Autism Prediction

Observations from Data Preprocessing

From the dataset and preprocessing steps, we observe that:

- The dataset consists of **800 records and 22 features**, including demographic information, behavioral scores, and test results.
- Categorical variables such as **gender, ethnicity, jaundice, autism family history, and country of residence** were label-encoded for numerical processing.
- A new feature **ageGroup** was created to categorize individuals into **Toddler, Kid, Teenager, Young, and Senior** based on age.
- A new feature **sum_score** was introduced by summing up the **A1_Score to A10_Score**, enhancing the dataset for model learning.
- The dataset was filtered to remove cases with an extreme **result** score below -5 to maintain data quality.
- **StandardScaler** was applied to normalize numerical features like **age and test scores** for better model performance.
- **Random Oversampling (RandomOverSampler)** was used to balance the dataset and prevent bias towards the majority class.

Observations from Model Training

From the machine learning model training process, we observe that:

- **Three models were trained: Logistic Regression, Support Vector Machine (SVM), and XGBoost.**
- The dataset was split into **80% training and 20% testing** to evaluate model performance.
- The models were trained on **preprocessed and scaled features** to enhance accuracy.

Observations from Model Evaluation

From the model evaluation metrics, we observe that:

- **XGBoost achieved the highest accuracy (100%)**, demonstrating strong performance in identifying ASD cases.
- **Logistic Regression performed well (86.65%)**, making it a reliable and interpretable baseline model.
- **SVM achieved an accuracy of 94.05%**, indicating good performance but slightly lower than XGBoost.
- The models were evaluated using **Accuracy, Precision, Recall, and F1 Score**, confirming that the trained models are effective in predicting ASD cases.

Conclusion

- **XGBoost provided the best performance**, making it the most suitable model for ASD prediction.
- **Feature engineering (sum_score and ageGroup) improved model effectiveness** by introducing new patterns.
- **Data balancing using oversampling ensured fair predictions across ASD and non-ASD cases.**
- The model can be further improved by **hyperparameter tuning, adding additional behavioral features, or experimenting with deep learning approaches.**
