JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Naive Bayes Classification for Predicting Student Academic Performance

A Repeated Cross-Validated Evaluation of Categorical Naive Bayes, Semi-Naive and Hybrid Variants on the Higher Education Students Performance Evaluation Dataset

Najma Barkat[1] · Dr. Chetna Bhavsar[2]
[1]Research Scholar, Statistics Department, Gujarat University, Ahmedabad, India
[2] Rtd. Professor, Statistics Department, Gujarat University, Ahmedabad, India
Correspondence: najmabarkat22@gmail.com

Abstract — This paper evaluates Naive Bayes (NB) classification for predicting student academic performance on the Higher Education Students Performance Evaluation dataset (Yilmaz & Sekeroglu, 2019): 145 students at Near East University, Cyprus, described by 30 categorical or ordinal predictors and a final grade on an eight-point scale. Categorical NB is compared with nine NB-family variants (feature-selection NB, AODE, tree-augmented NB, Bernoulli and Complement NB, class-prior and threshold adjustments, random oversampling, and a soft-voting ensemble) and five benchmark learners on four targets: the original eight-class grade, a three-class grade band, the original Fail/Pass split (8 versus 137 students), and a more balanced split (DD or lower versus above). Every result comes from 10 × 5 repeated stratified cross-validation with nested tuning and in-fold feature selection, and is reported with bootstrap 95% confidence intervals, Nadeau–Bengio corrected resampled t-tests, Holm adjustment, and permutation tests. On the Fail/Pass target, Categorical NB reaches a ROC-AUC of 0.806 (95% CI 0.62–0.96; permutation p = 0.002), which is clearly above chance but very imprecise because only 8 students failed; its accuracy (92.6%) does not exceed the 94.5% majority baseline. On the eight-class target its accuracy is 24.8% (95% CI 19.8–30.3), statistically indistinguishable from the 24.1% majority baseline (p = 0.832), although a quadratic weighted kappa of 0.24 (permutation p = 0.003) shows that it recovers part of the ordering of grades. A single five-fold split can return a considerably higher figure (29.0%) than any of the ten repeated-CV estimates (23.4–27.6%). On the three-class target accuracy is 49.3% against a 46.2% baseline; the best hybrid (feature-selection NB) reaches 54.3%, but this is not significant after multiple-comparison correction. No NB variant and no benchmark learner significantly outperforms the majority baseline on any target once Holm adjustment is applied. Pairwise feature associations are weak (mean bias-corrected Cramér's V = 0.089), and the gap between training and test accuracy (49 points on the eight-class target) shows that NB overfits at this sample size rather than resisting overfitting. Sex and transportation mode were the only features consistently ranked as informative when importance was recomputed inside the cross-validation folds. The paper concludes that Categorical NB is a transparent and cheap risk-ranking tool for this dataset, with limited and imprecisely estimated predictive value for exact grades.

Keywords — *Naive Bayes; Categorical Naive Bayes; Tree-Augmented Naive Bayes; AODE; educational data mining; student performance prediction; repeated cross-validation; class imbalance; permutation test; Cramér's V; Python; scikit-learn*

1. Introduction

Predicting student academic performance from demographic, socio-economic, and behavioral data is a well-established application area within educational data mining, motivated chiefly by the goal of identifying students at risk of underperformance or attrition early enough for institutional support to be directed where it is most needed (Baker & Yacef, 2009; Romero & Ventura, 2010; Shahiri, Husain, & Rashid, 2015). Achieving this goal requires classification techniques that are not only reasonably accurate but also practical to deploy and interpret in

real institutional settings, where computational resources, statistical expertise, and sample sizes are frequently limited.

Naive Bayes classification is a natural candidate under these constraints. Grounded directly in Bayes' theorem, it estimates the probability of each outcome class from the observed distribution of predictor values under a simplifying assumption of conditional independence between features given the class. The assumption keeps the number of estimated parameters small and makes the model cheap to fit and update, which suits datasets such as the one examined here, where observations are few relative to the number of predictors.

This paper applies Naive Bayes classification to the Higher Education Students Performance Evaluation dataset, 145 students at Near East University, Cyprus (Yilmaz & Sekeroglu, 2019). Small samples make performance estimates unstable, so the study is designed around three safeguards: repeated rather than single cross-validation, explicit uncertainty quantification (bootstrap confidence intervals, corrected significance tests and permutation tests), and the comparison of plain Categorical NB with hybrid and semi-naive variants and with standard benchmark learners, so that the limits of the technique can be separated from the limits of the data.

2. Literature Review

2.1 Educational data mining and student performance prediction

Educational data mining (EDM) applies data mining and machine learning to educational records in order to understand and predict learner outcomes (Baker & Yacef, 2009; Romero & Ventura, 2010). Systematic reviews identify decision trees, Naive Bayes, neural networks, support vector machines and ensemble methods as the most commonly used techniques for predicting grades and pass/fail status, with prior grades, demographic variables, and engagement indicators as the most frequent predictors (Shahiri et al., 2015). Early applied studies already contrasted several learners on small institutional datasets: Kotsiantis, Pierrakeas, and Pintelas (2004) predicted performance in a distance-learning programme from demographic data and assignment marks and found Naive Bayes competitive with more elaborate learners; Cortez and Silva (2008) predicted secondary-school grades in Portugal using decision trees, random forests, neural networks and support vector machines; Bhardwaj and Pal (2011) applied Bayesian classification to student records; and Amrieh, Hamtini, and Aljarah (2016) showed that behavioral engagement features and ensemble methods improved prediction of academic performance.

A recurring feature of this literature is that datasets are small and outcomes imbalanced, yet results are often reported as a single accuracy figure from one cross-validation or hold-out run, without confidence intervals or significance tests. That practice is risky precisely in the regime this paper studies, where a handful of students can change an estimate by several percentage points.

2.2 The dataset analyzed in this paper

The dataset was collected and described by Yilmaz and Sekeroglu (2019) from students of the Faculty of Engineering and the Faculty of Educational Sciences at Near East University, Cyprus. It contains 145 students, 31 attributes (30 predictors plus a course identifier) spanning personal characteristics, family background, socio-economic indicators and study habits, and each student's final course grade on an eight-point scale (Fail, DD, DC, CC, CB, BB, BA, AA). It was released through the UCI Machine Learning Repository under a Creative Commons Attribution 4.0 International licence. Related work from the same research group applied several machine learning algorithms to student datasets and reported that pre-processing of the raw data improved classification performance (Sekeroglu, Dimililer, & Tuncal, 2019). Other studies have also used this dataset; the present study differs in that it compares Naive Bayes variants and benchmark learners under a repeated, nested, uncertainty-quantified protocol.

