Applied Computer Science, vol. 19, no. 1, pp. 47–63
doi: 10.35784/acs-2023-04

Submitted: 2023-02-20 | Revised: 2023-03-06 | Accepted: 2023-03-09

Keywords: Missingness, Predictor variable, Training datasets, heuristics, unidimensionality

Olufemi A. FOLORUNSO [0000-0002-0242-9316]*,
Olufemi R. AKINYEDE [0000-0002-5544-8529]**,
Kehinde K. AGBELE [0000-0002-4265-0314]***

ARDP: SIMPLIFIED MACHINE LEARNING PREDICTOR FOR MISSING UNIDIMENSIONAL ACADEMIC RESULTS DATASET

Abstract

In this paper, we present the Academic Results Datasets Predictor (ARDP), for missing academic results datasets, based on chi-squared expected calculation, positional clustering, progressive approximation of relative residuals, and positional averages of the data in a sampled population. Academic results datasets are data originating from inside academic institutions' results repositories. It is a technique designed specifically for predicting missing academic results. Since the whole essence of data mining is to elicit useful information and gain knowledge-driven insights into datasets, ARDP positions data explorer at this advantageous perspective. ARDP is committed to solve missing academic results dataset problems more quickly over and above what currently obtains. PARD is computed by leveraging on the averages of neighbouring values. The predictor was implemented using Python, and the results show that it is admissible in a minimum of up to 85 percent accurate predictions of the sampled cases. It has been verified that ARDP shows a tendency toward greater precision in providing the best solution to the problems of predictions of missing academic results datasets in universities.

1. INTRODUCTION AND MOTIVATION

Data mining, also known as knowledge discovery, is the process of extracting useful insights and patterns from large amounts of data. It has become a crucial tool for businesses, researchers, and governments in order to make informed decisions and improve their operations. The concept of data mining can be traced back to the 1960s, when computer scientists started to explore the potential of using computers to analyze and process large amounts of data. In the 1970s, the relational database was created, which made it easier to store and find data, which was the first step toward data mining techniques.

* Computer Department, Elizade University, Ilara Mokin, Nigeria, olufemi.folorunso@elizadeuniversity.edu.ng
** Information Systems Dept., Federal University of Technology, Akure, Nigeria, femi_akinyede@yahoo.com
*** Computer Department, Elizade University, Ilara Mokin, Nigeria, kehinde.agbele@elizadeuniversity.edu.ng

In the 1980s, the term ‘data mining’ was coined by computer scientist and statistician J. Ross Quinlan, who developed the decision tree algorithm for data classification. This algorithm, known as Iterative Dichotomizer-3 (ID3), was one of the first data mining techniques used for predictive modeling. In the 1990s, data mining saw significant advancements with the development of new algorithms and technologies such as artificial neural networks (ANN), support vector machines (SVM), and association rule mining (ARM). These methods made it possible to look at unstructured data, like text and pictures, and find hidden patterns and trends.

The 21st century has seen a rapid expansion of data mining with the explosion of big data and the widespread use and adoption of the internet. The advancement of machine learning techniques and the proliferation of data analytics tools have made it easier for businesses and organizations to mine and analyze large datasets. Today, data mining is used in a variety of industries, including finance, healthcare, marketing, and e-commerce. It has become an essential part of decision making and has the potential to revolutionize how we understand and interact with the world around us. Therefore, data mining has come a long way since its inception in the 1960s. From simple decision tree algorithms to advanced machine learning techniques, it has become a crucial tool for extracting valuable insights and patterns from large amounts of data. As the amount of data continues to grow, it is likely that data mining will continue to play a vital role in shaping the way we make decisions and understand the world around us.

Undoubtedly, the invasions of the internet, the World Wide Web (WWW), and other educational resources have birthed enormous, occasionally uncontrollable databases for academic institutions worldwide. In order to forecast the future behavior or anticipated performance of prospective students and staff, researchers are constantly looking for connections between these similar, related, but disparate pieces of information, as mentioned by Breve et al. (2022) and Petropoulos et al. (2022). As a result, educational data mining (EDM) has recently attracted much scholarly attention. Educational data mining (EDM) aims to mine these unique types of datasets (Baker and Yacef, 2009, Baker, 2010; Romero & Ventura, 2006, 2013; Bucos & Drăgulescu, 2018). Due to their volume and uniqueness, conventional data mining methods and techniques would be unable to effectively predict and correctly visualize missing cases in these unique and special types of datasets specific to educational institutions. A major issue confronting most academic institutions today is how to accurately predict students' academic results while keeping the data left behind free of missingness. Another challenge during the visualization of Educational Data Mining (EDM) is to enable a data miner to get a win-win situation for both the users (the data generators) and the management (the data keepers) despite the missingness or incompleteness of the datasets. The introduction of PARD, by leveraging on the averages of neighbouring values could address the elusiveness of EDM, the gradient boosting algorithm, and the XGBoost algorithm as data mining tools.

