# Introduction

Autism or Autism Spectrum Disorder (ASD), is a neurological disease either in children or adult (Nurul Amirah Mashudi et al., 2021), which is also a wide range of condition characterized by challenges with speech, social skills, repetitive behavior and non-verbal communication(*Autism Spectrum Disorder (ASD) | Autism Speaks*, 2024). The signs can be detected as early as 1 year old *(Pierce et al., 2011)*, includes no response to name, poor skills in imitation, problems with eye contact and joint attention (*When Do Children Usually Show Symptoms of Autism?*, 2017).

Meanwhile, adult autism is way more complex than toddler as the symptoms can be present differently for adults of different genders. The common signs adult ASD includes difficulties in making friends, anxiety in social situations, problems in understanding what others are thinking or feeling, struggle in describing own feelings (NHS Choices, 2024). Women are often diagnosed later in life due to the development of compensatory strategies to mask their challenges, such as strong imitation skills to “camouflage” themselves into understanding social norms, therefore make their ASD traits undetectable in everyday interactions (*How Is Autism Different in Women?*, 2023).

In the data set, the score of each question in AQ-10 is recorded along with details of the person such as age, sex, and ethnicity. The Autism Spectrum Quotient (AQ) is a questionnaire designed for adults (16 years old and above) who do not have learning disability, AQ-10 is used as a screening tool to determine whether a person should be referred for an autism assessment *(Engelbrecht, 2020)*.

The models used for predicting autism/ASD involved Logistic Regression (LR), Random Forest (RF), K-Nearest Neighbour (KNN) and VotingClassifier in Ensemble Learning. The result from each model and their evaluation would be recorded as a comparison to determine which is the best model.

# Problem Formulation

## Predict autism based on the AQ-10 answer

The AQ-10 score of 6 and above indicates autism or number of autistic traits, while lower than 6 means the person might not be autistic. Hence, the accuracy of the AQ-10 test is the subject of study.

## Predict autism from all the features given

Further details such as age, sex, gender, ethnicity, and presence of jaundice are added to enhance the result of prediction. The additional variables would potentially increase the accuracy of prediction on top of the AQ-10 answers, as there is a chance to discover the hidden patterns in the matrix of data points.

# Problem and Data Understanding

## Problem Understanding

### Ideal solution for the problem

The model is expected to be able to predict ASD or autism up to accuracy of 80%.

### Characteristics for the ideal solution

Use a classifier to classify whether the case is autistic or not.

### What type of data mining modelling must perform?

Classification models which are K-Nearest Neighbour (KNN), ensemble learning methods, RF and LR are used to predict the result in target variable “Class/ASD”, on whether autism is “Yes” or “No”.

### Range of estimates

The output in target variable is “Yes” for positive in autism, and “No” for not autistic.

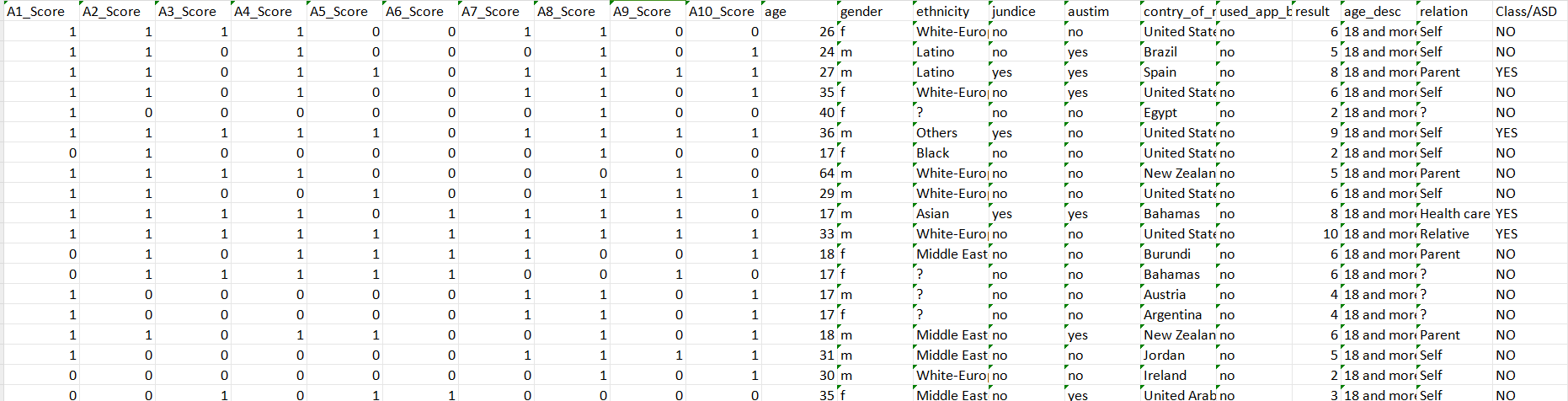
## Data Understanding

### Data collection

Data descriptions (Umar, 2024):

|  |  |  |  |
| --- | --- | --- | --- |
| **Features** | **Type** | **Description** | **Expected value** |
| A1\_Score | Binary | Answer code of question in AQ-10 | 0 or 1 |
| A2\_Score | Binary | Answer code of question in AQ-10 | 0 or 1 |
| A3\_Score | Binary | Answer code of question in AQ-10 | 0 or 1 |
| A4\_Score | Binary | Answer code of question in AQ-10 | 0 or 1 |
| A5\_Score | Binary | Answer code of question in AQ-10 | 0 or 1 |
| A6\_Score | Binary | Answer code of question in AQ-10 | 0 or 1 |
| A7\_Score | Binary | Answer code of question in AQ-10 | 0 or 1 |
| A8\_Score | Binary | Answer code of question in AQ-10 | 0 or 1 |
| A9\_Score | Binary | Answer code of question in AQ-10 | 0 or 1 |
| A10\_Score | Binary | Answer code of question in AQ-10 | 0 or 1 |
| age | Number | Age of case | Minimum of 17 years old |
| gender | String | Male or Female | f or m |
| ethnicity | String | Ethnicity of participants | White-European, Latino, Black, Asian, Middle Eastern, Pasifika, South Asian, Hispanic, Turkish, Others |
| jaundice | Boolean | If the case was born with jaundice | Yes or no |
| austim | Boolean | If immediate family member has been diagnosed with autism | Yes or no |
| contry\_of\_res | String | Where the participant resides | United States, Bahamas, etc |
| used\_app\_before | Boolean | ASDTest app or not | Yes or no |
| result | Integer | Score for AQ1-10 screening test | Value ranged from 1 to 10 |
| age\_desc | String | Age category of the case | 18 and more |
| relation | String | Relation of patient who completed the test | Self, Parent, Health Care Professional, Relative, Others |
| Class/ASD | Boolean | Target variable, whether the case has autism or not | Yes or No |

### Data sample



### What is the data size

The data set “autism\_screening.csv” consists of 21 columns and 704 records.

