Unsupervised Learning for Grading Students

Gautam Singh, G. V. V. Sharma
Indian Institute of Technology, Hyderabad

***Abstract*—This document is a report which compares the grade distribution obtained by using a method based on machine learning as compared to fitting a normal curve to the scores of the students.**

1 Introduction

In recent times, especially during and after the COVID-19 pandemic, there has been an increase in the amount of e-learning resources. Many educational institutions have embraced online learning by using platforms such as Moodle and have also invested in infrastructure to facilitate hybrid (both online and in-person) classrooms. There has also been a marked increase in the number of and amount of traffic on learning platforms such as Coursera, Udemy and edX [1]. This boom in e-learning has led to the collection of unique and large-scale data related to educational settings. This has led to the emergence of the two closely related research areas of educational data mining (EDM) and learning analytics (LA). EDM is mainly concerned with developing methods for exploring this data, while LA is mainly concerned with the analysis and interpretation of data about learners [2]. Both fields strive to gain actionable insights for improving learning experiences in the form of personalized feedback, prompt academic support, better learning strategies and holistic evaluation methods.

In this paper, we explore the utility of unsupervised machine learning (ML) methods for fair grading of students in a course where students performances are skewed. We test the utility of the K-means clustering algorithm in assigning letter grades as compared to grading using the Z score. We use the scores of 94 students who have taken a course in our institute. The main contributions of this paper are as follows.

1) Evaluate the utility of unsupervised learning methods as compared to standard methods for grading students such as using Z scores.
2) Analyze the impact of skewed performance distributions on the fairness of grading.

The remainder of this paper is structured as follows. Section 2 provides an overview of prior works related to grading of students and assignments. Section 3 describes the methods used to solve this problem. Section 4 describes the implementation details of our approach. Section 5 presents the results of our approach. Finally, Section 6 concludes the paper and discusses future work.

2 Related Work

EDM and LA encompass a wide range of fields such as education research, educational technology, statistics, data science and artificial intelligence [3]. We present a brief overview of some of the works related to grading of students using traditional methods and newer ML methods.

The most common statistic used for grading students is the *Z score* [4], which is defined as the number of standard deviations a score is from the mean of the population. Apart from scoring in standardized tests [5], the Z score is also used to predict the financial status of companies [6] and outlier detection [7], [8]. However, the Z score is not a very effective measure in situations where there are fewer students or when the performance is skewed. This leads to unfair grading where students with similar scores end up with different grades just because they lie on either side of a predefined boundary. Further, the Z score can adversely impact learning experiences by intensifying academic competition [9].

To overcome the limitations of using the Z score, researchers have appealed to supervised and unsupervised ML methods for grading. Many studies [10]–[13] have compared the performance of various supervised and unsupervised ML algorithms for grading students in various course, assignment and examination settings. In recent times, there has been a marked increase in the use of unsupervised classification methods for grading since there may not be enough labeled data for a particular course or exam collected to sufficiently train a supervised model. In particular, clustering methods have become a key technology for EDM and LA [14]. A particularly popular clustering method used in many studies is the K-means clustering algorithm [15]. The authors of [16] were able to use the K-means algorithm to identify students at risk of failing to complete coursework by collecting data pertaining to a single course offered over many years. The authors of [17] used the K-means algorithm to promptly identify and guide students at the edge of the grading classification. Variants of the original K-means algorithm have also been used for grading students. The K-means++ algorithm was used by the authors of [18] to enhance the quality of assignments given to students. The authors of [19] evaluated the utility of the K-medians algorithm along with the K-means algorithm for determining the rankings of schools in a common examination. However, the aforementioned studies do not evaluate the performance of clustering algorithms in scenarios where the performance of students is skewed. We aim to fill this gap by comparing the utility of K-means clustering in assigning grades in such a scenario.

