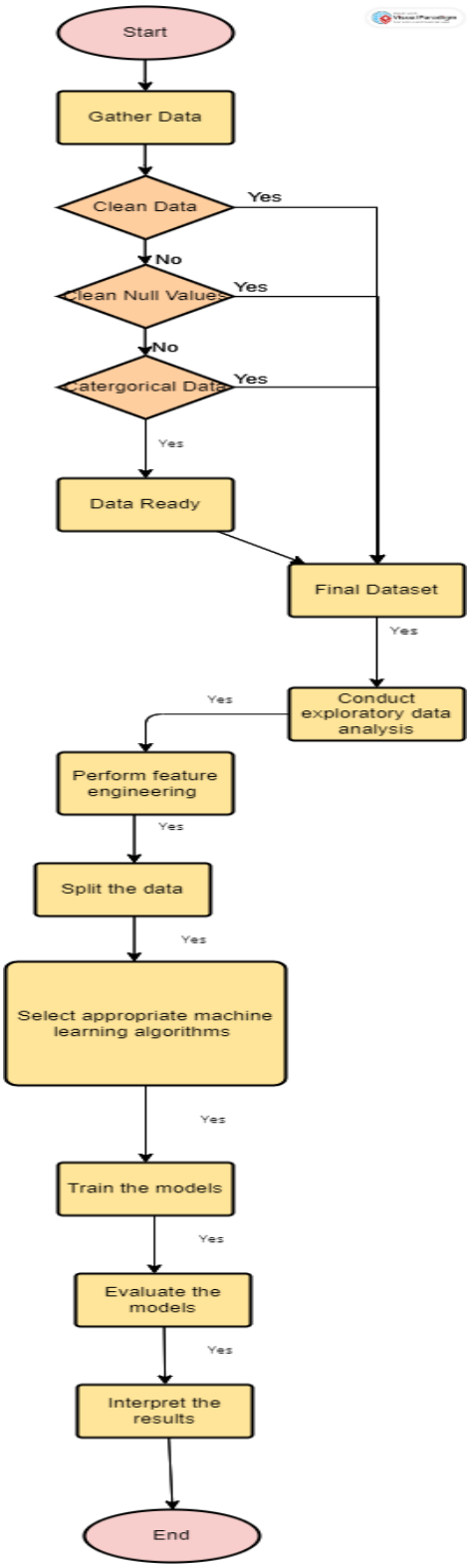
Name:-Ankan Dutta ID No:-E21BCAU0084

Batch:-1 Project:-AI in Healthcare

• Flowchart/ Arch Diagrams:-



• Flowchart explanation

Here is a flowchart for the Drug Consumption Data Set:

1. Gather the raw data on drug consumption, which includes information about the demographic characteristics of the participants and their self-reported drug use.
2. Clean and preprocess the data to remove any errors, inconsistencies, or missing values.
3. Conduct exploratory data analysis to gain insights into the distribution and patterns of drug use in the sample.
4. Perform feature engineering to create new variables that capture meaningful aspects of the data, such as the total number of drugs used, the frequency of use, and the age of first use.
5. Split the data into training and testing sets to develop and evaluate machine learning models.
6. Select appropriate machine learning algorithms to predict drug use based on the demographic and behavioral variables in the dataset.
7. Train the models using the training data and tune hyperparameters to improve their performance.
8. Evaluate the models using the testing data and metrics such as accuracy, precision, recall, and F1 score.
9. Interpret the results of the models to gain insights into the factors that contribute to drug use and identify potential intervention strategies.
10. Share the findings with stakeholders and the broader community to inform policy and public health initiatives aimed at reducing drug use and its harmful effects.

• Algorithm/technique/model related explanation.

There are several algorithms, techniques, and models that can be used to analyze the Drug Consumption (Quantified) Data Set. Here are some examples:

1. Logistic Regression: This is a statistical model that can be used to predict the probability of drug use based on the demographic and behavioral variables in the dataset. Logistic regression is a popular method for binary classification problems, such as predicting whether an individual uses drugs or not.
2. Decision Trees: This is a machine learning algorithm that can be used to build a model that predicts drug use based on a set of decision rules. Decision trees are easy to interpret and can handle both categorical and continuous variables.
3. Random Forest: This is an ensemble learning algorithm that combines multiple decision trees to improve the accuracy and robustness of the model. Random forests can handle high-dimensional data and are resistant to overfitting.

4. Support Vector Machines (SVM): This is a machine learning algorithm that can be used for both classification and regression tasks. SVMs are particularly useful for datasets with a small sample size and a large number of features.
5. Neural Networks: This is a deep learning technique that involves building a model composed of multiple layers of artificial neurons. Neural networks can learn complex patterns in the data and are capable of handling large datasets.
6. K-Nearest Neighbors (KNN): This is a non-parametric algorithm that can be used for both classification and regression tasks. KNN involves finding the k-nearest neighbors of a given data point and using their values to make a prediction.

Overall, the choice of algorithm or model depends on the specific research question, the characteristics of the data, and the desired level of interpretability and accuracy.

- Dataset description

Database contains records for 1885 respondents. For each respondent 12 attributes are known: Personality measurements which include NEO-FFI-R (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness), BIS-11 (impulsivity), and ImpSS (sensation seeking), level of education, age, gender, country of residence and ethnicity. All input attributes are originally categorical and are quantified. After quantification values of all input features can be considered as real-valued. In addition, participants were questioned concerning their use of 18 legal and illegal drugs (alcohol, amphetamines, amyl nitrite, benzodiazepine, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, nicotine and volatile substance abuse and one fictitious drug (Semeron) which was introduced to identify over-claimers. For each drug they have to select one of the answers: never used the drug, used it over a decade ago, or in the last decade, year, month, week, or day.
Database contains 18 classification problems. Each of independent label variables contains seven classes: "Never Used", "Used over a Decade Ago", "Used in Last Decade", "Used in Last Year", "Used in Last Month", "Used in Last Week", and "Used in Last Day".

Problem which can be solved:
* Seven class classifications for each drug separately.
* Problem can be transformed to binary classification by union of part of classes into one new class. For example, "Never Used", "Used over a Decade Ago" form class "Non-user" and all other classes form class "User".
* The best binarization of classes for each attribute.
* Evaluation of risk to be drug consumer for each drug.
Detailed description of database and process of data quantification are presented in E. Fehrman, A. K. Muhammad, E. M. Mirkes, V. Egan and A. N. Gorban, "The Five Factor Model of personality and evaluation of drug consumption risk.," arXiv [Web Link], 2015
Paper above solve binary classification problem for all drugs. For most of drugs sensitivity and specificity are greater than 75