### How the data is stored

The data set is stored in Comma separated values (CSV) and Attribute-Relation File Format (ARFF) format.

### Data source

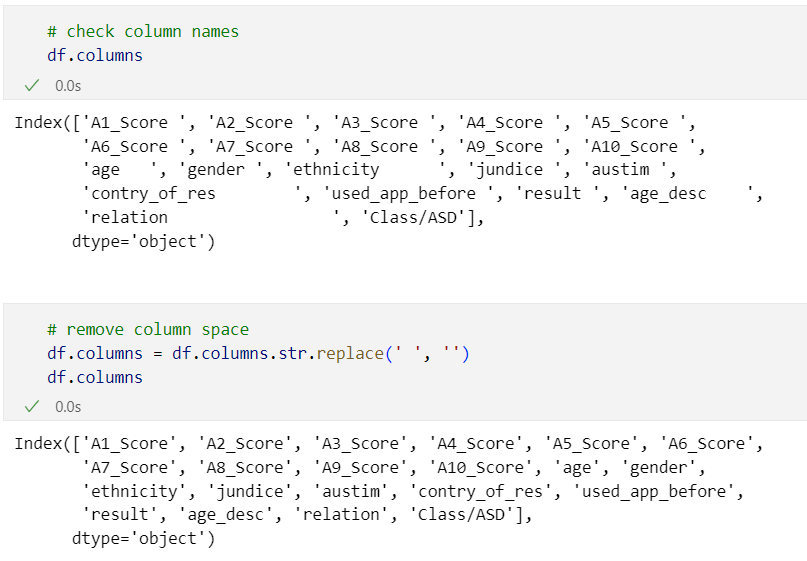
The Autism Screening on Adults (Larxel, 2017) dataset is obtained from Kaggle. Similar dataset with title Autism Screening Adult (UCI Machine Learning Repository, 2017) is found on another source. Features of anonymous participants details are from ASDTests screening app (Fadi, 2018), while the 10 AQ answer are based on AQ-10 adult (Allison et al., 2011).

# Data Pre-processing Methods

The issues found in the data set and steps taken for preprocessing are shown:

### Blank space in columns and values

1. Column: Each of the columns has a variety of blank spaces behind the last character. Therefore, *replace()* method is used for removing those spaces, by replacing each space with an empty string.

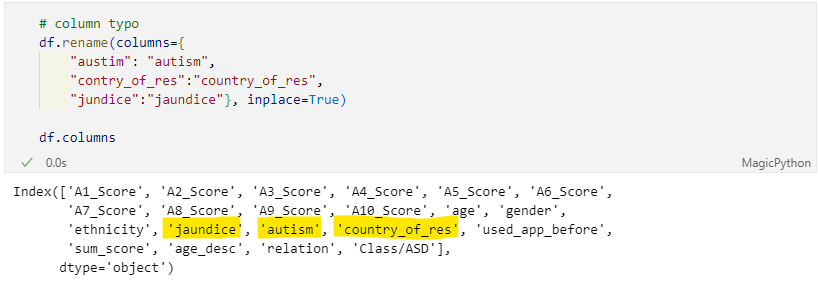


1. Values: There are also spaces behind values under the columns, such as “ethnicity”. Effect of before and after the cleaning is shown.



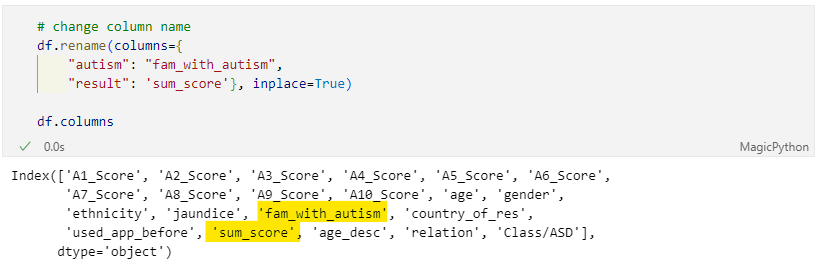
### Typo on columns’ name

Column name: *rename()* method is used for renaming the column names with typo, which are “austim”, “contry\_of\_res” and “jundice”. The names are corrected after running commands.



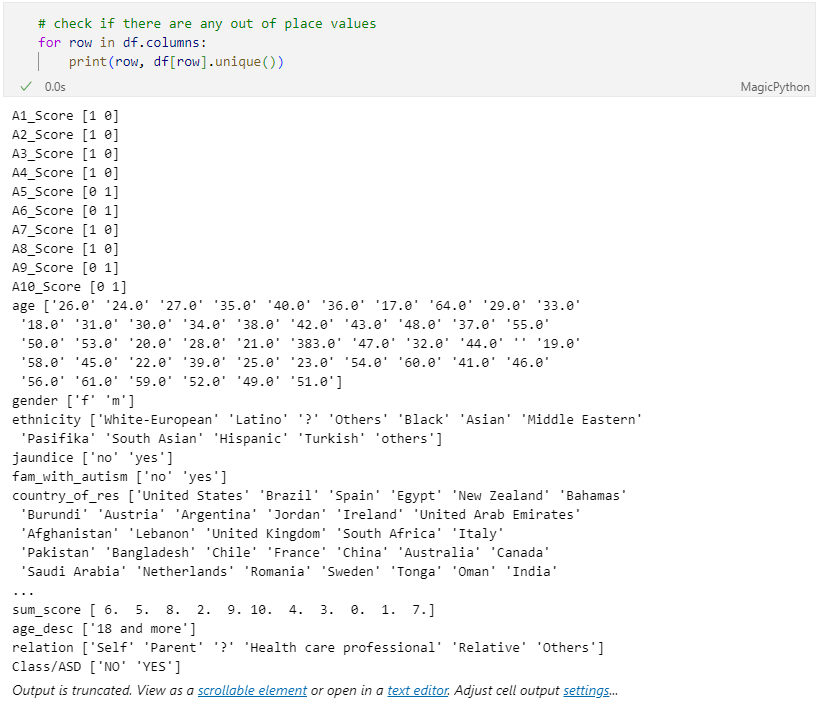
### Rename column to avoid confusion:

“result” column refers to sum of AQ score, and “autism” refers to family members diagnosed with autism, hence the columns shall be named accordingly.