2.3 Naive Bayes and its relatives

Naive Bayes has a long history in text categorization and medical diagnosis (Rish, 2001; Han, Kamber, & Pei, 2011). Although the conditional-independence assumption rarely holds exactly, the classifier is often competitive with more sophisticated methods: Domingos and Pazzani (1997) showed that the classifier can be optimal under zero-one loss even when independence is violated, and Zhang (2004) attributed this to the fact that classification depends on the decision boundary the model implies rather than on the accuracy of its probability estimates. The same body of work also records the technique's weaknesses: probability estimates are poorly calibrated because correlated features are double-counted (Niculescu-Mizil & Caruana, 2005), and performance degrades when the sample is too small to estimate class-conditional distributions.

Two families of remedies have been proposed. The first reduces the influence of redundant or irrelevant attributes through feature selection, using filter criteria such as mutual information or chi-square, correlation-based selection (Hall, 2000), or wrapper search (Langley & Sage, 1994; Guyon & Elisseeff, 2003). The second relaxes the independence assumption in a controlled way. Tree-Augmented Naive Bayes (TAN) allows each feature one additional parent feature chosen by conditional mutual information (Friedman, Geiger, & Goldszmidt, 1997); Averaged One-Dependence Estimators (AODE) average all models in which one feature is the parent of the others (Webb, Boughton, & Wang, 2005); and NBTree places Naive Bayes classifiers at the leaves of a decision tree (Kohavi, 1996). Other approaches combine Naive Bayes with other learners, for example hidden Naive Bayes (Jiang, Zhang, & Cai, 2009) or voting ensembles. These semi-naive methods add parameters, so their benefit on a sample of 145 students is an empirical question rather than a foregone conclusion; it is one the present study tests directly.

2.4 Evaluating classifiers on small, imbalanced samples

Several methodological points are central here. First, single hold-out or single k-fold estimates have high variance; repeated stratified cross-validation and bootstrap-based intervals give more stable and more honest estimates (Kohavi, 1995; Efron & Tibshirani, 1993). Second, any step that uses labels, including hyper-parameter tuning and feature selection, must be nested inside the resampling loop, otherwise performance is optimistically biased (Ambroise & McLachlan, 2002; Varma & Simon, 2006; Cawley & Talbot, 2010). Third, differences between classifiers evaluated on overlapping training sets should be tested with variance corrections: the corrected resampled t-test of Nadeau and Bengio (2003) addresses the dependence between folds that invalidates the naive paired t-test (Dietterich, 1998), and multiple comparisons should be controlled, for example with Holm's (1979) procedure. Fourth, permutation tests establish whether a cross-validated score exceeds what the same learner achieves on shuffled labels (Ojala & Garriga, 2010).

Fifth, with imbalanced targets accuracy is dominated by the majority class, so balanced accuracy, the Matthews correlation coefficient (Chicco & Jurman, 2020), Cohen's kappa (Cohen, 1960) and the precision–recall curve (Saito & Rehmsmeier, 2015) are more informative. When outcomes are ordered grades, exact-match accuracy penalises a near miss as heavily as a gross error; quadratic weighted kappa (Cohen, 1968) and within-one-grade accuracy respect the ordinal structure (Frank & Hall, 2001). Finally, class imbalance can be addressed by prior adjustment, threshold tuning or resampling such as SMOTE and SMOTE-NC (Chawla, Bowyer, Hall, & Kegelmeyer, 2002), always inside the training folds. These principles shape the protocol described in Section 4.5.

3. Research Objectives

This study pursues five objectives:

- To estimate, with repeated cross-validation and uncertainty intervals, how well Categorical Naive Bayes predicts student grade outcomes from the 30 demographic, socio-economic and study-habit predictors, under four target formulations (eight-class grade, three-class band, Fail/Pass, and DD-or-lower versus above).
- To test whether Naive Bayes hybrids that address small samples and dependent features (feature selection, AODE, TAN, Bernoulli and Complement NB, soft voting) and imbalance remedies (prior adjustment, threshold tuning, oversampling) improve on plain Categorical NB, and how they compare with logistic regression, support vector machines, k-nearest neighbours, random forest and gradient boosting.
- To establish by significance testing (corrected resampled t-tests, permutation tests) which of the observed differences and which of the observed performance levels exceed what chance alone would produce.
- To measure, rather than assume, the dependence between predictors, and to assess whether the conditional-independence assumption is the practical bottleneck.
- To identify which predictors carry signal, using importance measures computed inside the cross-validation folds so that no information from held-out students leaks into the ranking.

4. Data and Methodology

4.1 Dataset Description

The dataset comprises 145 students and 33 columns: a student identifier (excluded from modeling), 30 categorical or ordinal predictor attributes, a course identifier (also excluded), and the target grade. The 30 predictors include personal characteristics (age, sex), family background (parental education and occupation, parental status, number of siblings), socio-economic indicators (scholarship type, additional

work status, total salary, accommodation type, transportation mode), and study habits and engagement (weekly study hours, class attendance, note-taking, seminar attendance, reading frequency, examination-preparation methods, and previous and expected GPA). All predictors are encoded as small-integer category codes, and there are no missing values. Note that the workbook accompanying this paper renders some "None" answers as blank text when read with default spreadsheet-import settings; these are valid category labels, not missing data.

The target grade takes eight values (0 = Fail through 7 = AA), distributed as Fail = 8, DD = 35, DC = 24, CC = 21, CB = 10, BB = 17, BA = 13, AA = 17. Section 4.4 describes the four targets derived from it.

4.2 The Naive Bayes Algorithm and Hybrid Variants

Naive Bayes applies Bayes' theorem to estimate the posterior probability of each class given the observed predictor values, under the assumption that predictors are conditionally independent given the class. For a class c and predictors x1…xn, the posterior P(c | x1,…,xn) is proportional to the prior P(c) multiplied by the product of the likelihoods P(xi | c); the predicted class maximizes this posterior (Han et al., 2011). **Categorical NB**, the reference model, estimates each P(xi | c) as a categorical distribution with Laplace smoothing parameter α. The reference configuration uses α = 1 ("Categorical NB, α = 1"); a tuned-α version is also evaluated.

The following variants address the weaknesses of plain NB at small N:

- **Feature selection + NB.** The k most informative features are retained using mutual information (MI) or chi-square, computed on the training fold only ($k \in \{5, 8, 12, 18\}$; $\alpha \in \{0.5, 1, 3\}$).
- **AODE** (Webb et al., 2005): the average of 30 one-dependence estimators in which each feature in turn is the parent of all others ($\alpha \in \{0.3, 1, 3, 10\}$), implemented without a frequency limit.
- **Tree-Augmented NB (TAN)** (Friedman et al., 1997): a maximum-weight spanning tree over features using conditional mutual information given the class, so that each feature has at most one additional feature parent.
- **Bernoulli NB and Complement NB** on a one-hot encoding of the predictors, as substitutes for Gaussian NB on ordinal codes.
- **Imbalance remedies** (binary targets only): uniform class priors; a decision threshold tuned on inner out-of-fold probabilities to maximize balanced accuracy; and random oversampling of the minority class inside the training fold. SMOTE-NC (Chawla et al., 2002) was not available in the computing environment, so random oversampling is used as the simpler analogue.
- **Soft-voting ensemble** averaging the class probabilities of Categorical NB, a random forest and an L2-regularized logistic regression.
- **Benchmarks:** L2 logistic regression, linear support vector machine and k-nearest neighbours (all on one-hot features, with the regularization strength or k tuned), a random forest with 200 trees (Breiman, 2001), and gradient boosting (60 depth-2 trees) as a stand-in for XGBoost, which was not installed.
- **Gaussian NB on ordinal codes** is retained only for completeness (Appendix B). Modeling ordinal integers as normally distributed variables is a known mismatch to the data, so its poor results are not interpreted as evidence about Naive Bayes.

