REPORT ON AUTISM PREDICTION

A MACHINE LEARNING APPROACH

**A PROJECT REPORT**

***Submitted by***

Rupali (23MCI10137)

***in partial fulfilment for the award of the degree of***

**Master of Computer Applications**

**IN**

Artificial Intelligence And Machine Learning

**Chandigarh University**

****

OCT 2024

****

**BONAFIDE CERTIFICATE**

Certified that this project report **“AUTISM PREDICTION”** is

the bonafide work of **“Rupali”**

who carried out the project work under my/our supervision.

**SIGNATURE HEAD OF THE DEPARTMENT**

**SIGNATURE SUPERVISOR**

**INTERNAL EXAMINER EXTERNAL EXAMINER**

**Table Of Content**

**CHAPTER 1. INTRODUCTION .......................................................................4**

1. Introduction to the project

2. Problem Definition and /solution Overview

**Chapter 2: Literature Review/Background Study...........................................................5**

1. Review Summary

2. Goals and Objectives

**CHAPTER 3. SETUP AND DESIGN ..........................................................................6**

1. Methodologies and Tools and Technologies

**CHAPTER 4. IMPLEMENTATION AND RESULT............................9**

**CHAPTER 5. DEPLOYMENT AND PUBLICATION...................................15**

1. Publishing on Github

**CHAPTER 6. FUTURE SCOPE...................................18**

**CHAPTER 7. CONCLUSION...................................19**

**Chapter – 1: INTRODUCTION**

Autism prediction involves using data analysis, machine learning, and sometimes genetics to identify individuals who may be at higher risk of having autism spectrum disorder (ASD). Autism is a developmental disorder characterized by challenges in social interaction, communication, and repetitive behaviors. Early diagnosis is key because timely interventions can greatly benefit a child's development and improve long-term outcomes.

**Why Predict Autism?**

1. **Early Intervention**: The earlier autism is detected, the more effective interventions can be. Interventions in language, behavior, and social skills are more successful during the critical developmental periods in early childhood.
2. **Personalized Treatment**: Predicting autism early allows families and healthcare providers to create customized treatment plans.
3. **Resource Allocation**: Early prediction can help governments and health systems allocate resources more effectively, targeting those most at risk.

**Methods in Autism Prediction**

1. **Screening Tools**: Tools like the Modified Checklist for Autism in Toddlers (M-CHAT) are widely used to screen young children for autism risk. These tools typically involve parent-reported surveys and behavioral observations.
2. **Genetic Analysis**: Studies show that genetics play a significant role in autism. Specific genetic markers and mutations have been associated with a higher likelihood of autism. Genomic data can be used alongside other factors for predictive analysis.
3. **Machine Learning**: Data-driven models use various data points, like genetic, demographic, behavioral, and neuroimaging data, to predict autism. Algorithms analyze these data types, looking for patterns associated with ASD.
4. **Behavioral Data**: Observing early social behaviors and responses to stimuli can be insightful. For example, reduced eye contact or slower response to social cues can be early indicators of autism.
5. **Brain Imaging and Biomarkers**: Neuroimaging techniques, like MRI and EEG, help researchers study brain structure and function differences in individuals with autism. Machine learning models trained on these imaging data can help predict ASD by identifying atypical brain patterns.

**Challenges in Autism Prediction**

1. **Data Diversity**: Autism is a spectrum, so symptoms vary widely among individuals. Models must be trained on diverse datasets to make accurate predictions across different cases.
2. **Ethical Concerns**: Predicting autism in early infancy or even pre-birth raises ethical questions about diagnosis and intervention, especially in cases with inconclusive results.
3. **Data Availability and Privacy**: Autism prediction models often require vast amounts of sensitive data, and obtaining and securing this data can be challenging.

**Current Research and Future Directions**

Recent research aims to improve the accuracy of prediction models by combining genetic, behavioral, and environmental data. Some researchers focus on deep learning methods to analyze complex patterns in behavioral or neuroimaging data. In the future, personalized prediction models that adapt to individual differences in genetic and environmental risk factors could offer more precise predictions and interventions for each child.