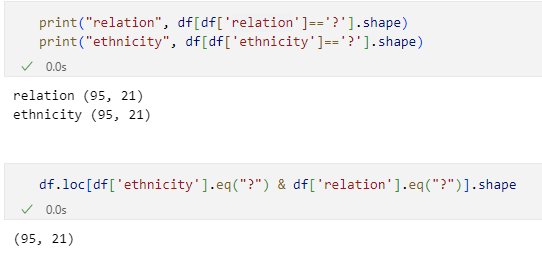


### Mismatch values

Through running the *unique()* method for each column, it is found that “ethnicity” and “relation” columns have odd value of “?”.



By using *shape* method, 95 records in total have “?”. Then, filter by *loc* shows that “ethnicity” and “relation” has overlap in the occurrence of “?”, that means if “ethnicity” has “?”, “relation” surely would have it too, or vice versa.



Since there are only 95 out of 704 records with this unknown value, it is safe to use *drop()* method to remove these records. After the removal, there are 609 records left.



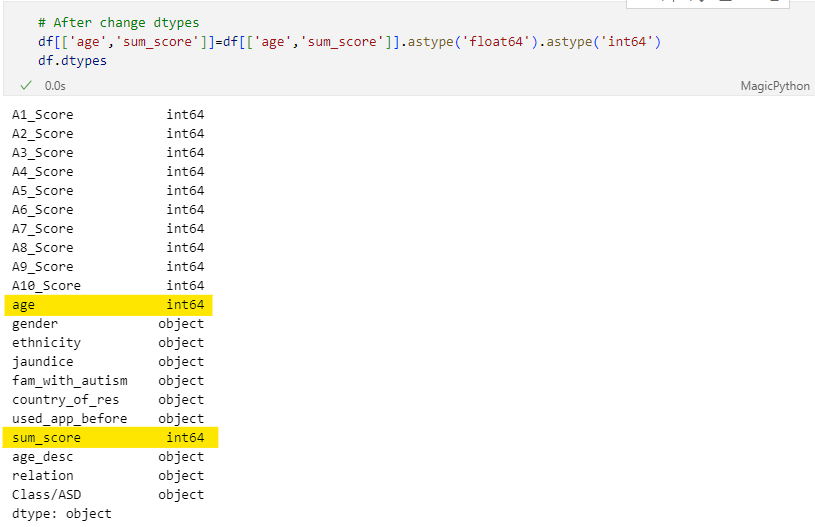
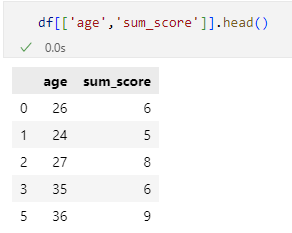
### Standardize “ethnicity” values

From *unique()* method, the “Others” and “others” differ by a capital letter; “Turkish” is considered as “Middle Eastern”, while “Latino” and “Hispanic” do have some share traits but can be geographically distributed, same goes to “Asian” and “South Asian” whom might have contrasting genetics. Thus, only “others” and “Turkish” are being replaced. Also, a hyphen is removed from “White-European”.



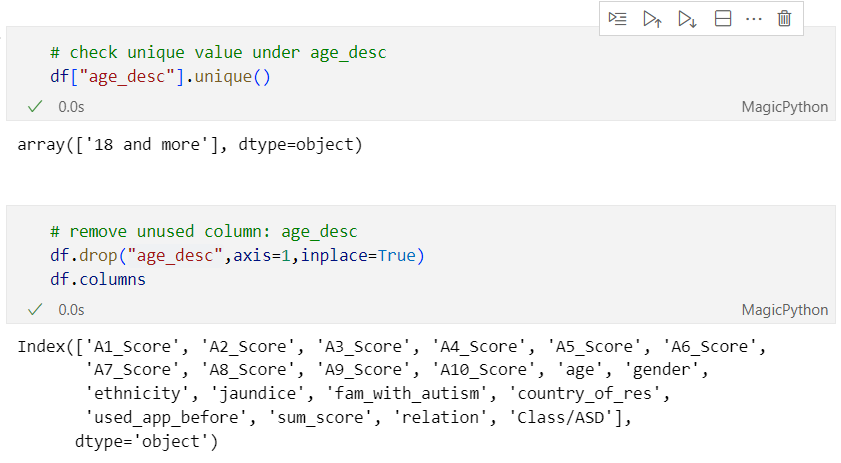
### Wrong data types

The values under “age” and “sum\_score” column should be integer as there is only 0 after decimal. Data types of these columns are changed to “float64” first, then to “int64”.



The “age” and “sum\_score” do not have decimal place now.

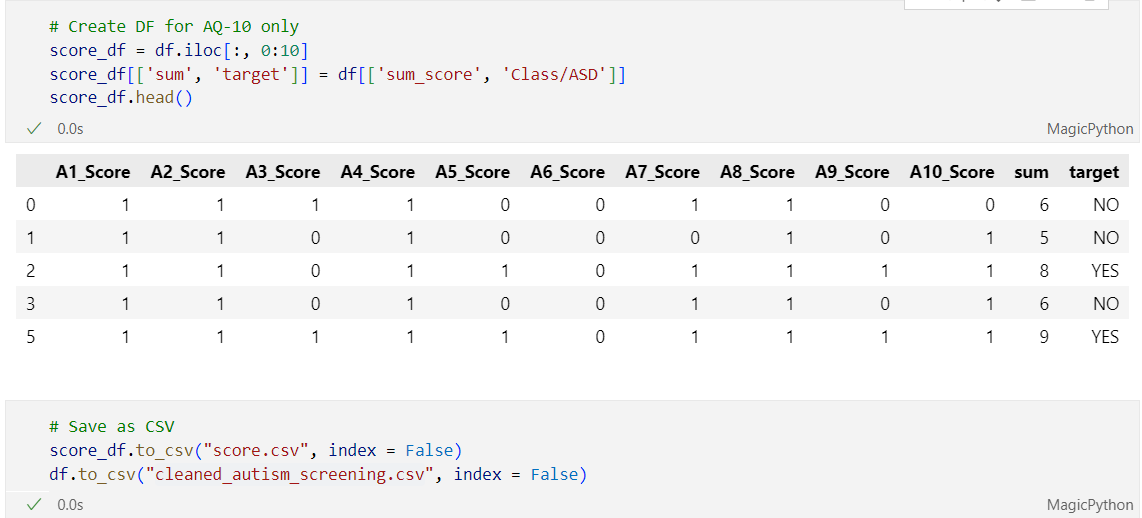
### Remove unused column

Under “age\_desc” column, “18 and more” is the only value, therefore this column does not have any variety. So, this column can be removed by using *drop()* method 

### Create new data sets

New data frame containing AQ-10 results and target variable is created, and saved as a new data set with CSV format, under name “score.csv”

Cleaned version of data frame containing all features except “age\_desc” also saved as “cleaned\_autism\_screening.csv”



