# Explainable AI as a Social Microscope: A Case Study on Academic Performance

Anahit Sargsyan[1,2(✉)], Areg Karapetyan[1,3(✉)], Wei Lee Woon[1], and Aamena Alshamsi[1,4]

[1] Department of Computer Science, Masdar Institute, Khalifa University, Abu Dhabi, UAE
akarapetyan@masdar.ac.ae

[2] Division of Social Science, New York University Abu Dhabi, Abu Dhabi, UAE
as12831@nyu.edu

[3] Research Institute for Mathematical Sciences (RIMS), Kyoto University, Kyoto, Japan

[4] The MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA

**Abstract.** Academic performance is perceived as a product of complex interactions between students' overall experience, personal characteristics and upbringing. Data science techniques, most commonly involving regression analysis and related approaches, serve as a viable means to explore this interplay. However, these tend to extract factors with wide-ranging impact, while *overlooking variations specific to individual students*. Focusing on each student's peculiarities is generally impossible with thousands or even hundreds of subjects, yet data mining methods might prove effective in devising more targeted approaches. For instance, subjects with shared characteristics can be assigned to clusters, which can then be examined separately with machine learning algorithms, thereby providing a more nuanced view of the factors affecting individuals in a particular group. In this context, we introduce a data science workflow allowing for fine-grained analysis of academic performance correlates that captures the *subtle differences in students' sensitivities to these factors*. Leveraging the Local Interpretable Model-Agnostic Explanations (LIME) algorithm from the toolbox of Explainable Artificial Intelligence (XAI) techniques, the proposed pipeline yields groups of students *having similar academic attainment indicators*, rather than similar features (e.g. familial background) as typically practiced in prior studies. As a proof-of-concept case study, a rich longitudinal dataset is selected to evaluate the effectiveness of the proposed approach versus a standard regression model.

**Keywords:** Explainable AI · LIME · Data science · Machine learning · Computational social science · Academic performance · GPA prediction

**Electronic supplementary material** The online version of this chapter (https://doi.org/10.1007/978-3-030-64583-0_24) contains supplementary material, which is available to authorized users.


G. Nicosia et al. (Eds.): LOD 2020, LNCS 12565, pp. 257–268, 2020.
https://doi.org/10.1007/978-3-030-64583-0_24

## 1 Introduction

With far-ranging consequences on young people's lives and careers, academic performance is susceptible to various types and forms of influence. It is often path-dependent, correlating with an individual's past performance [3,18]. Furthermore, factors with significant impact on academic performance can be specific to the subject in question, such as intelligence and determination [2,11,22], or exogenous, resulting from the social, emotional and socioeconomic environment in which the individual was raised [5,10,14]. Therefore, in general, investigation of academic performance predictors is attained through longitudinal studies [5,8].

Mainly, two directions are evidenced in this line of research. The first, followed in [6,21], relies on statistical models to measure the correlation of a few variables that were premised on prior results in the literature. The second, attended in [1,12,13,17], resorts to data science and machine learning techniques for extracting informative predictors from large datasets with thousands of candidate features.

While both approaches recognize the uniqueness of the students' backgrounds, the effect of the correlates is still determined as an aggregate over the entire study population/group, leaving the *subtle variations between individuals* largely overlooked. In particular, the former assumes that all the subjects are impacted by the same set of selected correlates in the same manner, whereas the latter seeks to derive a predictive model with high accuracy that generalizes to the overall population. However, no two *subjects are identical*, and factors profoundly affecting one individual might have a merely negligible impact on another person, even under comparable circumstances.

Against this backdrop, we introduce a novel data science approach in which (i) the predictors of academic performance for each student *are identified and quantified* (ii) the study population is segmented into clusters based on the obtained values. The proposed pipeline allows the groups with *similar success indicators* to be analyzed collectively, which should enable their effects to reinforce each other and be more readily discoverable.