Hyper-parameters were tuned by an inner three-fold stratified cross-validation within each training fold, minimizing log-loss; all feature selection and resampling were likewise confined to the training fold (nested validation; Varma & Simon, 2006; Cawley & Talbot, 2010). NBTree, full Bayesian-network structure learning and boosted or bagged NB were not evaluated and are listed as future work.

4.3 Preprocessing

The student identifier and course identifier were dropped. For all NB models on ordinal codes, predictors were converted to zero-indexed integer category codes, with the number of categories per feature fixed to the full observed range so that categories absent from a training fold do not cause errors. Logistic regression, the support vector machine, k-nearest neighbours and the Bernoulli and Complement NB variants used a one-hot encoding fitted with the same fixed category ranges; random forest and gradient boosting used the integer codes. No missing-value imputation was required.

4.4 Target Variable Formulations

Four targets were evaluated, because the original two cannot support firm conclusions at N = 145:

- **Eight-class grade** (original): Fail, DD, DC, CC, CB, BB, BA, AA, treated as ordinal. Classes average about 18 students and the smallest has 8.

- **Three-class band**: Low (Fail, DD, DC; 67 students), Medium (CC, CB, BB; 48) and High (BA, AA; 30).
- **Fail versus Pass** (original binary): Fail = 8, Pass = 137 (5.5% positives). The at-risk (Fail) class is treated as the positive class for sensitivity, PR-AUC and ROC-AUC.
- **DD or lower versus above** (balanced binary): 43 students at DD or lower versus 102 above (29.7% positives), a cut-off that gives enough positives for meaningful evaluation.

4.5 Evaluation Protocol and Statistical Methods

Resampling. Performance was estimated by 10 × 5 repeated stratified cross-validation (50 train/test partitions, random seed 42), using identical partitions for all models on a given target so that models can be compared pairwise. Each test fold holds 29 students and each training fold 116.

Metrics. Accuracy; balanced accuracy; macro-averaged F1; Matthews correlation coefficient (MCC); Cohen's kappa; for the ordinal targets, quadratic weighted kappa (QWK) and, for the eight-class target, accuracy within one grade; and for the binary targets, ROC-AUC, PR-AUC, sensitivity and specificity for the at-risk class, and F1. In the binary tables "F1 (Pass)" and "F1 (>DD)" refer to the F1 of the majority (not-at-risk) class; this is why a classifier that predicts Pass for everyone scores 0.972 on the Fail/Pass target. Point estimates are means over the 10 repeats of each metric computed on that repeat's pooled out-of-fold predictions. Fold-level "±" values quoted in earlier single-split results are the population standard deviation (ddof = 0) of the five fold accuracies; this was verified by reproducing the majority baseline (94.5% ± 1.7%, 24.1% ± 0.0%). Fold-level SDs are given explicitly where used in this paper.

Confidence intervals. 95% percentile intervals from 600 stratified bootstrap resamples of the students, applied to the pooled out-of-fold predictions of all 10 repeats (Efron & Tibshirani, 1993). These intervals reflect sampling variability of the test students; they do not capture variability from refitting on different training sets, and are therefore, if anything, somewhat narrow.

Significance tests. Models were compared with the majority baseline and with Categorical NB ($\alpha = 1$) by the corrected resampled t-test of Nadeau and Bengio (2003) on the 50 paired fold scores (variance inflated by 1/50 + 29/116; accuracy for the multi-class targets and balanced accuracy for the binary targets). Holm's (1979) step-down adjustment was applied across the models compared within each target. Wilcoxon signed-rank p-values on the same paired scores are provided in the accompanying workbook but are treated as anti-conservative, since fold scores are not independent. To test whether a score exceeds chance, 1,000 label permutations were run with Categorical NB ($\alpha = 1$) on fixed 3 × 5-fold partitions; $p = (1 + \text{number of permuted scores} \geq \text{observed})/(1 + 1{,}000)$ (Ojala & Garriga, 2010).

Feature dependence and importance. Pairwise association among the 30 predictors was measured with the bias-corrected Cramér's V (Bergsma, 2013). Feature importance was computed inside each of the 50 training folds by (i) the log-probability separation score between the two classes, (ii) mutual information, (iii) chi-square, and (iv) permutation importance, the increase in held-out log-loss when a feature is shuffled in the test fold (five shuffles) (Altmann, Toloşi, Sander, & Lengauer, 2010). Rankings are summarized by mean rank and by the share of folds in which a feature is in the top five.

Software. Python with pandas, NumPy, SciPy and scikit-learn (Pedregosa et al., 2011); AODE, TAN and the wrapper models were implemented directly (Appendix A).

5. Results

5.1 Stability of Single-Split Estimates

Before the main results, Table 1 shows how much a single five-fold split can vary. For Categorical NB ($\alpha = 1$), the ten repeats of 5-fold cross-validation give eight-class accuracies of 23.4–27.6% and Fail/Pass accuracies of 91.7–93.1%. A single-split eight-class accuracy of 29.0% obtained in a preliminary analysis of this dataset lies above every one of the ten repeated-CV estimates; it is best read as a favorable split rather than as the expected performance of the model. All estimates below use the full 10 × 5 protocol.

Table 1. Variability of single 5-fold cross-validation estimates for Categorical NB (α = 1), across ten repeats

Target	Mean of 10 repeats (%)	Min–max (%)	SD across repeats (pts)	Majority baseline (%)
Eight-class grade	24.8	23.4–27.6	1.5	24.1
Three-class band	49.3	45.5–52.4	1.9	46.2
Fail vs Pass	92.6	91.7–93.1	0.5	94.5
≤DD vs above	59.2	55.9–62.1	1.8	70.3

Each repeat is one complete 5-fold stratified cross-validation with a different random partition; accuracy is pooled over its five test folds.

