**Research Proposal: Identifying Factors Influencing the Occurrence of Autism Spectrum Disorder (ASD) in Children Using R and Machine Learning**

**1. Introduction**

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by challenges in social communication, interaction, and repetitive behaviors. Despite extensive research, the specific factors contributing to the development of ASD remain unclear. This study aims to identify and analyze the factors influencing the occurrence of ASD in children by leveraging data from the National Survey of Children’s Health (NSCH) and applying various machine learning techniques using R.

**2. Objectives**

The primary objective of this study is to identify key factors that correlate with an increased risk of ASD in children. The specific objectives include:

* To explore the socio-economic, prenatal, perinatal, environmental, and familial factors associated with ASD.
* To develop predictive models that can accurately estimate the risk of ASD based on these factors.
* To use statistical and machine learning techniques to rank the importance of these factors.

**3. Research Questions**

* What socio-economic factors are associated with an increased risk of ASD in children?
* How do prenatal and perinatal factors influence the likelihood of ASD?
* What environmental factors are significantly correlated with ASD development?
* How does family history of neurodevelopmental disorders contribute to the risk of ASD?

**4.1 Data Source**

This study will utilize data from **the National Survey of Children’s Health (NSCH)**, which includes comprehensive information on various factors such as socio-economic status, prenatal and perinatal conditions, family health history, environmental exposures, and child health outcomes.

**4.2 Data Preprocessing**

* **Data Cleaning:** Handle missing values using appropriate imputation methods, remove duplicates, and standardize the data formats.
* **Feature Engineering:** Convert categorical variables to numerical values where necessary (e.g., one-hot encoding for categorical data).
* **Data Splitting:** The dataset will be split into training (70%) and testing (30%) sets to validate the models' performance.

**4.3 Steps to Identify Correlated Factors**

1. **Data Cleaning and Preprocessing:**
   * **Purpose:** Ensure the dataset is ready for analysis by handling missing values, removing duplicates, and converting categorical variables to a numerical format if necessary.
   * **Tools in R:** na.omit(), mice, missForest, dummyVars() from the caret package.
2. **Exploratory Data Analysis (EDA):**
   * **Purpose:** Gain a preliminary understanding of the data by summarizing its main characteristics and identifying potential patterns or anomalies.
   * **Steps:**
     + **Descriptive Statistics:** Compute mean, median, standard deviation, etc., to summarize each variable.
     + **Visualization:** Use plots (e.g., scatter plots, box plots) to visually inspect relationships between variables and ASD diagnosis.
     + **Correlation Matrix:** Calculate the correlation coefficients between numeric variables to identify linear relationships.
   * **Tools in R:** summary(), cor(), ggplot2, corrplot.
3. **Statistical Testing:**
   * **Purpose:** Determine if relationships between variables are statistically significant.
   * **Steps:**
     + **Chi-Square Test:** For categorical variables to see if there's an association with ASD diagnosis.
     + **T-tests/ANOVA:** For continuous variables to compare means between groups (e.g., ASD vs. non-ASD).
   * **Tools in R:** chisq.test(), t.test(), aov().
4. **Feature Selection:**
   * **Purpose:** Identify the most relevant features (variables) that should be included in the predictive models.
   * **Steps:**
     + **Recursive Feature Elimination (RFE):** Iteratively remove less important features based on model performance.
     + **Lasso Regression:** Regularization technique that helps in feature selection by penalizing the coefficients of less important features.
   * **Tools in R:** caret::rfe, glmnet for Lasso regression.

**4.4 Machine Learning Models and Feature Selection**

The following machine learning models will be used, each incorporating a set of key factors identified through EDA and correlation analysis:

1. **Logistic Regression:**
   * **Features:** Parental age, maternal education level, household income, prenatal care quality, family history of neurodevelopmental disorders.
   * **Purpose:** Serve as a baseline model to understand linear relationships.
2. **Random Forest:**
   * **Features:** All significant variables from correlation analysis, including environmental exposures, access to healthcare, and socio-economic status.
   * **Purpose:** Capture non-linear interactions and provide feature importance ranking.
3. **Support Vector Machine (SVM):**
   * **Features:** Parental age, prenatal exposure to smoking/alcohol, birth complications, household income, family history.
   * **Purpose:** Classify data with an emphasis on maximizing the margin between ASD and non-ASD cases.
4. **Gradient Boosting Machines (GBM):**
   * **Features:** Key socio-economic, prenatal, perinatal, and environmental factors.
   * **Purpose:** Improve prediction accuracy through boosting weak learners.
5. **Neural Networks:**
   * **Features:** Comprehensive set of variables including socio-economic, prenatal, perinatal, environmental, and familial factors.
   * **Purpose:** Capture complex patterns and interactions among multiple variables.

**4.5 Model Training and Evaluation**

* **Cross-Validation:** Perform k-fold cross-validation on all models to prevent overfitting and validate model performance.
* **Evaluation Metrics:** Use accuracy, precision, recall, F1-score, and AUC-ROC to evaluate and compare model performance.
* **Permutation Importance and SHAP Values:** For advanced models like Random Forest and GBM, use these techniques to quantify the contribution of each feature to the model's predictions.

**4.6 Model Interpretation and Reporting**

* **Feature Importance Ranking:** Identify the most significant factors contributing to ASD risk based on model outputs.
* **Partial Dependence Plots:** Visualize the relationship between important features and ASD probability.
* **Interpretation of SHAP Values:** Provide a detailed understanding of how each feature influences individual predictions.

**5. Expected Outcomes**

* Identification of the most significant socio-economic, prenatal, perinatal, environmental, and familial factors influencing ASD risk.
* Development of accurate predictive models for estimating the risk of ASD in children.
* Insights that could guide public health interventions and inform future research aimed at early detection and prevention of ASD.

