SUPERVISED LEARNING

TITANIC

AI - A2_55
01/05/2024

GONÇALO SANTOS
GONÇALO PINTO
RUI CARVALHO

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

PROBLEM DESCRIPTION

The goal of this project is to predict whether a passenger aboard the Titanic survived or not during the tragic sinking on April 15, 1912. This is a binary classification problem where the target variable is "Survived" (1 for survived, 0 for not survived).

Goal: The goal is to develop a machine learning model that accurately predicts whether a passenger survived or not during the Titanic disaster based on available attributes such as age, gender, and socio-economic class. The aim is to achieve a high level of prediction accuracy to understand the factors influencing survival outcomes.

TOOLS / ALGORITHMS

Algorithms

- Decision Tree Classifier
- Random Forest
- XGBoost
- SVM
- K-Nearest Neighbour

Libraries

- Pandas
- Seaborn
- Scikit-Learn
- NumPy
- Anaconda
- SciPy
- MatPlotLib
- XGBoost

DATASET

The dataset contains information about passengers, including socio-economic status, age, gender, number of siblings/spouses aboard, number of parents/children aboard, ticket details, fare, cabin, port of embarkation, and other related attributes.

	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
0	1	1	Allen, Miss. Elisabeth Walton	female	29.00	0	0	24160	211.3375	B5	S	2	NaN	St Louis, MO
1	1	1	Allison, Master. Hudson Trevor	male	0.92	1	2	113781	151.5500	C22 C26	S	11	NaN	Montreal, PQ / Chesterville, ON
2	1	0	Allison, Miss. Helen Loraine	female	2.00	1	2	113781	151.5500	C22 C26	S	NaN	NaN	Montreal, PQ / Chesterville, ON
3	1	0	Allison, Mr. Hudson Joshua Creighton	male	30.00	1	2	113781	151.5500	C22 C26	S	NaN	135.0	Montreal, PQ / Chesterville, ON
4	1	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.00	1	2	113781	151.5500	C22 C26	S	NaN	NaN	Montreal, PQ / Chesterville, ON

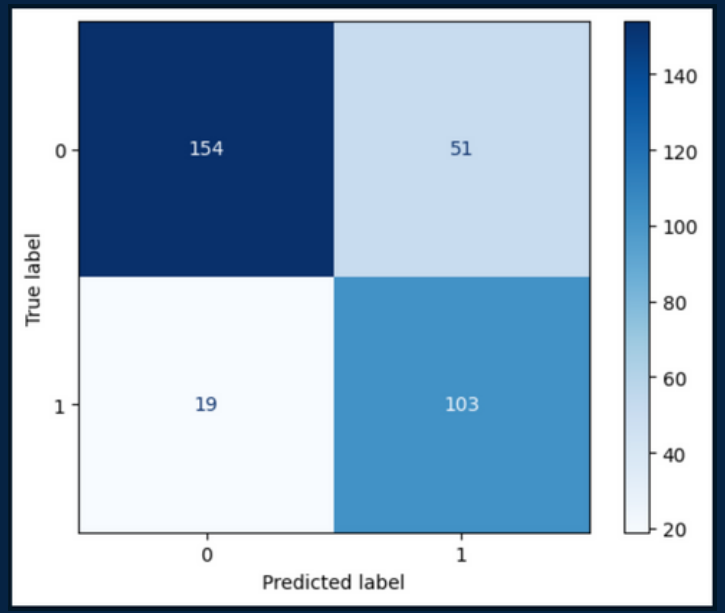
PRE-PROCESS DATA

After analyzing the dataset, we considered removing some columns from the original dataset once its values wouldn't make any significant difference to the results obtained. After removing those, we felt the need to reavaliate the data due to some missing values: in some columns, we decided to assign the mean value of the existing values to the missing ones; in other cases, we simply removed the rows.
Completed this stage, we started to work on a more useful dataset.

	pclass	survived	age	fare	co_pas	female	male	C	Q	S
0	1	1	29.00	211.3375	0	1	0	0	0	1
1	1	1	0.92	151.5500	3	0	1	0	0	1
2	1	0	2.00	151.5500	3	1	0	0	0	1
3	1	0	30.00	151.5500	3	0	1	0	0	1
4	1	0	25.00	151.5500	3	1	0	0	0	1

DECISION TREES

Uses a tree-like structure to represent decisions and possible consequences. To construct the tree, the algorithm seeks to divide the dataset into subsets. The result is a decision tree that can be used to classify new objects.



