**Final Report**

1. **Raw Data Preprocessing**
   1. **Data Discretization and Analysis**

* An optional task is to do statistic analysis of data and try different subset of attributes as features.
  1. **Missing Data Filling**
* Explain your reasons for the way you group the users.

Firstly, we drop two attributes, that is, “votes” and “happy”. “votes” refers to the number of questions and “happy” is actually the label.

1. **Classification**
   1. **Logistic Regression**

* **Running time**

Scikit-learn:

Self-implemented:

**Training accuracy**

Scikit-learn:

Self-implemented:

* 1. **Naïve Bayes**
* We choose a **BernoulliNB** classifier, because data in most of our features is distributed according to multivariate Bernoulli distributions; i.e., there may be multiple features but each one is assumed to be a binary-valued feature vectors. This is the result of one-of-K encoding during data preprocessing.
* **Running time**

Scikit-learn:

Self-implemented:

**Training accuracy**

Scikit-learn:

Self-implemented:

* 1. **SVM**
* SVC accepts different kernel functions. Kernel functions could be one of ‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’, ‘precomputed’ or a callable function defined by yourself. Please choose a kernel function  and state your reasons

**3. Write a Report**

* **What are the characteristics of each of the four classifiers?**

1. **Logistic Regression**

Logistic regression is a [regression](https://en.wikipedia.org/wiki/Regression_analysis) model where the dependent variable is categorical and it can be used for classification of the data. It estimates the probability(y/x) directly from the training data by minimizing error. As a discriminative classifier, it tries to  model class boundary and membership directly. For example, in a simple two feature dimension case, this means trying to finding the line that best separates the classes and in >3 feature dimensions cases, it would be looking for the hyperplane that best separate classes.

1. **Naïve Bayes**

Naïve Bayes classifiers are based on applying Bayes' theorem with strong (which means naive) independence assumptions between the features. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. The different naive Bayes classifiers differ mainly by the assumptions they make regarding the distribution of P(xi | y). In spite of their apparently over-simplified assumptions, naive Bayes classifiers have worked quite well in many real-world situations, famously document classification and spam filtering. They require a small amount of training data to estimate the necessary parameters. Naive Bayes learners and classifiers can be extremely fast compared to more sophisticated methods. On the flip side, although Naive Bayes is known as a decent classifier, it is known to be a bad estimator. In addition, Naïve Bayes are relatively simple, you just need to do a bunch of counts.

1. **SVM**

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. Support vector machines are efficient in high dimensional space, but If the number of features is much greater than the number of samples, the method is likely to give poor performances

1. **Random Forests**

Random forests are an [ensemble learning](https://en.wikipedia.org/wiki/Ensemble_learning) method for [classification](https://en.wikipedia.org/wiki/Statistical_classification), [regression](https://en.wikipedia.org/wiki/Regression_analysis) and other tasks, that operate by constructing a multitude of [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning) at training time and outputting the class that is the [mode](https://en.wikipedia.org/wiki/Mode_(statistics)) of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of [overfitting](https://en.wikipedia.org/wiki/Overfitting) to their training set. They can be used to rank the importance of variables in a regression or classification problem in a natural way.

* **Different classification models can be used in different scenarios. How do you choose classification models for different classification problems? Please provide some examples.**

