Supervised Machine Learning Models

Autism Dataset for Toddlers

Bernardo Campos - up202006056
Davide Teixeira - up202109860
Emanuel Maia - up202107486

Project Specification

Autistic Spectrum Disorder (ASD) is a neurodevelopmental condition with a lengthy and inefficient diagnosis, that results in significant healthcare costs and worse quality of life for the patients and those related to them.

The work to be developed is a classification problem, supported by machine learning models, with the goal of obtaining a fast and accurate ASD diagnosis on toddlers.

The dataset to be analyzed contains the answers to the *Q-Chat-10* behavioural questionnaire and its final score, alongside other characteristics that have proved to be useful in the diagnosis of ASD.

The questionnaire had as possible answers: “always”, “usually”, “sometimes”, “rarely” and “never”.
For questions A1-A9 “sometimes”, “rarely” and “never” were mapped to 1, others to 0.
For question A10 “always”, “usually”, “sometimes” were mapped to 1, others to 0.

Related Work

The dataset used in the project:
https://www.kaggle.com/datasets/vaishnavisirigiri/autism-dataset-for-toddlers

Projects using the same or similar datasets:
https://www.kaggle.com/code/biyawalavaibhav/asd-case-study
https://www.kaggle.com/code/vaishnavisirigiri/detection-of-autism-in-toddlers-using-ml
https://www.nature.com/articles/s41598-023-35910-1

Analysis Tools and Algorithms

- **Software**: Python (via Jupyter notebook) and Scikit-learn and TensorFlow packages

- **Target concept**: Presence/absence of ASD traits in toddlers

- **Data preprocessing**: Feature subset selection - Recursive Feature Elimination using SVM to determine feature importance

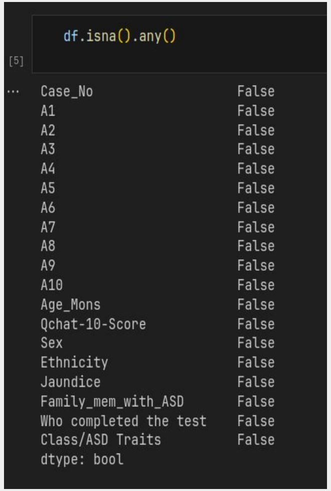
- **Classification algorithms:** Neural Networks (*TensorFlow/Keras*), K-Nearest Neighbours (*Scikit-learn/KNeighborsClassifier*), Random Forest (decision tree based method also using *Scikit-learn/GridSearchCV*), Naive Bayes

- **Note**: All of these algorithms were used in our project, despite some not being present in this slideshow.

Data Preprocessing

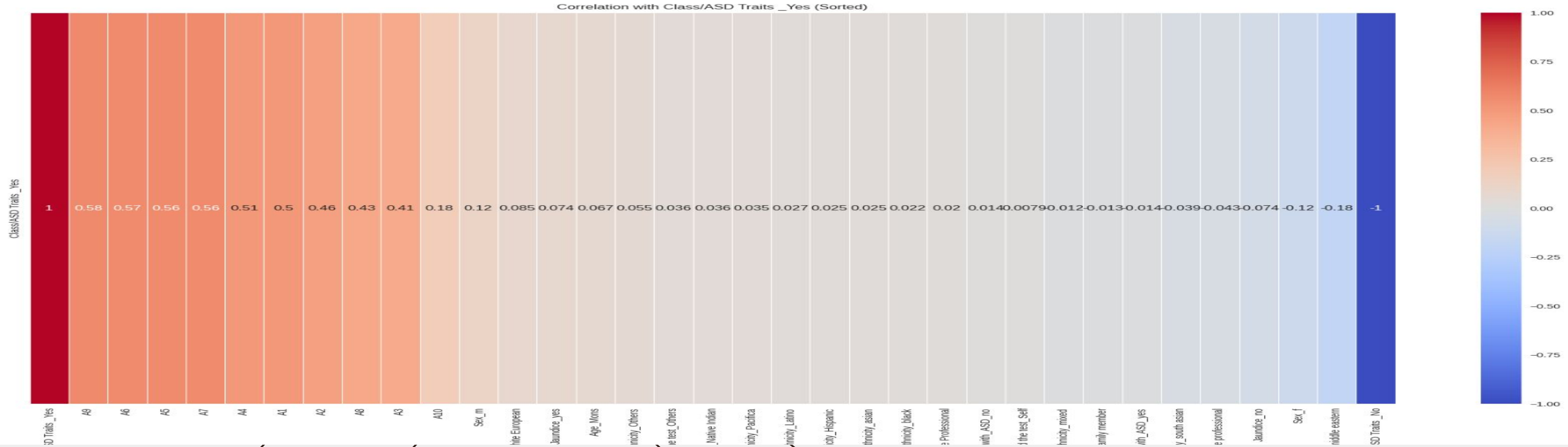
Filtering out outliers - We discarded the “Case_no” column since it is not relevant, and the “Qchat-10-Score” because it would serve as a direct indicator of an ASD diagnosis, rendering the other variables, and the purpose of this project useless.