RANDOM FOREST

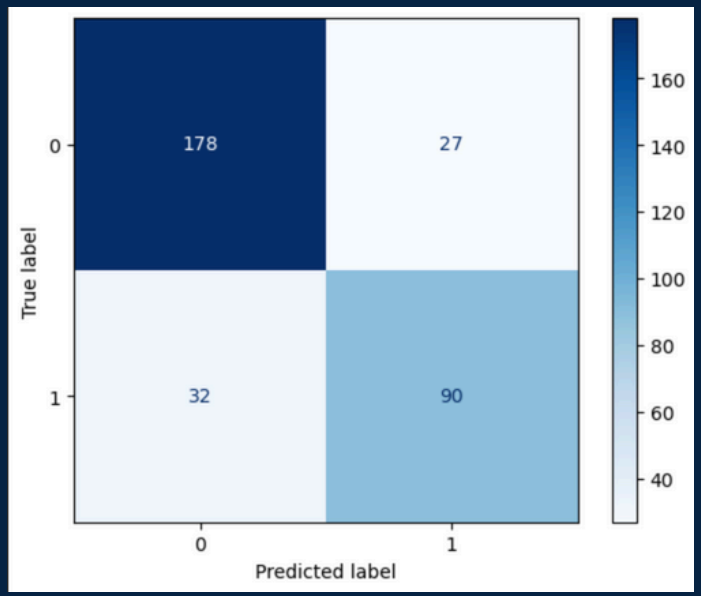
An ensemble learning method that constructs multiple decision trees during training. The final output is determined by aggregating the predictions of these trees, typically through averaging (regression) or majority voting (classification), to improve accuracy and control overfitting.

XGBOOST

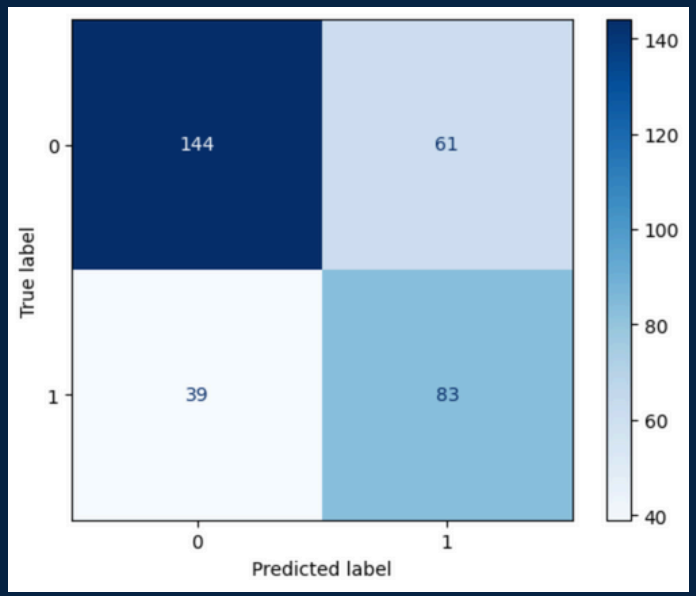
An optimized gradient boosting algorithm designed for speed and performance. It builds an ensemble of trees sequentially, where each new tree corrects the errors of the previous ones, enhancing the overall model accuracy.

S V M

A supervised learning algorithm used for classification and regression tasks. It works by finding the hyperplane that best separates the data into classes, maximizing the margin between different classes to achieve better classification results.



K - N N



Calculates the distance between a new data point and all existing data points, and the K nearest data points are used to determine the class of the new data point, based on the most common class among the K neighbors.

ALGORITHMS COMPARISON

Algorithm	Accuracy	Precision		Recall		F1-score	
		0	1	0	1	0	1
Decision Trees	79%	89%	67%	75%	84%	81%	75%
Random Forest	80%	85%	72%	82%	76%	84%	74%
XGBoost	81%	86%	73%	83%	77%	84%	75%
SVM	81%	85%	77%	87%	74%	86%	75%
KNN	82%	86%	71%	81%	77%	83%	74%

CONCLUSION

- To conclude we achieved similar values across all the algorithms used.
- Due to the weak dataset it was not possible to achieve better results of accuracy.

REFERENCES

- https://www.kaggle.com/datasets/sakshisatre/titanic-dataset
- IA slides and exercises