To quantify academic performance predictors specific to individual students, we avail of a recently developed XAI algorithm, known as LIME [16]. The outputs from LIME serve as "explanations", which are *localized* in that a unique explanation is generated for each subject. Though descriptive on an individual case basis, these explanations are intrinsically disassociated, and thus their direct interpretation (one by one) becomes intractable with a growing number of subjects. This paper presents an efficient solution by grouping the students according to LIME coefficients, as detailed in Sect. 3. Distinctively, under such clustering criterion, the subsequent analysis is explicitly centered at groups of students who *share similar academic performance correlates*, as opposed to the classical approach of clustering in the feature space (i.e., based on observable characteristics such as gender, familial background or financial class). In a sense, with this scheme in place, it proves possible for subjects with fairly diverse backgrounds and needs to be grouped together, provided they share common markers of academic attainment.

As one demonstration, the proposed approach is benchmarked on a longitudinal dataset, released for the Fragile Families Challenge (FFC) competition, against a standard regression model. The results reveal a striking difference in the depth of insights gained, with the devised pipeline featuring prominently. While intended as a proof-of-concept, this preliminary study unveils findings on academic performance indicators that could serve social and data science communities. Furthermore, the workflow proposed herein can potentially pave the way towards more efficient and targeted intervention strategies by providing insights that would be inconceivable to achieve with traditional methods.

## 2 Data and Pre-processing

### 2.1 FFC Dataset

The examined dataset stems from the Fragile Families and Child Wellbeing study that documented the lives of over 4000 births occurring between 1998 and 2000 in U.S. cities with at least 200,000 population. As such, the study was carried out in the form of questionnaire surveys and interviews with parents shortly after the children's birth, and when the infants were 1, 3, 5 and 9 years old (*overall five waves*). The elicited data covered essential temporal information on the children's attitudes, parenting behavior, demographic characteristics, to name a few (further details of the dataset can be consulted in [15]). The data was released within the scope of FFC competition which sought to predict 6 life outcomes, including Grade Point Average (GPA), based on these data records. Analysis of the overall results and ensuing findings of the competition are summarized in [19].

In total, the dataset comprises 4,242 rows (one per child) and 12,943 columns, including the unique numeric identifier. During FFC, however, only half of the data rows were released as a training set. Of these, 956 entries had GPA values missing, and therefore the final dataset analyzed in this study totalled 1,165 subjects, as appears in Fig. 6 in the Appendix.

### 2.2 Pre-processing and Feature Selection

Before proceeding with this step, we remark that it is stipulated exclusively by nature (e.g., missing values) and properties (e.g., dimensionality) of the dataset under study and per se is not a principal constituent of the developed pipeline. Indeed, it is tailored specifically for the FFC dataset and might very well be substituted by any other appropriate routine yielding a sufficiently informative feature subset (i.e., *with a decent predictive accuracy*) of reasonable cardinality. Thus, for clarity of exposition, the respective particulars are deferred to the Appendix.

The target subset of optimally descriptive features, as revealed through extensive experimentation and *validated by its predictive accuracy*[1], contained 65 fea-

[1] A mean squared error (MSE) of approximately 0.359 was achieved under 3-fold cross-validation (a result of comparable fidelity, submitted during the FFC, *secured a place in the top quartile* of the scoreboard).

tures, tabulated in Table 1 in [20]. This pool of features forms the input for the proceeding analysis laid out in Sect. 3. Figure 1 depicts the spread of these 65 features across the 5 waves (i.e., over the trajectory of children's lives). For each wave, the features are arranged into the following six categories: {*familial, financial, academic, social, personality, other*} and their respective counts are illustrated in Fig. 1 as a stacked bar chart.

As deduced from Fig. 1, the distribution of features in familial and financial categories is skewed towards the early span of children's lives. For the correlates falling in academic, social and personality categories, the opposite trend is evidenced.