3 Methodology

3.1 Z Score

For a given dataset $\{X_1, X_2, \ldots, X_N\}$ of size N, we can compute the population mean and population variance using the following equations.

$$\mu = E[X] = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{1}$$

$$\sigma^2 = E\left[(X-\mu)^2\right] = \frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)^2 \tag{2}$$

Using the central limit theorem [4], we assume that the scores $X \sim N\left(\mu, \sigma^2\right)$. Thus, the Z-score of X is given by

$$Z = \frac{X-\mu}{\sigma}. \tag{3}$$

The letter grades are assigned as per Table 1.

Interval	**Grade**
$(-\infty$, -3]	F
(-3, -2]	D
(-2, 1]	C
(-1, 0]	B-
(0, 1]	B
(1, 2]	A-
(2, 3]	A
$(3,\infty)$	A+

TABLE 1: Grading Scheme.

3.2 K-Means Clustering

K-Means clustering is an unsupervised classification model, which attempts to cluster unlabeled data in order to gain more structure from it.

To find the optimum means for a fixed number of letter grades as per Table 1, we frame the problem of assigning letter grades as an optimization problem. For a set of data points $\{\mathbf{X}_i\}_{i=1}^{N}$ and means $\{\boldsymbol{\mu}_i\}_{i=1}^{K}$, we define for $1 \le n \le N,\ 1 \le k \le K$,

$$r_{nk} \triangleq \begin{cases} 1 & \arg\min_j \left\lVert \mathbf{X}_n - \boldsymbol{\mu}_j \right\rVert = k \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Thus, we need to find points $\boldsymbol{\mu}_k$ minimizing the cost function

$$J \triangleq \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left\lVert \mathbf{X}_n - \boldsymbol{\mu}_k \right\rVert^2 \tag{5}$$

Clearly, (5) is a quadratic function of $\boldsymbol{\mu}_k$. Differentiating with respect to $\boldsymbol{\mu}_k$ and setting the derivative to zero, we get

$$\sum_{n=1}^{N} 2\boldsymbol{\mu}_k r_{nk}\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right) = 0 \tag{6}$$

$$\implies \boldsymbol{\mu}_{\mathbf{k}} = \frac{\sum_{n=1}^{N} r_{nk}\mathbf{x}_n}{\sum_{n=1}^{N} r_{nk}} = \frac{\mathbf{X}\mathbf{r}_k}{\mathbf{1}^{\top}\mathbf{r}_k} \tag{7}$$

where

$$\mathbf{X} \triangleq \begin{pmatrix} \mathbf{X}_1 & \mathbf{X}_2 & \dots & \mathbf{X}_n \end{pmatrix} \tag{8}$$

$$\mathbf{r}_k \triangleq \begin{pmatrix} r_{1k} & r_{2k} & \dots & r_{nk} \end{pmatrix}^{\top} \tag{9}$$

$$\mathbf{1} \triangleq \begin{pmatrix} 1 & 1 & \dots & 1 \end{pmatrix}^{\top} \tag{10}$$

From (7), we see that the optimum is attained when $\boldsymbol{\mu}_k$ is set to the expectation of the $\mathbf{X}_n$ with respect to r_{nk}.

Thus, the K-means algorithm is an *expectation maximization* (EM) algorithm where each iteration consists of two steps until convergence.

1) *Expectation Step* (E Step): For $1 \le k \le K$, calculate the expected means using (11).

$$\tilde{\boldsymbol{\mu}_k} \triangleq \frac{\sum_{n=1}^{N} r_{nk}\mathbf{X}_n}{\sum_{n=1}^{N} r_{nk}} \tag{11}$$

2) *Maximization Step* (M Step): Set $\boldsymbol{\mu}_k \leftarrow \tilde{\boldsymbol{\mu}_k}$ for $1 \le k \le K$.

4 Implementation

We implemented the Z score algorithm as well as the EM algorithm for the K-means clustering algorithm in Python using the `numpy` and `pandas` libraries. The raw data is collected in the `marks.xlsx` file. The source codes and dataset are openly available on GitHub.