5.2 Fail/Pass Classification

Table 2 reports repeated cross-validated results for the original binary target (8 Fail, 137 Pass), for the reference model, the imbalance remedies and selected hybrids; all 18 models are in Appendix B. Categorical NB (α = 1) attains an accuracy of 92.6% against 94.5% for the majority baseline, so it does not exceed the baseline on raw accuracy (fold-level SD 4.0 points). With only 8 Fail cases the baseline reaches 94.5% by predicting Pass for everyone, so accuracy is uninformative here. At the default decision rule the model detects 45% of Fail students (sensitivity 0.45, about 3.6 of 8) with specificity 0.95 (about 6.4 false alarms among 137 Pass students; Figure 1), a balanced accuracy of 70.2% (95% CI 53.3–86.4) and an MCC of 0.36 (0.07–0.63).

Table 2. Fail vs Pass (positive class = Fail): 10 × 5 repeated stratified CV, selected models

Model	Acc. (%)	Bal. acc. (%)	F1 (Pass)	Sens.	Spec.	MCC	ROC-AUC [95% CI]	PR-AUC [95% CI]
Majority baseline	94.5	50.0	0.972	0.00	1.00	0.00	0.500	0.055
Categorical NB (α = 1)	**92.6**	**70.2**	**0.960**	**0.45**	**0.95**	**0.36**	**0.806 [0.62–0.96]**	**0.405 [0.15–0.72]**
Categorical NB (tuned α)	95.2	65.7	0.975	0.33	0.99	0.43	0.809 [0.64–0.95]	0.417 [0.16–0.73]
NB, uniform priors	92.3	72.4	0.959	0.50	0.95	0.38	0.825 [0.67–0.96]	0.502 [0.21–0.80]
NB + threshold tuning	85.7	71.8	0.920	0.56	0.87	0.28	0.806 [0.62–0.96]	0.405 [0.15–0.72]
NB + random oversampling	91.1	60.0	0.953	0.25	0.95	0.19	0.724 [0.53–0.90]	0.222 [0.09–0.46]
FS(chi2)+NB	92.8	55.6	0.962	0.14	0.97	0.14	0.717 [0.53–0.88]	0.240 [0.09–0.48]
AODE	95.0	56.8	0.974	0.14	1.00	0.33	0.815 [0.65–0.95]	0.447 [0.18–0.75]
Complement NB (one-hot)	94.8	70.7	0.972	0.44	0.98	0.45	0.825 [0.66–0.96]	0.499 [0.21–0.80]
Soft voting NB+RF+LR	94.4	52.9	0.971	0.06	1.00	0.14	0.814 [0.63–0.96]	0.334 [0.15–0.61]
Random forest	94.5	50.0	0.972	0.00	1.00	0.00	0.799 [0.63–0.93]	0.252 [0.12–0.50]

F1 (Pass) is the F1-score of the Pass class. Sensitivity and specificity refer to the Fail class. 95% CIs: stratified bootstrap over students. The shaded row is the reference model.

ROC-AUC is 0.806 (95% CI 0.62–0.96) and PR-AUC 0.405 (0.15–0.72) against a no-skill PR-AUC of 0.055. The permutation test (Table 3, Figure 2) shows that this AUC is far above what the same learner obtains on shuffled labels (observed 0.812 in the 3 × 5-fold design, null mean 0.452, null 95th percentile 0.672; p = 0.002). The result supports the weaker statement that Categorical NB ranks Fail students above Pass students more often than chance would; it does not support the stronger statement of "genuine, moderate discriminative ability" without qualification, because the 95% CI for the AUC spans 0.62–0.96: with 8 positive cases the true discriminative ability could be anywhere from modest to very good.

Imbalance remedies change the operating point more than the ranking. Uniform class priors raise balanced accuracy to 72.4% and AUC to 0.825, and a tuned threshold gives 71.8% balanced accuracy at the cost of overall accuracy (85.7%). The most favorable comparison with the majority baseline (uniform-prior NB, balanced accuracy)

has an uncorrected p of 0.048, which becomes 0.866 after Holm adjustment across the 17 models; no model is distinguishable from the baseline once the number of comparisons is taken into account. The 70/30 hold-out confusion matrix reported in earlier work contained only 2 Fail students and is dropped; Figure 1 instead pools out-of-fold predictions so that all 8 Fail students contribute.

[Insert figure image here]

Figure 1. Pooled out-of-fold confusion matrices for Categorical NB (α = 1), averaged over 10 repeats of 5-fold CV (cells show mean counts; colour shows row proportions). Left: Fail vs Pass. Right: eight-class grade.

Table 3. Permutation tests for Categorical NB (α = 1): 1,000 label permutations, 3 × 5-fold partitions

Target	Statistic	Observed	Null mean	Null 95th pct.	p-value
Fail vs Pass	ROC-AUC	0.812	0.452	0.672	0.002
Fail vs Pass	PR-AUC	0.382	0.081	0.178	0.002
≤DD vs above	ROC-AUC	0.555	0.491	0.597	0.163
Three-class	Accuracy	0.499	0.377	0.444	0.003
Three-class	Quadratic weighted kappa	0.266	-0.007	0.129	<0.001
Eight-class	Accuracy	0.241	0.159	0.207	0.009
Eight-class	Quadratic weighted kappa	0.237	-0.007	0.117	0.003
Eight-class	Macro-F1	0.203	0.112	0.156	0.002

For accuracy and macro-F1 the permutation null is lower than the majority-class rate, because a model trained on shuffled labels does not learn to predict the most frequent class. These tests therefore show that the model beats random-label learning, not that it beats the majority baseline; the latter comparison is made by the corrected t-test in Section 5.6.

[Insert figure image here]

Figure 2. Permutation null distributions (1,000 shuffles) with the observed statistic (red line).

5.3 Balanced Binary Classification (DD or Lower versus Above)

Moving the cut-off to DD-or-lower versus above gives 43 positives and removes the extreme imbalance, but it does not improve performance (Table 4). Categorical NB (α = 1) obtains an accuracy of 59.2% against 70.3% for the majority baseline, a balanced accuracy of 49.7% (95% CI 43.4–55.9), and a ROC-AUC of 0.560 (0.47–0.65); the permutation p-value for the AUC is 0.163. The best AUC belongs to chi-square feature selection + NB (0.645, 0.56–0.73), and none of the models differs significantly from the baseline on balanced accuracy (uncorrected p ≥ 0.116). This target thus reveals that, with these predictors, no model has demonstrable ability to separate the two lowest grade levels from the rest. The Fail/Pass AUC of 0.806 should therefore be read with caution: it rests on 8 students.