2. A REVIEW OF EXISTING MISSING DATA PREDICTION TOOLS AND TECHNIQUES

Missing data prediction in data mining and visualization has become an increasingly important topic in recent years as datasets become larger and more complex. We explored the current state of the field and considered the future direction that missing data prediction is likely to take. Currently, missing data prediction techniques are used in a variety of applications, including predictive modeling, data visualization, and data analytics. These techniques are typically based on statistical methods such as imputation, interpolation, and extrapolation, which aim to fill in missing data points: an example was demonstrated by Jolani et al. (2015); Daberdaku et al. (2020). However, these techniques can be limited in their ability to accurately predict missing data, particularly when dealing with large or complex data sets.

One promising direction for the future of missing data prediction is the use of machine learning techniques. These techniques have the potential to learn patterns in data and make more accurate predictions, particularly when combined with advanced visualization techniques. For example, deep learning algorithms have been used successfully to recognize images, and it is likely that they will be used in the future to predict what data is missing. Another promising direction is the use of domain-specific knowledge to inform missing data prediction. In many cases, data sets are collected within a specific domain, such as finance, healthcare, or marketing. By incorporating domain-specific knowledge into missing data prediction algorithms, it may be possible to improve the accuracy of predictions. In addition to these technical advancements, it will also be important for the missing data prediction algorithms to be transparent and explainable. This will be particularly important for applications in which the results of the algorithm will be used to make important decisions, such as in healthcare or finance.

The future of missing data prediction in data mining and visualization looks promising, with the potential for significant advances in both technical capabilities and transparency. As data sets continue to grow in size and complexity, the need for effective missing data prediction techniques will continue to increase, making it an important area of research and development. This school of thoughts is otherwise known as heuristics. Therefore, missing data has become the focus of much recent data science research. Some situations arise in a university where a student(s) (especially, final-year students) is/are prevented from writing their examinations due to factors beyond their immediate control. Because of the course unit system, which is currently used by most universities worldwide; such students are deemed to have an additional year(s) of study in most cases. The situation is more daunting if the affected student is a finalist. We believe that requiring such students to repeat the affected courses in subsequent years is a waste of valuable resources. It has been established that such students go through emotional imbalance, isolation, rejection, and humiliation while going through this process, over which they have no single control. According to by Baepler and Murdoch (2010) such students go through emotional imbalance, isolation, rejection, and humiliation while going through this process, over which they have no single control.

Searching for missing academic data using relevant keywords in databases or search engines that specialize in academic research is one method. These databases and search engines can help locate articles, papers, and other academic resources that may contain the data being sought. Another way to find missing academic data is to reach out to other

researchers who may have access to such data. A lot of approaches have been applied for missing data imputation using a variety of algorithms and techniques that fit a value for the missing case(s) based on the overall behavior and pattern of the data population (Donders et al., 2006; Nadimi-Shahraki et al., 2021). Finding missing academic data can therefore require a combination of persistence, networking, and creative problem-solving using a variety of methods and resources.

In academic setting, data are generated daily by teachers, university administrators, and other stakeholder groups. Several analytical methods have also been applied to these data sets at various times. However, because these data are about humans, whose behavior is largely unpredictable due to other difficult-to-predict factors, using them as a barometer for prediction yields very few or no correlations. These were highlighted by McCalla (2004), Castro et al. (2007), Koedinger et al. (2008), Baepler & Murdoch (2010) and Zhou (2021). For example, a large volume of data is discarded because users don't really know how to get the best meaning out of it. In the five universities chosen for this study, up to 120 terabytes of data are discarded annually because it no longer makes sense to the universities. One can just imagine the enormous amount of information lost from these vast datasets! Even with the opportunity of secured cloud archiving, a good number of universities still discard data or save it in unusable formats.

Researchers like Brown et al. (2018) have linked missing and inconsistent data to human or machine error Some authors like Batista & Monard (2003), Choudhury & Kosorok (2020), believe that missing data are a normal part of any database. However, for good reasons, results for some students may be missing, not because of machine or human error but because of problems associated with other factors beyond the immediate control of the students. Machine learning algorithms such as the Gradient Boosting Algorithm and the XGBoost Algorithm cannot be applied successfully because of the obvious limitation of their dependencies on non-parametric statistical assumptions that have a lot to do with human and natural factors (Joel et al., 2022). Missing values can occur as a result of many human or machine problems, ranging from a deliberate attitude towards a questionnaire to an absence from a survey. Whichever case applies, missing data constitutes a bigger threat to today's databases. EDM tried to address this issue in many ways, ranging from the use of heuristics to the usage of algorithms, with very little success. It is a common opinion that no single method is perfect for missing data predictions, but some give decent performance that can be relied upon.