Encoding the target variable/creating correlation matrix - It is important to encode the target variable so as to make it usable for our models.



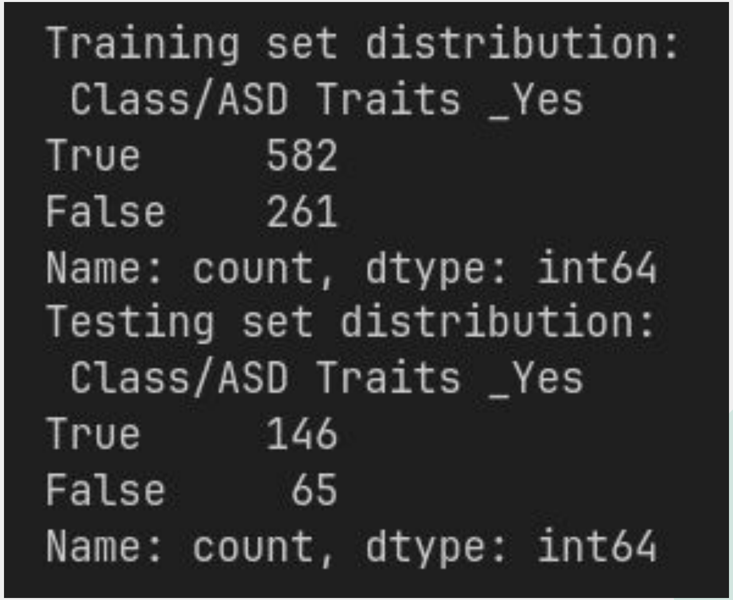
```
[5]   df.isna().any()

...  Case_No                     False
     A1                          False
     A2                          False
     A3                          False
     A4                          False
     A5                          False
     A6                          False
     A7                          False
     A8                          False
     A9                          False
     A10                         False
     Age_Mons                    False
     Qchat-10-Score              False
     Sex                         False
     Ethnicity                   False
     Jaundice                    False
     Family_mem_with_ASD         False
     Who completed the test      False
     Class/ASD Traits            False
     dtype: bool
```