**Chapter 2: Literature Review/Background Study**

**Review Summary**

**Summary of Literature Review on Autism Prediction**

The review of autism prediction literature highlights significant progress in identifying factors and developing models to predict autism spectrum disorder (ASD) early in life. Studies reveal that genetic, environmental, and neurobehavioral markers are strongly linked to ASD. Notably, machine learning (ML) and artificial intelligence (AI) are transforming autism prediction, with models utilizing diverse data sources like genetic information, neuroimaging, and behavioral assessments. Despite these advances, several challenges persist, including the need for high-quality and diverse data, model interpretability issues, and ethical concerns around privacy and bias.

Current autism prediction models show promising accuracy but face generalization limitations due to biased or imbalanced datasets and limited sample sizes. Single-modal approaches tend to miss the broader spectrum of autism markers, leading to the adoption of multi-modal models that combine various data types. However, balancing interpretability with predictive accuracy remains challenging, especially with complex deep learning models that lack transparency.

Future research should prioritize more inclusive datasets, improved feature selection techniques, and collaborative approaches involving clinicians, researchers, and ethicists. By addressing these challenges, predictive models can become more reliable and clinically applicable, ultimately contributing to timely, personalized interventions for individuals at risk of ASD.

**Chapter 3: Setup and Design**

**3.1 Introduction**

* Briefly introduce the purpose of this chapter, emphasizing how the setup and design will help achieve reliable autism prediction.
* Outline the key components of this chapter, including data sources, methodologies, tools, and the overall system design.

**3.2 Research Methodology**

* **Overview of Methodology**: Provide an overview of the research design, specifying if it is experimental, observational, or based on secondary data analysis.
* **Data Collection Process**:
  + Describe the sources of data (e.g., public autism datasets, clinical records, or surveys).
  + Include details about data acquisition, including any ethical considerations (e.g., consent, privacy).
* **Selection Criteria**: Explain any inclusion/exclusion criteria for data (e.g., age, diagnostic status, data quality) and how the dataset is prepared for analysis.

**3.3 Data Preprocessing**

* **Data Cleaning**: Describe techniques for handling missing data, outliers, and irrelevant features.
* **Feature Engineering**:
  + Explain any feature extraction and transformation processes, such as converting categorical variables to numerical, normalizing values, or creating new features from existing ones.
* **Feature Selection**: Outline methods to select the most relevant features, such as correlation analysis, recursive feature elimination, or principal component analysis (PCA), particularly if using large datasets with many features.
* **Data Splitting**: Describe how the data is split into training, validation, and test sets to avoid overfitting and ensure model generalization.

**3.4 Machine Learning Models and Techniques**

* **Algorithm Selection**: List and explain the machine learning algorithms used in the prediction model (e.g., logistic regression, support vector machines (SVM), random forests, decision trees, neural networks).
* **Model Architecture**:
  + If deep learning models are used, specify the architecture (e.g., CNNs for image data or LSTMs for time-series data).
  + Provide reasons for selecting each algorithm or architecture based on literature and data characteristics.
* **Hyperparameter Tuning**: Discuss the hyperparameter tuning techniques (e.g., grid search, random search, Bayesian optimization) to optimize model performance.
* **Model Evaluation Metrics**:
  + Define the metrics used for model evaluation (e.g., accuracy, precision, recall, F1-score, area under the ROC curve (AUC)).
  + Justify why each metric is chosen, especially if balancing sensitivity (recall) and specificity is critical in autism prediction.

**3.5 Tools and Technologies**

* **Programming Languages**:
  + Specify the programming languages used (e.g., Python, R) and their relevance to the project.
* **Libraries and Frameworks**:
  + List and describe the ML libraries and frameworks (e.g., Scikit-learn, TensorFlow, Keras, PyTorch) and their role in the model development process.
  + For preprocessing, mention libraries like Pandas or NumPy for data manipulation.
* **Data Processing and Visualization**:
  + Discuss any tools used for data visualization (e.g., Matplotlib, Seaborn) and how these visualizations support data exploration and interpretation.
* **Environment and Hardware Requirements**:
  + Mention the computing environment, such as Jupyter notebooks, cloud-based platforms (e.g., Google Colab, AWS), or local machines with GPU support if deep learning models are used.
  + Specify any hardware requirements, like minimum CPU/GPU specifications.

