Guiding Questions

- Report the training and testing accuracy of the Naive Bayes classifier. (A correct implementation should have testing accuracy above 70%). (1 point)

- What strong assumption about the features/attributes of the data does Naive Bayes make? Comment on this assumption in the context of credit scores. (3 points)
  
  **Solution:** Every feature is independent given the class. Probably doesn’t hold true in the context of credit scores.

- This dataset was originally structured as follows:

<table>
<thead>
<tr>
<th>Month</th>
<th>Credit Amount</th>
<th>Number of credits</th>
<th>...</th>
<th>Credit</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>1169</td>
<td>2</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>48</td>
<td>5951</td>
<td>1</td>
<td>...</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>2096</td>
<td>1</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>2134</td>
<td>3</td>
<td>...</td>
<td>1</td>
</tr>
</tbody>
</table>

  For each of the above attributes, describe what transformations to the original dataset would need to occur for it to be usable in a Bernoulli Naive Bayes model. (hint: every attribute must take on the value of 0 or 1) (5 points)

  **Solution:** Discretize “month” and “credit amount”
  
  Binarize “month,” “credit amount,” “number of credits”
  
  Switch “credit” encoding to 0-1

- Restate the definition of Disparate Impact from lecture (also included in code comments); make sure to notate what each variable (e.g. $S$) represents. Why might this be a useful measure of model performance? What are some limitations of this measure? (5 points)

  **Solution:**
  
  $P(\hat{Y} = 1|S = 1) \over P(Y = 1|S = 0)$

  $\hat{Y} = 1$ indicates the “good” decision, e.g., “good credit.”

  $S$ is the sensitive attribute, and $S = 1$ indicates membership in the "disadvantaged" group while $S = 0$ means “privileged.”

  Useful – provides one lens into why a model can be interpreted as discriminatory (outcome-based); often we do care that the outcomes of two groups “match” to an extent. (Also, since this can also be calculated on the data and not just the model output, we have some sense of if the model is improving/worsening the discrimination already in the dataset.)

  Limitations – Doesn’t provide the full picture: if the underlying distributions of the two groups are very different, then a "perfect" DI score could actually be very undesirable – e.g. giving one group a lot of false positives and another a lot of false negatives.
A different way to think about fairness is based on the errors the model makes. We define the false positive rate (FPR) as $\Pr(\hat{Y} = 1 | Y = 0)$, and the false negative rate (FNR) as $\Pr(\hat{Y} = 0 | Y = 1)$. Suppose we calculate FPR and FNR for each group. In words, what does the false positive rate and false negative rate represent in the context of credit ratings? What are the implications if one group’s FPR is much higher than the other’s? What are the implications if one group’s FNR is much higher than the other’s? (6 points)

**Solution:**

**FPR:** fraction of the time someone was given "good credit" when they actually had “bad credit”

**FNR:** fraction of the time someone was given "bad credit" when they actually had “good credit”

**FPR disparity:** one group gets more access to credit/loans/etc than they should, unfairly rewarding those who are “undeserving” in the group with higher FPR

**FNR disparity:** one group gets less access to credit/loans/etc than they should, unfairly punishing those who are “deserving” in the group with higher FNR