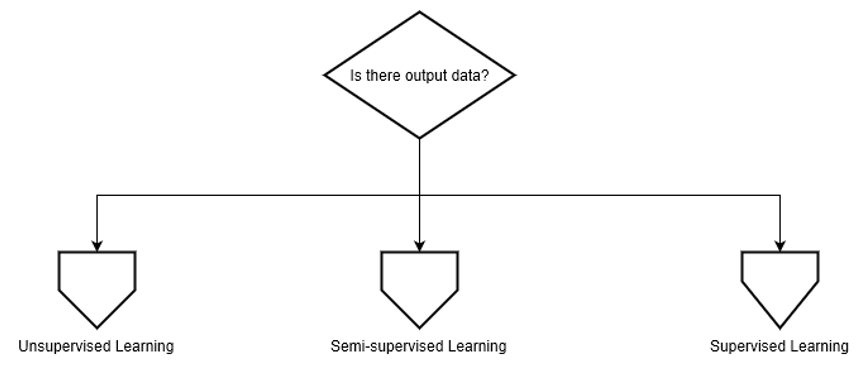
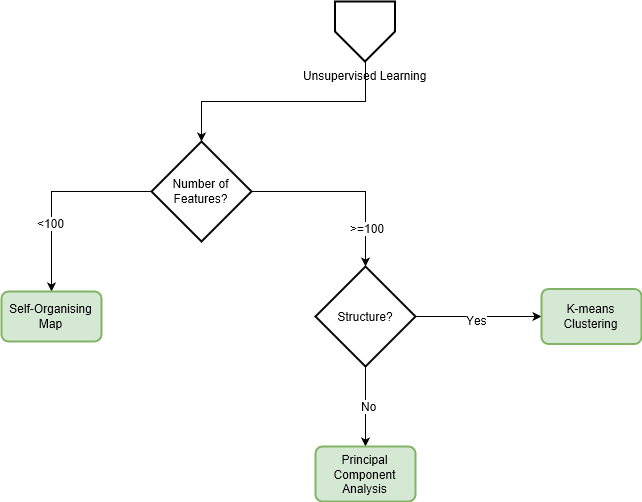
**Initial Design Ideas-Decision Tree and Algorithms**

The long-term goal for the team project is to use decision algorithm to determine which is the best analyzing tool for datasets. First of all, determine whether the dataset is supervised. According to whether there is output data in the dataset, we divide the dataset into three categories: Unsupervised Learning, Semi-supervised Learning and Supervised Learning.

* Unsupervised Learning

If the dataset has no output data, divide it into Unsupervised Learning. If there are lots of features in the dataset, suggest using Self-Organizing Map to analyze. Otherwise judge whether the dataset have structure or not. If have, suggest K-means Clustering to analyze, otherwise suggest using Principal Component Analysis.



Definition of Self-Organizing Map (SOM):

A self-organizing map (SOM) or self-organizing feature map (SOFM) is a type of artificial neural network (ANN) that is trained using unsupervised learning to produce a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples, called a map, and is therefore a method to do dimensionality reduction. Self-organizing maps differ from other artificial neural networks as they apply competitive learning as opposed to error-correction learning (such as backpropagation with gradient descent), and in the sense that they use a neighborhood function to preserve the topological properties of the input space. [1]

Definition of K-means Clustering:

k-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. [2]

Definition of Principal Component Analysis:

Principal Component Analysis is a dimension-reduction tool that can be used advantageously in such situations. Principal component analysis aims at reducing a large set of variables to a small set that still contains most of the information in the large set. [3]