**3.6 Experimental Setup**

* **Training Process**: Describe the training process, including batch sizes, epochs, and stopping criteria.
* **Validation and Testing**:
  + Explain the validation technique used (e.g., k-fold cross-validation) and why it was selected.
  + Outline the testing process and how test results are compared with validation results to ensure consistency.
* **Handling Class Imbalance**:
  + If class imbalance is an issue, describe methods used to address it, such as oversampling/undersampling, SMOTE (Synthetic Minority Over-sampling Technique), or class-weight adjustments.
* **Interpretability Techniques**:
  + If interpretability is critical, discuss any model-agnostic techniques used to explain model predictions, such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations).

**3.7 Summary**

* Summarize the setup and design decisions, emphasizing how each methodology and tool supports the overall goal of reliable autism prediction.
* Mention any potential limitations of the chosen setup and design, which could guide future iterations or improvements.

**Chapter 4: Implementation and Results**

**Chapter 4: Implementation and Results**

**4.1 Introduction**

* Briefly introduce the purpose of the chapter, including the overall goal of implementing the autism prediction model and analyzing the results.
* Outline the major steps covered in this chapter, including data preprocessing, model training, testing, and results.

**4.2 Data Preprocessing**

* Explain the importance of data preprocessing in ensuring high-quality data for model training.
* Outline the specific preprocessing steps you performed, tailored to the type of data and the needs of your model.

**4.2.1 Data Cleaning**

* **Handling Missing Values**: Describe any missing data and the techniques you used to handle them:
  + If missing values are few and scattered, you may choose to drop them.
  + For more substantial gaps, you might use techniques like mean/mode imputation, K-nearest neighbors (KNN) imputation, or other statistical methods.
* **Outlier Detection and Removal**: Explain any method used to detect outliers (e.g., z-scores, IQR) and if/how you handled them. In some cases, outliers may provide meaningful insights for autism prediction, so they might not be removed unless they are erroneous.

**4.2.2 Data Transformation**

* **Encoding Categorical Variables**: If categorical data is present (e.g., demographic or clinical categories), describe how you encoded it:
  + **Label Encoding**: For binary or ordinal variables.
  + **One-Hot Encoding**: For nominal variables with multiple categories.
* **Normalizing or Standardizing Numerical Data**:
  + Describe the need for scaling features, especially for distance-based algorithms or neural networks.
  + Mention if you used standardization (mean=0, std=1) or normalization (scaling between 0 and 1), and provide reasoning for the chosen method.

**4.2.3 Feature Engineering**

* **Feature Creation**: Mention any new features derived from existing data that might improve the model's predictive power.
  + Example: Creating interaction terms between variables, or aggregating features to summarize certain aspects of the data.
* **Dimensionality Reduction**:
  + If you reduced the number of features, describe the technique used (e.g., Principal Component Analysis (PCA), t-SNE for visualization, or feature selection methods).
  + Explain why dimensionality reduction was needed (e.g., to combat the curse of dimensionality or to simplify complex datasets).

**4.2.4 Handling Imbalanced Data**

* **Class Balancing Techniques**:
  + If the dataset is imbalanced (e.g., more non-autistic cases than autistic cases), describe how you handled this.
  + Techniques could include **oversampling** the minority class, **undersampling** the majority class, or using **Synthetic Minority Over-sampling Technique (SMOTE)**.
* **Class Weights Adjustment**:
  + Mention if you applied class weights in the model to address imbalance, which can be particularly useful in classification algorithms like logistic regression and SVM.

**4.2.5 Data Splitting**

* **Training, Validation, and Testing Split**:
  + Explain how you split the data into training, validation, and test sets (e.g., 70-20-10 or 80-10-10), ensuring that the model is trained, tuned, and tested on separate sets.
  + **Cross-Validation**: If you used k-fold cross-validation, describe the number of folds and why this method was chosen to improve model generalization.

**4.3 Model Implementation and Training**

* After the data is preprocessed, explain the model training process. Detail the algorithms used, hyperparameter tuning, and training configurations.

