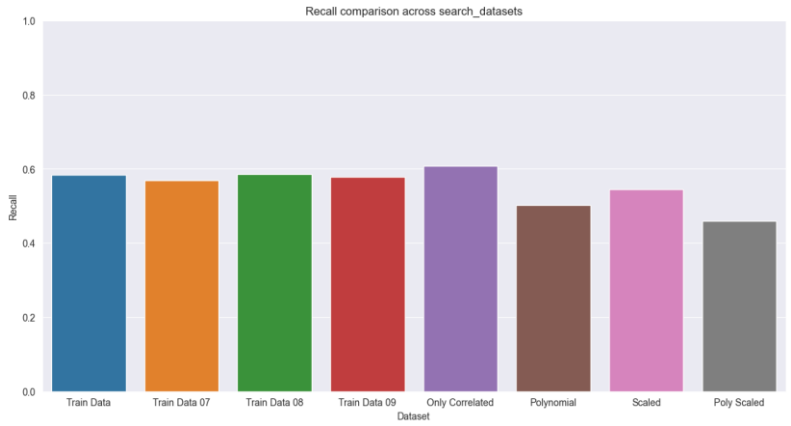


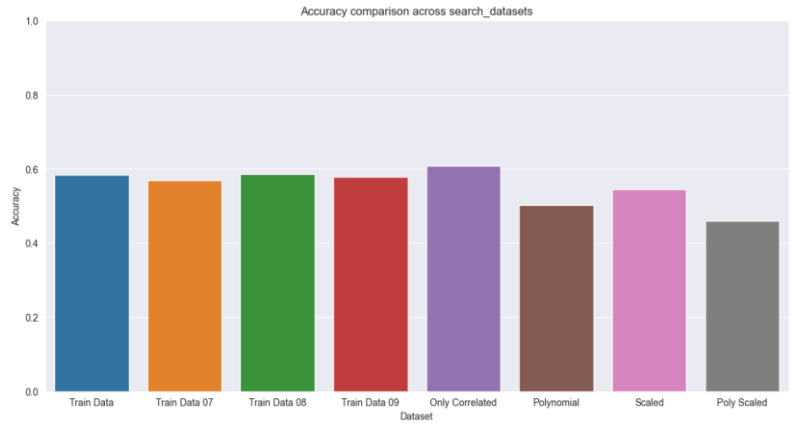




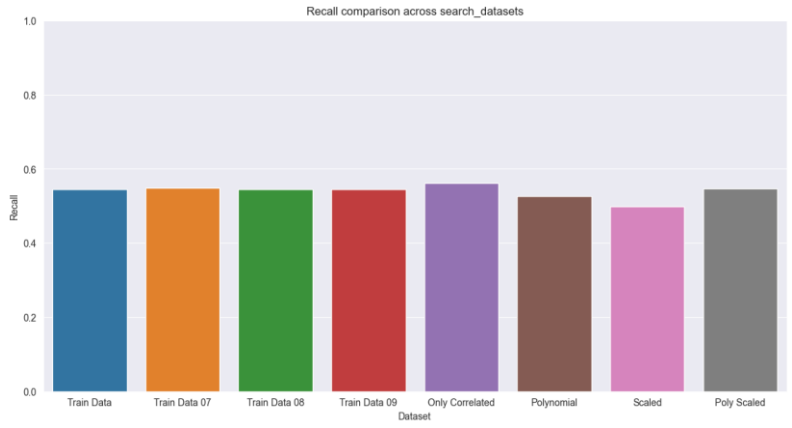
Decision Tree

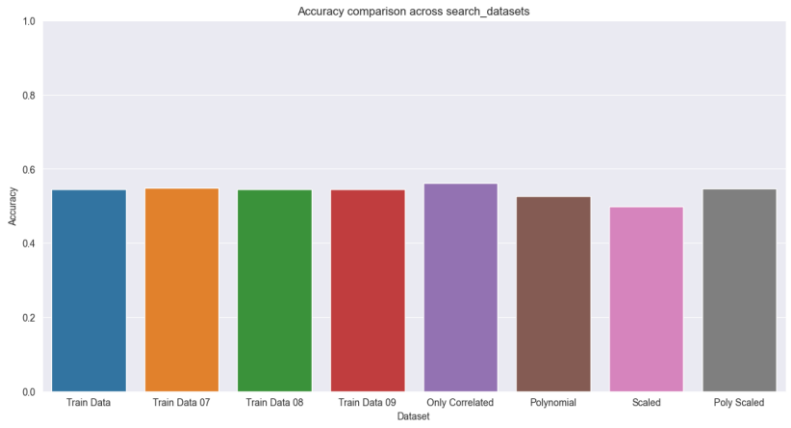


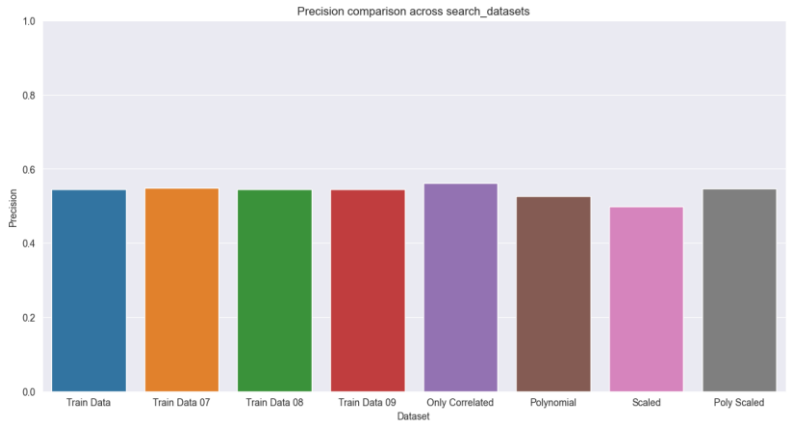