# Data Modelling & Evaluation

## Models

The prediction of autism / ASD from data set is a classification problem, therefore, the techniques used are:

1. Ensemble learning - VotingClassifier with Logistic regression, Decision Tree, and Support Vector as sub models

* Advantages: Has more resistance towards incorrect or impacts from individual model's errors through taking various viewpoints into account. Also, accuracy can be improved via reduced variance and bias to enhance the overall performance by leveraging the collective knowledge of multiple models, therefore promoting higher accuracy in prediction. (*SoulPage IT Solutions*, 2023)
* Disadvantages: There is a risk of information loss as the confidence of individual models is being ignored. (Syed Wahaj, 2023)

1. K-Nearest Neighbours (KNN)

* Advantages: Due to the simplicity of algorithm, it has high accuracy with less hyper-parameter, which are k value and distance metric. In terms of adaptability, this algorithm could adjust itself for new data as training data already stored into memory.
* Disadvantages: KNN has low scalability as it requires more memory, thus needs longer time for computation. The algorithm is also guilty of performance issues with high-dimensional data input, called peaking phenomenon. (IBM, 2021)
* The disadvantages of KNN can be ignored as the size of the data set is small (21 columns, 608 records), and there would be no additional features or records to be added for extra training and testing.

1. Random Forest (RF)

* Advantages: Parallel processing is supported as the trees are created in parallel, speeding up training time since the iterations are independent. The model can handle vast amount data with high dimension and is valuable in understanding the hidden patterns by providing information about the importance of each feature in data.
* Disadvantages: Possess a degree of challenge in making interpretation as many decision trees are involved. Computational complexity is also another factor as it needs lots of memory when comes to large datasets. (AIML.com, 2022)

1. Logistic regression (LR)

* Advantages: Fast and able to handle data sets of large dimensions. Due to its simplicity, the result in the form of coefficients is not complex for interpretation.
* Disadvantages: The sensitivity towards outliers could affect the coefficients and predictions. The model also has risk of over-fitting when it comes to feature rich data set. (AIML.com, 2022)

## Evaluation

Regression and classification have different sets of metrics in evaluating the performance of their models. The set of metrics included for classification are:

1. Confusion Matrix:

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Actual | |
| Y | N |
| Predicted | Y | True Positive (TP) | False Positive (FP) |
| N | False Negative (FN) | True Negative (TN) |

* Accuracy: correct predictions out of total observations.
* Precision: positive prediction among all positively predicted actual cases.
* Recall or sensitivity: true positive rate (Accuracy vs. Precision vs. Recall in Machine Learning: What’s the Difference?, 2024), positive prediction among all actual positive cases.
* Specificity: true negative rate, negative prediction among all actual negative cases.
* Misclassification: is the complement of accuracy, calculated by 1-accuracy. Lower value is preferred

1. Receiver Operator Characteristic (ROC curves): This method works by plotting true positive versus false positive rate. The maximum accuracy is 1.0 or 100%. The models would be plotted on the same ROC curve for comparison. The best performed model would be close to 100% or have the largest area under the curve (AUC) value.

# Data Modelling and Evaluation phases

Before the modelling process, the features with string values would undergo *one-hot encoding* technique to convert categorical variables into numerical values (GeeksforGeeks, 2019). While the “Yes” and “No” values would be translated into 1 for “Yes”, and 0 for “No”. Next, each set of data is splitted into 75% for training and 25% for testing. Those training and testing data sets would undergo feature scaling to standardise the range of data in features (Christos Goumopoulos & Stergiopoulos, 2022).

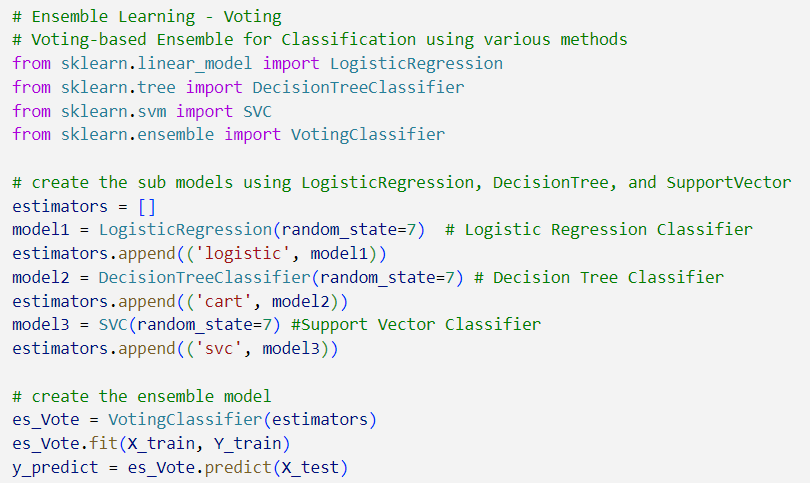
## score.csv

Repeated steps for all models: library import, data splitting and feature scaling

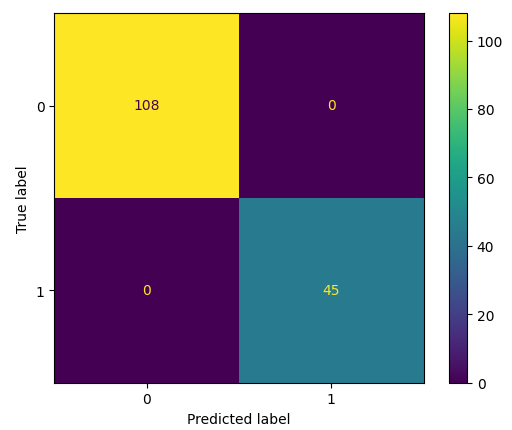


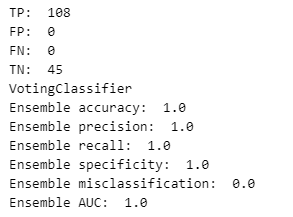
### Ensemble Learning - VotingClassifier

Use LR, decision tree, and support vector classifier as the models for voting.