The organization of this paper is as follows: First, an exploration and definition of educational data mining (EDM), was presented alongside reasons why conventional data mining methods and techniques would be unable to effectively predict and correctly visualize missing cases in these unique and special types of datasets specific to educational institutions. Next, the historical perspectives and current state of data prediction models and tools were discussed as well as the future of missing data prediction in data mining and visualization. The significance of data preprocessing for PARD application was discussed in the next section. The next section talks about data mining tools, techniques, and associated algorithms, with specific tools and algorithms used for varying types of datasets. The peculiarities of academic results datasets are then discussed, emphasizing why and how they are different from other types of educational datasets in the following section. The following section discusses the peculiarities of academic results datasets followed by the formulation and ideology behind PARD. The formulation and the rationale behind the PARD predictor

are presented using an example with an emphasis on data preparation, cleaning, and usability using life samples of unidimensional datasets. A comparison of academic results data mining (PARD) techniques to conventional educational data mining approaches is presented. The results obtained after implementation of PARD with Python programming language was discussed and analyzed using tables. In the conclusion part, the basis for PARD application is presented by listing conditions that are acceptable for an PARD application. PARD was subjected to litmus tests by applying it to different kinds of academic results obtained from carefully selected universities across Nigeria. The results of the comparisons were analyzed and discussed. And in the last section, we provided a summary and suggested a few recommendations and modalities for the adoption of the predictor. A few indications for other possible future sub-branches of EDM were also suggested.

3. DATA PREPROCESSING – CRUCIAL STEP FOR PARD APPLICATION

Data preprocessing is a crucial step in the process of predicting missing data. It involves cleaning, transforming, and preparing the data for analysis and modeling. In this article, we will explore some of the key considerations for data preprocessing before predicting missing data. First, it is important to check for missing values in the dataset. Missing values can occur for a variety of reasons, such as errors in data collection, data entry mistakes, or data that was not collected in the first place. It is important to identify and address missing values before attempting to predict missing data, as they can affect the accuracy and reliability of the predictions.

One common approach to dealing with missing values is to simply remove the rows or columns that contain missing data. However, this can also be a limitation, as it may result in a significant loss of data. Alternatively, missing values can be imputed using techniques such as mean imputations, median imputation, or multiple imputations. These techniques involve estimating the missing values based on the values of other variables in the dataset. Another important consideration in data preprocessing is the scaling of the data. Data values may be on different scales, which can affect the performance of certain machine learning algorithms. For example, if one variable is measured in dollars and another in euros, the values of the latter will be much larger, which can distort the results. Scaling the data can help ensure that all variables are on the same scale, which can improve the accuracy of the predictions.

Data preprocessing helps to ensure that the data is clean, accurate, and ready for predictions and analysis. By following best practices for data preprocessing, researchers and analysts can improve the reliability and accuracy of their predictions, and gain valuable insights into their data. Please note the careful choice of usage of the words ‘missing value’ and ‘missing dataset’.

For most machine learning algorithms, cleaning is a major preprocessing activity. Data goes through a number of stages before it can be used, and the training samples for this predictor are no exception. Data values could be described as “missing” for so many reasons, ranging from users’ refusal to select appropriate options, user or machine errors, the putting in of guess values, to poor data archiving and maintenance problems. Several approaches have become acceptable in the literature for dealing with cases of missing values; one quick way is outright deletion of rows and/or columns. Removing rows and columns with multiple cases of missingness or zeros often provides a quicker fix for many datasets. However, this

is only good when the desired output does not depend on the deleted data or when the data is extremely large, in which case the deletion of a few rows makes little or no significant difference. In PARD, outright deletion of empty rows and columns is the first step. This is followed by removal of outliers. However, repeated values cannot be removed due to the sensitivity and nature of the datasets, same scores can be scored by contiguous students.

Another approach is using forward and backward fills, whose major disadvantages are obvious – data pollution as reported by Omri (2019). Backward fills cannot also be applied for obvious reasons. However, extreme outliers such as scores less that 15 out of 100 are removed because they could pose a significant threat to the rows average or corresponding positional values. The interpolation method can also be used for missing data value imputation. One example is using Panda's interpolation methods such as linear, polynomial, and quadratic. In this instance, interpolation of values will only lead to greater confusion because of averages. Hence, fixing by interpolation method will equally fail for academic datasets. It has also been said that regression analysis offers the most preferred option when performing predictions, simply because of its ability to establish some dependencies between dependent and independent variables. However, in the case of academic data, the dependent variables fall into a class of unpredictable phenomena – the students are never the same. This makes the regression analysis inapplicable for this illustration.

4. TOOLS, TECHNIQUES, AND BASIC ASSUMPTIONS

A fundamental assumption in this paper is that all data presented in any of the study samples are unidimensional, univariate, and rated in percentages. All the datasets used in this paper are live (based on examinations taken within the universities).

Missing data prediction tools are a class of statistical models that aim to fill in missing values in a dataset. These tools are widely used in various fields, including economics, finance, the social sciences, and health care, where data collection is often incomplete or prone to errors. One popular method for predicting missing data is the use of multiple imputations, which involves creating multiple datasets by imputing different values for the missing observations. These datasets are then analyzed separately, and the results are combined to produce a final estimate. Multiple imputations are better than other methods because they take into account the uncertainty that comes with missing values and give more accurate results.

Another popular method is the use of predictive modeling, which involves building a statistical model to predict the missing values based on the available data. This method is particularly useful when the missing data is not randomly distributed and is correlated with other variables in the dataset. Some common techniques for predictive modeling include linear regression, logistic regression, and decision trees.

A newer approach to missing data prediction is the use of machine learning algorithms, which can handle large and complex datasets with a high degree of accuracy. Machine learning algorithms like random forests, gradient boosting, and deep learning models are often used to make predictions with missing data. One important consideration when using missing data prediction tools is the choice of imputation method, as different methods can produce significantly different results. It's important to choose a method that works well with the data set and research question at hand.