* Supervised Learning

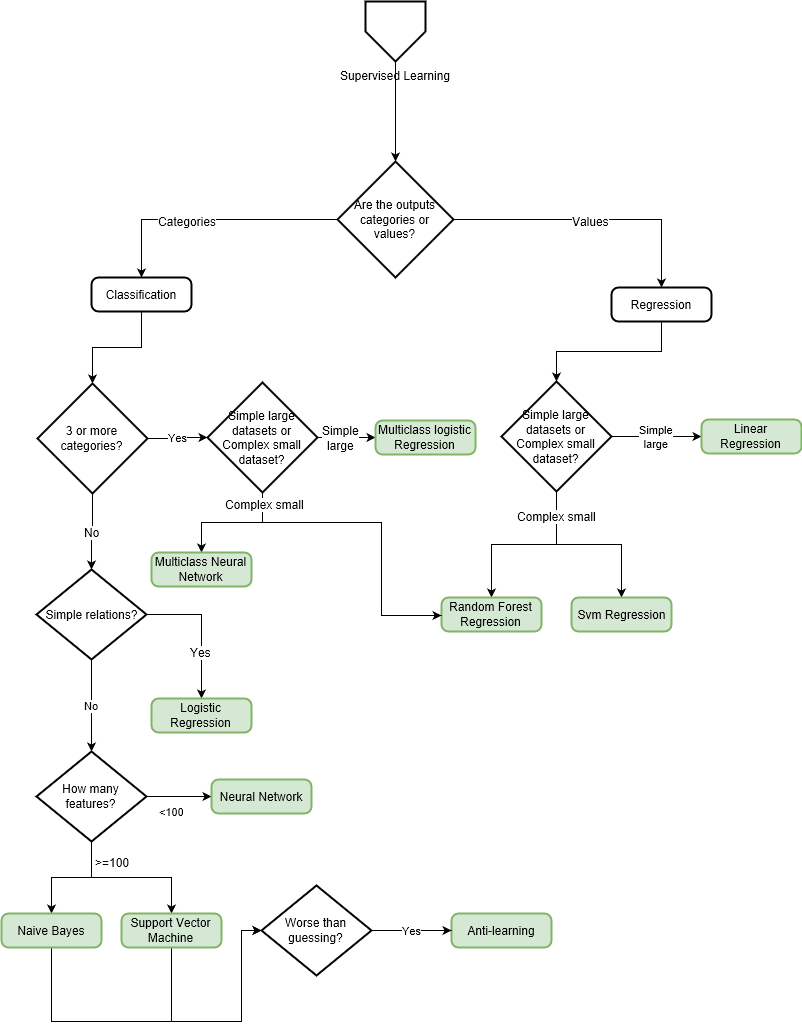
If the dataset has output data, divide it into Supervised Learning. After determining the dataset to Supervised Learning, if the outputs are categories divide the dataset to Classification. Else if the outputs are Values, divide the dataset to Regression.

After dividing it to classification, determine how many categories the dataset contains. If dataset contain 3 or more categories, discriminate the dataset simple or not.

If the dataset is a complex small dataset, suggest Multiclass Neural Network or Random Forest Regression. Else if the dataset is a simple large dataset, suggest using Multiclass logistic Regression to analyze.

If the dataset only contains less than 3 categories, determine whether the relation in the dataset is simple. If the dataset has simple relationships, suggest Logistic Regression.

Otherwise determine how many features in the dataset. If the dataset has less than 100 features, suggest Neural Network, otherwise using Naïve Bayes or Support Vector Machine. After that, judge if the Correct rate is less than guessing, suggest Anti-learning.

If the dataset is composed of values, also determine the dataset simple or not, if it is a simple large dataset, suggest Linear Regression. Else if a complex small dataset, suggest Random Forest Regression and Sum Regression.

Definition of Linear Regression:

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model. [4]

Definition of Multiclass Neural Network:

A neural network is a set of interconnected layers. The inputs are the first layer, and are connected to an output layer by an acyclic graph comprised of weighted edges and nodes.

Between the input and output layers you can insert multiple hidden layers. Most predictive tasks can be accomplished easily with only one or a few hidden layers. However, recent research has shown that deep neural networks (DNN) with many layers can be very effective in complex tasks such as image or speech recognition. The successive layers are used to model increasing levels of semantic depth.