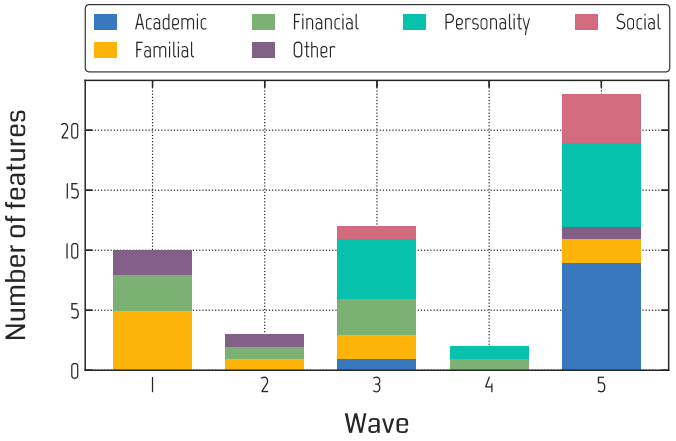


**Fig. 1.** The distribution of selected 65 features, categorized into 6 major factor types, among the 5 waves (i.e., over the course of the children's lives).

## 3 Comparative Analysis

This section contrasts the proposed approach against a conventional regression model, as illustrated in Fig. 2, and discusses the results.

We cast the problem as a classification task by discretizing the GPA scores into three classes, `Low, Middle and Top`, defined respectively by the following ranges: $[1, 2.5], (2.5, 3.25), [3.25, 4]$. Consequently, only the subjects falling into the `Top` and `Low` categories were retained (861 in total). The motivation is to steer the focus of classification algorithms towards the aspects discerning high and low performers. Indeed, the factors responsible for "borderline" performances are likely the ones with negligible impact, hence inferring them might obscure the results. On the other hand, omitting a large group of subjects could lead to the loss of pertinent data. Thus, the above thresholds were set according to the top and bottom 30% percentiles of GPA score records. This ensured a solid number of participant students while retaining a sizable gap between the two classes.

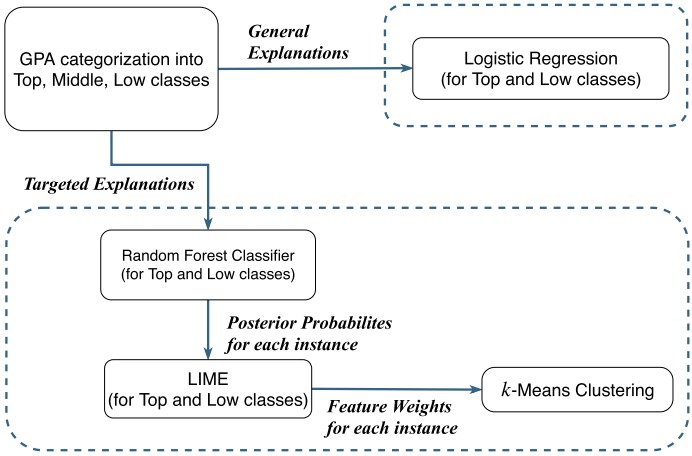


**Fig. 2.** Flowchart of the conducted comparative analysis including the featured methodology for obtaining targeted explanations.

### 3.1 General Indicators

Following the common practice of previous works, in this initial phase, we employed the logistic regression algorithm to screen the selected features that broadly correlate with academic performance. The subjects from the `Top` and `Low` categories were fit to the model and the resulting coefficients, under L1 regularization, are presented in Fig. 3.

As observed from Fig. 3, test grades, along with other early metrics of academic performance, are imperative, and so are the factors associated with the child's social background. In particular, the indicators in familial and financial categories appear to influence children's academic performance predominantly at an early age. Whereas the correlates associated with scholastic aptitude manifest their effect mostly at later stages of subjects' lives. These observations are consistent with prior findings in the literature, providing some measure of validation. Overall, the most influential predictors are listed below.

1. The two most important factors relate to the parents' education [4].
2. The *Peabody Picture Vocabulary Test* (PPVT) percentile rank correlated with academic performance. PPVT is a standardized test designed to measure an individual's vocabulary and comprehension and provide a quick estimate of verbal ability or scholastic aptitude. Another standard test's (known as Woodcock Johnson Test) percentile rank was identified as a significant indicator as well.
3. Interestingly, the fact that the father has been incarcerated[2] - a proxy for family support - affects children's performance negatively. Contrariwise, the

[2] Note that in Fig. 3 the positive correlation of this feature is due to the reversed order of values (i.e., the highest value indicates the father has not spent time in jail).

complexity/rank of the mother's job, a surrogate for the financial situation, conduced to enhanced academic performance.