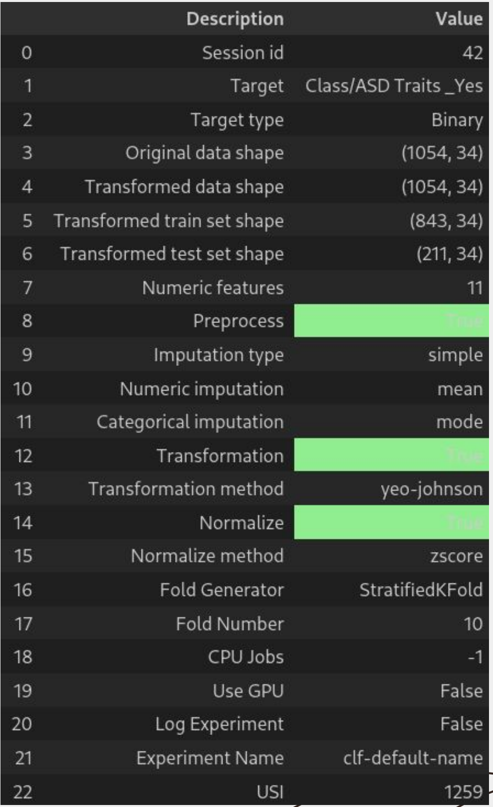
Data Preprocessing

Training/test set splitting - Our training test makes up for 80% of the total dataset. The remaining 20% are reserved for testing.

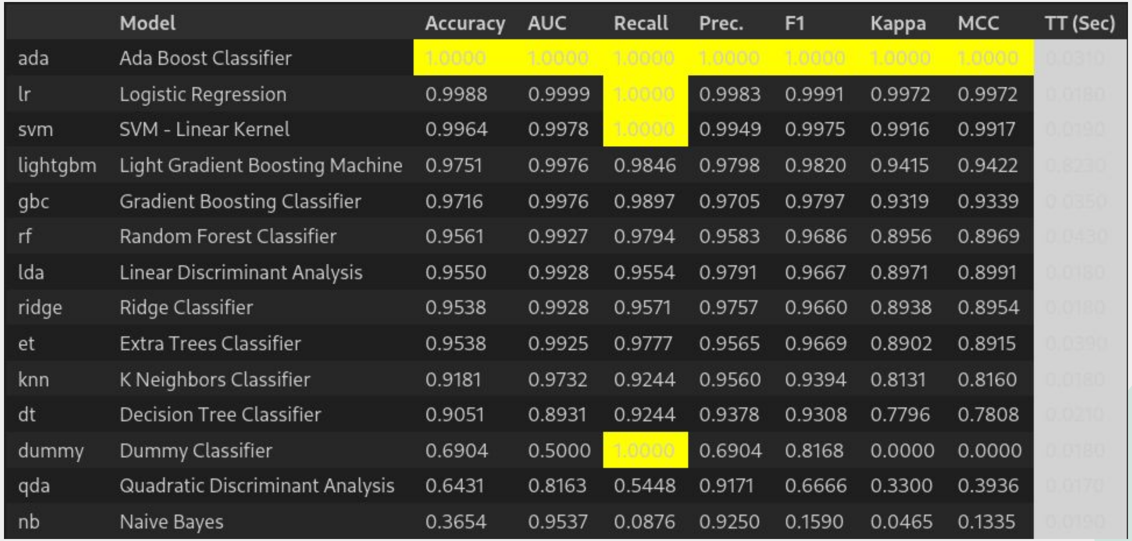


```
Training set distribution:
 Class/ASD Traits _Yes
True     582
False    261
Name: count, dtype: int64
Testing set distribution:
 Class/ASD Traits _Yes
True     146
False     65
Name: count, dtype: int64
```

Training Result Comparison



	Description	Value
0	Session id	42
1	Target	Class/ASD Traits _Yes
2	Target type	Binary
3	Original data shape	(1054, 34)
4	Transformed data shape	(1054, 34)
5	Transformed train set shape	(843, 34)
6	Transformed test set shape	(211, 34)
7	Numeric features	11
8	Preprocess	True
9	Imputation type	simple
10	Numeric imputation	mean
11	Categorical imputation	mode
12	Transformation	True
13	Transformation method	yeo-johnson
14	Normalize	True
15	Normalize method	zscore
16	Fold Generator	StratifiedKFold
17	Fold Number	10
18	CPU Jobs	-1
19	Use GPU	False
20	Log Experiment	False
21	Experiment Name	clf-default-name
22	USI	1259