The relationship between inputs and outputs is learned from training the neural network on the input data. The direction of the graph proceeds from the inputs through the hidden layer and to the output layer. All nodes in a layer are connected by the weighted edges to nodes in the next layer.

To compute the output of the network for a particular input, a value is calculated at each node in the hidden layers and in the output layer. The value is set by calculating the weighted sum of the values of the nodes from the previous layer. An activation function is then applied to that weighted sum.

The website will provide the most suitable machine learning tool for each dataset uploaded by client. Besides, the website will analyze the dataset by using the best machine learning tool and visualize the output. [5]

Definition of Random Forest Regression:

A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement if bootstrap=True (default). [6]

Definition of Multiclass logistic Regression:

In statistics, multinomial logistic regression is a classification method that generalizes logistic regression to multiclass problems, i.e. with more than two possible discrete outcomes.[1] That is, it is a model that is used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable, given a set of independent variables (which may be real-valued, binary-valued, categorical-valued, etc.). [7]

Definition of logistic Regression:

Logistic Regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. [8]

Definition of Neural Network:

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. ANNs, like people, learn by example. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons. This is true of ANNs as well. [9]

Definition of Naïve Bayes:

A naive Bayes classifier is an algorithm that uses Bayes' theorem to classify objects. Naive Bayes classifiers assume strong, or naive, independence between attributes of data points. Popular uses of naive Bayes classifiers include spam filters, text analysis and medical diagnosis. These classifiers are widely used for machine learning because they are simple to implement. [10]

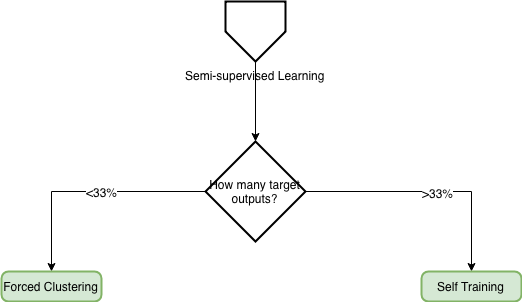
Definition of Support Vector Machine:

“Support Vector Machine” (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well. [11]

Definition of Anti-learning:

Anti-learning is the term given for the situation where performance of a trained computational intelligence technique is significantly worse than random guessing and is not overfitting or overtraining. Anti-learning has been observed in a range of synthetic and real-world datasets. [12]

* Semi-Supervised Learning

Semi-Supervised Learning is a class of machine learning tasks and techniques that also make use of unlabeled data for training – typically a small amount of labeled data with a large amount of unlabeled data. Choose the unlabeled data to be added to the labeled data. Divide the dataset into Semi-Supervised Learning. If the target outputs in the dataset less than 33%, suggest Forced Clustering. Otherwise using self-Training to analyze.

Definition of Forced Clustering:

UNKNOWN[13]

Definition of self-Training:

The basic assumption of the self-learning model is that when the classifier predicts the sample, the sample with high confidence is highly likely to be correctly classified. For example, when a SVM classifies a sample, those samples that are far away from the classification interface can be considered to be correctly classified. Based on this assumption, the self-learning model is extremely simple. [14]

Reference list:

[1] <https://en.wikipedia.org/wiki/Self-organizing_map>

[2] <https://en.wikipedia.org/wiki/K-means_clustering>

[3] <https://www.itl.nist.gov/div898/handbook/pmc/section5/pmc55.htm>

[4] <http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm>

[5]<https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/multiclass-neural-network>

[6]<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>

[7] <https://en.wikipedia.org/wiki/Multinomial_logistic_regression>

[8] <https://www.statisticssolutions.com/what-is-logistic-regression/>

[9]<https://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html#What%20is%20a%20Neural%20Network>

[10] <https://www.techopedia.com/definition/32335/naive-bayes>

[11]<https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/>

[12] <https://arxiv.org/ftp/arxiv/papers/1607/1607.06190.pdf>

[13]

[14] <http://www.leexiang.com/self-training-and-co-training>