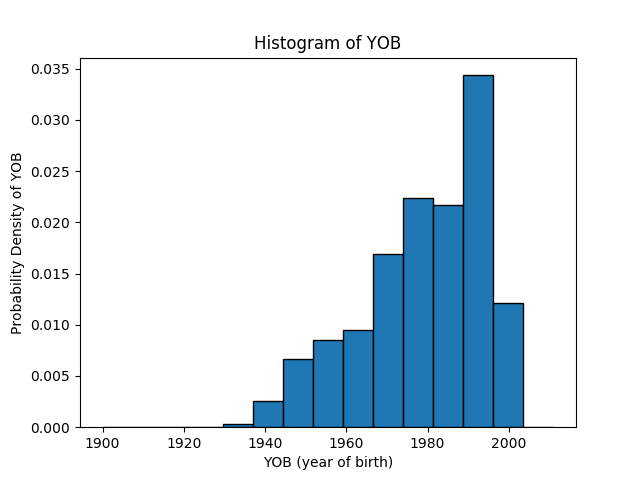
1. If features are independent or we want something fast and easy, we will choose **Naïve Bayes**, because Naïve Bayes classifiers are relatively simple-we just need to do a bunch of counts. If the NB conditional independence assumption actually holds, a Naive Bayes classifier will converge quicker than discriminative models like logistic regression, so that we need less training data. For example, in spite of their apparently over-simplified assumptions, naive Bayes classifiers have worked quite well in many real-world situations, famously document classification and spam filtering.
2. If we want a probabilistic framework (e.g., to easily adjust classification thresholds, to say when we’re unsure, or to get confidence intervals) or if we expect to receive more training data in the future that we want to be able to quickly incorporate new data into our model, we will choose **Logistic Regression**. Because there are lots of ways to regularize our model and we don’t have to worry as much about our features being correlated like we do in Naïve Bayes. We also have a nice probabilistic interpretation and can easily update our model to take in new data. For example, Logistic Regression is applied very widely in the medical and social sciences.
3. If we want a non-parametric classifiers so that we don’t have to worry about tuning a bunch of parameters, we will choose **Random Forests**. Results of random forests classifiers tend to be each to interpret and explain, and fast and scalable at the same time. For example, from physicochemical properties such as alcohol, acidity, and sulphates, we can apply random forests to learn a model and predict wine taste preferences levels (for example, for 0 to 10).
4. If we want a classifier with nice theoretical guarantees regarding overfitting and a high accuracy, we will choose **SVMs**. With an appropriate kernel, SVMs can work well even if our data isn’t linearly separable in the base feature space. For example, SVMs can be used to solve various real world problems, such as  text and hypertext categorization, as their application can significantly reduce the need for labelled training instances in both the standard inductive and transductive settings. Other examples include hand-written characters recognition, proteins classification and so on.

* **How do the cross validation techniques help in avoiding overfitting?**

The problem with residual evaluations is that they do not give an indication of how well the learner will do when it is asked to make new predictions for new data it has not already seen. One way to overcome this problem is to not use the entire data set when training a learner. Some of the data is removed before training begins. Then when training is done, the data that was removed can be used to test the performance of the learned model on “new” data. This is the basic idea for cross validation.

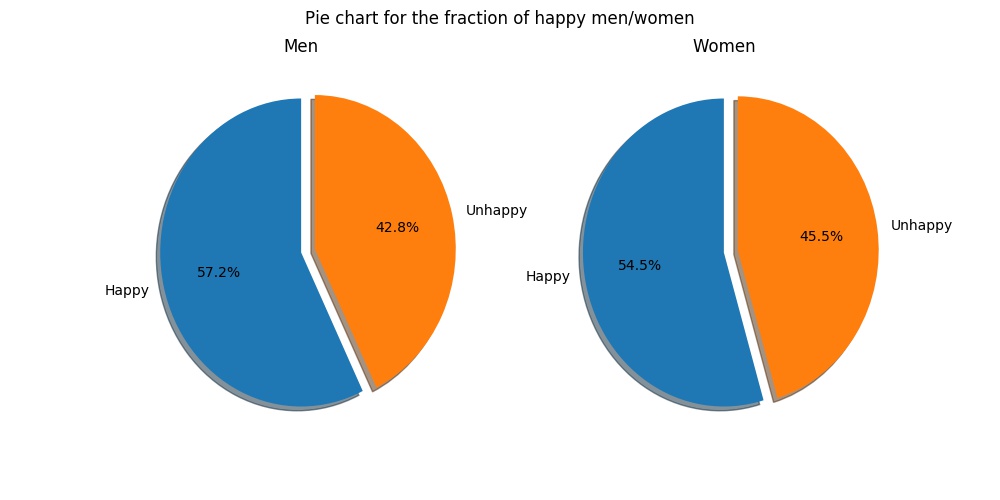
**4. Visualization**

1. **Histogram of YOB**



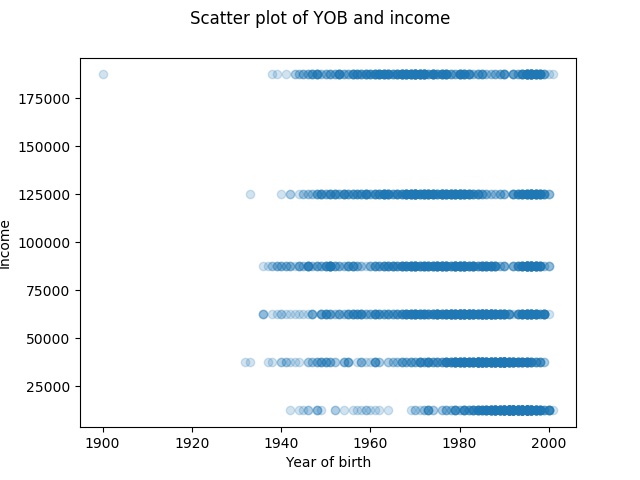
We can see that among all people who participated in the votes, the majority of them are young people and more specifically speaking, the number of voters who were born around 1990s is the most.