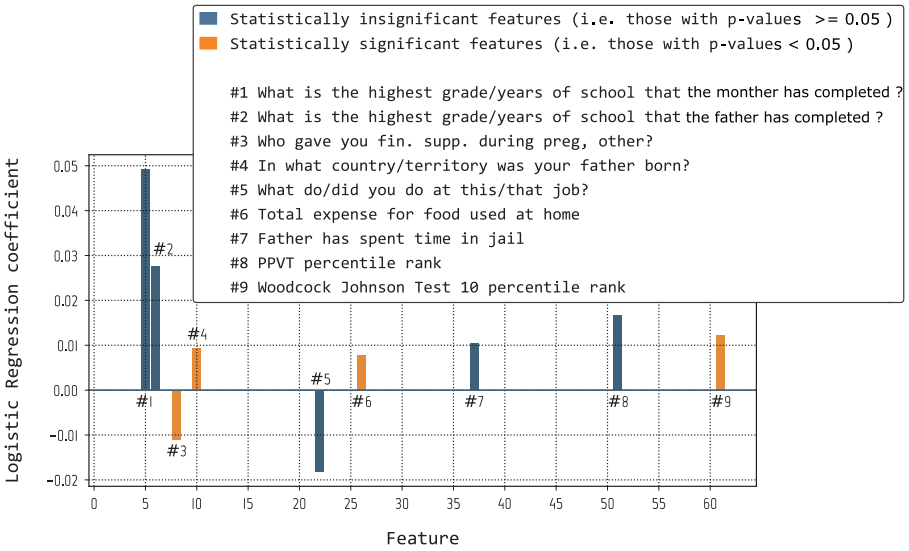


**Fig. 3.** Logistic Regression coefficients of the selected 65 features (ordered in a non-decreasing value of their waves) for `Top` and `Low` classes.

The emerging picture is compelling and multifaceted. On the one hand, test scores and academic aptitude occupy a central role, which is to be expected. Yet, there are indications that, beyond this, other features reflective of social and financial stability could also play a part, which strongly motivates the second, targeted part of this study.

### 3.2 Proposed Methodology: Targeted Indicators

While the insights highlighted in Sect. 3.1 were illuminating, they were extracted from the entire dataset, and the perspectives obtained were thus quite broad. To further extract targeted or localized indicators of academic success, we resorted to LIME [16]. For each instance, LIME produces a localized explanation of the classifier output by perturbing the feature values to generate a set of synthetic data points in the vicinity of the true instance. The posterior probability for each data point is estimated using the trained classifier, and a linear regression model is trained using the synthetic points as the inputs, and the posterior probabilities as the targets. The localized regression coefficients obtained in this way can then be interpreted as the *importance of a feature*, and are estimated separately for each subject.