Table 4. DD or lower (positive) vs above: 10 × 5 repeated stratified CV, selected models

Model	Acc. (%)	Bal. acc. (%)	F1 (>DD)	Sens.	Spec.	MCC	ROC-AUC [95% CI]	PR-AUC [95% CI]
Majority baseline	70.3	50.0	0.826	0.00	1.00	0.00	0.500	0.297
Categorical NB (α = 1)	**59.2**	**49.7**	**0.716**	**0.26**	**0.73**	**-0.01**	**0.560 [0.47–0.65]**	**0.318 [0.27–0.39]**
Categorical NB (tuned α)	64.5	51.4	0.768	0.19	0.84	0.03	0.568 [0.47–0.67]	0.343 [0.28–0.44]
NB, uniform priors	65.0	52.4	0.770	0.21	0.83	0.06	0.572 [0.47–0.67]	0.358 [0.29–0.48]
NB + threshold tuning	47.0	55.8	0.471	0.77	0.34	0.12	0.560 [0.47–0.65]	0.318 [0.27–0.39]
FS(MI)+NB	62.1	51.0	0.744	0.24	0.78	0.02	0.607 [0.52–0.70]	0.352 [0.30–0.43]
FS(chi2)+NB	64.8	53.8	0.763	0.27	0.81	0.09	0.645 [0.56–0.73]	0.400 [0.33–0.49]
AODE	67.0	52.8	0.788	0.18	0.88	0.08	0.591 [0.50–0.68]	0.379 [0.31–0.47]
TAN	68.8	49.5	0.814	0.02	0.97	-0.03	0.545 [0.46–0.63]	0.334 [0.28–0.42]
Complement NB (one-hot)	65.9	52.2	0.780	0.18	0.86	0.06	0.572 [0.47–0.67]	0.359 [0.29–0.48]
Soft voting NB+RF+LR	62.6	49.4	0.755	0.17	0.82	-0.01	0.573 [0.48–0.67]	0.342 [0.28–0.43]

Model	Acc. (%)	Bal. acc. (%)	F1 (>DD)	Sens.	Spec.	MCC	ROC-AUC [95% CI]	PR-AUC [95% CI]
Random forest	69.0	51.9	0.810	0.10	0.94	0.07	0.633 [0.54–0.72]	0.403 [0.33–0.50]

F1 (>DD) is the F1-score of the not-at-risk class (grade above DD). Sensitivity and specificity refer to the at-risk class (DD or lower).

5.4 Three-Class Grade Band

Collapsing the grade into three bands gives the most usable target (Table 5). Categorical NB (α = 1) reaches 49.3% accuracy (95% CI 42.5–55.5) against 46.2% for the majority class, with macro-F1 0.444 (baseline 0.211) and QWK 0.24 (0.11–0.36). The permutation tests are strongly significant (accuracy p = 0.003; QWK p = 0.001), so the model does learn something about the grade band. The advantage over the majority baseline is, however, not established: the corrected t-test gives p = 0.355. The best hybrid is mutual-information feature selection + NB (54.3% accuracy; uncorrected p = 0.024; Holm-adjusted p = 0.332), followed by chi-square selection (53.8%). With 17 models compared, this is suggestive rather than conclusive.

Table 5. Three-class grade band (Low / Medium / High): 10 × 5 repeated stratified CV, selected models

Model	Accuracy (%) [95% CI]	Macro-F1 [95% CI]	QWK [95% CI]	Bal. acc. (%)	MCC	Train acc. (%)
Majority baseline	46.2	0.211	0.000	33.3	0.00	46.2
Categorical NB (α = 1)	**49.3 [42.5–55.5]**	**0.444 [0.37–0.51]**	**0.24 [0.11–0.36]**	**44.4**	**0.18**	**75.8**
Categorical NB (tuned α)	50.6 [44.2–56.2]	0.384 [0.32–0.44]	0.14 [0.03–0.25]	41.1	0.16	66.0
FS(MI)+NB	54.3 [48.7–59.9]	0.455 [0.40–0.51]	0.22 [0.09–0.34]	46.4	0.25	64.7
FS(chi2)+NB	53.8 [48.7–58.3]	0.423 [0.38–0.47]	0.20 [0.09–0.30]	45.0	0.24	64.2
AODE	50.2 [43.5–56.3]	0.393 [0.32–0.46]	0.16 [0.03–0.26]	41.3	0.16	74.1
TAN	46.9 [42.7–51.3]	0.337 [0.30–0.38]	0.09 [0.01–0.18]	37.2	0.09	81.5
Bernoulli NB (one-hot)	49.4 [42.5–56.5]	0.441 [0.37–0.52]	0.22 [0.08–0.34]	44.0	0.18	71.9
Complement NB (one-hot)	51.3 [44.5–58.1]	0.451 [0.38–0.53]	0.21 [0.08–0.33]	45.1	0.20	72.3
Soft voting NB+RF+LR	50.5 [44.0–56.3]	0.446 [0.37–0.52]	0.23 [0.10–0.36]	44.5	0.19	84.4
Logistic regression	51.7 [46.1–56.9]	0.402 [0.33–0.47]	0.19 [0.08–0.30]	42.3	0.18	69.0
Random forest	52.1 [46.8–57.2]	0.416 [0.36–0.47]	0.20 [0.09–0.29]	43.3	0.20	98.5
Gradient boosting	53.0 [47.0–59.0]	0.473 [0.41–0.54]	0.27 [0.16–0.39]	47.2	0.24	95.4

Train acc. = accuracy on the training folds (see Section 5.7). QWK = quadratic weighted kappa with equally spaced ordinal weights.

5.5 Eight-Class Grade

On the original eight-class target (Table 6), Categorical NB (α = 1) attains 24.8% accuracy (95% CI 19.8–30.3) against 24.1% for the majority class (DD); the difference is not significant (corrected t-test p = 0.832). Its macro-F1 of 0.208 (0.16–0.26) is far above the baseline's 0.049, but that comparison is uninformative: a classifier that always predicts one class has near-zero macro-F1 by construction, whereas the model spreads its predictions across classes, correctly or not. The more telling ordinal metrics are a QWK of 0.24 (0.10–0.39; permutation p = 0.003) and a within-one-grade accuracy of 48.5% against 46.2% for the baseline. Taken together, Categorical NB captures a modest part of the ordering of grades but does not improve exact-match accuracy beyond the majority rate. The confusion matrix (Figure 1, right) shows predictions scattered around the diagonal rather than concentrated on it.