1. **Pie chart for the fraction of happy men/women**



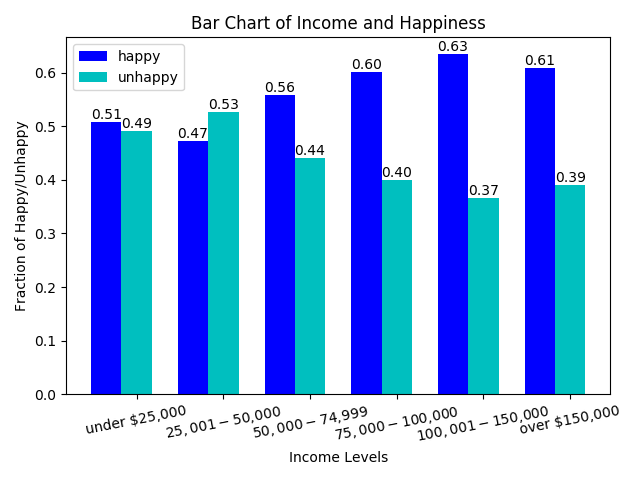
Generally speaking, there are more people who are happy than those who are unhappy and men who are happy are slightly more than women who are happy (in terms of the same sample size).

1. **Scatter plot of YOB and income**



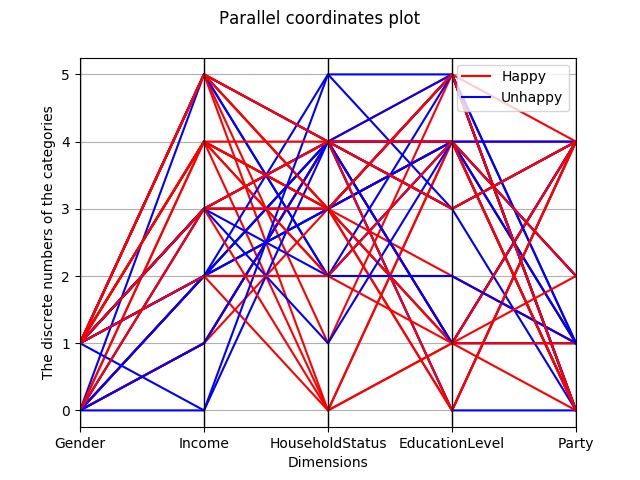
We can see from the graph that young people tend to get lower income and as the age increases, people tend to get higher income.

1. **Bar chart of income and happiness**

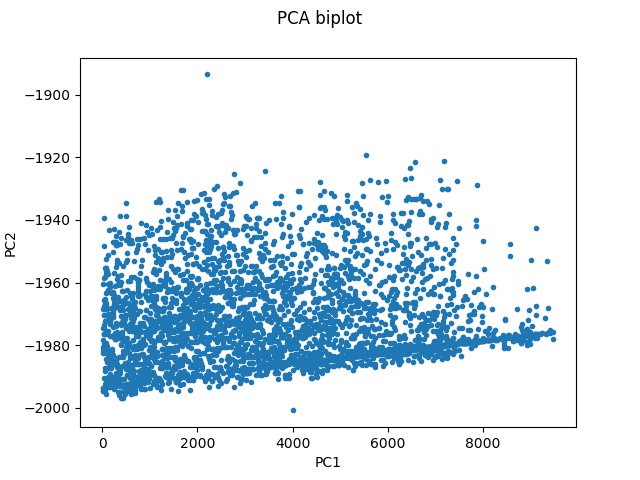


People who get higher income are tend to be more likely to be happy.

1. **Parallel Coordinates Plot**



1. **PCA and biplot**



**Q1 What’s the physical meaning the vector corresponded to each variable? Explain it in one sentence.**

The vectors corresponding to the two variables represent the directions, onto which all the data points are projected to having the largest and the second largest variances.

**Q2 What are the factors closely related to happiness according to this biplot? Write down your answer and use one more sentence to explain why.**

1. **Visualize SVM**