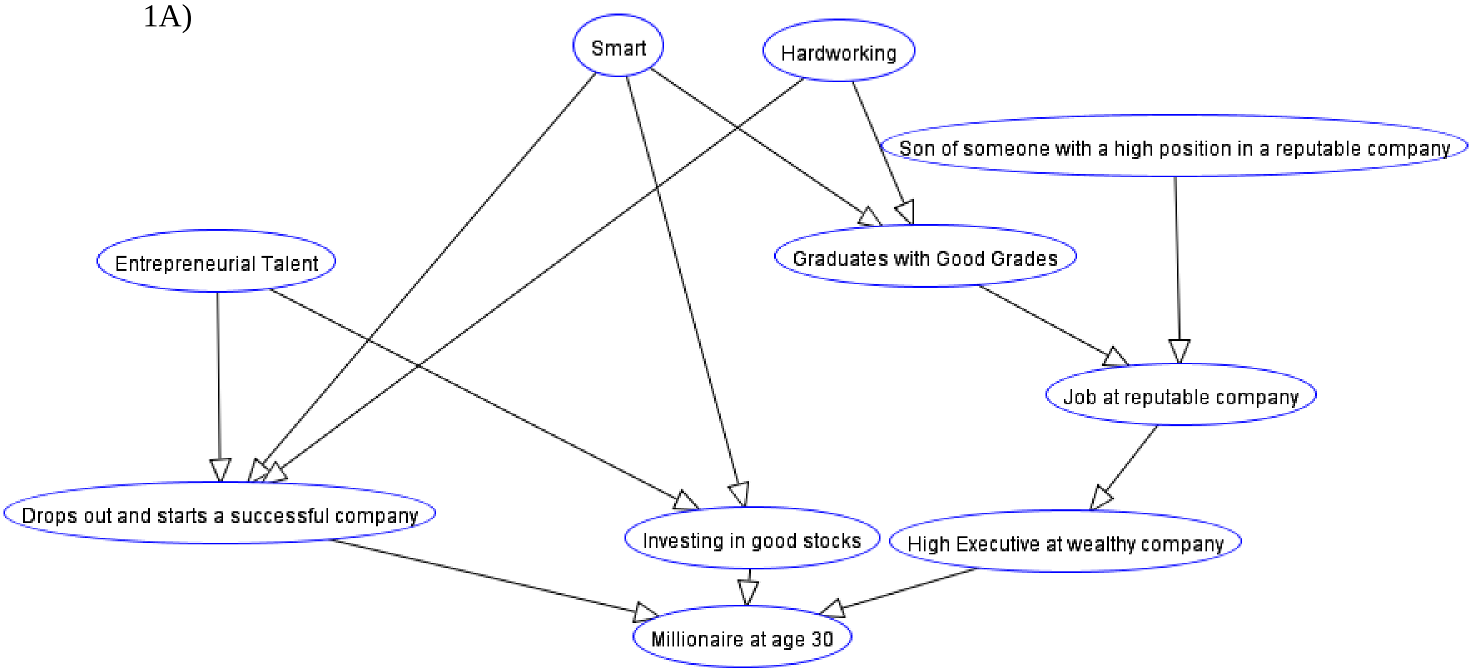
1A)


Smart
Hardworking
Son of someone with a high position in a reputable company
Entrepreneurial Talent
Graduates with Good Grades
Job at reputable company
Drops out and starts a successful company
Investing in good stocks
High Executive at wealthy company
Millionaire at age 30

1B) Scenario 1: Given that the node Son of someone with high position in a reputable company is true and the node investing in good stocks is true, the probability that Millionaire at age 30 is true is .75182 and false is .24818. This makes sense because the probability of becoming a millionaire at age 30 given investing in good stocks is high. The probability of having a job at a reputable company given the son of someone in a high position is high. Having a job at a reputable company directly affects being a high executive at a wealthy company, and the probability of being a Millionaire at age 30 given investing in good stocks and being a high executive is high.

Scenario 2: Give that the node Entrepreneurial Talent is false and the node Hardworking is true, the probability that Millionaire at age 30 is true is .43782 and false is .56218. This makes sense because the probability of investing in good stocks given no entrepreneurial talent is lower, and the probability of being a millionaire given not investing in good stocks is also low. Because hardworking is true however, the probability of graduating with good grades is high, and getting a job at a reputable company and then becoming a high executive are also high. The probability of becoming a millionaire given being a high executive but not investing in good stocks is not very high however, so our result makes sense.

1C) Given scenario 1, Drops out and starts a successful company and job at reputable company are conditionally independent given investing in good stocks. This is because the only path from drops out -> millionaire -> high executive -> job at reputable company is blocked. However, in real life they are not conditionally independent. In fact, if you know that the person has a job at a reputable company, he probably did not drop out and start a successful company because if he did, he is not likely to leave his successful company as a CEO to get a job at a reputable company. As such, the probability changes as you know one of these values.

2.3

Precision Training: .946154

Recall Training: 1

Precision Test: .903846

Recall Test: .723077

2.4

1. The evaluation results for training are much higher than those for precision. This is because we used the training set for the naïve bayes learning, so testing its precision and recall would be much higher (in fact its recall was 1 for training). The evaluation results would diminish and be less accurate and precise on a different set, the test set.

2. Based on the probabilities, the 3 words that are most likely to occur in Ham are:

Enron, please, subject

The 3 words that are most likely to occur in Spam are:

http, email, more

3. To improve the precision, we can change how the learner is labelling whether an email is spam or ham. Rather than comparing if the P(Spam|e) > P(Ham|e), I found the average P(Spam|e) for all of the spam training emails, and the average P(Ham|e) for all the ham training emails. Then, I went through all the emails and found each P(Spam|e) and P(Ham|e), and made a conditional: if the absolute value of P(Spam|e) – averageP(Spam|e) < threshold and P(Ham|e) – averageP(Ham|e) > threshold (for a given threshold), then mark the email as spam. Otherwise mark it as ham. This greatly improved the precision of the learner.

OS: Windows

Language: C++

To run the program, simply compile proj3.cpp with

```
g++ -g –o proj3 proj3.cpp
```

and then run with

```
./proj3
```

The output of the program is the precisions and recalls for training and test and the 3 most common words of both.