Table 6. Eight-class grade: 10 × 5 repeated stratified CV, selected models

Model	Accuracy (%) [95% CI]	Macro-F1 [95% CI]	QWK [95% CI]	Within ±1 grade (%)	Bal. acc. (%)	MCC	Train acc. (%)
Majority baseline	24.1	0.049	0.000	46.2	12.5	0.00	24.1
Categorical NB (α = 1)	**24.8 [19.8–30.3]**	**0.208 [0.16–0.26]**	**0.24 [0.10–0.39]**	**48.5**	**21.1**	**0.10**	**73.8**
Categorical NB (tuned α)	23.6 [20.6–26.3]	0.099 [0.06–0.14]	0.04 [-0.04–0.11]	44.7	14.1	0.02	41.3
FS(MI)+NB	25.8 [22.0–29.5]	0.145 [0.12–0.17]	0.11 [-0.00–0.21]	48.3	16.9	0.08	45.9
FS(chi2)+NB	27.9 [24.5–31.3]	0.151 [0.11–0.19]	0.07 [0.01–0.14]	52.1	18.3	0.11	39.6
AODE	24.7 [21.4–27.8]	0.120 [0.07–0.17]	0.10 [0.00–0.20]	46.5	15.4	0.05	56.1
TAN	23.8 [21.4–26.3]	0.094 [0.07–0.12]	0.03 [-0.05–0.10]	47.5	13.6	0.03	68.5
Bernoulli NB (one-hot)	25.0 [20.5–29.9]	0.177 [0.13–0.23]	0.23 [0.10–0.36]	48.5	18.2	0.08	64.2
Complement NB (one-hot)	25.6 [20.9–30.6]	0.196 [0.15–0.24]	0.20 [0.07–0.33]	46.3	20.9	0.10	68.1
Soft voting NB+RF+LR	24.3 [19.5–29.9]	0.195 [0.14–0.24]	0.23 [0.08–0.37]	48.1	19.6	0.08	81.5
Logistic regression	24.9 [21.1–28.6]	0.133 [0.10–0.16]	0.14 [0.04–0.24]	47.8	15.8	0.06	53.4
Random forest	25.4 [21.2–29.7]	0.162 [0.12–0.20]	0.20 [0.08–0.31]	52.8	17.6	0.08	99.3
Gradient boosting	23.4 [18.7–28.6]	0.180 [0.14–0.21]	0.30 [0.17–0.42]	53.0	18.4	0.09	99.7

Within ±1 grade = share of students whose predicted grade is within one grade of the true grade. Majority baseline = always predict DD.

5.6 Hybrid and Benchmark Models

Table 7 summarizes each model's primary metric and its corrected t-test p-value against the majority baseline on every target. None of the 17 models differs significantly from the baseline after Holm adjustment on any target; the only highly significant contrasts are in the opposite direction (Gaussian NB on ordinal codes is significantly worse than the baseline on the multi-class targets). Differences among the NB variants and the benchmark learners are small relative to their confidence intervals (Figure 3). On the eight-class target the best accuracy is 27.9%; the benchmark learners reach 25.4% (random forest), 24.9% (logistic regression) and 23.4% (gradient boosting), so Categorical NB is not inferior to more flexible learners, but nothing is superior to it.

The hybrids behave as follows. Feature-selection NB most often retained only the k = 5 highest-scoring features (the modal choice in 46 of 50 folds on the three-class target) and gives the largest gain on the three-class band (above). AODE (24.7% eight-class, 50.2% three-class) and TAN (23.8%, 46.9%) do not improve on plain NB: with 116 training students, 8 classes and conditional probability tables conditioned on a parent feature, the extra parameters cost about as much as the dependence they model. Tuning α by log-loss selected α = 10 in most folds, which reduces the train–test gap (Section 5.7) but pulls predictions toward the class priors, so accuracy and macro-F1 fall (eight-class: 23.6% and 0.099). The soft-voting ensemble (24.3%, 50.5%) does not beat its best member. For the binary targets, Bernoulli and Complement NB produce the highest Fail/Pass balanced accuracy and AUC among the non-prior-adjusted models (Complement NB: 70.7% and 0.825), again within the wide intervals of the reference model.

Table 7. Primary metric (% and p-value of the corrected resampled t-test versus the majority baseline)

Model	8-class acc. % (p)	3-class acc. % (p)	Fail/Pass bal. acc. % (p)	≤DD/above bal. acc. % (p)
Majority baseline	24.1 (ref.)	46.2 (ref.)	50.0 (ref.)	50.0 (ref.)
Categorical NB (α = 1)	**24.8 (0.832)**	**49.3 (0.355)**	**69.7 (0.074)**	**49.7 (0.936)**
Categorical NB (tuned α)	23.6 (0.754)	50.6 (0.180)	65.9 (0.129)	51.4 (0.690)
FS(MI)+NB	25.8 (0.667)	54.3 (0.024)	53.5 (0.592)	51.1 (0.748)
FS(chi2)+NB	27.9 (0.231)	53.8 (0.057)	55.7 (0.472)	53.9 (0.226)
AODE	24.7 (0.802)	50.2 (0.212)	56.9 (0.359)	52.8 (0.497)

Model	8-class acc. % (p)	3-class acc. % (p)	Fail/Pass bal. acc. % (p)	≤DD/above bal. acc. % (p)
TAN	23.8 (0.877)	46.9 (0.804)	50.9 (0.721)	49.5 (0.754)
Bernoulli NB (one-hot)	25.0 (0.758)	49.4 (0.366)	71.0 (0.066)	52.4 (0.503)
Complement NB (one-hot)	25.6 (0.632)	51.3 (0.153)	71.4 (0.077)	52.2 (0.519)
Logistic regression	24.9 (0.827)	51.7 (0.095)	50.5 (0.787)	50.0 (0.996)
SVM (linear, one-hot)	23.1 (0.755)	50.8 (0.138)	53.0 (0.609)	48.8 (0.721)
k-NN (one-hot)	23.1 (0.702)	48.0 (0.619)	50.0 (1.000)	50.1 (0.949)
Random forest	25.4 (0.701)	52.1 (0.186)	50.0 (1.000)	51.9 (0.322)
Gradient boosting	23.4 (0.828)	53.0 (0.099)	49.1 (0.195)	53.0 (0.449)
Soft voting NB+RF+LR	24.3 (0.959)	50.5 (0.267)	53.3 (0.581)	49.5 (0.875)
NB, uniform priors	—	—	72.4 (0.048)	52.4 (0.511)
NB + threshold tuning	—	—	71.7 (0.059)	55.8 (0.116)
NB + random oversampling	—	—	59.0 (0.269)	49.5 (0.915)

Primary metric: accuracy for the 8-class and 3-class targets; balanced accuracy for the two binary targets (computed fold by fold). p-values are uncorrected for multiple comparisons; after Holm adjustment none is below 0.05 for any model shown. Gaussian NB (Appendix B) is significantly worse than the baseline on the multi-class targets. Bernoulli, Complement, imbalance-remedy and ensemble models are as described in Section 4.2; "—" = not applicable.

[Insert figure image here]

Figure 3. Model comparison with 95% bootstrap confidence intervals. Dark blue: reference Categorical NB (α = 1); blue: other NB-family models; grey: benchmark learners. Dashed line: majority baseline (accuracy panels) or chance (AUC panels).

5.7 Overfitting

Naive Bayes is often described as resistant to overfitting because it estimates few parameters per feature. Table 8 tests that statement by comparing accuracy on the training folds with accuracy on the test folds. For Categorical NB (α = 1) the eight-class training accuracy is 73.8% against 24.8% on test folds, a gap of 49.1 points; on the three-class target the gap is 26.4 points. The gap is large because the model has 30 features × 8 classes of categorical probabilities to estimate from 116 students. Tree ensembles show the same behavior to a greater degree (random forest: 99.3% train, 25.4% test), whereas tuned smoothing narrows the gap to 17.7 points by shrinking predictions toward the priors. NB is therefore not unusually robust against overfitting on this sample; what it does offer is low computational cost and interpretability. The earlier statement that NB is "resistant to overfitting" is withdrawn.

