**1. Introduction**

* 1. **Background on Autism Spectrum Disorder (ASD)**

Autism Spectrum Disorder (ASD) is a neurological condition that affects how a person thinks, communicates, interacts socially, and processes sensory information. The term "spectrum" reflects the wide range of characteristics and abilities individuals with autism may have. While some may face significant challenges in their day-to-day existence, others may have exceptional skills and live independently.

ASD usually starts to show up in early childhood, and by the time a child is two or three years old, there are often clear signs. Some common signs are trouble communicating verbally and nonverbally, doing the same things over and over, having few interests, and having trouble getting along with others. It's important to understand and help people with ASD because they are all different.

Early diagnosis is very important. If kids with ASD get the right kind of help and support, like occupational therapy, behavioural therapy, speech therapy, or educational support, they are more likely to learn important skills and become independent over time. Support is still very important in adolescence and adulthood; it helps people with their mental health, careers, and social lives.

Although the exact cause of ASD is still unknown, researchers think that a combination of environmental and genetic factors contributes to the disorder. Parenting styles or individual decisions are not to blame. As ASD has become more widely recognised and diagnosed in the past few decades, there are now more resources and support networks available.

Early assistance in the development of vital life skills is very beneficial for children with ASD. Personalised treatment plans and educational support can make a significant difference. Importantly, the increasing acceptance of the abilities and contributions of people on the spectrum is promoting an open culture that values neurodiversity. Acceptance and inclusion are becoming increasingly important in today's society. Numerous communities and advocacy organisations are trying to change the conversation away from "curing" autism and towards embracing neurodiversity. Individuals with ASD can contribute unique perspectives, innovative concepts, and useful abilities, and they can lead fulfilling lives in the correct setting.

In conclusion, autism spectrum disorder is a collection of differences that impact how individuals view and engage with the world instead of a single illness. People with ASD can succeed in every aspect of life if they are understood, accepted, and given the proper support. Making the world more inclusive and supportive begins with understanding ASD. We can all help people on the autism spectrum live better lives if we have patience, empathy, and knowledge.

**1.2 Importance of Early Detection**

Early detection of Autism Spectrum Disorder (ASD) is one of the most critical factors in ensuring a better quality of life for individuals on the spectrum. Identifying signs of autism at an early age, typically between 18 months and 3 years, allows for timely intervention, which is essential during the critical period of brain development. Children can get the help they require during the most critical phases of brain development if the early symptoms of autism, such as delayed speech, limited eye contact, or repetitive behaviours, are identified.

During early childhood, the brain is especially adaptable. A child's brain is most capable to learn new behaviours and skills during this time, which is frequently referred to as the "critical window." If ASD is identified early, targeted therapies can begin when they are most effective. According to research, learning, social interactions, and communication skills can all significantly improve with early intervention.

Additionally, early detection gives carers and parents more power. It allows them to learn how to support their child's special requirements, access the right services, and understand their child's condition. Children with ASD may not receive the necessary support if they are not diagnosed early, and parents may feel overwhelmed or confused if they are unaware of the causes of their child's behaviour.

Schools, medical professionals, and therapists all gain from early identification. Early diagnosis allows professionals to create specialised education and support plans for the child. The child benefits from these specialised methods in both educational and social environments.

It's also critical to remember that early detection decreases long-term difficulties. Early support increases a child's likelihood of becoming self-sufficient, confident, and emotionally resilient. As they age, they might also need less intensive support.

In conclusion, early identification of ASD is transformative rather than just helpful. It creates the basis for a happy life, boosts the likelihood of positive development, and provides access to early support.

Early identification of ASD can alter a child's entire developmental trajectory in addition to providing a head start. It makes it possible for professionals and families to interact right away, setting the foundation for development, education, and deep connections with society.

* 1. **Motivation for Using Machine Learning**

Machine learning (ML) has emerged as an increasingly essential asset in the healthcare sector, enabling the analysis of extensive and intricate datasets, uncovering concealed patterns, and making highly accurate predictions. In relation to Autism Spectrum Disorder (ASD), ML offers a distinctive opportunity to revolutionize the conventional diagnostic approach, which is frequently subjective, labour-intensive, and time-consuming.

One significant reason for implementing ML in ASD detection is its capacity to facilitate early screening through automation and standardization. Traditional techniques heavily depend on expert interpretation of behavioural symptoms, which can differ significantly among clinicians. In contrast, machine learning can analyze objective data—such as responses to questionnaires, developmental history, and even video or audio signals—to produce predictions based on learned patterns from previously classified cases.

