**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.

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:

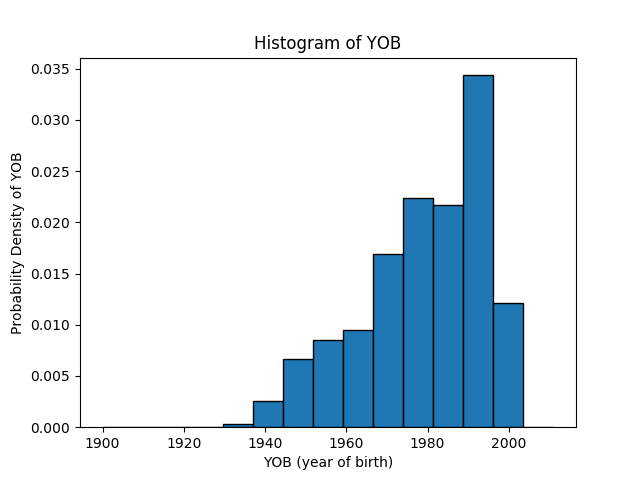
* 1. **SVM**
* SVC accepts di erent kernel functions. Kernel functions could be one of ‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’, ‘precomputed’ or a callable function defined by yourself. Please choose a kernel function  and state your reasons

**3. Write a Report**

* What are the characteristics of each of the four classifiers?
* Di erent classification models can be used in di erent scenarios. How do you choose classification models for di erent classification problems? Please provide some examples.
* How do the cross validation techniques help in avoiding overfitting?

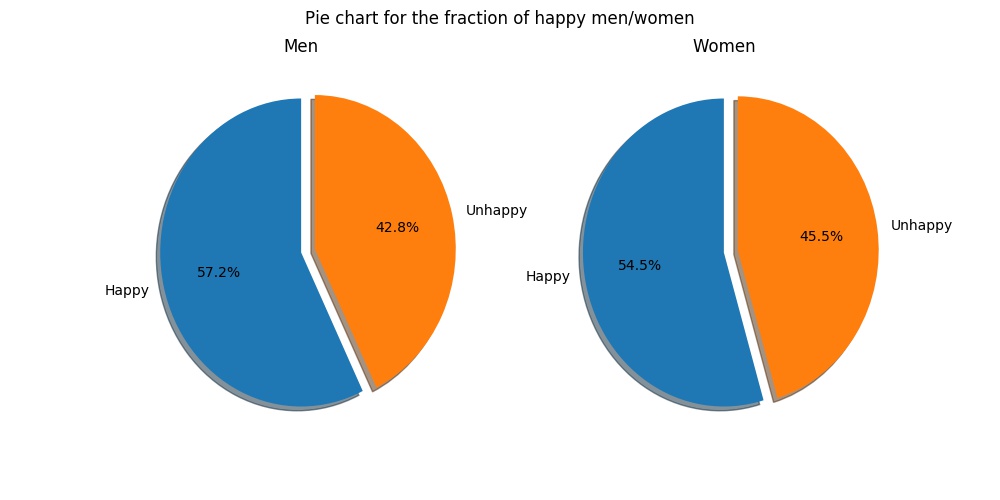
**4. Visualization**

1. **Histogram of YOB**



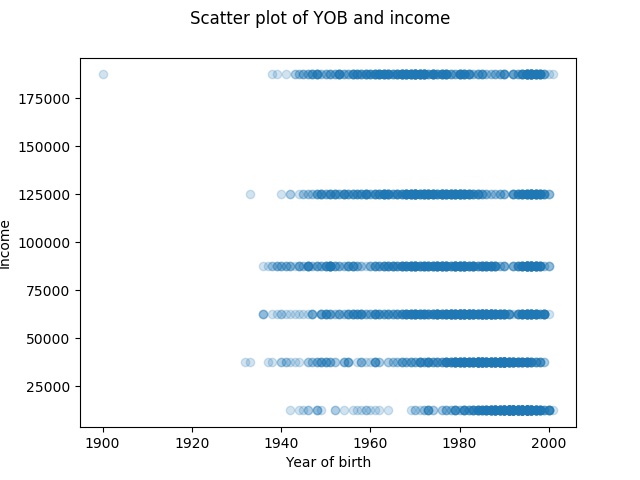
We can see that among all people who participated in the votes, the majority of them are young people and more specifically speaking, the number of voters who were born around 1990s is the most.

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



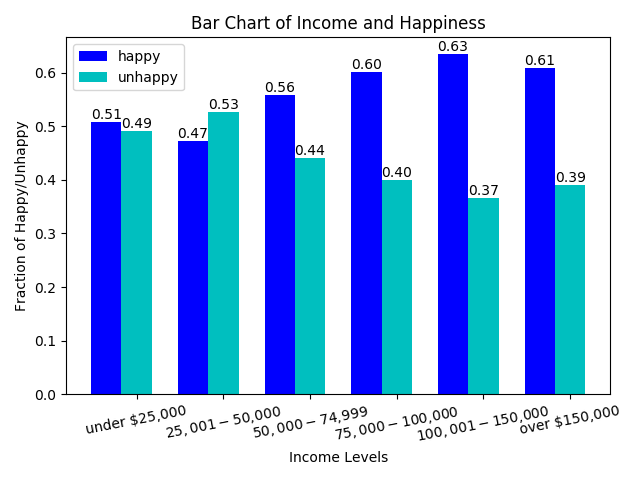
Generally speaking, there are more people who are happy than those who are unhappy and men who are happy are slightly more than women who are happy (in terms of the same sample size).

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.

4.1.4 **Bar chart of income and happiness**



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

4.2 Visualizing High-dimensional Data

4.2.1 Parallel Coordinates Plot

**4.2.2 PCA and biplot**

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

**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.

4.3 **Visualizing Classification Result**

**4.3.1 Visualize SVM**