Missing data prediction tools are therefore a valuable resource for analysts and researchers working with incomplete datasets. These tools can help fill in the gaps and produce more accurate estimates, enabling more robust and reliable conclusions to be drawn from the data. Data mining algorithms are used to analyze and extract useful insights from large datasets. These algorithms can identify patterns, trends, and relationships in data that may not be immediately apparent to humans. There are several types of data mining algorithms, including decision tree algorithms, clustering algorithms, and neural networks.

One of the main limitations of data mining algorithms is that they are limited by the quality and completeness of the data they are given. If the data is biased, incomplete, or incorrect, the results of the data mining algorithms will also be biased, incomplete, or incorrect. Additionally, data mining algorithms may not be able to identify all relevant patterns or trends in the data, as they rely on statistical analysis and may not be able to account for more complex or nuanced relationships. Another limitation of data mining algorithms is that they may be computationally intensive, requiring significant processing power and time to analyze large datasets. This can be hard for organizations with limited resources or when analysis needs to be done right away.

It has been noted by Abugroon (2018) that there are different educational data mining algorithms and approaches, so a specific comparison would depend on which algorithms and approaches are being considered. However, some common methods used in educational data mining include decision trees; neural networks, clustering, and association rule mining have been mentioned. These methods have been applied to a variety of educational data sets and have been shown to have varying levels of accuracy and usefulness depending on the specific context and data set. For example, decision trees have been shown to be effective at predicting student performance as reported by Anupama Kumar & Vijayalakshmi (2011), Coelho & Silveira (2017), while neural networks have been used to analyze student interactions with educational technology Fiore (2019). Clustering methods have been used to group students with similar characteristics or learning needs (Pasina et al., 2019), whereas Wang et al. (2022), reported on the use of association rule mining to identify patterns in student behavior and performance.

On the whole, data mining algorithms may raise ethical and privacy concerns, as they may extract sensitive or personal information from data. It is important for organizations, especially academic ones, to be open about how they use data mining and to have the right safeguards in place to protect people's privacy as much as possible.

5. ACADEMIC RESULTS DATASETS

Academic datasets greatly differ from other types of datasets in many respects. Academic results datasets, also known as student performance data, are an important resource for educators, researchers, and policymakers. They are a unique type of data that is often used in research and analysis. These datasets typically include information about students' grades, test scores, and other measures of academic achievement. However, there are a few peculiarities about these datasets that make them particularly unique and challenging to work with.

First, academic results datasets are often highly personal and sensitive. They contain information about students' academic abilities, which can have a significant impact on their future prospects. This means that it is important to protect the privacy of students when working with these datasets and to ensure that any data sharing or analysis is conducted in an ethical and responsible manner. Another peculiar aspect of academic results datasets is that they are often large and complex. These datasets may include information about thousands or even millions of students across a wide range of subjects, grades, and schools. This can make it difficult to identify patterns or trends in the data and to accurately interpret the results of any analysis. One of their peculiarities is that they are often highly structured and standardized. This allows for easy comparison of students across different schools, grades, and subjects. However, this standardization can also be a limitation, as it may not fully capture the complexity and diversity of students' learning experiences and achievements.

A third peculiarity of academic results datasets is that they are often dynamic and constantly changing. Another peculiar aspect of academic results datasets is that they are often collected over a long period of time, sometimes spanning decades. This can provide valuable insights into the long-term trends and patterns in academic performance, but it also requires careful consideration of changes in educational policies, curricula, and other factors that may affect the results over time. They often contain sensitive and personal information about students. This raises ethical and privacy concerns, as the data may be used to make decisions that have significant consequences for students' futures. It is important for researchers and analysts to be transparent about their data collection and use practices and to have appropriate safeguards in place to protect the privacy of students.

Students' grades and test scores can fluctuate over time, and new data may be added as students' progress through their education. This can make it challenging to accurately track students' academic progress and to identify areas of concern or potential improvement. Despite these peculiarities, academic results datasets can be an incredibly valuable resource for educators, researchers, and policymakers. By analyzing these datasets, we can gain insights into students' academic performance, identify areas of strength and weakness, and develop strategies to improve student outcomes. But it is important to be careful and thoughtful about these datasets and make sure that any analysis or sharing of data is done in an ethical and responsible way.

6. FORMULATION OF PARD

The computation of missing value involves formulating a machine learning algorithm for the prediction based on the averages of neighbouring values' rows and columns in the available data. The algorithm relies on the chi-squared model for computation of the expected values in missing datasets, and the progressive approximation of columns and rows averages, by simply computing the average of the duo.

There are several approaches to formulating and classifying missing value models. A very common approach is the regression method. In this approach, missing data is predicted based on a linear or nonlinear relationship with other variables in the dataset. There are several drawbacks to predicting missing data based on a linear or nonlinear relationship. Linear relationships might not be able to capture complex or nonlinear patterns in the data

correctly, which could lead to wrong predictions. Further, nonlinear relationships may be more difficult to model and may require more complex algorithms, which can be computationally intensive and time-consuming. In all cases, both linear and nonlinear relationships can be affected by outliers, which can change the results in a big way. Predicting missing data based on a linear or nonlinear relationship may not take into account the influence of other variables on the missing data, leading to incomplete or biased predictions. These approaches may not be suitable for predicting missing data in datasets with high levels of noise or variability, as they may not be able to accurately capture the underlying patterns in the data. On the whole, predicting missing data based on a linear or nonlinear relationship can be very useful, but it is important to carefully consider the limitations and drawbacks of these methods in order to ensure accurate and reliable predictions.