ML models, including decision trees, support vector machines (SVM), random forests, and deep learning networks, can be trained on labelled ASD datasets to differentiate between individuals with ASD and those without. These models can also pinpoint which characteristics (e.g., lack of eye contact, delayed speech) are most indicative of an ASD diagnosis, thereby enhancing the refinement of current screening tools.

The scalability of ML represents another significant benefit. Once trained and validated, these models can be integrated into mobile applications, online platforms, or diagnostic software, thereby making screening more accessible in under-resourced or rural regions. Additionally, ML models function efficiently, providing predictions within seconds and alleviating the workload of healthcare professionals.

Moreover, ML paves the way for personalized diagnostics by continuously learning from new data and adjusting to specific population traits. As more data becomes available, models can be updated to reflect evolving diagnostic trends and enhance generalizability across diverse populations.

In conclusion, the rationale for employing machine learning in ASD detection is rooted in its ability to support early and accurate diagnosis.

**1.4 Project Objectives**

The main aim of this initiative is to utilize machine learning algorithms for the early identification of Autism Spectrum Disorder through the analysis of real-world datasets. This encompasses the design, implementation, and assessment of predictive models capable of categorizing individuals as either likely or unlikely to have ASD based on behavioural, demographic, or developmental characteristics.

The project is directed by the following specific goals:

* To gather and pre-process dependable ASD datasets that encompasses pertinent features such as age, gender, responses to screening questionnaires, and developmental history.
* To conduct exploratory data analysis (EDA) to comprehend the structure, distribution, and correlations within the data, which will guide feature selection and model development.
* To train and evaluate various machine learning algorithms, including logistic regression, decision trees, random forests, and support vector machines, and to compare their effectiveness using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.
* To pinpoint the most critical predictive features that influence ASD classification, which may assist in enhancing current screening tools or inform the creation of new ones.
* To assess the practical implications and constraints of employing machine learning in a real-world clinical environment, taking into account ethical considerations, data privacy, and interpretability.

By fulfilling these goals, the project aims to advance the integration of artificial intelligence into healthcare, particularly in domains that can benefit from early intervention and improved accessibility, such as autism diagnosis.

**1.5 Organization of the Report**

This document is organized to deliver a thorough overview of the project, encompassing its conceptualization, implementation, and evaluation phases. Each section builds on the preceding one to create a unified narrative regarding the utilization of machine learning for the early detection of Autism Spectrum Disorder (ASD).

NEEDS CHANGE

Chapter 2: Literature Review

This section examines existing literature on Autism Spectrum Disorder, focusing on current diagnostic methods and the application of machine learning in the healthcare sector. It emphasizes previous research that utilized machine learning for ASD prediction, analyzing their findings, methodologies, and limitations.

Chapter 3: Methodology

This section outlines the procedures undertaken to complete the project. It includes a description of the dataset, preprocessing techniques, feature selection, model selection, and the justification for the chosen machine learning methods. Additionally, it details the tools and frameworks employed during implementation.

Chapter 4: Results and Discussion

This section showcases the results of model training and testing, featuring performance metrics and a comparative analysis of various algorithms. It also explores the implications of the findings, potential real-world applications, and insights derived from the analysis of feature importance.

Chapter 5: Conclusion and Future Work

The concluding section summarises the project's accomplishments and limitations. It reflects on the project's contribution to the wider field of ASD diagnostics and proposes future research directions, such as the integration of deep learning, multi-modal data, or real-time screening tools.

The report concludes with a list of references and relevant appendices, including code snippets, supplementary graphs, or extended evaluation tables, to ensure complete transparency and reproducibility.

**5. Experimental Results and Analysis**

**5.1 Evaluation Metrics**

Evaluation metrics are used to assess the performance of trained models. We employed the following evaluation metrics on our models.

1. **Accuracy**: Accuracy measures the ratio of correct predictions made by the model to the total number of predictions. It gives us a general idea of how often the model is correct, but can be unreliable for unbalanced datasets.
2. **Precision**: Precision measures the accuracy of positive predictions. It is the ratio of true positive predictions to the total number of positive predictions that was made by the model. Models having a high precision have a very low rate false positives.
3. **Recall (or Sensitivity)**: Recall measures the ability of a model to correctly identify all relevant positive cases. It is the ratio of correctly predicted positive instances to the total actual positive instances. A high recall indicates a low false negative rate.
4. **F1-Score**: F1 Score is the harmonic mean of precision and recall. Its value lies in the range of 0 and 1, both included. It balances the trade-off between precision and recall and tells us how precise (correctly classifies how many instances) and robust (doesn’t miss any significant number of instances) our classifier is. It is particularly useful when the class distribution is imbalanced.
5. **Confusion Matrix**: Confusion Matrix is the visual representation of the relation between actual values and predicted values.