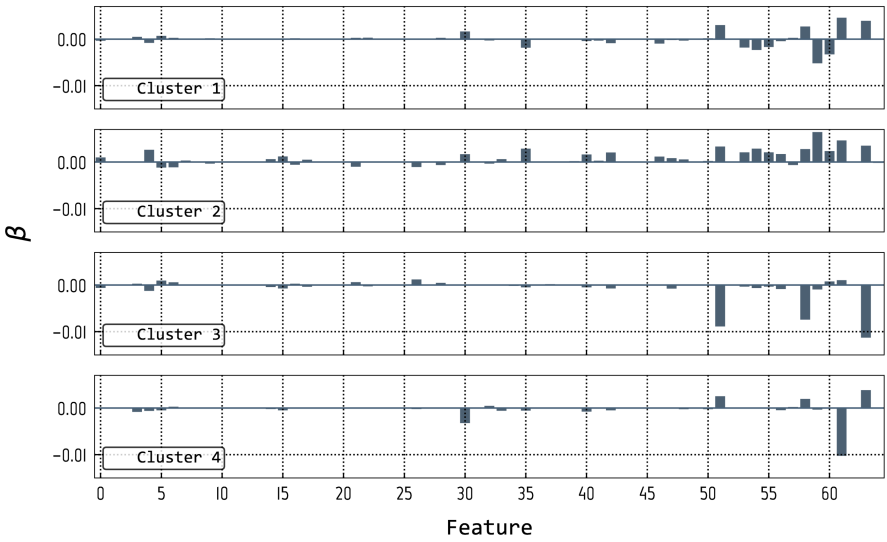


**Fig. 4.** Clustering of students in `Top` and `Low` classes based on LIME coefficients. The horizontal axis represents the selected 65 features sorted by their wave values in a non-decreasing order (i.e., over the course of children's lives). The vertical axis ($\beta$) depicts the deviations of the mean LIME coefficients for each cluster from the overall mean values of each LIME coefficient across the population.

This technique was adapted for the present context as follows. First, Random Forest classifier is trained on the data and is invoked to estimate the posterior probability for each instance. Then, LIME is applied to subjects falling into the `Top` and `Low` groups to produce feature weights specific to each subject, which are then clustered with the $k$-means algorithm. Each cluster is then characterized by the centroid of the LIME coefficients for the instances therein.

### 3.3 Results and Discussion

The $k$-means clustering results, with parameter $k = 4$ (i.e., four clusters), appear in Fig. 4. This choice of $k$ enables a compact yet expressive representation of the underlying insights. In general, the higher the $k$, the more likely the groups are to resemble each other closely. On the other hand, when $k$ is small, one might possibly overlook factors critical to some subset of subjects.

In Fig. 4, to emphasize the differences between clusters, we focus on each feature's relative weights. That is, a cluster is represented in terms of the difference between its centroid and population means. As the figure suggests, the characteristics of the subjects vary significantly between the clusters. In particular, the salient patterns observed were as follows:

– **Cluster 1** (*189 subjects*): The subjects in this cluster appear to be strongly influenced by features 51, 58, 61 and 63, which are all linked to test scores.

Whereas features 59 and 60, related to the ability to pay attention, appeared to be less important.

- **Cluster 2** (*141 subjects*): This cluster was similar to Cluster 1 except that features 59 and 60 were more important than average.
- **Cluster 3** (*297 subjects*): In this cluster, features 51, 58 and 63 (all test score related), were all less important than average.
- **Cluster 4** (*234 subjects*): Here, feature 61 (the Woodcock Johnson Test score) was significantly less important, while features 51, 58 and 63 all had slightly stronger impacts on performance.

Overall, the features with values close to zero are the ones with a uniform effect on all individuals regardless of a cluster.

These results are deeply compelling in a number of ways. While Fig. 3 (results of the standard regression model) provided a broad overview of the overall success factors, the relative importance of each of these factors differed substantially among subjects. For example, in clusters 1 and 2, test scores appeared to have a substantial impact on future academic performance, while the opposite was observed in cluster 3. Also, features 59 and 60, which measure a student's general attentiveness, seemed relatively peripheral in cluster 1 and crucial in cluster 2 (and close to the average in clusters 3 and 4).