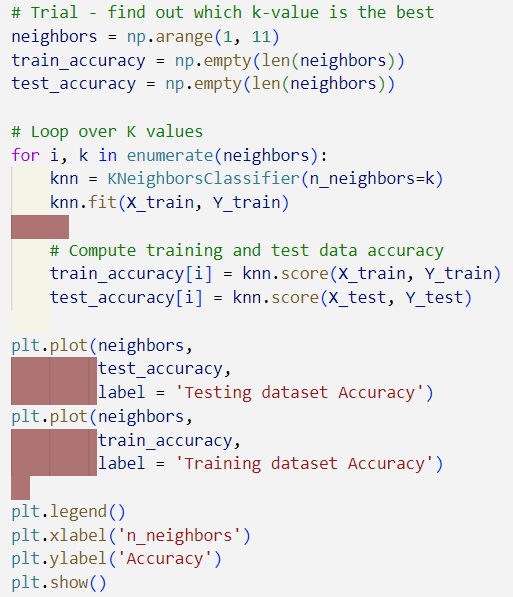
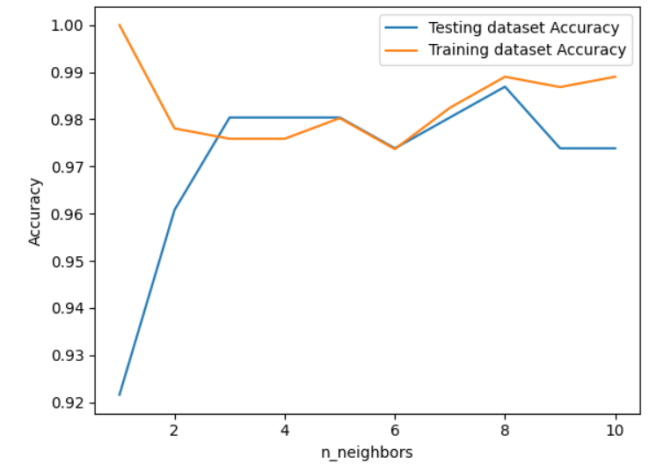
Run Confusion Matrix to get values for accuracy, precision, recall, specificity, and misclassification. Also, use *roc\_auc\_score()* to find AUC.



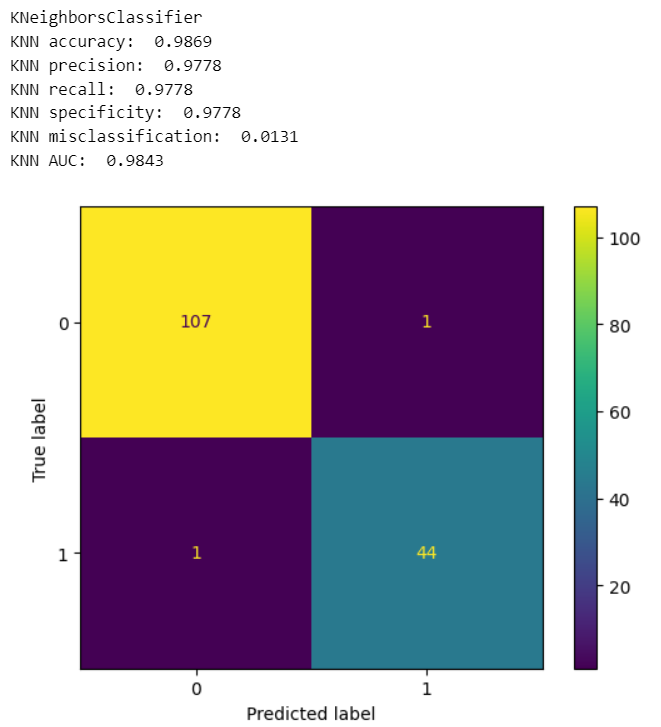


### KNN

Set a range of k-Neighbours from 1 to 10 to find out the k-value for highest accuracy in testing data set. From the graph, the overlap accuracy between testing and training set is at k=5 and k=6, while the accuracy for testing is highest at 8.

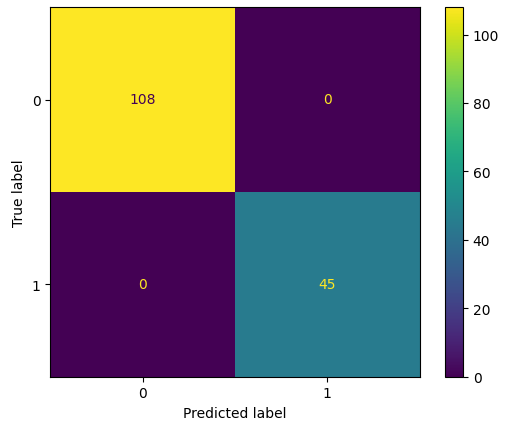


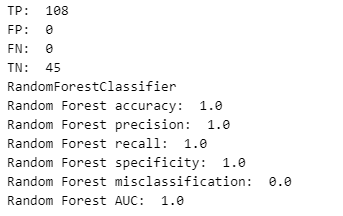
Therefore, k=8 is used for getting the values in Confusion matrix, and AUC.



### RF

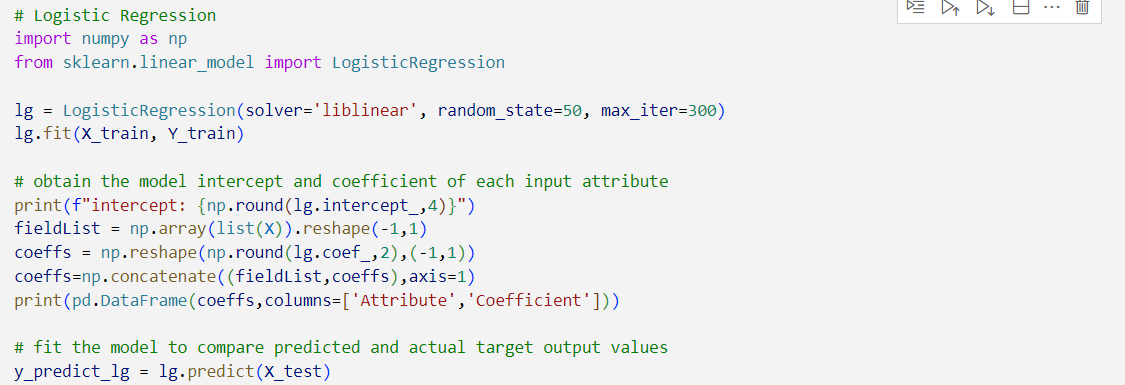
The result of using *RandomForestClassifier()* is the same as *VotingClassifier(),* where the accuracy and AUC have the ideal value of 1.0. Hence, there is no improvement needed to enhance the model’s performance.



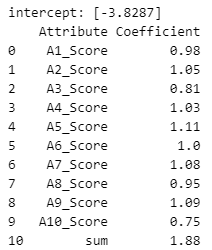


### LR

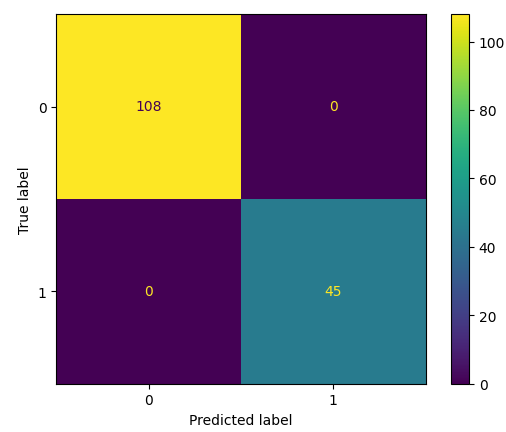
*LogisticRegression()* is applied with *solver = ‘liblinear’* for binary classification with small dataset.