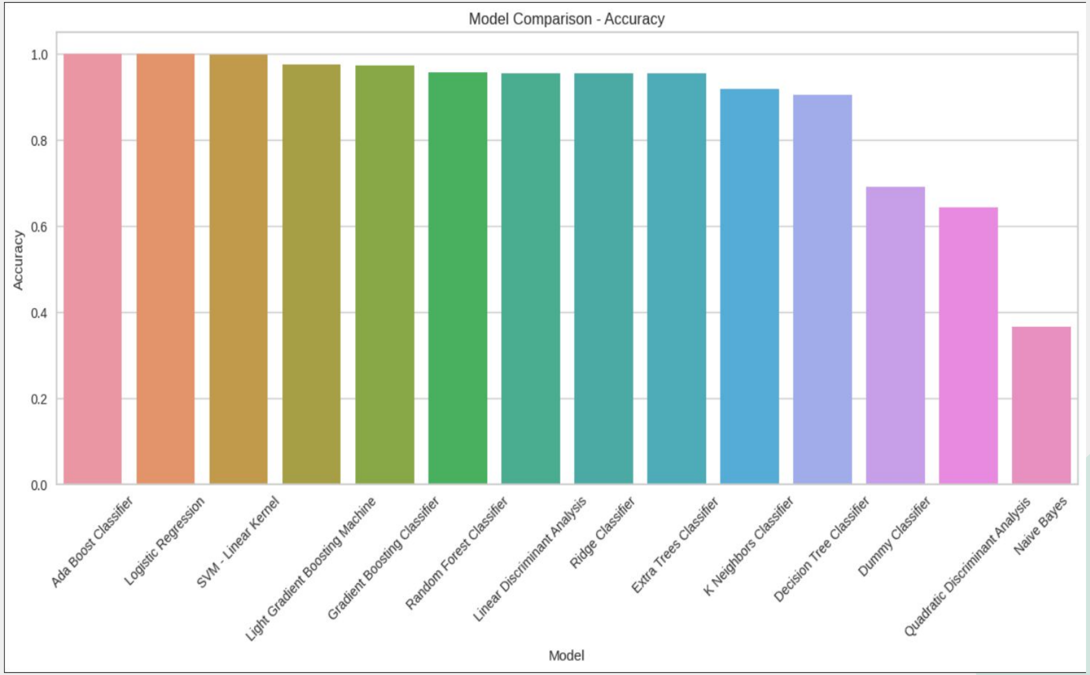


	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
ada	Ada Boost Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0310
lr	Logistic Regression	0.9988	0.9999	1.0000	0.9983	0.9991	0.9972	0.9972	0.0180
svm	SVM - Linear Kernel	0.9964	0.9978	1.0000	0.9949	0.9975	0.9916	0.9917	0.0190
lightgbm	Light Gradient Boosting Machine	0.9751	0.9976	0.9846	0.9798	0.9820	0.9415	0.9422	0.8230
gbc	Gradient Boosting Classifier	0.9716	0.9976	0.9897	0.9705	0.9797	0.9319	0.9339	0.0350
rf	Random Forest Classifier	0.9561	0.9927	0.9794	0.9583	0.9686	0.8956	0.8969	0.0430
lda	Linear Discriminant Analysis	0.9550	0.9928	0.9554	0.9791	0.9667	0.8971	0.8991	0.0180
ridge	Ridge Classifier	0.9538	0.9928	0.9571	0.9757	0.9660	0.8938	0.8954	0.0180
et	Extra Trees Classifier	0.9538	0.9925	0.9777	0.9565	0.9669	0.8902	0.8915	0.0390
knn	K Neighbors Classifier	0.9181	0.9732	0.9244	0.9560	0.9394	0.8131	0.8160	0.0180
dt	Decision Tree Classifier	0.9051	0.8931	0.9244	0.9378	0.9308	0.7796	0.7808	0.0210
dummy	Dummy Classifier	0.6904	0.5000	1.0000	0.6904	0.8168	0.0000	0.0000	0.0180
qda	Quadratic Discriminant Analysis	0.6431	0.8163	0.5448	0.9171	0.6666	0.3300	0.3936	0.0170
nb	Naive Bayes	0.3654	0.9537	0.0876	0.9250	0.1590	0.0465	0.1335	0.0190