Another one is the classification method. In this approach, missing data is predicted based on a classification algorithm that categorizes data into distinct groups based on shared characteristics. There are several drawbacks to using a classification algorithm to predict missing data. First and foremost, classification algorithms rely on the availability of labeled data, which may not always be available or may be limited in quantity. They may not be able to accurately predict missing data if the data does not fit into a clear category or if the categories are not well defined. Most classification algorithms are sensitive to imbalanced data, where one class is much larger or more prevalent than the others. This can lead to biased or skewed results. These kinds of algorithms may not be able to accurately predict missing data if there are significant differences between the training and testing datasets. Finally, classification algorithms may be computationally intensive and time-consuming, especially for large datasets or complex classification tasks. Although classification algorithms can be a very useful way to predict missing data, it is important to think carefully about their limitations and drawbacks to make sure that their predictions are accurate and reliable.

A recent approach is the decision tree. In this approach, missing data is predicted based on a series of decision rules that split the data into smaller subsets based on specific criteria. However, decision tree algorithms may be sensitive to the quality and completeness of the data, as they rely on the data to make decisions. If the data is biased or incomplete, the predictions may also be biased or inaccurate. They may also be prone to “overfitting”, where the model becomes too complex and does not generalize well to new data. This can lead to poor performance on unseen data. They are also sensitive to the parameters of the model, such as the minimum number of samples required to make a split or the maximum depth of the tree. It is also a fact that decision tree algorithms could be computationally intensive and time-consuming, especially for large datasets or complex decision rules, and they are not be able to accurately predict missing data if there are significant differences between the training and testing datasets.

Another recent approach is the neural network, in which missing data is predicted using a complex network of artificial neurons that can learn and adapt to patterns in the data, similar to artificial intelligence (AI). However, neural network algorithms may be sensitive to the quality and completeness of the data, as they rely on the data to learn and adapt. If the data is biased or incomplete, the predictions may also be biased or inaccurate. They are also sometimes difficult to design and tune, as they require careful selection of the number and size of the layers, the type of activation functions, and the learning rate. Further, neural network algorithms are also computationally intensive and time-consuming, especially for

large datasets or complex neural networks. They may also be prone to "overfitting", where the model becomes too complex and does not generalize well to new data. This can lead to poor performance on unseen data. Finally, neural network algorithms may not be able to accurately predict missing data if there are significant differences between the training and testing datasets.

An ancient method called clustering is another approach, where missing data is predicted based on the patterns and relationships within a group of data points. However, it also comes with its own limitations. First, clustering algorithms are not be able to accurately predict missing data if the data does not clearly fit into distinct groups or if the groups are not well defined. In most of the cases, clustering algorithms are extremely sensitive to the initial conditions of the algorithm, which can significantly affect the resulting clusters. Again, these algorithms may be computationally intensive and time-consuming, especially for large datasets or complex clustering tasks. They are also unable to accurately predict missing data if there are significant differences between the training and testing datasets. These algorithms may not be able to accurately predict missing data if there are significant amounts of noise or variability in the data. Since academic datasets, of course, have these salient features of linear interdependencies largely because each tuple talks about a particular student (individual). This makes this method or approach useful and forms the basis of the PARD.

It is important to carefully select the appropriate missing value model based on the characteristics of the data and the goals of the analysis. Formulating the correct missing value model can greatly improve the accuracy and reliability of the predictions. The PARD is largely a mixture of the strengths of some of these major approaches. Using progressive approximation of relative residuals and positional averages of the data in the sampled population, it took advantage of the good things about clustering and neural network approaches and carefully avoided their biggest problems.

Consider a set of unidimensional unbiased datasets with n tuples and $k\alpha$ instances as shown in Table 1 below, where: $n, k \geq 1$, $\alpha \in N+$.

Tab. 1. Summary of predictive functions

	k_1	k_2	k_3	k_4	.	.	.	$k_{\alpha-1}$	k_α	
1			Q_i							*rs*
2			Q_{i+1}							*rs*
3			Q_{i+2}							*rs*
4			Q_{i+3}							*rs*
5	P_i	P_{i+1}	$\Omega_i = P_{i+2}$ (or Q_{i+4})	P_{i+3}				$P_{i+\alpha-2}$		*rs*
…										
n										
	cs	*cs*	*cs*	*cs*						*GT*

where: P_i – the predicted missing data (the Predictor),
Ω_i – the (chi-squared) expected value for data element in the i-th position,

$$\Omega_i - P_{i+2} \text{ or } (Q_{i+4}) = \text{MAX} \begin{cases} \text{Avg } Pi \text{ s less } Pi + 2 \\ \text{Avg } Qi \text{ s less } Qi + 4 \end{cases}$$

Rs – sum of data elements along row,
rs′ – sum of data elements along row less Ω_i,
cs – sum of data elements along column,
cs′ – sum of data elements along column less Ω_i,
GT – Grand total of either the rows or the columns,
GT′ – Grand total of either the rows or the columns less Ω_i.