The Python code `codes/grades_norm.py` takes the given input population dataset `marks.xlsx` and assigns grades appropriately. The grades are output to `grades_norm.xlsx`. A similar task is performed by the `codes/grades_kmeans.py` code, which uses the K-means algorithm to assign grades and outputs the results to `grades_kmeans.xlsx`.

5 Results

The grade distribution using each method is shown in Figure 1 and Figure 2. Based on the results, we can make the

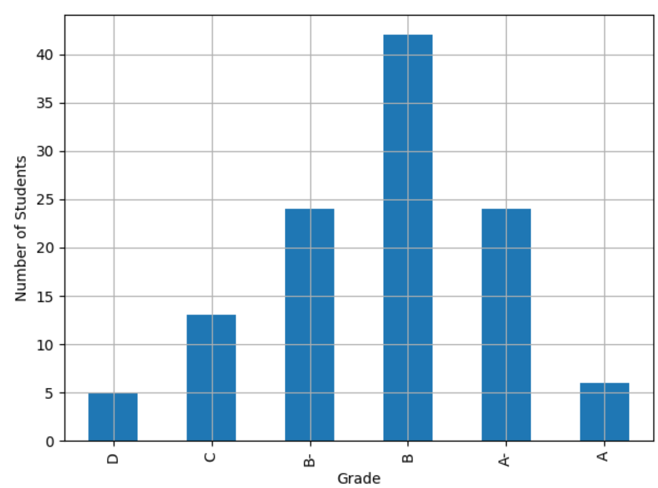

Fig. 1: Grade distribution using Z scores.

following observations:

1) Using the Gaussian distribution is quite unfair, since there could be students with quite similar marks but with a

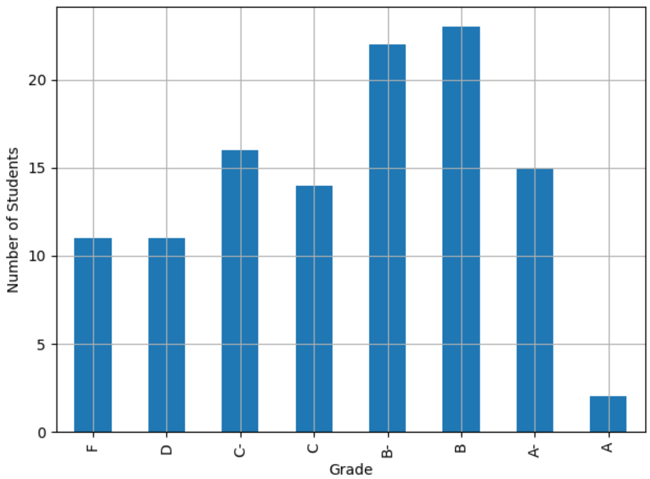

Fig. 2: Grade distribution using the K-means algorithm.

difference in grade, just because they lie on either side of a predefined boundary.

2) The K-means algorithm allows for better decision boundaries, depending on how skewed the performance of the students is, accordingly to the difficulty of the course.
3) Unlike the Gaussian distribution, the K-means algorithm can be used for a fairer assignment of the grades, no matter how skewed the performance of students in a course is.

6 Conclusion and Future Work

In this paper, we analyzed the utility of unsupervised learning methods for grading students in a course where the performance of students is skewed. We compared the performance of the K-means algorithm with using Z scores in assigning grades to students. Our results indicate that the K-means algorithm is more effective in handling skewed distributions and providing fairer grade assignments. A possible future extension can be to explore the utility of other unsupervised learning algorithms with datasets containing more features collected over a period of time. Another possible extension could be to integrate these methods with existing content management systems to provide continuous and personalized feedback to students.

References

[1] S. Roy and S. N. Singh, "Emerging trends in applications of big data in educational data mining and learning analytics," in *2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence*, Jan. 2017, pp. 193–198.

[2] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 3, p. e1355, May 2020.

[3] "What is Learning Analytics," https://www.solaresearch.org/about/what-is-learning-analytics/.