Table 8. Training versus test accuracy (mean over 50 folds)

Model	8-class train (%)	8-class test (%)	Gap (pts)	3-class train (%)	3-class test (%)	Gap (pts)
Categorical NB (α = 1)	**73.8**	**24.8**	**49.1**	**75.8**	**49.3**	**26.4**
Categorical NB (tuned α)	41.3	23.6	17.7	66.0	50.6	15.4
FS(MI)+NB	45.9	25.8	20.1	64.7	54.3	10.3
AODE	56.1	24.7	31.4	74.1	50.2	23.9
TAN	68.5	23.8	44.7	81.5	46.9	34.6
Logistic regression	53.4	24.9	28.5	69.0	51.7	17.4
SVM (linear, one-hot)	88.4	23.1	65.3	74.4	50.8	23.5
k-NN (one-hot)	32.6	23.1	9.4	55.1	48.0	7.1

Model	8-class train (%)	8-class test (%)	Gap (pts)	3-class train (%)	3-class test (%)	Gap (pts)
Random forest	99.3	25.4	74.0	98.5	52.1	46.4
Gradient boosting	99.7	23.4	76.3	95.4	53.0	42.4
Soft voting NB+RF+LR	81.5	24.3	57.2	84.4	50.5	33.9

5.8 Feature Dependence

The conditional-independence assumption was examined by measuring dependence among the 30 predictors. The bias-corrected Cramér's V averages 0.089 across the 435 feature pairs (median 0.075); only 9 pairs reach 0.30 and none reaches 0.50 (Figure 4, Table 9). The strongest association is between previous-semester and expected graduation GPA (V = 0.468); mother's and father's education, the pair given as an example of dependence in earlier work, has V = 0.315, moderate and not exceptional. Dependence is thus concentrated in a few pairs rather than widespread. This is consistent with the finding that TAN and AODE, which model such pairs explicitly, do not outperform plain NB. Two caveats apply: Cramér's V measures marginal association, whereas the NB assumption concerns association given the class; and with N = 145 and sparse contingency tables, V is estimated noisily even after bias correction.

[Insert figure image here]

Figure 4. Bias-corrected Cramér's V between the 30 predictors (colour scale clipped at 0.5).

Table 9. Ten most strongly associated predictor pairs (bias-corrected Cramér's V)

Feature 1	Feature 2	Cramér's V
GPA Last Semester	Expected Graduation GPA	0.47
Transportation	Accommodation Type	0.38
Weekly Study Hours	Midterm Prep Timing	0.38
Parental Status	Midterm Prep Timing	0.37
Sex	Seminar Attendance	0.36
Reading NonScientific Freq	Reading Scientific Freq	0.34
Mothers Education	Fathers Education	0.32
Sex	Expected Graduation GPA	0.31
Student Age	Midterm Prep Timing	0.30
Mothers Education	Midterm Prep Method	0.29

5.9 Feature Importance Computed Inside the Folds

Importance for the Fail/Pass target was recomputed within each training fold. On the full data, the separation score reproduces the ranking used previously (sex, transportation, mother's education, accommodation type and parental status as the top five). Computed inside the folds, however, only some of these are stable (Table 10): sex appears in the top five in 100% of folds and transportation in 96%, whereas mother's education drops to 44% and is replaced by project-activity impact (46%). Mutual information agrees on sex, transportation and mother's education, and chi-square points to a partly different set (project-activity impact, midterm preparation method). Permutation importance on held-out folds, which has only one or two Fail students per test fold to work with, is essentially uncorrelated with the other methods (Spearman ρ between mean separation and mean permutation importance = -0.03). Only sex is picked up consistently by all four methods. The earlier conclusion that family background and logistics carry more signal than classroom-engagement variables should be treated as tentative.

Table 10. Fail vs Pass: share of the 50 training folds in which each feature ranks in the top five

Feature	Separation score, mean (SD)	Separation (%)	Mutual info (%)	Chi-square (%)	Permutation (%)	Full-data rank
Sex	1.27 (0.05)	100	100	50	56	1
Transportation	1.24 (0.13)	96	90	68	28	2
Accommodation Type	1.08 (0.14)	62	34	38	30	4
Parental Status	1.04 (0.19)	58	0	62	38	5
Project Activity Impact	1.03 (0.14)	46	4	86	20	7
Mothers Education	1.02 (0.12)	44	94	0	30	3
Mothers Occupation	1.03 (0.12)	40	22	10	12	6
Scholarship Type	0.91 (0.10)	16	2	0	8	9

Full-data rank = rank of the separation score computed once on all 145 students (the original procedure).

For the balanced binary target, previous-semester GPA and expected graduation GPA are consistently the top two features (previous GPA is in the top five in 100% of folds by separation, 100% by mutual information, 100% by chi-square and 80% by permutation importance). For the eight-class grade, previous-semester GPA is top-ranked by mutual information (100%) and chi-square (98%), while mother's education and weekly study hours are also stable under mutual information. Agreement across methods is weaker on this target (Spearman ρ between 0.09 and 0.33). Prior GPA is therefore the best-supported predictor of grade level, which is plausible but also a reminder that it may partly proxy for the outcome.

6. Discussion

The repeated, nested evaluation changes the picture in three ways. First, performance on this dataset is modest and imprecisely estimated. Eight-class accuracy for Categorical NB (24.8%, 95% CI 19.8–30.3) cannot be distinguished from the majority rate (24.1%), and the single-split figure of 29.0% reported earlier was a favorable draw (Section 5.1). What the model does capture is ordering: its quadratic weighted kappa (0.24), its within-one-grade accuracy and its Fail/Pass AUC (0.806) are all above what shuffled labels produce. A reasonable summary is that Naive Bayes is a weak risk-ranking tool for this dataset rather than an accurate grade predictor.

Second, the explanation offered earlier for the ceiling, violation of the independence assumption by correlated predictors, is not supported by the measurements. Pairwise association is weak on average (mean Cramér's V 0.089), is concentrated in a few pairs, and models that explicitly relax independence (TAN, AODE) do not outperform plain NB. The more parsimonious explanation is that the 30 predictors carry limited information about the exact grade and that N = 145 is too small to estimate richer models; the benchmark learners, which model interactions flexibly, also fail to beat the baseline significantly. This does not prove that independence is harmless (V is a marginal measure), but it removes it as the default explanation.

Third, the claims of robustness and stability need to be qualified. NB's training accuracy is roughly three times its test accuracy on the eight-class target (Section 5.7), so it overfits like other learners at this sample size; smoothing reduces the gap only by shrinking toward the priors. The feature ranking is also less stable than a single full-data computation suggests: only sex and, to a lesser extent, transportation mode remain consistently high-ranked once importance is computed inside the folds.