Normally, in Chi-Squared computation, the expected value (*E*) for each data value on the table (matrix) is defined by

$$E = (rs * cs)/GT$$

However, because Ω_i is unknown and must be removed, the expected Ω_i of the *i*-th position by chi-squared formula normally changes to:

$$\Omega_i = ((rs' + \Omega_i) * (cs' + \Omega_i))/(GT' + \Omega_i)$$

or

$$\Omega_i = ((rs'cs' + rs'\,\Omega_i + cs'\Omega_i + (\Omega_i\text{^}2))/(GT' + \Omega_i)$$

$$\Omega_i GT' + \Omega_i\text{^}2 = rs'cs' + rs'\,\Omega_i + cs'\Omega_i + (\Omega_i\text{^}2)$$

$$\Omega_i GT' = rs'cs' + \Omega_i(rs' + cs')$$

$$\Omega_i(GT' - rs' - cs') = rs'cs'$$

$$\Omega_i = \frac{rs'cs'}{GT' - rs' - cs'}$$

Now, let the positional average of each row and column be defined as, and could be computed easily

$$P_j = \text{MAX}\{\text{Avrg. } P_i s \text{ less } \Omega_i \text{ or Avrg. } Q_i s \text{ less } \Omega_i\}$$

Now, we define Pi the predicted missing data value as the average of Ωi and the positional expected average Pj as:

$$P_i = (\Omega\text{i} + \text{Pj})/2$$

or

$$P_i = \frac{\frac{rs'cs'}{GT' - rs' - cs'} + p_j}{2}$$

$$P_i = \frac{(rs'cs' + Pj(GT' - rs' - cs'))}{2(GT' - rs' - cs')}$$

Since all variables in the formula are known (or could be easily evaluated), Pi is determined using equation 2.

7. RESULTS

In order to generate the machine learning predictor, we used the PARD based on chi-squared expected calculation, positional clustering and progressive approximation in a sampled population. The results were implemented using Python programming language.

A schematic example is given below for illustration using PARD.

Tab. 2. Example extracted from one of the sample data

Matric number	xxH 101	xxH x103	xxH 105	xxM 101	xxY 103	xxS 001
…./FT/2340	56	65	70	67	68	72
…./FT/2440	63	65	73	67	67	68
…./FT/2441	45	44	56	45	54	65
…./FT/2443	55	60	67	55	54	63
…./FT/2444	67	70	65	60	71	60
…./FT/2445	72	68	63	67	70	56
…./FT/2340	44	55	50	60	72	66

Assuming the score in the second course for ../FT/2443, i.e. "60," is missing; we can use the predictor in equation 2 to determine what the score should be, thus:

Tab. 3. Extracted from the schematic example from Table 2 above

Matric number	xxH1 01	xxHx1 03	xxH 105	xxM1 01	xxY 103	xxS0 01
…./FT/2443	55	60	67	55	54	63

$$\text{Now, } P_j = \text{MAX}\{\text{Avrg. } P_i \text{ s less } \Omega_i \text{ or Avrg. } Q_i \text{ s less } \Omega_i\}$$

$$P_j = \text{Max}\{58.8 \text{ and } 61.2 \text{ , using equation 1}$$

$$P_j \approx 61$$

$$\text{Now, } \Omega_i = 57.42 \quad P_j \approx 61;\ rs' = 294;\ cs' = 367, \text{ and } GT' = 2540$$

The adjusted Table 1 now appears like Table 4 below with the missing data “Pi”:

Tab. 4. Schematic example with the missing Pi*

56	**65**	**70**	**67**	**68**	**72**	**398**
63	65	73	67	67	68	403
45	44	56	45	54	65	309
55	**Pi**	67	55	54	63	**294**
67	70	65	60	71	60	393
72	68	63	67	70	56	396
44	55	50	60	72	66	347
402	**367**	444	421	456	450	**2540**

Using equation 2 above:

$$P_i = \frac{(\text{rs}\square\text{cs}\square + Pj(GT\square - \text{rs}\square - \text{cs}\square))}{2(GT\square - \text{rs}\square - \text{cs}\square)}$$

$$P_i = \frac{(367 * 294 + 61(2540 - 294 - 367))}{2(2540 - 294 - 367)}$$

$$P_i = \frac{222517}{3758} \text{ or } P_i \approx 59.21$$

8. DISCUSSION

The difference between the predicted 59.2 and the actual 60 is 0.8; the percentage error is 1.33%. Despite the insignificance of the difference, both scores end up with approximately same grades according to the examined universities' grading systems. This resulted in a prediction that was 100% accurate. Notwithstanding their seeming level of accuracy, all data predictors, including PARD, are imperfect. The results could be highly erroneous based on some factors. However, the level of imperfection could be mitigated to the bare minimum if the following were carefully considered and taken into consideration.

- The academic data must be live, unbiased and at same rating level.
- Only one case must be treated at a time.
- Reuse of predicted results or data should be avoided.