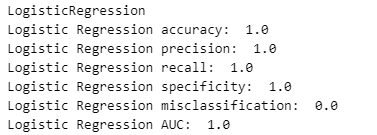


LR result:

From the coefficients of each attribute, the influence of AQ-10 scores on target variable is almost the same. Meanwhile “sum” attribute carried the heaviest weight among all the attributes, hence it gives the most effect on the target variable.

Confusion Matrix and AUC: Similar result as VotingClassifier and RF





## cleaned\_autism\_screening.csv

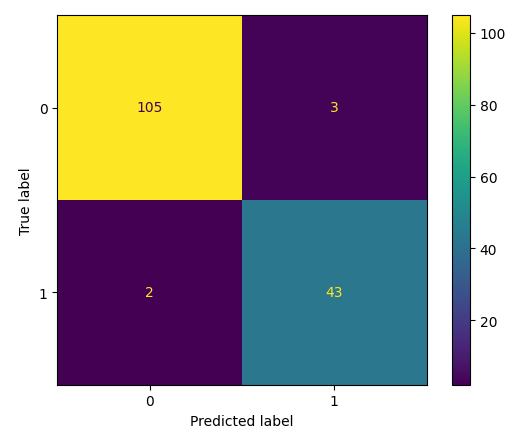
Before the modelling process start, a few steps are taken:

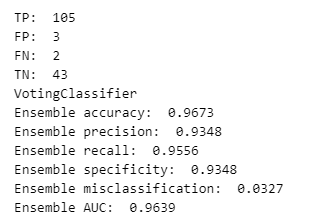
* Libraries import
* Import and read the CSV
* Convert the “Yes” and “No” values into 1 and 0
* Apply one-hot encoding to convert all string objects into numeric
* Split data into 75% training and 25% testing
* Feature scaling



Note: the snippets for modelling and performance measure are same as the images above, hence current snippets are not displayed.

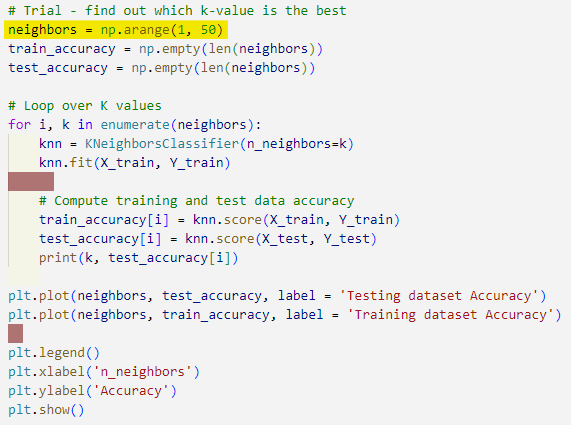
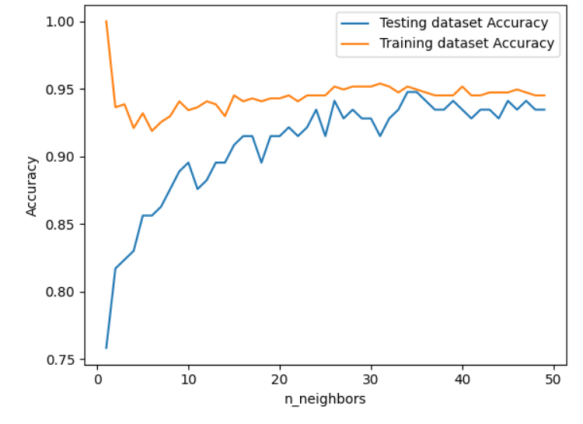
### Ensemble Learning - VotingClassifier

The accuracy and AUC of VotingClassifier is close to ideal value of 1.0.

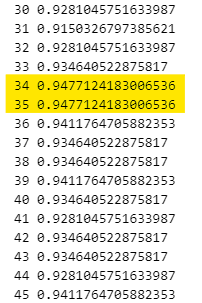
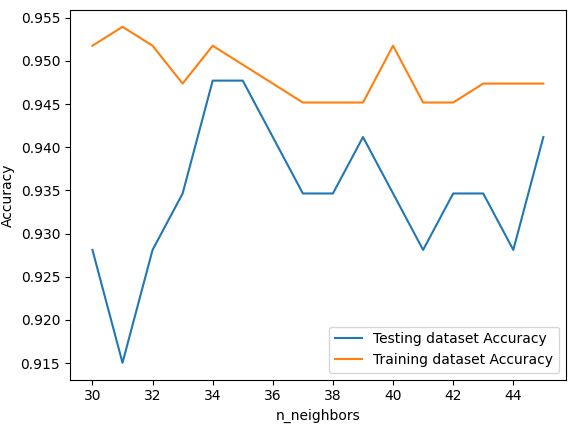


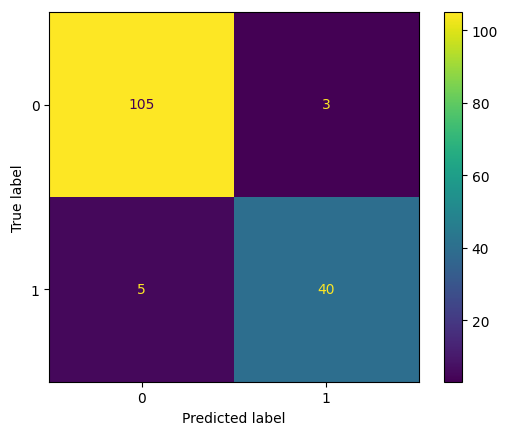
### KNN

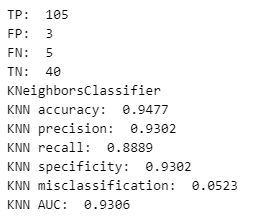
The k-value range of 1 to 50 is run to find the highest accuracy for both training and testing data. From the graph, accuracy of both data sets is closest between 30 to 45.



Hence, the range is set between 30 to 45: the highest accuracy of test data is at k=34 and 35, with accuracy at 0.945.