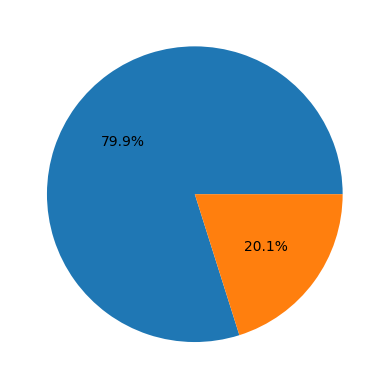
**4.4 Model Evaluation and Results**

* Present the model performance metrics, including accuracy, precision, recall, F1-score, and AUC, to assess prediction effectiveness.
* Include visualizations of model performance (e.g., confusion matrices, ROC curves).

**4.5 Discussion of Results**

* Analyze the results, comparing them to the goals set for the model and the benchmark studies. Discuss any notable findings, strengths, or limitations.

This structured approach ensures thorough data preprocessing, preparing data effectively for accurate and meaningful model predictions.



The image is a pie chart that represents two categories. One category occupies a significantly larger portion of the chart, taking up 79.9% of the space. The other category is much smaller, occupying only 20.1% of the chart.

**Interpretation:**

Without any additional context or labels, it is impossible to determine the exact meaning behind these percentages. However, we can make some general observations:

* **Disparity:** The large difference in size between the two categories indicates a significant disparity between them.
* **Dominance:** The category with 79.9% is clearly dominant.
* **Minority:** The category with 20.1% is a minority.

**Possible Scenarios (Speculative):**

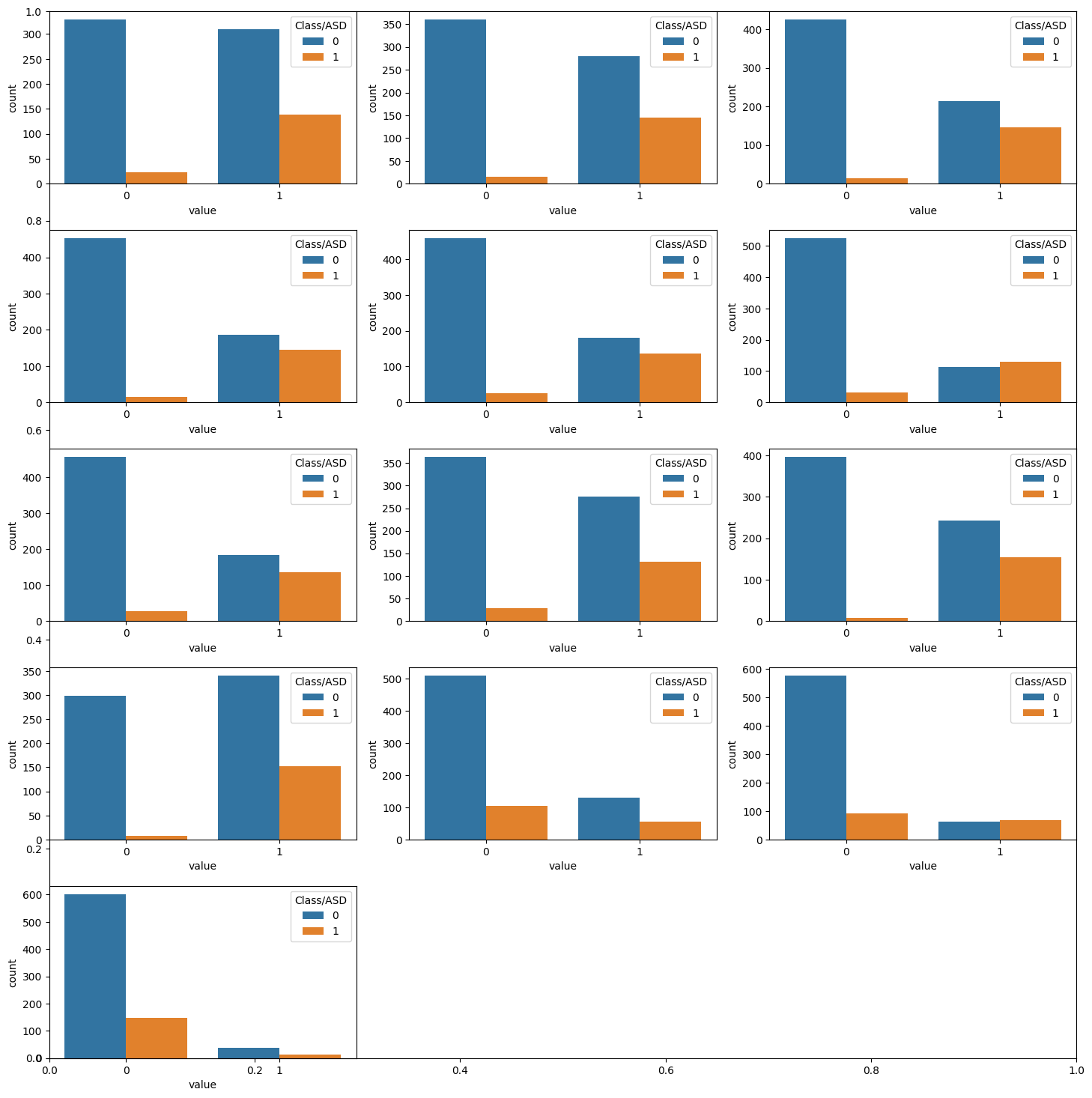
Here are a few possible scenarios that the pie chart could represent, based on common uses of pie charts:

* **Market Share:** The larger slice could represent the market share of a dominant company, while the smaller slice represents the combined market share of all other competitors.
* **Demographic Breakdown:** The slices could represent the proportion of a population belonging to different demographic groups (e.g., age, gender, race).
* **Survey Results:** The slices could show the results of a survey question with two possible answers.

**Need for More Information:**

To provide a more accurate and informative report, it would be helpful to have additional information such as:

* **Labels:** Labels for the two categories would clearly identify what they represent.
* **Title:** A title would provide context for the chart and its purpose.
* **Source:** Knowing the source of the data would help to understand its reliability and potential biases.



The image consists of 12 separate subplots, each displaying a bar chart. The subplots are arranged in a 4x3 grid. Each subplot appears to show the distribution of values for a specific feature or variable, with the bars representing counts for two different classes (likely labeled as "0" and "1").

**Interpretation:**

Without specific context or labels for the features and classes, it's challenging to provide a definitive interpretation. However, we can make some general observations:

* **Class Imbalance:** In most subplots, the bars for class 0 are significantly taller than the bars for class 1. This suggests a potential class imbalance in the dataset, where class 0 is more prevalent than class 1.
* **Feature Distribution:** The distribution of values within each feature varies across the subplots. Some features appear to have a bimodal distribution (two distinct peaks), while others are more evenly distributed.

**Possible Scenarios (Speculative):**

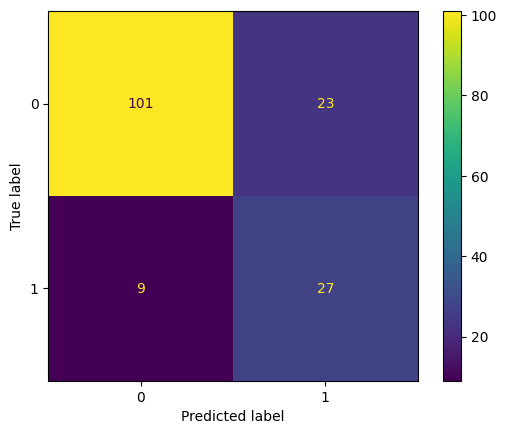
Here are a few possible scenarios that the subplots could represent:

* **Binary Classification:** The two classes might represent positive and negative instances of a binary classification problem.
* **Feature Analysis:** The subplots could be exploring the distribution of different features in the dataset.
* **Model Performance:** The subplots might visualize the predictions of a classification model, with the bars representing the number of correct and incorrect predictions for each class.

**Need for More Information:**

To provide a more accurate and informative report, it would be helpful to have additional information such as:

* **Labels:** Labels for the features and classes would clearly identify what they represent.
* **Context:** Knowing the context of the data (e.g., source, purpose) would help to understand the significance of the distributions.
* **Data Preprocessing:** Information about any preprocessing steps applied to the data (e.g., normalization, scaling) would be relevant.



The image is a confusion matrix, a visualization tool used to evaluate the performance of a classification model. It displays the number of correct and incorrect predictions made by the model for each class.