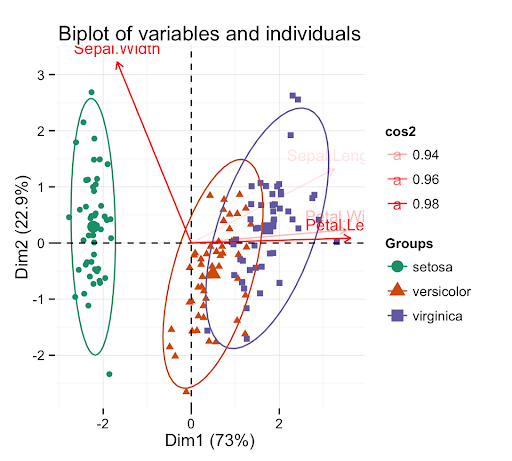
**Mock tables**

**1.Baseline Table: Comparison of Children with ASD vs. Non-ASD**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | ASD (n = X) | Non-ASD (n = Y) | p-value |
| Parental Age at Birth (years) | **Mean (SD)** | **Mean (SD)** | **p-value** |
| Maternal Education Level | **Median (IQR) or % in categories** | **Median (IQR) or % in categories** | **p-value** |
| Household Income (USD) | **Median (IQR)** | **Median (IQR)** | **p-value** |
| Birth Weight (grams) | **Mean (SD)** | **Mean (SD)** | **p-value** |
| Gestational Age (weeks) | **Mean (SD)** | **Mean (SD)** | **p-value** |
| Prenatal/Perinatal Complications | **% with complications** | **% with complications** | **p-value** |
| Maternal Smoking During Pregnancy | **% smoked** | **% smoked** | **p-value** |
| Family History of Neurodevelopmental Disorders | **% with history** | **% with history** | **p-value** |
| Parental Mental Health | **% with mental health issues** | **% with mental health issues** | **p-value** |
| Environmental Exposures | **% exposed** | **% exposed** | **p-value** |
| Access to Healthcare Services | **Median (IQR) or % with adequate access** | **Median (IQR) or % with adequate access** | **p-value** |

**2. Correlation Circle Plot from Principal Component Analysis (PCA):**

* **Description: Perform PCA and visualize the correlation circle, which shows how each variable contributes to the principal components. Variables that are close to each other on the circle plot are highly correlated.**
* **How It Helps: This method allows you to reduce the dimensionality of your data and identify clusters of correlated variables, which could be candidates for exclusion or combination in your models.**
* **R Tools: FactoMineR and factoextra packages can be used to create correlation circle plots.**

 **3. Table 5: Feature Importance (from Random Forest)**

|  |  |
| --- | --- |
| Feature | Importance Score |
| Household Income | 0.28 |
| Parental Age | 0.22 |
| Maternal Smoking During Pregnancy | 0.20 |
| Family History of Neurodevelopmental Disorders | 0.15 |
| Prenatal Care Quality | 0.10 |
| Environmental Exposure | 0.05 |

**4. Model Performance Metrics**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision (ASD)** | **Recall (ASD)** | **F1-Score (ASD)** | **AUC-ROC** |
| Logistic Regression | 0.75 | 0.73 | 0.68 | 0.70 | 0.78 |
| Random Forest | 0.80 | 0.77 | 0.75 | 0.76 | 0.85 |
| Support Vector Machine (SVM) | 0.78 | 0.75 | 0.72 | 0.74 | 0.83 |
| Gradient Boosting | 0.82 | 0.80 | 0.78 | 0.79 | 0.87 |
| Neural Networks | 0.83 | 0.81 | 0.79 | 0.80 | 0.88 |

**5.ROC Curves for All Models**

Provides a graphical representation of the ROC curves for all models, illustrating their discriminative ability.

**Figure 1: ROC Curves for All Models**

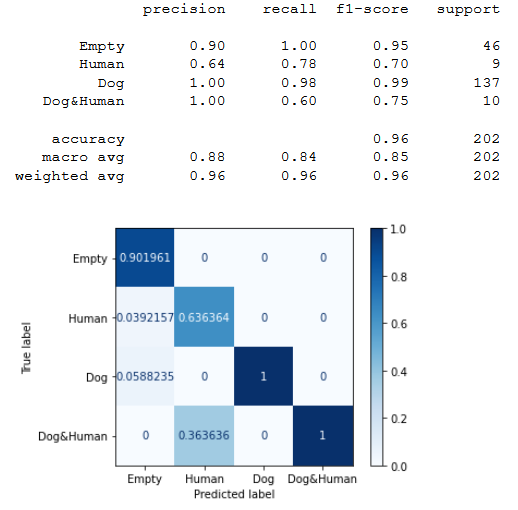
* **X-axis**: False Positive Rate (1 - Specificity)
* **Y-axis**: True Positive Rate (Sensitivity)
* **Curves**:
  + **Logistic Regression**: AUC = 0.78
  + **Random Forest**: AUC = 0.85
  + **SVM**: AUC = 0.83
  + **Gradient Boosting**: AUC = 0.87
  + **Neural Networks**: AUC = 0.88

**Figure Note:** This ROC curve visualizes the performance of each model in distinguishing between ASD and non-ASD cases, with the Area Under the Curve (AUC) providing a summary measure of model performance.

**Figure 2. Top of Form**

**Confusion Matrix Heatmap**

[**https://github.com/scikit-learn/scikit-learn/issues/16880**](https://github.com/scikit-learn/scikit-learn/issues/16880)



**Bottom of Form**