**Naïve Bays:**

**Justification for Model Choice**

For this project, I chose to use the Naive Bayes classifier because it’s a simple yet powerful algorithm, especially for binary classification problems like mine, predicting whether a user is addicted or not. My dataset included several numerical features such as App\_Usage\_Time, Battery\_Drain, and Data\_Usage, which are fit for the Gaussian Naive Bayes variant since it handles continuous data under the assumption of normal distribution.

**What I Learned from Using Naive Bayes**

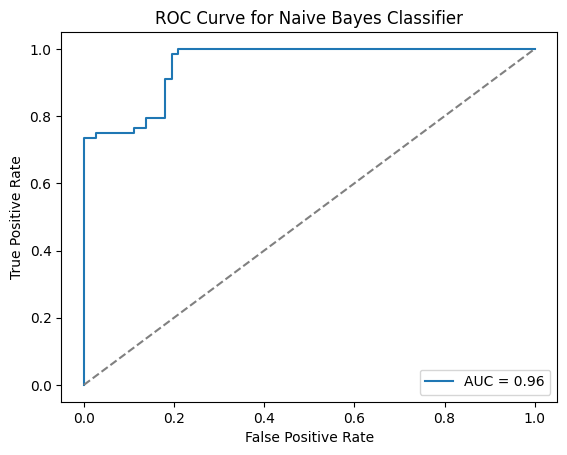
After training the Naive Bayes model, I got a training accuracy of 90% and a testing accuracy of approximately 83.6%. These results indicate that the model generalizes fairly well and isn't simply memorizing the training data.

To evaluate the quality of the predictions, I looked at the F1 scores, which were:

* **0.84 for “No” (Not Addicted)**
* **0.83 for “Yes” (Addicted)**

These scores are balanced, suggesting that the model doesn’t favor one class over the other, which is important in a behavioral prediction task where both false positives and false negatives can have implications.

The ROC Curve (fig 1.1) also showed promising results with an AUC score of 0.96, indicating that the model is very good at distinguishing between addicted and non-addicted users.

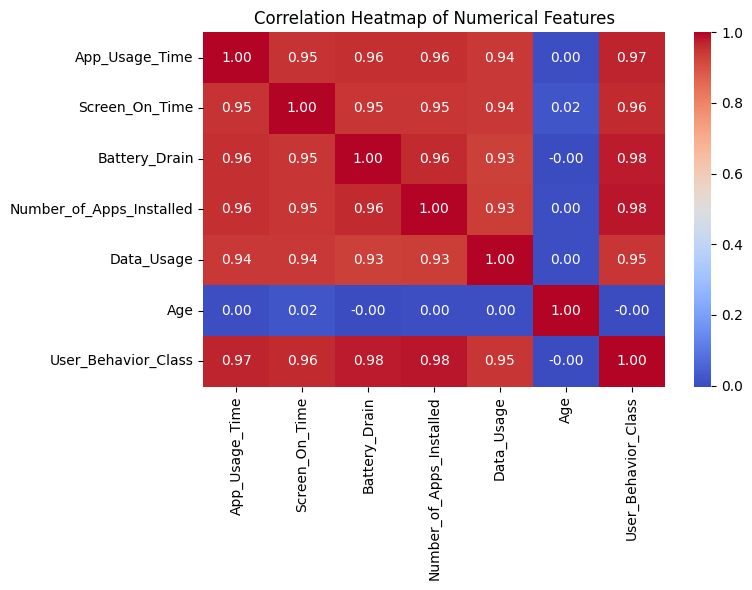


**Fig 1.1** This plot shows how well the model distinguishes between the two classes. A curve close to the diagonal indicates weak performance.

**Insights from the Correlation Matrix**

When I looked at the correlation matrix (fig 1.2), I noticed that many features were highly correlated with each other and with the target class. For example:

* Battery\_Drain, App\_Usage\_Time, and Number\_of\_Apps\_Installed had correlations above 0.95 with each other and with User\_Behavior\_Class.

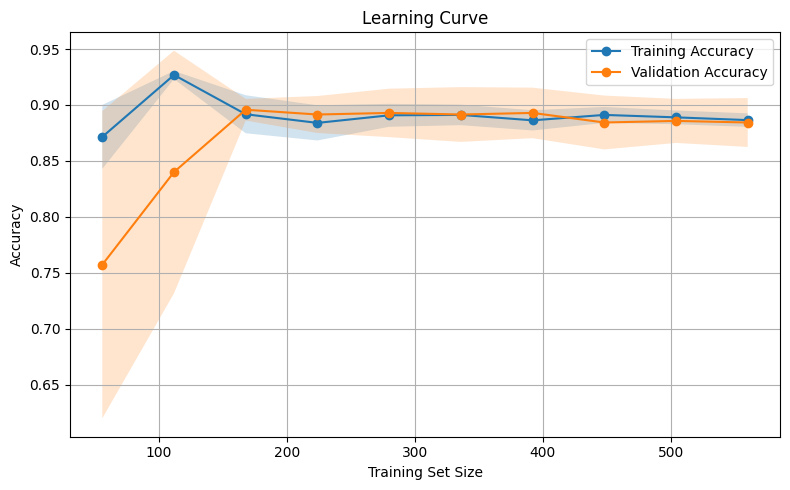
****This is important because Naive Bayes assumes all features are independent, which clearly isn’t the case here. The strong correlations between features likely affected the model’s accuracy. In theory, Naive Bayes may overestimate the influence of certain features when they carry overlapping information. This could explain why the accuracy wasn’t higher, despite the model doing reasonably well overall.

**Fig 1.2**

**Is the Model Overfitting or Underfitting?**

The learning curve (fig 1.3) gave me more insight into the model's behavior. Both the training and validation accuracy start off lower with small dataset sizes but converge to around 89% as the size increases. This is a good sign that the model is not overfitting, since the gap between training and validation is small.

However, since the accuracy levels off and doesn’t improve much with more data, it also shows some signs of underfitting. This makes sense because Naive Bayes is a fairly simple model and doesn’t capture complex relationships or feature interactions, especially when features are dependent.



**Fig 1.3**