These results still leave practical value for Naive Bayes. It is cheap, transparent, and produces ranked probabilities that can be used to prioritize students for review; the three-class band, where the best hybrid reaches about 54%, is a more realistic target than the exact grade. Any deployment should be framed as screening with human follow-up, and should be validated on additional cohorts before decisions rely on it.

7. Limitations and Future Work

- **Sample size.** With N = 145 (8 Fail students; about 18 per grade class), all estimates are imprecise. The bootstrap intervals capture test-sample variability only, not variability from retraining, so they are somewhat

optimistic; they are nevertheless wide, and the ranking among models is not reliable.

- **Multiple comparisons and dependence.** Seventeen models were compared on four targets. Repeated-CV folds are not independent, so the corrected t-test is approximate and Wilcoxon p-values are anti-conservative. The three-class result for feature-selection NB (uncorrected $p = 0.024$) is exploratory.
- **Tuning asymmetry and missing tools.** Hyper-parameters were tuned by log-loss, which favors calibration over accuracy (it selected heavy smoothing); other criteria could change the ordering. Random forest and gradient boosting used fixed settings, XGBoost was replaced by scikit-learn gradient boosting, and SMOTE-NC by random oversampling, because those tools were unavailable here.
- **Dependence analysis.** Cramér's V is marginal and noisy at this N; conditional association given the class was not tested directly, and NBTree, Bayesian networks, and boosted or bagged NB were not evaluated.
- **Single institution.** All students come from one university and two faculties; there is no external validation. Prior GPA, the strongest predictor, may partly proxy for the outcome.
- **Future work.** Collect more students or combine with other student-performance datasets; evaluate ordinal-aware Naive Bayes and cost-sensitive thresholds for the at-risk class; test NBTree and Bayesian-network learners; and apply the same repeated, nested protocol to external cohorts.

8. Conclusion

This study evaluated Categorical Naive Bayes and nine NB-family variants, alongside five benchmark learners, on four target formulations of the Higher Education Students Performance Evaluation dataset, using repeated nested cross-validation with confidence intervals and significance tests. On the Fail/Pass target Categorical NB ranks at-risk students above others better than chance (AUC 0.806, 95% CI 0.62–0.96), although its accuracy does not exceed the majority baseline and the interval is wide because only 8 students failed. On the eight-class grade its accuracy (24.8%) is not distinguishable from the majority baseline, although it recovers part of the grade ordering (QWK 0.24). On a three-class band it reaches 49.3% (about 54% with feature selection), an improvement that is suggestive but not significant after multiple-comparison correction.

Hybrid and benchmark models did not reliably improve on plain Naive Bayes, feature dependence is weak, and Naive Bayes overfits at this sample size. Naive Bayes is therefore best described as a transparent, inexpensive baseline and risk-ranking tool whose predictive value on this dataset is limited and imprecisely estimated, rather than as a validated predictor of individual grades. Larger samples and external validation are needed before stronger claims are justified.

References

1. Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. Bioinformatics, 26(10), 1340–1347.
2. Ambroise, C., & McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. Proceedings of the National Academy of Sciences, 99(10), 6562–6566.
3. Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining educational data to predict student's academic performance using ensemble methods. International Journal of Database Theory and Application, 9(8), 119–136.
4. Baker, R. S. J. d., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining, 1(1), 3–17.
5. Bergsma, W. (2013). A bias-correction for Cramér's V and Tschuprow's T. Journal of the Korean Statistical Society, 42(3), 323–328.
6. Bhardwaj, B. K., & Pal, S. (2011). Data mining: A prediction for performance improvement using classification. International Journal of Computer Science and Information Security, 9(4), 136–140.
7. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
8. Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research, 11, 2079–2107.

9. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.
10. Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21, 6.
11. Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37–46.
12. Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 70(4), 213–220.
13. Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. In Proceedings of the 5th Future Business Technology Conference (FUBUTEC 2008) (pp. 5–12). EUROSIS.
14. Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation, 10(7), 1895–1923.
15. Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, 29(2–3), 103–130.
16. Efron, B., & Tibshirani, R. J. (1993). An introduction to the bootstrap. Chapman & Hall.
17. Frank, E., & Hall, M. (2001). A simple approach to ordinal classification. In Machine Learning: ECML 2001 (LNCS 2167, pp. 145–156). Springer.
18. Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. Machine Learning, 29(2–3), 131–163.
19. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 1157–1182.
20. Hall, M. A. (2000). Correlation-based feature selection for discrete and numeric class machine learning. In Proceedings of the 17th International Conference on Machine Learning (pp. 359–366).
21. Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann.
22. Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6(2), 65–70.
23. Jiang, L., Zhang, H., & Cai, Z. (2009). A novel Bayes model: Hidden naive Bayes. IEEE Transactions on Knowledge and Data Engineering, 21(10), 1361–1371.
24. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of IJCAI-95 (pp. 1137–1143).
25. Kohavi, R. (1996). Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid. In Proceedings of KDD-96 (pp. 202–207).
26. Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2004). Predicting students' performance in distance learning using machine learning techniques. Applied Artificial Intelligence, 18(5), 411–426.
27. Langley, P., & Sage, S. (1994). Induction of selective Bayesian classifiers. In Proceedings of UAI-94 (pp. 399–406).
28. Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. Machine Learning, 52(3), 239–281.
29. Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In Proceedings of the 22nd International Conference on Machine Learning (pp. 625–632).
30. Ojala, M., & Garriga, G. C. (2010). Permutation tests for studying classifier performance. Journal of Machine Learning Research, 11, 1833–1863.
31. Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
32. Rish, I. (2001). An empirical study of the naive Bayes classifier. In IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence (Vol. 3, No. 22, pp. 41–46).
33. Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C, 40(6), 601–618.
34. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE, 10(3), e0118432.
35. Sekeroglu, B., Dimililer, K., & Tuncal, K. (2019). Student performance prediction and classification using machine learning algorithms. In Proceedings of the 8th International Conference on Educational and Information Technology (ICEIT 2019). ACM. https://doi.org/10.1145/3318396.3318419

36. Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A review on predicting student's performance using data mining techniques. Procedia Computer Science, 72, 414–422.

37. Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics, 7, 91.

38. Webb, G. I., Boughton, J. R., & Wang, Z. (2005). Not so naive Bayes: Aggregating one-dependence estimators. Machine Learning, 58(1), 5–24.

39. Yilmaz, N., & Sekeroglu, B. (2019). Higher Education Students Performance Evaluation [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C51G82. Also published as: Student performance classification using artificial intelligence techniques. In Advances in Intelligent Systems and Computing (Vol. 1095, pp. 596–603). Springer.

40. Zhang, H. (2004). The optimality of naive Bayes. In Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004).