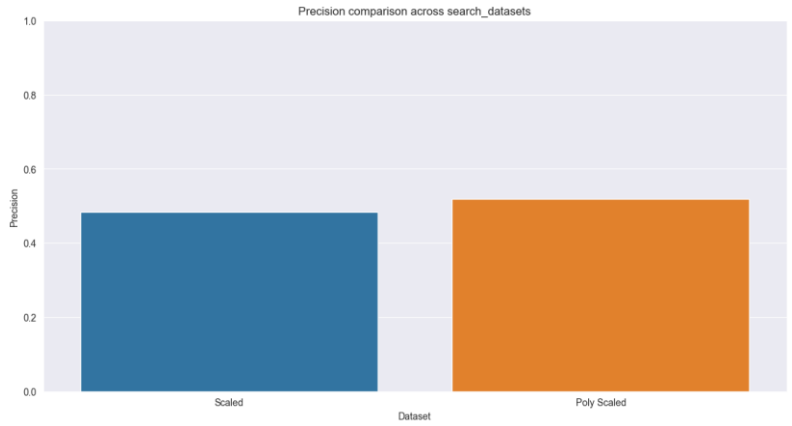
K-Nearest Neighbors

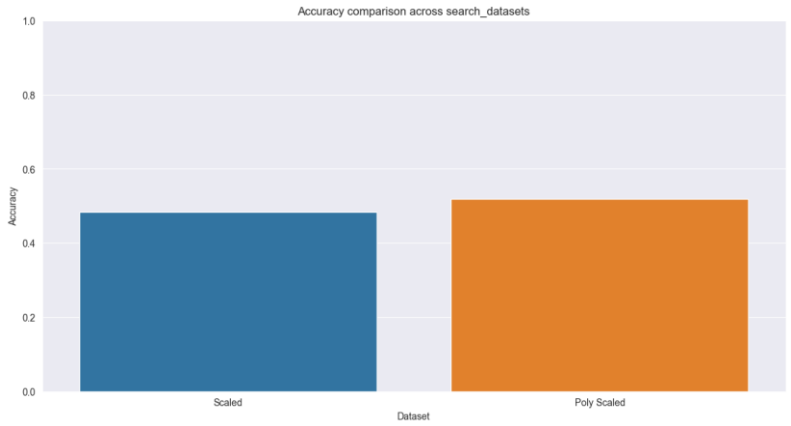


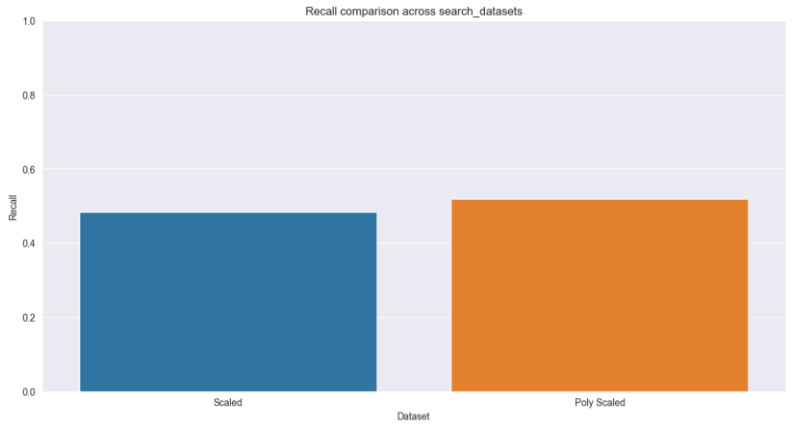




Support Vector Machine







Stochastic Gradient Descent