PROJECT REPORT







































## 4. Performance

## 4.1 Reusability

## 4.2 Application

## 4.3 Resource utilization

## 5. deployment

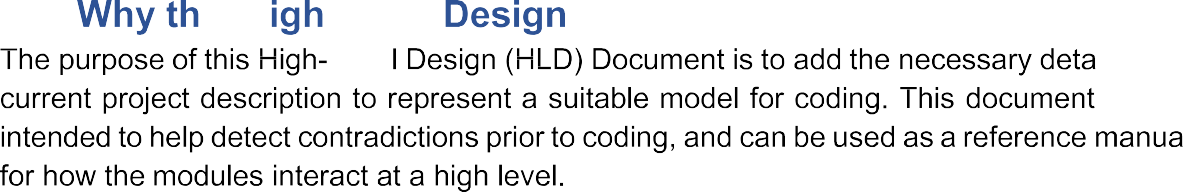
## 6. Conclusion

**Abstract:**

## Down syndrome is a chromosomal abnormality related to intellectual disabilities that affects 0.1% of live births worldwide. It occurs when an individual has a full or partial extra copy of chromosome 21. This chromosome trisomy results in the overexpression of genes that is believed to be sufficient to interfere normal pathways and normal responses to stimulation, causing learning and memory deficiency.

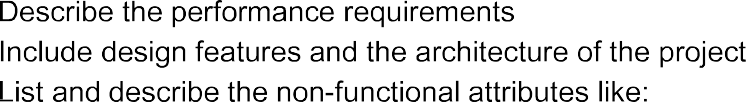
## Therefore, by studying these proteins and the disturbance in pathways that are involved in learning and memory, we can consider drugs that would correct the observed perturbations, and therefore assist in enhancing the memory and learning. Here, from genes based on an earlier study that identified 77 proteins differentially expressed in normal and trisomic wild mice exposed to context fear conditioning (CFC)

## 77 protein expression levels are analyzed from both control and trisomic (Ts65Dn) genotype mice, both with and without treatment from the drug memantine. Result suggest that decision tree and random forest classification approach can identify the most important proteins which may help to identify more effective drug to help learning process in people with DS.



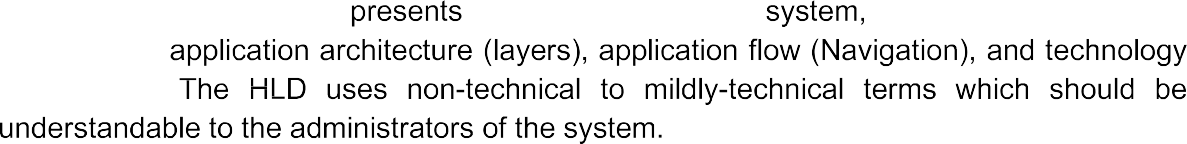
The HLD will:







1.2 Scope



1.3 Definitions





2 General Description:

2.1 Product Perspective:

## The Classification Model is a tree based model which will heip us to identify class from 8 classes.

2.2 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

2.3 Further Improments:

## In the further improvements improving the model when large data set coming by

## Reducing the rate of protein synthesis. ...

## Changing the growth medium: ...

## Co-expression of chaperones and/or foldases. ...

## Periplasmic expression: ...

## Using specific host strains: ...

## Addition of a fusion partner: ...

## Expression of a fragment of the protein:

2.4 Technical Requirements:

## 1. Applied Mathematics = Maths is quite an important skill in the arsenal of a Machine Learning engineer. It is also one of the basic subject to calculate accuracy and other calculations.