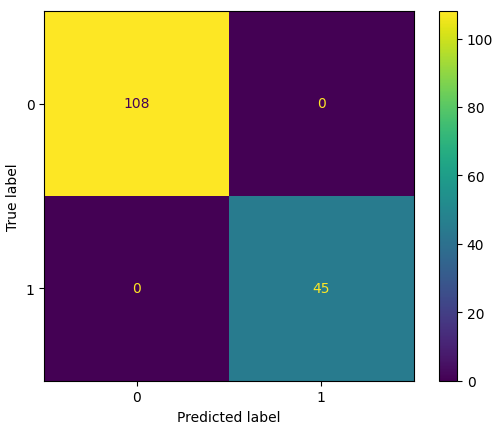


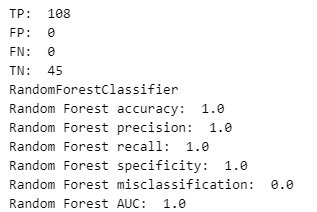
By setting k=35, the Confusion Matrix and AUC result are displayed as below:



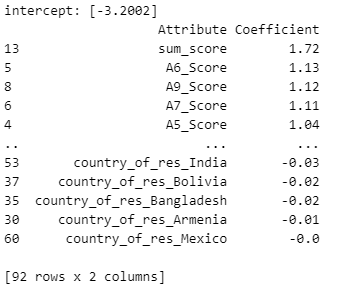
The accuracy stood firm at 0.9477, where the FP and FN did not exceed 5.

### RF

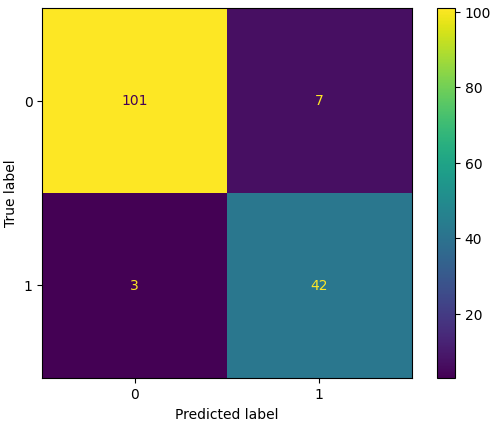
This model produced the ideal result with accuracy of 1.0

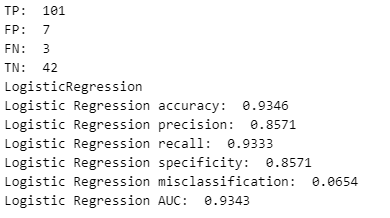


### LR

From the result, the equation of the LR is:

Therefore, “sum\_score” has the most ability among 92 features in affecting the target variable, “Class/ASD”; while “country\_of\_res\_Mexico” carries the least or almost none.

For the evaluation, accuracy is 0.9346 and AUC=0.9343



## ROC

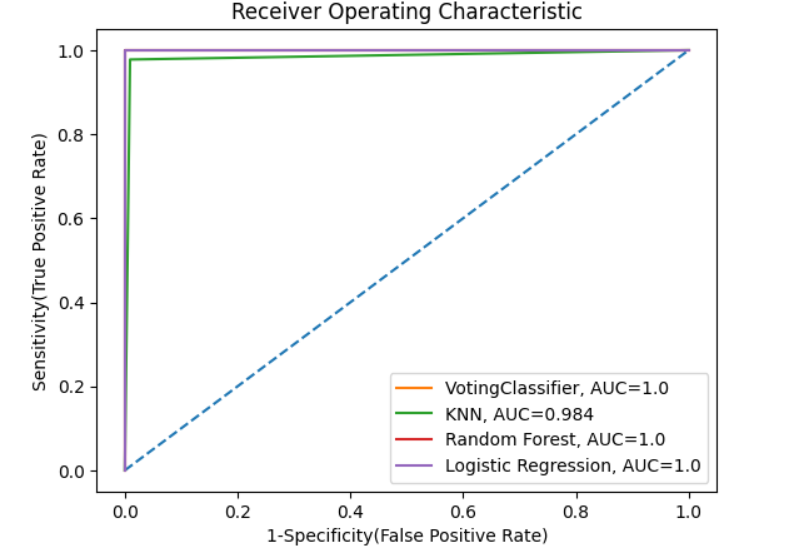
The snippet of ROC for VotingClassifier, KNN, RF, and LR:



# Model Evaluation results

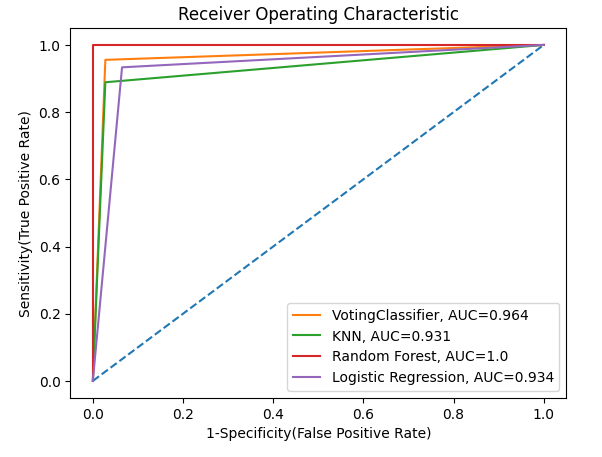
Data set: score.csv

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Metrics | |
| Confusion Matrix | ROC-AUC |
| Models | Ensemble Learning: Voting Classifier | Accuracy = 1.0  Precision = 1.0  Recall = 1.0  Specificity = 1.0  Misclassification = 0.0 | AUC = 1.0 |
| KNN | Accuracy = 0.9869  Precision = 0.9778  Recall = 0.9778  Specificity = 0.9778  Misclassification = 0.0131 | AUC = 0.9843 |
| RF | Accuracy = 1.0  Precision = 1.0  Recall = 1.0  Specificity = 1.0  Misclassification = 0.0 | AUC = 1.0 |
| LR | Accuracy = 1.0  Precision = 1.0  Recall = 1.0  Specificity = 1.0  Misclassification = 0.0 | AUC = 1.0 |



Data set: cleaned\_autism\_screening.csv

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Metrics | |
| Confusion Matrix | ROC-AUC |
| Models | Ensemble Learning | Accuracy = 0.9673  Precision = 0.9348  Recall = 0.9556  Specificity = 0.9348  Misclassification = 0.0327 | AUC = 0.9639 |
| KNN | Accuracy = 0.9477  Precision = 0.9302  Recall = 0.8889  Specificity = 0.9302  Misclassification = 0.0523 | AUC = 0.9306 |
| RF | Accuracy = 1.0  Precision = 1.0  Recall = 1.0  Specificity = 1.0  Misclassification = 1.0 | AUC = 1.0 |
| LR | Accuracy = 0.9346  Precision = 0.8571  Recall = 0.9333  Specificity = 0.8571  Misclassification = 0.0654 | AUC = 0.9343 |



# Conclusion

### score.csv

From the ROC, since KNN has the lowest AUC value although the value is close to 1.0, while VotingClassifer, LR and RF produce the same confusion matrix and ROC result of 1.0. Therefore, KNN is the least preferred model compared to the other 3.