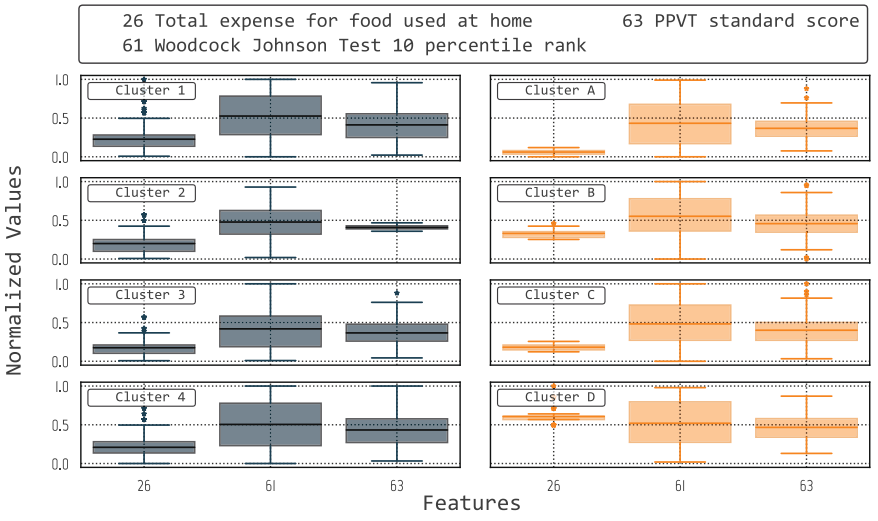


**Fig. 5.** Distributions of values of the selected numerical features within the clusters of students in `Top` and `Low` classes. Clusters 1 to 4 were produced with the proposed scheme (i.e., based on LIME coefficients), while clusters A to D with the conventional method (i.e., based on feature values).

Another interesting observation was the spread of feature values *within* clusters, which is depicted in Fig. 5 (representing three selected numerical features).

In particular, note the spread of values for feature 26, which is *the total expense for food used at home*. Apparently, the feature values in the LIME clusters exhibit a far greater range compared to the conventional clusters, which tend to group people with similar financial situations. For the other features, the results are slightly less straightforward, but LIME clusters 3 and 4, in particular, also exhibit a much broader spread of values. This underscores that clustering with the LIME coefficients does not merely group students based on their personal or social circumstances, but rather in terms of the factors which affect their future academic performance.

## 4 Future Work

The present findings indicate the potential effectiveness of the proposed methodology in analyzing causal relationships in datasets alike FFC. However, this work was intended as a proof-of-concept case study, leaving open several avenues for future investigations. In particular, below listed are several promising directions.

1. Analyze comparable data sets (e.g., The Millennium Cohort Study [7]), to test whether certain aspects of the observations are data-specific, or reflect true underlying patterns.
2. Perform a thorough sensitivity analysis on the chosen subset of features as well as incorporate XAI techniques alternative to LIME (e.g., SHapley Additive exPlanations [9]).
3. Apply the proposed approach to more diverse and larger datasets to demonstrate its generalizability to other use cases/study domains.

## 5 Concluding Remarks

In this study, a novel data science pipeline is proposed, which conduces the identification of the specific features associated with academic performance in different groups of students. A clustering algorithm was employed to group these subjects, then targeted success indicators were extracted from each of these groups and scrutinized. We note that the present findings rely on a technique (LIME), which was developed relatively recently and should be treated as preliminary. However, if and when superior methods are proposed, they can similarly be incorporated into the devised workflow.

The key point is that such localized models are vital if we are to obtain a more nuanced view of the actual success indicators for specific children and families. The findings suggest that the children of fragile families can be given the best chance of success through interventions that are tailored to their individual needs, e.g. in some families, a small home loan could be the difference between a star student and a dropout, whereas in others a free mentoring scheme might turn more valuable.

## Appendix

This section details the employed pre-processing and feature selection steps, portrayed as a flowchart in Fig. 6.