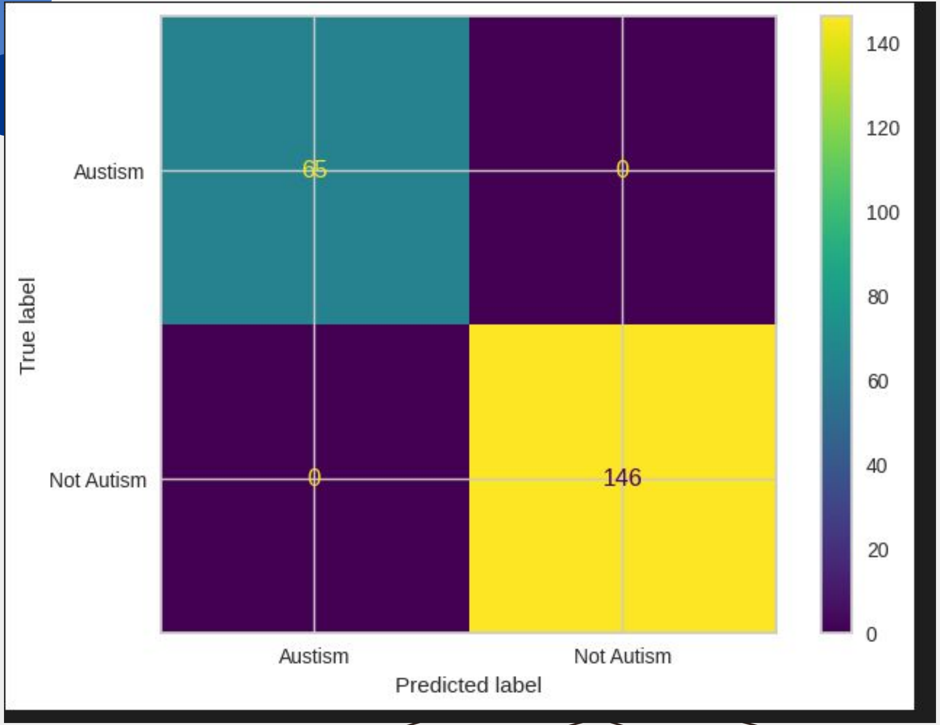
Training Results Comparison - Graph

Using this graph we can represent the accuracy of the models tested.

From it, we can infer that Ada Boost Clarifier, Logistic Regression and SVM - Linear kernel are the most accurate, with an accuracy of 100%.



Results Comparison



Using a confusion matrix, we can represent predicted and true labels.

From this information, we can assess that the algorithm was 100% accurate in predicting the presence/absence of autism in toddlers.

Results - Conclusion

When predicting the diagnosis of Autism Spectrum Disorder (ASD), the most crucial metric is recall. High recall ensures that actual ASD cases are correctly identified, allowing for timely intervention and support to address their challenges effectively.

Based on our results, if we were to choose one algorithm, we would select the AdaBoost classifier due to its 100% recall rate. This perfect recall means that it successfully identifies all true ASD cases, making it the best choice for our purposes.

Appendix - Questionnaire

- Q1 - Does your child look at you when you call his/her name?
- Q2 - How easy is it for you to get eye contact with your child?
- Q3 - Does your child point to indicate that s/he wants something? (e.g. a toy that is out of reach)
- Q4 - Does your child point to share interest with you? (e.g. pointing at an interesting sight)
- Q5 - Does your child pretend? (e.g. care for dolls, talk on a toy phone)
- Q6 - Does your child follow where you're looking?
- Q7 - If you or someone else in the family is visibly upset, does your child show signs of wanting to comfort them? (e.g. stroking hair, hugging them)
- Q8 - Would you describe your child's first words as unusual?
- Q9 - Does your child use simple gestures? (e.g. wave goodbye)
- Q10 - Does your child stare at nothing with no apparent purpose?