[4] M. R. Spiegel and L. J. Stephens, *Schaum's Outline of Statistics*, 6th ed. McGraw-Hill Education, 2018.

[5] K. Santhanakrishnan and V. Senthooran, "A Parallel distributed Cluster Computing Model for Z-Score Computation in Respect of Sri Lankan University Admissions," in *2022 2nd International Conference on Advanced Research in Computing (ICARC)*, Feb. 2022, pp. 344–348.

[6] S. K. Jagannathan, G. Bizel, and H. Alpagu, "Predicting Bankruptcy of Companies in the Pharmacy and Technology Sectors Using Altman's Z-score model," in *2023 6th International Conference on Recent Trends in Advance Computing (ICRTAC)*, Dec. 2023, pp. 610–616.

[7] A. S. Yaro, F. Maly, P. Prazak, and K. Malý, "Outlier Detection Performance of a Modified Z-Score Method in Time-Series RSS Observation With Hybrid Scale Estimators," *IEEE Access*, vol. 12, pp. 12 785–12 796, 2024.

[8] V. Aggarwal, V. Gupta, P. Singh, K. Sharma, and N. Sharma, "Detection of Spatial Outlier by Using Improved Z-Score Test," in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, Apr. 2019, pp. 788–790.

[9] J. Klein, "Assessing university students' achievements by means of standard score (Z score) and its effect on the learning climate," *Studies in Educational Evaluation*, vol. 40, pp. 63–68, Mar. 2014.

[10] Y. T. Badal and R. K. Sungkur, "Predictive modelling and analytics of students' grades using machine learning algorithms," *Education and Information Technologies*, vol. 28, no. 3, pp. 3027–3057, Mar. 2023.

[11] S. Palkhiwala, M. Shah, and M. Shah, "Analysis of Machine learning algorithms for predicting a student's grade," *Journal of Data, Information and Management*, vol. 4, no. 3, pp. 329–341, Dec. 2022.

[12] S. D. A. Bujang, A. Selamat, R. Ibrahim, O. Krejcar, E. Herrera-Viedma, H. Fujita, and N. A. M. Ghani, "Multiclass Prediction Model for Student Grade Prediction Using Machine Learning," *IEEE Access*, vol. 9, pp. 95 608–95 621, 2021.

[13] H. Gull, M. Saqib, S. Z. Iqbal, and S. Saeed, "Improving Learning Experience of Students by Early Prediction of Student Performance using Machine Learning," in *2020 IEEE International Conference for Innovation in Technology (INOCON)*, Nov. 2020, pp. 1–4.

[14] W. Zhang and S. Qin, "A brief analysis of the key technologies and applications of educational data mining on online learning platform," in *2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA)*, Mar. 2018, pp. 83–86.

[15] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. University of California Press, Jan. 1967, vol. 5.1, pp. 281–298.

[16] M. Bucos and B. Drăgulescu, "Student cluster analysis based on Moodle data and academic performance indicators," in *2020 International Symposium on Electronics and Telecommunications (ISETC)*, Nov. 2020, pp. 1–4.

[17] H. Wang, K. Zhang, and Y. Hu, "The Algorithm of Student Grade Evaluation Based on Clustering," in *2022 3rd International Conference on Information Science and Education (ICISE-IE)*, Nov. 2022, pp. 183–187.

[18] A. Jiménez-Macías, P. J. Muñoz-Merino, and C. D. Kloos, "Student Clustering Through KMeans to Enhance Teaching," in *2024 XVI Congreso de Tecnología, Aprendizaje y Enseñanza de La Electrónica (TAEE)*, Jun. 2024, pp. 1–5.

[19] M. Bharti, G. Pradhan, and D. Rout, "Centroid Based Clustering Models for Designing Dense Ranking of Schools Focusing on Total Marks of Individual Students in a Common Examination," in *2024 1st International Conference on Cognitive, Green and Ubiquitous Computing (IC-CGU)*, Mar. 2024, pp. 1–6.