DETAILED PROJECT REPORT

Adult Census Income Prediction

Project Title	**Mice Protein**
Author Name	**D.Jagannath**
Technologies	**Machine Learning Technology**
Domain	**Healthcare**
Project Difficulties	**level Intermediate**

Problem Statement:

Protein Expression classification models are frequently viewed not only as a difficult task, but also as a classification problem that, in some cases, requires a trade-off between accuracy and efficiency in analysis validation due to the large amount of data available.

Expression levels of 77 proteins measured in the cerebral cortex of 8 classes of control and Down syndrome mice exposed to context fear conditioning, a task used to assess associative learning.

The aim is to identify subsets of proteins that are discriminant between the classes.

Basically, this is multi-class classification problem

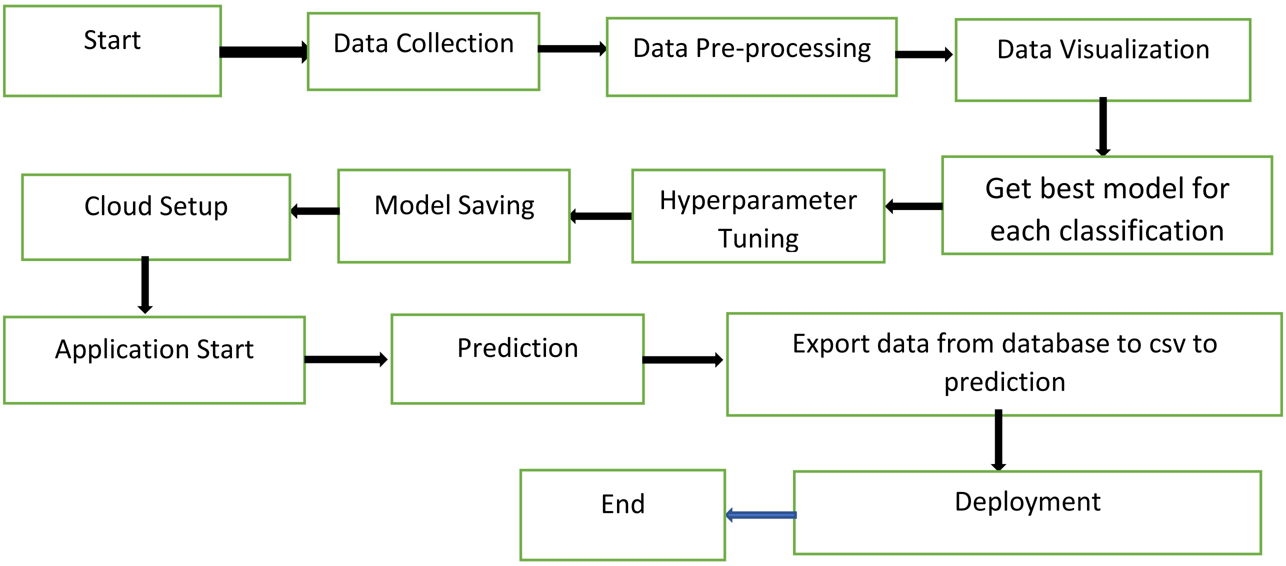
Objective:

Mice Protein project is a Machine Learning based model which will help us in prediction of body developments in mice

Benefits:

Benefits of this model is it will help us to about body developments two different type (Control mice and Trisomic mice) of Mice.

Architecture



Architecture Description:-

Data Collection: Data is collected from the source provided.

Data pre-processing

Performing EDA to get insight of data like identifying distribution , outliers ,trend among data etc.

Check for null values in the columns. If present impute the null values.

Encode the categorical values with numeric values.

Perform Standard Scalar to scale down the values.

Data Visualization: Visualization is done to understand the data set

Clustering :

Machine Learning algorithm is used to create clusters in the pre processed data. The optimum number of clusters is selected by plotting the elbow plot, and for the dynamic selection of the number of clusters . The idea behind clustering is to implement different algorithms on the structured data

The ML model is trained over pre processed data, and the model is saved for further use in prediction

Model Selection :

After the clusters are created, we find the best model for each cluster. By using Machine Learning algorithm. For each cluster both the hyper tunned algorithms are used. We calculate the AUC scores for both models and select the model with the best score. Similarly, the model is selected for each cluster. All the models for every cluster are saved for use in prediction

Prediction:

- The testing files are shared in the batches and we perform the same Validation operations ,data transformation and data insertion on them.
- The accumulated data from db is exported in csv format for prediction
- We perform data pre-processing techniques on it.
- Machine learning model created during training is loaded and clusters for the pre processed data is predicted
- Based on the cluster number respective model is loaded and is used to predict the data for that cluster.
- Once the prediction is done for all the clusters. The predictions are saved in csv format and shared.

Question & Answers

Q1) What's the source of data?

Ans) The Dataset was taken from iNeuron's Provided Project Description Document.

2: Is data cleaned before using?

Ans.: Yes

Q 3) What was the type of data?

The data was the combination of numerical and Categorical values.

Q 4) What techniques were you using for data pre-processing?

- Removing unwanted attributes
- Visualizing relation of independent variables with each other and output variables
- Checking and changing Distribution of continuous values
- Removing outliers
- Cleaning data and imputing if null values are present.
- Converting categorical data into numeric values.

Q 5) How training was done or what models were used?

- Before diving the data in training and validation set we performed clustering over fit to divide the data into
- clusters.
- As per cluster the training and validation data were divided.
- The scaling was performed over training and validation data
- Algorithms Decision Trees was used based on the recall final model was used for each cluster and we saved that model .