All these were carefully considered while applying PARD.

The tables below show the results obtained when PARD was applied to 536 final-year (400) level students across nine departments in three faculties. This resulted in a prediction accuracy of approximately 94%.

Tab. 5: Showing the summary of results obtained using Python

	LEVEL = 400								
	ENGINEERING			**SCIENCES**			**SOCIAL SCIENCES**		
	Civil Engineering	**Electrical Engineering**	**Mechanical Engineering**	**Medical Lab. Science**	**Nursing**	**Computer Science**	**Accounting**	**Business Administration**	**Economics**
No. of Students Compared	57	65	48	81	78	67	68	60	12
No. of Courses Compared	7	8	8	7	7	8	8	8	7
No. of Data Elements Examined	399	520	384	567	546	536	544	480	84
No. Acceptable Predicted Data	388	499	351	500	501	512	522	467	75
% Acceptance	97.2	96	91.4	88.2	91.8	95.5	96	97.3	89.3

9. CONCLUSIONSN AND RECOMMENDATIONS

The research paper "PARD – Academic Results Datasets Predictor" presents a new method for predicting missing data in academic results datasets. The approach is based on chi-squared expected computation and progressive approximation of relative residuals and positional averages of the data in the sampled population. This paper discusses the algorithm used by the PARD system and reports on the results of tests carried out using real-world academic results datasets. The results show that the PARD system is able to accurately predict missing data in these datasets up to about 93.6 percent accuracy level approximately, making it an acceptable and useful tool for researchers and administrators working with such educational datasets and for direct usage and application by academic institutions who may need it.

Testing the PARD system on a wider range of academic results datasets, including datasets from different universities in different countries and cultures, would provide a better understanding of the system's performance and generalizability. Incorporating additional data sources, such as student demographic information, into the PARD system could improve the accuracy of predictions. Developing a user-friendly interface for the PARD system would make it more accessible to researchers and administrators who are not familiar with programming. It was also believed that incorporating machine learning techniques such as neural networks or ensemble methods could improve the performance of the PARD system. A comparison of the performance of the PARD system with existing methods and algorithms for handling missing data in academic results datasets provides valuable insights into the relative strengths and weaknesses of the system. While collaborating with educational institutions to integrate the PARD system into their data management systems, it would allow for real-time missing data prediction and improve the accuracy and timeliness of data analysis.

It is also expected that evaluating the effect of the PARD system on the quality of research and decision-making by academic institutions would provide valuable feedback on the system's overall impact. The development of a similar tool for dealing with multidimensional datasets could definitely pose a greater challenge, and we believe that the application of PARD is a pointer for the possibilities of future predictors and formulae that could significantly undermine the problems of missingness and incompleteness in data mining. Finally, we see a great prospect in the application of the PARD for academic result validation.

Unlike other predictors and missing data imputation methods, PARD should be applied with utmost discretion. The advantages should be carefully weighed against needs and necessities before application. The following situations are recommended for its application:

- Sudden death of the student’s sponsor(s) within the examination period.
- A medical condition that has been determined to have a negative impact on the student's overall performance.
- A condition considered by the university management admissible for unavoidable absence from organized examinations.

10. FUNDING AND ACKNOWLEDGEMENTS

No grants were received for this research work. Consequently, the authors wish to thank the management of Elizade University, Ilara Mokin, Nigeria, for providing the necessary enabling environment for carrying out the research. Consequently, there are no financial or non-financial interests that are directly or indirectly related to this work submitted for publication.

Author Contributions

** Formulated the basic ideology behind PARD, provided guiding algorithm's basics, and implemented the results.*
*** Provided solutions to al mathematical functions and approximations*
****Developed the python programs used for testing the PARD.*

Funding

No grants were received for this research, consequently, there are no financial or non-financial interests that are directly or indirectly related to this work submitted for publication.

Acknowledgments

The authors wish to thank the management of Elizade university, Ilara Mokin, Nigeria for providing the necessary atmosphere for carrying out the research.

Conflicts of Interest

The authors declare that there are no conflicts of interest as applicable to this work.

REFERENCES

Abugroon, M. A. S. (2018). Comparison of Educational Datamining algorithms for Supporting the Decision in Sudanese Higher Education Institutions. *GCNU Journal*, *7*, 123-140.

Anupama Kumar, S., & Vijayalakshmi, Dr. M. N. (2011). Efficiency of decision trees in predicting student's academic performance. In D. C. Wyld, et al. (Eds.), *CCSEA 2011, CS & IT 02* (pp. 335–343). https://doi.org/10.5121/csit.2011.1230

Baepler, P., & Murdoch, C. J. (2010). Academic analytics and data mining in higher education. *International Journal for the Scholarship of Teaching & Learning, 4*(2), 1–9. https://doi.org/10.20429/ijsotl.2010.040217

Baker, R. S. J. D. (2010). Data mining for education. In B. McGaw, P. Peterson & E. Baker (Eds.), *International Encyclopedia of Education* (3rd ed, vol. 7, pp. 112–118). Elsevier.

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.

