Problem 1

The train accuracy: 0.7742857142857142
The test accuracy: 0.7257142857142858

Problem 2

What strong assumption about the features/attributes of the data does Naive Bayes make? Comment on this assumption in the context of credit scores.

We assume that each input variable is independent, which is unrealistic for real data. This means that the presence of a particular feature in a class is unrelated to the presence of any other feature. In the context of credit scores, we know that certain attributes depend on each other or the existence of others, we assume that they all independently contribute to the probaility that whether a person has good credit or not.

Problem 3

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)

We should let every attribute take on the value of 0 or 1. For the attribute 'month', we can set up a 3-bin binary variable. For instance, with 48 months as maximum, we can let month 1-16 be (1, 0, 0), month 17-32 be (0, 1, 0) and month 33-48 be (0, 0, 1). For the attribute 'credit amount', we can use the same way of categorizing them again to 3 bins, and setting up a binary indicator. For example, see '5951' as the maximum, we can let 1-2000 be (1, 0, 0), 2001-4000 be (0, 1, 0) and 4001-6000 be (0, 0, 1). For the attribute 'number of credits', we can see that the value only takes on 1, 2, or 3. Therefore, we can again use a size-3 binary indicator to indicate each value, (1, 0, 0) means 1, (0, 1, 0) means 2 and (0, 0, 1) means 3. For the attribute "credit", since the value only takes on 1 or 2, we can simply let 0 denote value '1', and 1 denote value '2'.

Problem 4

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?

Disparate Impact is a metric to evaluate fairness. It compares the proportion of individuals that receive a positive output for two groups: an unprivileged group and a privileged group. All unprivileged classes are grouped together as values of 0 and all privileged classes are given the class 1. Suppose we now have the dataset $D = (S, X, Y)$, where S represents the protected attribute (e.g., race, sex, religion, etc.), X represents the remaining attributes, and Y represents the class to be predicted, which is binary (e.g., "will hire"). we will say that D has disparate impact if: $P[Y\wedge = 1|S! = 1]/P[Y\wedge = 1|S = 1] <= (t = 0.8)$, which means that if the unprivileged group receives a positive outcome less than 80 percent of their proportion of the privileged group, this is a disparate impact violation.

Disparate impact can help to identify and address hidden biases or unintended consequences of policies or practices. By looking at the data and outcomes, it is possible to identify patterns of disparate impact and take action to mitigate them.

However, disparate impact analysis cannot determine whether discrimination was intentional or unintentional. It can only identify whether there is a statistically significant difference in outcomes. It does not provide a solution to the underlying problem. And it is only as effective as the data that is used to analyze it.

Problem 5

A different way to think about fairness is based on the errors the model makes. We define the false positive rate (FPR) as $P(Y\wedge = 1|Y = 0)$, and the false negative rate (FNR) as $P(Y\wedge = 0|Y = 1)$. 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?

The FPR represent the group of people that are predicted to have good credit but actually have bad credit. The FNR represent the group of people that are predicted to have bad credit but actually have good credit.

If one group's FPR is much higher than the other's, it means that members of that group are more likely to be wrongly denied credit or offered unfavorable terms, even though they are actually low credit risks. This could have significant implications for that group's access to credit and ability to build credit history, which in turn could limit their economic opportunities and perpetuate existing inequalities.

Similarly, if one group's FNR is much higher than the other's, it means that members of that group are more likely to be wrongly offered favorable credit terms, even though they are actually high credit risks. This could lead to higher default rates and losses for lenders, which could result in more stringent credit requirements for everyone and limit access to credit for members of that group. This is also in turn an discrimination towards those with low credit risks but were treated as with high credit risks.