## 2. Programming = Python is very suitable when it comes to machine-learning.

## 3. Knowledge of machine learning algorithms = It’s classification problem so it’s compulsory to be familiar with classification and also good bonding with decision trees, rfclassifier, naïve bayes classifier

## **4. Data Modeling and Evaluation=**

## As a machine learning engineer, you should be skilled in data modeling and evaluation.

2.5 Data Requirements:

## For this project we need some kind of data requirements:

## We need data that should be balanced and must have at least 1000 raws.

## **Large, diverse infrastructure for data management**

## data environment must provision large quantities of raw data for discovery-oriented analytics practices.

## Enough data to split into train and validation

## Related data to analysis the dependent feature relationship.

2.6 Tools used:

## Pycharm for first three steps of machine learning algo

## Google-Colab for building the machine learning models

## Aws is used for deployment

## Mysql is used to curd operations.

2.7 Constraints:

## The model must be user friendly, results and prediction can easily be a understandable by humans.

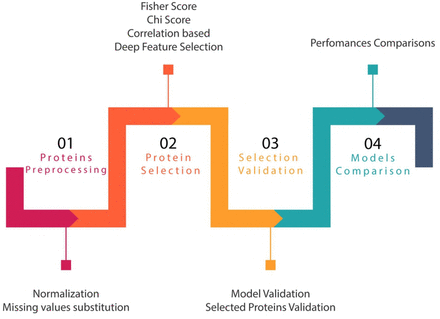
## 

2.8 Assumptions

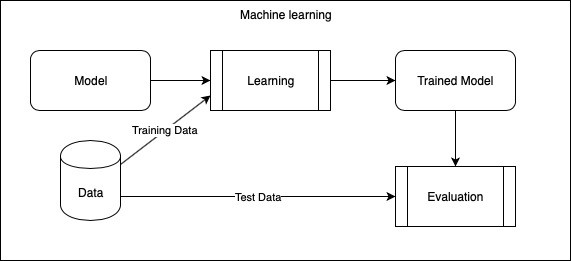
## The aim of the analysis is to identify those proteins that are discriminant for the different classes.

3 Design Details:

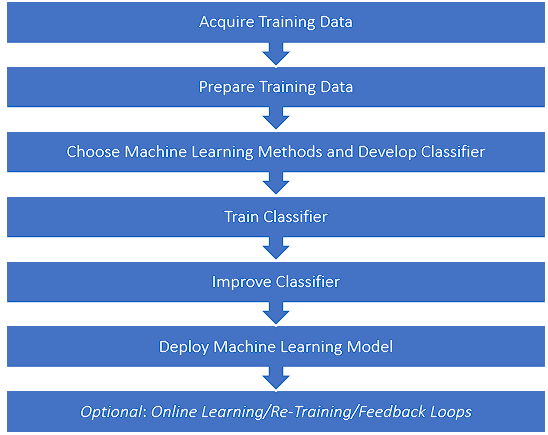
3.1 Process Flow:



3.1.1 Model Training and Evaluation



3.1.2 Deployment Process



3.2 Error Handling



4 Performance:

## Overall, we can see that the classification models we’ve used are useful in trying to infer which

## proteins are significant contributors to the learning of mice in Cognitive Stress.

## As expected, the Random Forest performed much better than the decision tree, which is a much less

## sophisticated model. The Random Forest had an almost perfect Recall, F1-Score and Precision scores

## of 0.99 for each, compared with the 0.84 for the decision tree. This was done by using exactly the

## same training and test data set by setting the seed of the random number generator for data

## selection. This does not affect the randomness within the Random Forest model.

4.1 Reusability:

## The code written and the components used should have the ability to be reused with no problems.

4.2 Application Compatibility

The different components for this project will be using python a an interface between them. Each component will have its own task to perform, and it is the job of the python to ensure proper transfer of information.

## 4.3 Resource Utilization

## When any task is performed ,it will likely use all the processing power available until that function is finished.

5 Deployment:

## Microsoft azure

## Google cloud

## Aws

6 Conclusion:

## In this project, random forest model behaves better than decision tree to predict which proteins are

## important for learning process of mice. Decision tree and random forest model shown the similar

## most crucial proteins in the metabolic pathways related to learning.