Batista, G. E. A. P. A., & Monard, M. C. (2010). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, *17*(5–6), 519–533. https://doi.org/10.1080/713827181

Breve, B., Caruccio, L., Deufemia, V., & Polese, G. (2022). RENUVER: A Missing Value Imputation Algorithm based on Relaxed Functional Dependencies. *Proceedings of the 25th International Conference on Extending Database Technology (EDBT)* (pp. 52-64). OpenProceedings.org.

Brown, A. W., Kaiser, A. K., & Allison, D. B. (2018). Issues with data and analyses: Errors, underlying themes, and potential solutions. *PNAS, 115*(11), 2563-2570. https://doi.org/10.1073/pnas.1708279115

Bucos, M., & Drăgulescu, B. (2018). Predicting Student Success Using Data Generated in Traditional Educational Environments. *TEM Journal*, *7*(3), 617-625. https://doi.org/10.18421/TEM73-19

Castro, F., Vellido, A., Nebot, A., & Mugica, F. (2007). Applying data mining techniques to e-learning problems. In: *Evolution of Teaching and Learning Paradigms in Intelligent Environment. Studies in Computational Intelligence* (vol. 62, pp. 183– 221). Springer. https://doi.org/10.1007/978-3-540-71974-8_8

Choudhury, A., & Kosorok, M. R. (2020), Missing data imputation for classification problems. *arXiv:2002.10709v1*. https://arxiv.org/pdf/2002.10709v1.pdf

Coelho, O. B., & Silveira, I. (2017). Deep Learning applied to Learning Analytics and Educational Data Mining: A Systematic Literature Review. Anais do SBIE 2017 (Proceedings of the SBIE 2017) (pp. 143-152). https://doi.org/10.5753/cbie.sbie.2017.143

Daberdaku, S., Tavazzi, E., & Di Camillo, B. A. (2020). Combined Interpolation and Weighted *K*-Nearest Neighbours Approach for the Imputation of Longitudinal ICU Laboratory Data. *Journal of Healthcare Informatics Research*, *4*(3), 174–188. https://doi.org/10.1007/s41666-020-00069-1

Donders, A. R. T., van der Heijden, G. J. M. G., Stijnen, T., & Moons, K. G. M. (2006). A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, *59*, 10, 1087–1091.

Fiore, U. (2019).Neural Networks in the Educational Sector: Challenges and Opportunities. *Balkan Region Conference on Engineering and Business Education*, *3*(1), 332-337. https://doi.org/10.2478/cplbu-2020-0039

Joel, L. O., Doorsamy, W., & Paul, B. S. (2022). A Review of Missing Data Handling Techniques for Machine Learning. *International Journal of Innovative Technology and Interdisciplinary Sciences*, *5*(3), 971–1005. https://doi.org/10.15157/IJITIS.2022.5.3.971-1005

Jolani, S., Debray, T.P., Koffijberg, H., van Buuren, S., & Moons, K. G. (2015). Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE. *Statistic in Medicine*, *34*(11), 1841-63. https://doi.org/10.1002/sim.6451

Koedinger, K., Cunningham, K., Skogsholm, A., & Leber, B. (2008). An open repository and analysis tools for finegrained, longitudinal learner data. In: *First International Conference on Educational Data Mining* (pp. 157–166).

McCalla, G. (2004). The ecological approach to the design of elearning environments: purpose-based capture and use of information about learners. *Journal of Interactive Media Education*, *1*, 3. https://doi.org/10.5334/2004-7-mccalla

Morales, C. R., Ventura, S. (2006). *Data Mining in E-learning*. Wit-Press.

Nadimi-Shahraki, M. H., Mohammadi, S., Zamani, H., Gandomi, M., & Gandomi, A. H. (2021). A Hybrid Imputation Method for Multi-Pattern Missing Data: A Case Study on Type II Diabetes Diagnosis, *Electronics*, *10*(24), 3167. https://doi.org/10.3390/electronics10243167

Omri, B.-S. (2019). Data Pollution. *Journal of Legal Analysis*, *11*, 104–159. https://doi.org/10.1093/jla/laz005.

Pasina, I., Bayram, G., Labib, W., Abdelhadi, A., & Nurunnabi, M. (2019). Clustering students into groups according to their learning style. *MethodsX*, *6*, 2189-2197. https://doi.org/10.1016/j.mex.2019.09.026

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben Taieb, S., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Cyrino Oliveira, F. L., De Baets, S., Dokumentov, A., … & Ziel, F. (2022). Forecasting: theory and practice. *International Journal of Forecasting*, *38*(3), 705–871. https://doi.org/10.1016/J.IJFORECAST.2021.11.001

Romero, C., & Ventura, S.(2013). Data Mining in Education. *WIREs Data Mining Knowledge Discovery*, *3*, 12–27. https://doi.org/10.1002/widm.1075

Wang, T., Xiao, B., & Ma, W. (2022). Student Behavior Data Analysis Based on Association Rule Mining. *International Journal of Computational Intelligence Systems*, *15*, 32. https://doi.org/10.1007/s44196-022-00087-4

Zhou, D. (2021). Financial Market Prediction and Simulation Based on the FEPA Model. *Journal of Mathematics*, 2021, 5955375. https://doi.org/10.1155/2021/5955375