* **True Positive**: The values which were actually positive and were predicted positive.
* **False Positive**: The values which were actually negative but falsely predicted as positive. Also known as Type I Error.
* **False Negative**: The values which were actually positive but falsely predicted as negative. Also known as Type II Error.
* **True Negative**: The values which were actually negative and were predicted negative.

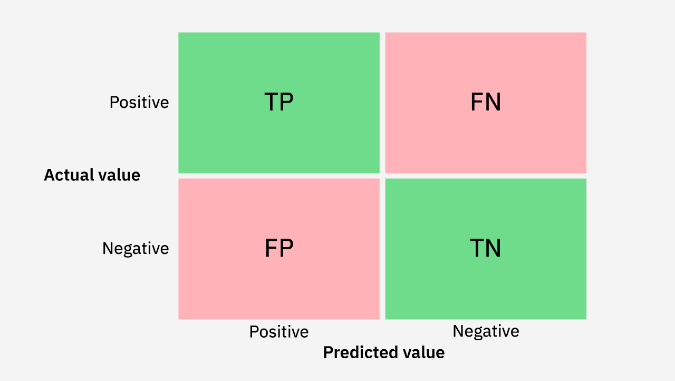


Figure 1: Confusion Matrix (<https://developer.ibm.com/tutorials/awb-confusion-matrix-r/>)

**5.2 Results on Text-Based Questionnaire Dataset**

The text-based dataset used in our project consists of responses to standard Autism Spectrum Disorder (ASD) screening questionnaire. This data was preprocessed and used to train six different machine learning models: Logistic Regression, Support Vector Machines (SVM), Decision Tree, Random Forest, k-Nearest Neighbor, and Naïve Bayes. The dataset was split in training and testing samples using 80:20 ratio.

**5.2.1 Performance Comparison**

The following table summarizes the performance metrics of the six models:

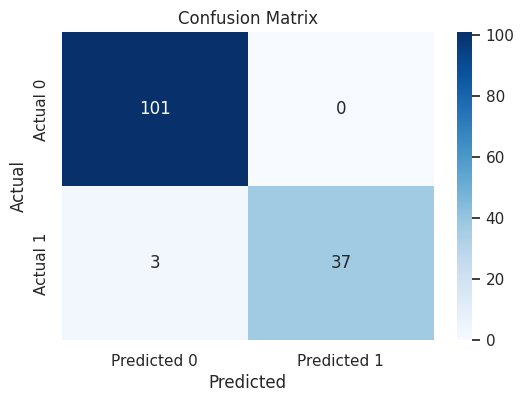
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| Decision Trees | 0.957 | 0.925 | 0.925 | 0.925 |
| Random Forest | 0.900 | 1.000 | 0.650 | 0.787 |
| Support Vector Machines | 0.978 | 1.000 | 0.925 | 0.961 |
| K-Nearest Neighbors | 0.950 | 0.923 | 0.900 | 0.911 |
| Naïve Bayes | 0.950 | 0.923 | 0.900 | 0.911 |
| Logistic Regression | 0.992 | 1.000 | 0.975 | 0.987 |

**5.2.2 Confusion Matrices**

A confusion matrix gives us a detailed analysis of a model's classification performance, displaying the number of correct and incorrect predictions made on each class. It includes four key components:

* **True Positives (TP)**: Correctly predicted ASD cases
* **True Negatives (TN)**: Correctly predicted non-ASD cases
* **False Positives (FP)**: Non-ASD cases incorrectly classified as ASD
* **False Negatives (FN)**: ASD cases incorrectly classified as non-ASD

Below are the confusion matrices for the two best performing models i.e. Logistic Regression and Support Vector Machines (SVM)-



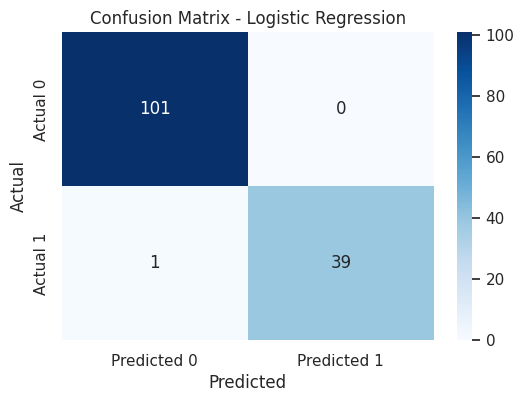


Figure 2: Confusion Matrix for Support Vector Machines

Figure 2: Confusion Matrix for Logistic Regression

**Interpretation**:

* Logistic Regression achieved a nearly perfect classification, with only 1 false positive and 0 false negatives.
* SVM also performed strongly, with three false positives but no false negatives.
* Both of these models, showed high recall and precision which makes them reliable for ASD detection.