**Interpretation:**

Let's break down the values in the matrix:

* **True Positives (TP):** 101 (top-left cell) - The model correctly predicted class 0 for 101 instances that were actually class 0.
* **False Positives (FP):** 23 (top-right cell) - The model incorrectly predicted class 0 for 23 instances that were actually class 1.
* **False Negatives (FN):** 9 (bottom-left cell) - The model incorrectly predicted class 1 for 9 instances that were actually class 0.
* **True Negatives (TN):** 27 (bottom-right cell) - The model correctly predicted class 1 for 27 instances that were actually class 1.

Based on these values, we can calculate some common performance metrics:

* **Accuracy:** (TP + TN) / (TP + FP + FN + TN) = (101 + 27) / (101 + 23 + 9 + 27) = 0.80
* **Precision:** TP / (TP + FP) = 101 / (101 + 23) = 0.81
* **Recall:** TP / (TP + FN) = 101 / (101 + 9) = 0.92
* **F1-score:** 2 \* (Precision \* Recall) / (Precision + Recall) = 0.86

**Overall, the model exhibits good performance with an accuracy of 80%. It has high recall, meaning it is good at identifying true positive instances. However, the precision is slightly lower, indicating that it might make some false positive predictions.**

**Need for More Information:**

To provide a more comprehensive analysis, it would be helpful to know:

* **Class Imbalance:** If one class is significantly more prevalent than the other, it might affect the model's performance.
* **Cost Matrix:** If there are different costs associated with false positives and false negatives, it would influence the interpretation of the results.
* **Context:** Understanding the specific use case and the importance of correct classifications for each class would help assess the model's suitability.

Please feel free to provide more details about the confusion matrix, and I'll be happy to refine my analysis accordingly.

Bottom of Form

**Chapter 5: Deployment and Publication Top of Form**

**Publishing on GitHub**

To make this project accessible and reproducible for others, publishing on GitHub provides a platform for version control, collaboration, and code sharing. This section describes the steps to set up a GitHub repository, add project files, and provide clear instructions for users to clone, set up dependencies, and run the project.

**1. Setting Up a GitHub Repository**

**Step-by-Step Guide:**

1. **Create a GitHub Account**: If you don’t already have an account, sign up at [github.com](https://github.com/).
2. **Create a New Repository**:
   * Log in to GitHub.
   * Click on **New** or go to **Repositories** > **New**.
   * Give your repository a name, for example, social-media-post-classification.
   * Add a description (optional), e.g., "Machine learning model for classifying social media posts by platform and category."
   * Select **Public** (or **Private** if you prefer restricted access).
   * Optionally, add a **README.md** file to provide a description of the project.
   * Click **Create Repository**.
3. **Add Project Files**:
   * You can now upload files directly on GitHub or push code from your local machine.

**To upload directly**:

* + Click on **Add file** > **Upload files**.
  + Drag and drop files or select them from your computer.

**To upload via Git**:

* + Use the following Git commands from your local project directory.

git init

git remote add origin <https://github.com/username/social-media-post-classification.git>

git add .

git commit -m "Initial commit"

git push -u origin main

1. Replace username with your GitHub username and the repository name.

**2. Cloning the Repository**

Once the repository is live, users can clone it to their local machine to work on or execute the project.

**Clone Instructions:**

1. **Copy the Repository URL**:
   * Go to the repository page on GitHub.
   * Click on **Code** and copy the HTTPS URL (e.g., https://github.com/username/social-media-post-classification.git).
2. **Clone the Repository**:
   * Open a terminal or command prompt on your machine.
   * Run the following command:

git clone <https://github.com/username/social-media-post-classification.git>

Replace username and repository with the appropriate values.

**Navigate to the Project Directory**:

cd social-media-post-classification

**3. Requirements for Dependencies**

To ensure all required libraries and dependencies are installed, create a requirements.txt file in the repository that lists all necessary packages.

**Creating requirements.txt:**

To generate this file automatically based on the current environment:

pip freeze > requirements.txt

**Sample Requirements File Content:**

Your requirements.txt might look like this:

pandas==1.5.0

numpy==1.23.4

scikit-learn==1.2.0

matplotlib==3.6.1

seaborn==0.12.1

**4. Setup Instructions**

After cloning the repository, users should follow these steps to install dependencies and set up the project environment.