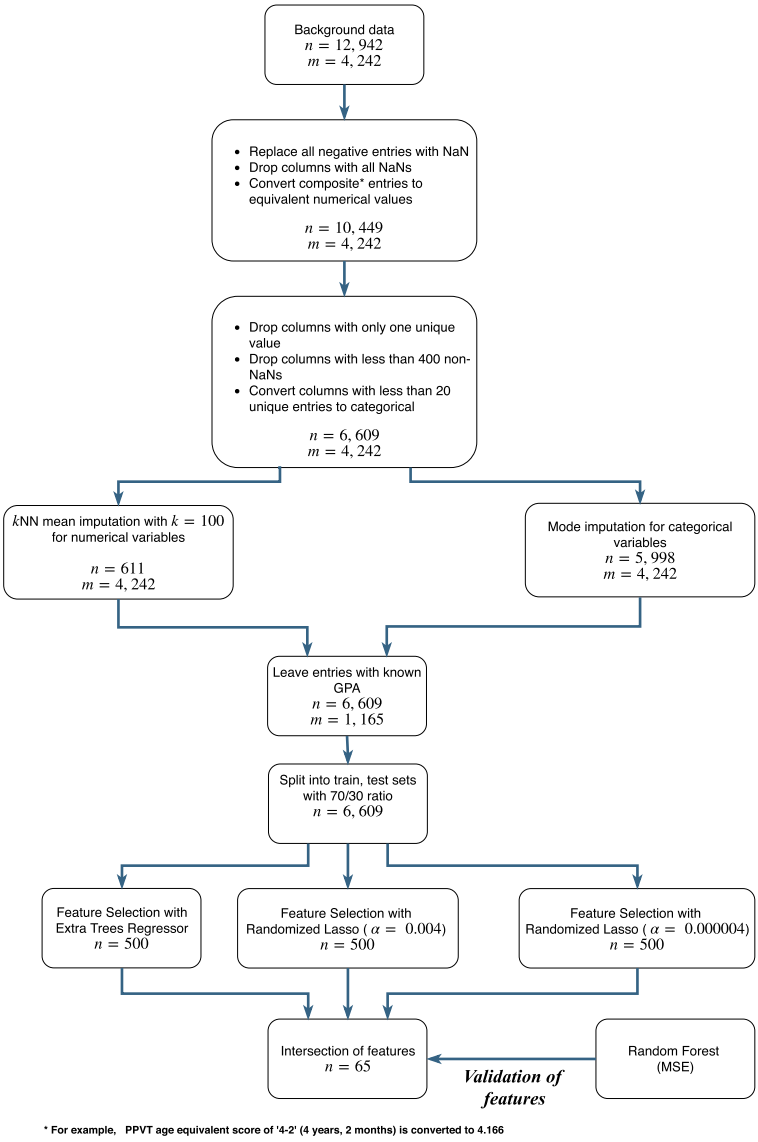


**Fig. 6.** A step-by-step illustration of the performed pre-processing and feature selection routines, depicted as a flowchart, with $n$ and $m$ standing for the number of features and entries in the given step, respectively.

Due to the nature of the dataset under study, there were numerous instances of missing values where respondents either refused or were unavailable to answer. This was resolved through the judicious combination of data transformation, reduction, and imputation techniques, as listed below.

- All missing and negative values were replaced by NaN and the columns with 0 variance were removed. Then, only the columns having at least 400 non-NaN values were retained.
- Features with less than 20 unique values were treated as categorical and a simple median imputation was applied to replace the missing values.
- The variant of $k$NN ($k$-Nearest Neighbors) imputation algorithm, implemented in the Python package Fancyimpute, was leveraged, with value of 100 for the parameter $k$, to estimate the remaining NaN values.

The number of features in the resulting dataset was reduced from over $12,900$ to $6,609$. However, this number of covariates was still exceedingly high, necessitating a subsequent feature selection phase. An array of filter- and wrapper-based methods, including Principal Component Analysis, Lasso, and Gradient Boosting Regression, were attempted in search of the most informative feature subset of reasonable cardinality. These methods were applied to the extracted pool of $6,609$ features, both recursively and explicitly, and probed under diverse parameter settings. The acquired subsets were then evaluated for their predictive accuracy across various models trained, effectively providing a means of validation. In essence, the latter step intends to establish the *overall validity of the model* underlying the proceeding analysis, thereby solidifying credibility of the explanations derived therein.

The target subset of optimally descriptive features was obtained by the following means. Feature importances were estimated by the Extra Trees Regressor algorithm (with 500 estimators) and Randomized Lasso, and the top 500 features were retained from each. For the latter, two different values were considered for the regularization parameter $\alpha$, namely 0.004 and 0.000004, thus resulting in two separate feature subsets. The intersection of these three subsets, containing 65 features, led to maximized GPA prediction accuracy. In particular, with the Random Forest algorithm, an MSE of approximately 0.359 was achieved under 3-fold cross-validation.