**5.2.3 Discussion**

* With the best accuracy and F1-score, Logistic Regression showed outstanding performance and reliability.
* SVM was also quite successful, particularly when it came to high recall and perfect precision.
* Recall was marginally lower for Decision Trees and k-NN than for the best performers, but they still performed admirably.
* Despite having excellent precision, Random Forest missed a lot of good cases because of its much weaker recall.
* Although it performed rather well, Naïve Bayes trailed somewhat behind the other models.

**5.3 Results on ABIDE Dataset**

The Autism Brain Imaging Data Exchange (ABIDE) dataset consists of resting-state functional MRI (fMRI) scans from people with and without ASD. In this project, we’ve only worked with data from the **NYU site** to maintain uniform imaging protocols and reduce inter-site variability.

Data was fetched using the “fetch\_abide\_pcp” function with the **‘rois\_aal’ derivative**, which extracts mean regional time-series features from brain regions defined by the **AAL (Automated Anatomical Labelling)** atlas. This approach reduces the high dimensionality of raw fMRI data by computing average signals from predefined anatomical regions.

**5.3.1 Preprocessing**

Since the data was already pre-processed by the Pre-processed Connectomes Project (PCP), standard pre-processing steps like motion correction, skull stripping, normalization, and nuisance regression had already been applied. The AAL-based ROI features were then directly used for training machine learning models.

**5.3.2 Performance Comparison**

Three classical machine learning models were evaluated on the extracted ROI-based features:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| Logistic Regression | 0.771 | 0.705 | 0.8 | 0.75 |
| Support Vector Machines (SVM) | 0.657 | 0.666 | 0.736 | 0.700 |
| k-Nearest Neighbors (KNN) | 0.628 | 0.571 | 0.285 | 0.380 |

**5.3.3 Confusion Matrix**

Below is the confusion matrix for the model which performed the best i.e. Logistic Regression.

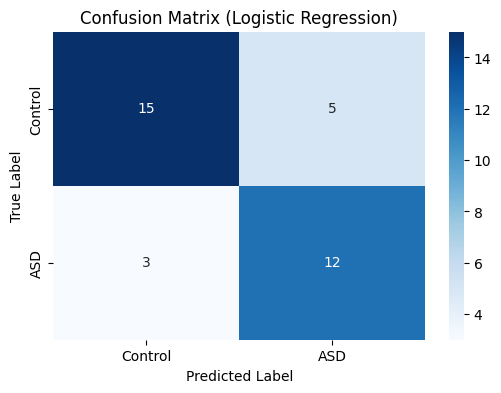


Figure 2: Confusion Matrix for Logistic Regression Model on the ABIDE dataset

**5.3.4 Discussion**

* **Logistic Regression** performed the best among the three models, with an accuracy of 77.1%, a high recall of 0.800, and an F1-score of 0.750. This suggests that Logistic Regression was able to identify a significant proportion of ASD cases, although its precision is slightly lower compared to its recall.
* **SVM** exhibited lower overall performance, with an accuracy of 65.7% and an F1-score of 0.700. While it performed reasonably well, it had lower precision and recall than Logistic Regression, indicating that it may have been more conservative in predicting ASD cases, potentially leading to more false positives and negatives.
* **KNN** performed the worst among the three models, with an accuracy of 62.8%, precision of 0.571, recall of 0.285, and an F1-score of 0.380. The low recall indicates that KNN struggled significantly to identify positive ASD cases, leading to many false negatives.

These results indicate that **Logistic Regression** is the most reliable model for ASD detection on the ABIDE dataset, though the overall performance of the models was lower compared to the text-based dataset. This highlights the challenges associated with working with neuroimaging data, which is complex and high-dimensional. Pre-processing and feature engineering are critical factors that can influence model performance in this domain.

**5.3.5 Challenges**

* **Pre-processing Complexity**: Neuroimaging data require extensive pre-processing including normalization, brain region extraction, and feature selection.
* **Data Heterogeneity**: Differences in scanner types, imaging protocols, and demographics across acquisition sites introduce variability.
* **Sample Size**: Despite the richness of each sample, the number of subjects remains limited, making generalization difficult.