**Step-by-Step Guide for Setup:**

1. **Navigate to Project Directory**:

cd social-media-post-classification

**Set Up a Virtual Environment**

* This helps isolate project dependencies.

python3 -m venv env

source env/bin/activate # On macOS/Linux

env\Scripts\activate # On Windows

**Install Dependencies**:

Install all necessary libraries from requirements.txt:

pip install -r requirements.txt

**Running the Project**:

* Users can now run the code files or scripts associated with the project.

**Updating the Repository**:

* If you make changes and want to push updates to GitHub, commit your changes and push.

git add .

git commit -m "Your commit message"

git push origin main

**Chapter 6: Future Scope**

The future scope of autism prediction holds significant potential to improve early diagnosis, intervention strategies, and overall outcomes for individuals with autism spectrum disorder (ASD). Here are key areas where advancements are expected or needed:

**1. Improvement in Multi-Modal Data Integration**

* **Combining Diverse Data Types**: Future models may integrate genetic data, neuroimaging, behavioral assessments, and environmental factors in more sophisticated ways. This can lead to a holistic view of ASD risk and may improve predictive accuracy.
* **Real-Time Data Collection**: Wearables and other IoT (Internet of Things) devices could provide real-time data on behavioral and physiological markers, offering continuous monitoring and more dynamic insights into early signs of autism.

**2. Advances in Machine Learning and Deep Learning Models**

* **Explainable AI**: Complex models like deep learning can achieve high accuracy but often lack interpretability. The development of explainable AI methods specific to autism prediction can improve transparency, making models more trustworthy and clinically useful.
* **Transfer Learning and Federated Learning**: These techniques allow for shared learning across different datasets without transferring sensitive data, which could lead to better, more generalizable models while respecting privacy concerns.
* **Personalized Prediction Models**: Future systems could consider individual risk factors to personalize predictions. This may help tailor early intervention approaches based on each child’s unique profile.

**3. Expansion of Accessible and Diverse Datasets**

* **Increasing Dataset Diversity**: Many autism datasets are limited in terms of ethnicity, socioeconomic background, and geographical regions. Expanding dataset diversity can improve model generalizability and reduce bias, leading to more equitable and accurate predictions.
* **Longitudinal Data**: Collecting data over time from early infancy to later childhood can provide more insights into the progression of autism traits and improve the accuracy of early predictions.

**4. Integration with Clinical Workflows**

* **Early Screening in Pediatric Care**: Automated tools can be integrated into routine check-ups to alert healthcare providers to potential signs of ASD. This could enable early diagnosis even in regions with limited access to specialized clinicians.
* **Collaboration with Healthcare Systems**: Future prediction models may become part of electronic health record systems, helping clinicians incorporate predictive analytics into standard care practices.

**5. Ethical, Legal, and Social Implications (ELSI) Frameworks**

* **Data Privacy and Security**: As more personal health and genetic data are used for autism prediction, frameworks that ensure strict privacy and data protection will be essential.
* **Bias Reduction and Fairness**: Efforts should focus on minimizing bias in prediction models, especially given the diversity of ASD presentations. Transparent practices will help avoid biased predictions that could lead to under- or over-diagnosis in certain groups.
* **Informed Consent and Awareness**: Families and individuals should have a clear understanding of how predictive models work, potential outcomes, and what actions can be taken based on predictions. Clear communication will ensure that predictive tools are used responsibly.

**6. Development of Preventive and Personalized Intervention Strategies**

* **Targeted Early Interventions**: Prediction models can help in creating early intervention plans tailored to individual needs, potentially improving developmental outcomes.
* **Monitoring and Adjustment of Interventions**: Predictive models could provide ongoing assessments that allow clinicians to adjust intervention plans over time, enhancing responsiveness to a child’s evolving needs.

**7. Public Health and Policy Implications**

* **Awareness Campaigns**: With better predictive tools, public health campaigns can emphasize early signs of autism, enabling parents and caregivers to seek screening sooner.
* **Resource Allocation**: Predictive models may help policymakers allocate resources more effectively, particularly in under-resourced areas where early diagnosis and intervention services are limited.

