Autism Prediction Analysis

Introduction: This project focuses on predicting **Autism Spectrum Disorder (ASD)** using machine learning models. The dataset includes demographic, behavioural, and medical history-related attributes to identify potential ASD cases. The workflow involves **data preprocessing, exploratory data analysis (EDA), feature engineering, model training, and evaluation** to achieve optimal predictions.

Skills used: Data Cleaning, Data modelling, Data visualization

Dataset Overview: The dataset consists of **800 records** with **22 attributes**, including autism screening test scores, age, gender, ethnicity, medical history, and the final diagnosis (Class/ASD).

Dataset Features

1. **A1_Score - A10_Score** – Responses to autism screening questions (Binary: 0 or 1).
2. **Age** – Age of the individual (Numerical).
3. **Gender** – Male/Female (Categorical).
4. **Ethnicity** – Ethnic background (Categorical).
5. **Jaundice** – Whether the individual had jaundice at birth (Yes/No).
6. **Austim** – Family history of autism (Yes/No).
7. **Country of Residence** – The country where the individual resides (Categorical).
8. **Used App Before** – Whether the individual has previously used an autism screening app (Yes/No).
9. **Result** – Autism screening test score (Numerical).
10. **Age Description** – Categorized age group (Categorical).
11. **Relation** – Relationship of the respondent to the individual (e.g., Self, Parent).
12. **Class/ASD (Target Variable)** – 1 for ASD, 0 for no ASD (Binary Classification).

Analysis of Data

1. Data Distribution & Preprocessing

- The dataset **contains no missing values** based on df.info().
- The **target variable (Class/ASD) is imbalanced**, requiring **oversampling** to handle class distribution.
- **Categorical Encoding:**
  - LabelEncoder is used to convert categorical values into numerical values.
- **Feature Scaling:**
  - StandardScaler is applied to normalize numerical variables like **age** and **result scores**.

Feature Engineering

1. **New Feature – Age Group**:
   - A function is used to categorize individuals into **Toddler, Kid, Teenager, Young, and Senior** based on age.
2. **New Feature – Sum Score**:
   - A new column, sum_score, is created by summing up the **A1_Score to A10_Score**, providing a stronger predictor for ASD.

Machine Learning Model Implementation

1. Handling Class Imbalance

- **Random Oversampling** (RandomOverSampler) is applied to ensure a **balanced dataset**, preventing the model from being biased towards the majority class.

2. Model Selection and Training

The following models are trained on the dataset:

1. **Logistic Regression** – A linear model for binary classification.
2. **Support Vector Machine (SVM)** – Efficient for high-dimensional data.
3. **XGBoost (XGBClassifier)** – A powerful ensemble learning model using gradient boosting.

- The dataset is split into **80% training and 20% testing** using train_test_split.
- Hyperparameters are **not explicitly tuned** in the extracted code but could improve model performance.

3. Model Performance Evaluation

The models are evaluated using:

- **Training Accuracy** – Overall correctness of predictions in the training sample.
- **Validation Accuracy –** Overall correctness of predictors in the validation data.

Model	Training Accuracy	Validation Accuracy
Logistic Regression	0.8665	0.7823
Support Vector Machine (SVM)	0.9405	0.8042
XGBoost (XGBClassifier)	1.0	0.7491

4. Key Insights from Model Performance

- **XGBoost performed the best (100% accuracy)** due to its strong ability to handle nonlinear relationships.
- **Logistic Regression is a strong baseline model (86.65%)**, offering high interpretability.
- **SVM provides good performance (94.05%)** but may require hyperparameter tuning for improvements.
- **Using oversampling improves fairness in predictions**, reducing bias towards the majority class.

Conclusion

This autism prediction project successfully applies **data preprocessing, feature engineering, and machine learning models** to predict ASD cases.

- **Feature engineering (age group and sum score) enhances model accuracy.**
- **XGBoost is the best-performing model, but hyperparameter tuning could further improve results.**
- **Future improvements could include deep learning models and additional behavioral features for better generalization.**