## References

1. Asif, R., Merceron, A., Pathan, M.K.: Predicting student academic performance at degree level: a case study. Int. J. Intell. Syst. Appl. **7**(1), 49 (2014)
2. Colom, R., Escorial, S., Shih, P.C., Privado, J.: Fluid intelligence, memory span, and temperament difficulties predict academic performance of young adolescents. Pers. Individ. Differ. **42**(8), 1503–1514 (2007)
3. Coyle, T.R., Pillow, D.R.: Sat and ACT predict college GPA after removing G. Intelligence **36**(6), 719–729 (2008)
4. Ermisch, J., Francesconi, M.: Family matters: impacts of family background on educational attainments. Economica **68**(270), 137–156 (2001)

5. Graziano, P.A., Reavis, R.D., Keane, S.P., Calkins, S.D.: The role of emotion regulation in children's early academic success. J. Sch. Psychol. **45**(1), 3–19 (2007)
6. Jackson, L.A., Von Eye, A., Biocca, F.A., Barbatsis, G., Zhao, Y., Fitzgerald, H.E.: Does home internet use influence the academic performance of low-income children? Dev. Psychol. **42**(3), 429 (2006)
7. Joshi, H., Fitzsimons, E.: The millennium cohort study: the making of a multipurpose resource for social science and policy. Longit. Life Course Stud. **7**(4), 409–430 (2016)
8. Laidra, K., Pullmann, H., Allik, J.: Personality and intelligence as predictors of academic achievement: a cross-sectional study from elementary to secondary school. Pers. Individ. Differ. **42**(3), 441–451 (2007)
9. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, vol. 30, pp. 4765–4774. Curran Associates, Inc. (2017)
10. McLoyd, V.C.: Socioeconomic disadvantage and child development. Am. Psychol. **53**, 185 (1998)
11. Pajares, F., Hartley, J., Valiante, G.: Response format in writing self-efficacy assessment: greater discrimination increases prediction. Meas. Eval. Counsel. Dev. **33**(4), 214 (2001)
12. Pal, A.K., Pal, S.: Analysis and mining of educational data for predicting the performance of students. Int. J. Electron. Commun. Comput. Eng. **4**(5), 1560–1565 (2013)
13. Pandey, M., Sharma, V.K.: A decision tree algorithm pertaining to the student performance analysis and prediction. Int. J. Comput. Appl. **61**(13) (2013)
14. Pritchard, M.E., Wilson, G.S.: Using emotional and social factors to predict student success. J. Coll. Stud. Dev. **44**(1), 18–28 (2003)
15. Reichman, N.E., Teitler, J.O., Garfinkel, I., McLanahan, S.S.: Fragile families: sample and design. Child Youth Serv. Rev. **23**(4), 303–326 (2001)
16. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144. ACM (2016)
17. Romero, C., Ventura, S.: Educational data mining: a review of the state of the art. IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.) **40**(6), 601–618 (2010)
18. Salanova, M., Schaufeli, W., Martínez, I., Bresó, E.: How obstacles and facilitators predict academic performance: the mediating role of study burnout and engagement. Anxiety Stress Coping **23**(1), 53–70 (2010)
19. Salganik, M.J., Lundberg, I., Kindel, A.T., Ahearn, C.E., Al-Ghoneim, K., et al.: Measuring the predictability of life outcomes with a scientific mass collaboration. Proc. Natl. Acad. Sci. **117**(15), 8398–8403 (2020)
20. Sargsyan, A., Karapetyan, A., Woon, W.L., Alshamsi, A.: Explainable AI as a social microscope: a case study on academic performance. CoRR abs/1806.02615 (2020). http://arxiv.org/abs/1806.02615
21. Tillman, K.H.: Family structure pathways and academic disadvantage among adolescents in stepfamilies. Sociol. Inq. **77**(3), 383–424 (2007)
22. Tross, S.A., Harper, J.P., Osher, L.W., Kneidinger, L.M.: Not just the usual cast of characteristics: using personality to predict college performance and retention. J. Coll. Stud. Dev. **41**(3), 323 (2000)