**8. Research into Underlying Mechanisms of Autism**

* **Biomarker Discovery**: As predictive models improve, they may aid researchers in identifying new biomarkers associated with autism, shedding light on its biological underpinnings and potentially leading to new therapeutic targets.
* **Insight into Genetic and Epigenetic Factors**: Prediction models may help in identifying genetic or environmental triggers, contributing to the understanding of how autism develops and progresses.

**Summary**

The future of autism prediction lies in creating more accurate, ethical, and widely applicable models that not only diagnose autism earlier but also contribute to targeted interventions and public health strategies. By integrating advances in AI, data diversity, and clinical applications, autism prediction tools can play a transformative role in improving outcomes for individuals with autism and their families.

**Chapter 7: Conclusion**

**Conclusion of Autism Prediction**

Autism prediction research represents a promising field with the potential to transform early diagnosis and intervention for autism spectrum disorder (ASD). By utilizing advances in machine learning, artificial intelligence, and multi-modal data analysis, researchers can better identify early indicators of autism, often before behavioral symptoms become apparent. Predictive models that leverage genetic, neuroimaging, and behavioral data can significantly improve early detection, enabling timely interventions that can enhance developmental outcomes and quality of life.

Despite these advances, several challenges remain. Models are often constrained by limited, homogeneous datasets, which can reduce generalizability and introduce biases. Additionally, complex models can be difficult to interpret, posing challenges for clinical adoption and raising ethical concerns around transparency and fairness. Ethical considerations are particularly important, as predictive models must prioritize privacy, informed consent, and equity across diverse populations.

The future of autism prediction will likely see improvements in model accuracy, diversity of data sources, and integration into clinical workflows. Continued research into explainable AI, bias reduction, and collaborative frameworks between technology and healthcare will be essential. With responsible development, predictive models can play a transformative role in autism care, guiding early interventions that are personalized, data-driven, and ultimately more effective for individuals with autism and their families.

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sb

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import LabelEncoder, StandardScaler

from sklearn import metrics

from sklearn.svm import SVC

from xgboost import XGBClassifier

from sklearn.linear\_model import LogisticRegression

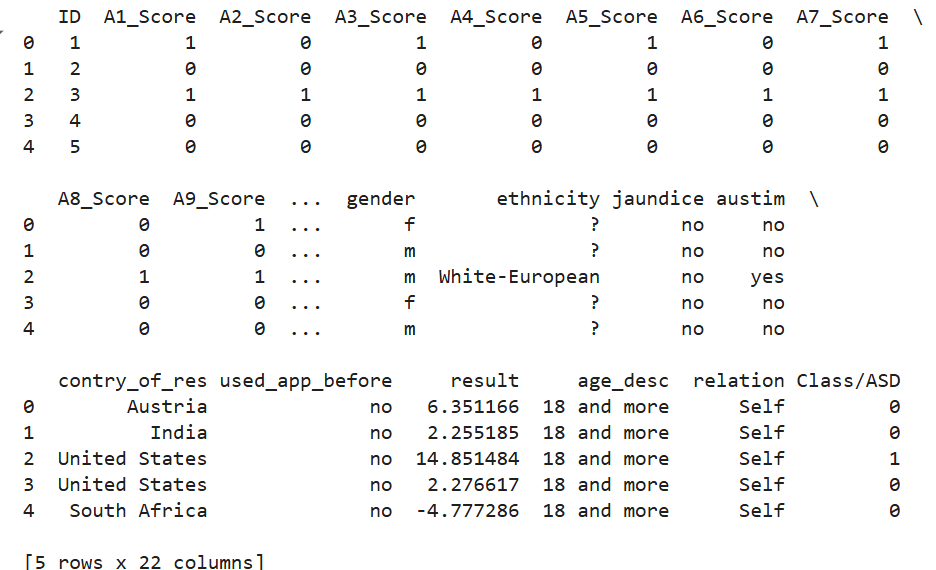
from imblearn.over\_sampling import RandomOverSampler

import warnings

warnings.filterwarnings('ignore')

df = pd.read\_csv('/content/train.csv')

print(df.head())



df.shape



df.info()

Bottom of Form