**Building a Student Intervention System**

**1. Classification vs. Regression**

The student intervention system is fundamentally a classification problem. We are attempting to identify at risk students and classify them as such in order to begin an intervention process. For this problem either the student is at risk or is not. We are not attempting to predict specific outcomes for each student, merely whether or not they are in need of intervention.

**2. Exploring the Data**

* Total number of students: 395
* Number of students who passed: 265
* Number of students who failed: 130
* Number of features: 31
* Graduation rate of the class: 0.67%

**4. Training and Evaluating Models**

The three models chosen, in order, are Logistic Regression, SVM, and K Nearest Neighbor.

* Logistic Regression

Table:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Training Set Size | | |
|  | 100 | 200 | 300 |
| Training Time (secs) | 0.002 | 0.003 | 0.004 |
| Prediction Time (secs) | 0.000 | 0.001 | 0.000 |
| F1 Score Training Set | 0.8740 | 0.86689 | 0.843267 |
| F1 Score Test Set | 0.78740 | 0.80315 | 0.794118 |

* SVM

|  |  |  |  |
| --- | --- | --- | --- |
|  | Training Set Size | | |
|  | 100 | 200 | 300 |
| Training Time (secs) | 0.002 | 0.003 | 0.006 |
| Prediction Time (secs) | 0.001 | 0.001 | 0.002 |
| F1 Score Training Set | 0.84768 | 0.8690 | 0.87212 |
| F1 Score Test Set | 0.76056 | 0.76712 | 0.73611 |

* K Nearest Neighbors

|  |  |  |  |
| --- | --- | --- | --- |
|  | Training Set Size | | |
|  | 100 | 200 | 300 |
| Training Time (secs) | 0.001 | 0.001 | 0.001 |
| Prediction Time (secs) | 0.002 | 0.002 | 0.005 |
| F1 Score Training Set | 0.78049 | 0.81595 | 0.824268 |
| F1 Score Test Set | 0.76510 | 0.741259 | 0.732394 |

**5. Best Model**

The best model given the constraints is Logistic Regression. The chosen model has the highest F1 score while simultaneously having the lowest training/prediction times.