### cleaned\_autism\_screening.csv

The shape of ROC curve suggests RF as the best model due to the highest accuracy, AUC value, followed by VotingClassifier, LR and lastly KNN.

(2654 words)

# Reference

Larxel. (2017). *Autism Screening on Adults*. Kaggle.com. <https://www.kaggle.com/datasets/andrewmvd/autism-screening-on-adults?resource=download>

*UCI Machine Learning Repository*. (2017). Uci.edu. <https://archive.ics.uci.edu/dataset/426/autism+screening+adult>

‌Afarin Bargrizan. (2023). *ASD questionnairs- Final*. Kaggle.com. <https://www.kaggle.com/datasets/afarinbargrizan/asd-final>

Engelbrecht, N. (2020, April 20). *The AQ-10*. Embrace Autism. <https://embrace-autism.com/aq-10/#The_AQ-10>

Nurul Amirah Mashudi, Ahmad, N., & Norliza Mohd Noor. (2021). Classification of adult autistic spectrum disorder using machine learning approach. *IAES International Journal of Artificial Intelligence*, *10*(3), 743–743. <https://doi.org/10.11591/ijai.v10.i3.pp743-751>

*Autism spectrum disorder (ASD) | Autism Speaks*. (2024). Autism Speaks. <https://www.autismspeaks.org/what-autism>

Pierce, K., Carter, C., Weinfeld, M., Desmond, J., Hazin, R., Bjork, R., & Gallagher, N. (2011). Detecting, Studying, and Treating Autism Early: The One-Year Well-Baby Check-Up Approach. *The Journal of Pediatrics*, *159*(3), 458-465.e6. <https://doi.org/10.1016/j.jpeds.2011.02.036>

*When do children usually show symptoms of autism?* (2017, January 31). Https://Www.nichd.nih.gov/. <https://www.nichd.nih.gov/health/topics/autism/conditioninfo/symptoms-appear#f4>

NHS Choices. (2024). *Signs of autism in adults*. <https://www.nhs.uk/conditions/autism/signs/adults/>

*How is Autism Different in Women?* (2023). Harvard.edu. <https://adult-autism.health.harvard.edu/resources/how-is-autism-different-in-females/>

Fadi. (2018). *Autism screening data for toddlers*. Kaggle.com. <https://www.kaggle.com/datasets/fabdelja/autism-screening-for-toddlers>

*ASD*. (2017). Asdtests.com. <https://www.asdtests.com/>

Allison, C., Auyeung, B., & Baron-Cohen, S. (2011). Toward Brief “Red Flags” for Autism Screening: The Short Autism Spectrum Quotient and the Short Quantitative Checklist in 1,000 Cases and 3,000 Controls. *Journal of the American Academy of Child & Adolescent Psychiatry*, *51*(2), 202-212.e7. <https://doi.org/10.1016/j.jaac.2011.11.003>

Umar, M. (2024). *Autism Spectrum*. Kaggle.com. <https://www.kaggle.com/datasets/umeradnaan/autism-screening>

*BaggingClassifier*. (2024). Scikit-Learn. <https://scikit-learn.org/1.5/modules/generated/sklearn.ensemble.BaggingClassifier.html>

Simplilearn. (2021, March 26). *Ensemble Learning: From Basics to Advanced Techniques!* Simplilearn.com; Simplilearn. <https://www.simplilearn.com/ensemble-learning-article>

*LinkedIn*. (2024). Linkedin.com. <https://www.linkedin.com/pulse/bagging-classifier-rupak-roy/>

GeeksforGeeks. (2019, April 9). *knearest neighbor algorithm in Python*. GeeksforGeeks. <https://www.geeksforgeeks.org/k-nearest-neighbor-algorithm-in-python/>

*SoulPage IT Solutions*. (2023, June 19). Soulpage IT Solutions. <https://soulpageit.com/ai-glossary/ensemble-voting-explained/#:~:text=Improved%20Accuracy%3A%20It%20leverages%20the,and%20achieve%20better%20overall%20performance.>

AIML.com. (2022, June 11). *What are the advantages and disadvantages of Random Forest?* AIML.com. <https://aiml.com/what-are-the-advantages-and-disadvantages-of-random-forest/>

*Accuracy vs. precision vs. recall in machine learning: what’s the difference?* (2024). Evidentlyai.com. <https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall#:~:text=Recall%20can%20also%20be%20called,research%20rather%20than%20machine%20learning.>

Verma, A. (2023, January 27). *Evaluation Metrics for Classification and Regression: A Comprehensive Guide*. DEV Community. <https://dev.to/anurag629/evaluation-metrics-for-classification-and-regression-a-comprehensive-guide-47hb#:~:text=The%20mean%20squared%20error%20is,the%20performance%20of%20regression%20models.>

Christos Goumopoulos, & Stergiopoulos, N. G. (2022). Mental stress detection using a wearable device and heart rate variability monitoring. *Elsevier EBooks*, 261–290. <https://doi.org/10.1016/b978-0-323-90585-5.00011-4>

Dr Ana Rojo-Echeburúa. (2024, June 26). *What Is One Hot Encoding and How to Implement It in Python*. Datacamp.com; DataCamp. https://www.datacamp.com/tutorial/one-hot-encoding-python-tutorial?utm\_source=google&utm\_medium=paid\_search&utm\_campaignid=19589720824&utm\_adgroupid=157098106775&utm\_device=c&utm\_keyword=&utm\_matchtype=&utm\_network=g&utm\_adpostion=&utm\_creative=716160943678&utm\_targetid=aud-1704732079567:dsa-2264919291989&utm\_loc\_interest\_ms=&utm\_loc\_physical\_ms=9066731&utm\_content=&utm\_campaign=230119\_1-sea~dsa~tofu\_2-b2c\_3-row-p2\_4-prc\_5-na\_6-na\_7-le\_8-pdsh-go\_9-nb-e\_10-na\_11-na-oct24&gad\_source=1&gclid=CjwKCAjw3624BhBAEiwAkxgTOmITa6qWv2vhmu3rpLNb-N0SLydR69zq\_Qbl51HOq87mBga8ESM1JRoCMDMQAvD\_BwE

GeeksforGeeks. (2019, June 12). *One Hot Encoding in Machine Learning*. GeeksforGeeks. <https://www.geeksforgeeks.org/ml-one-hot-encoding>

Syed Wahaj. (2023, August 8). *Hard vs. Soft Voting Classifiers - Techkluster*. Techkluster. <https://techkluster.com/technology/hard-vs-soft-voting-classifiers/#:~:text=Disadvantages%3A,be%20affected%20by%20outlier%20predictions.>‌