American University in Cairo

AUC Knowledge Fountain

Theses and Dissertations

6-1-2017

Visualization as a guidance to classification for large datasets

Heba Abdelfattah Atteya

Follow this and additional works at: https://fount.aucegypt.edu/etds

Recommended Citation

APA Citation

Atteya, H. (2017).*Visualization as a guidance to classification for large datasets* [Master’s thesis, the American University in Cairo]. AUC Knowledge Fountain.
https://fount.aucegypt.edu/etds/691

MLA Citation

Atteya, Heba Abdelfattah. *Visualization as a guidance to classification for large datasets*. 2017. American University in Cairo, Master's thesis. *AUC Knowledge Fountain*.
https://fount.aucegypt.edu/etds/691

This Thesis is brought to you for free and open access by AUC Knowledge Fountain. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of AUC Knowledge Fountain. For more information, please contact mark.muehlhaeusler@aucegypt.edu.

المنـــارة للاستشارات

www.manaraa.com

The American University in Cairo

School of Science and Engineering

Visualization as a guidance to classification for large datasets

A Thesis Submitted to

The Department of Computer Science and Engineering

In partial fulfillment of the requirements for the degree of Master of Science

By

Heba AbdelFattah Atteya

B.Sc. Computer Science, May 2017

Under The Supervision of

Prof. Mohamed Moustafa

Spring 2017

المنارة للاستشارات

www.manaraa.com

ACKNOWLEDGEMENTS

I would like to extend my sincere gratitude to people who always stood by my side through the thick and then and were the spark of inspiration and motivation whenever my fires went out.

First, I am grateful to my family who were always supportive all through my lengthy research period. I owe an enormous debt of gratitude to my mother who dedicated her life to making me who I am. My brother who has my back at all times. My beloved and supportive husband who was always a constant source of joy and motivation through my struggles and trials of this thesis. My father and mother in law who believed in the value of my research.

Second, I would like to express my gratitude to my thesis supervisor, Dr. Mohamed Moustafa for believing in my research point and challenging my idea for a more thorough and comprehensive research.

Third, I would like to thank AUC in general for the staff scholarship opportunity which made pursuing my Master's degree possible and my manager, Dr. Iman Megahed for believing in my potential and being a role model.

Lastly, and most of all, I would like to offer this endeavor to our GOD Almighty for blessing me with the strength, health and peace of mind to finish this research.

المنارة للاستشارات
www.manaraa.com

ABSTRACT

Data visualization has gained a lot of attention after the stressing need to make sense of the huge amounts of data that we collect every day. Lower dimensional embedding techniques such as IsoMap, Locally Linear Embedding and t-SNE help us visualize high dimensional data by projecting it on a two or three-dimensional space. t-SNE, or t-Distributed Stochastic Neighbor Embedding proved to be successful in providing lower dimensional data mappings that makes interpreting the underlying structure of data easier for our human brains.

We wanted to test the hypothesis that this simple visualization that human beings can easily understand will also simplify the job of the classification models and boost their performance. In order to test this hypothesis, we reduce the dimensionality of a student performance dataset using t-SNE into 2D and 3D and feed the calculated 2D and 3D feature vectors into a classifier to classify students according to their predicted performance. We compare the classifier performance before and after the dimensionality reduction.

Our experiments showed that t-SNE helps improve classification accuracy of NN and KNN on a benchmarking dataset as well as a user-curated dataset on performance of students at our home institution. We also visually compared the 2D and 3D mapping of t-SNE and PCA. Our comparison favored t-SNE's visualization over PCA's. This was also reflected in the classification accuracy of all classifiers used, scoring higher on t-SNE's mapping than on the PCA's mapping.

Keywords— t-SNE, t-Distributed Stochastic Neighbor Embedding, Classification, Data Visualization, Dimensionality Reduction, Clustering.

المنارة للاستشارات

www.manaraa.com

TABLE OF CONTENTS

ACKNOWLEDGEMENTS II

ABSTRACT III

TABLE OF CONTENTS IV

LIST OF FIGURES VI

LIST OF TABLES VII

LIST OF ABBREVIATIONS VIII

CHAPTER (1): INTRODUCTION 1

1.1. Background 1
- 1.1.1. High Dimensional Dataset 1
- 1.1.2. Data Mining and Business Success 2
- 1.1.3. The curse of Dimensionality and Dimensionality Reduction Techniques 3
- 1.1.4. Data Visualization 3
- 1.1.5. t-SNE: Student-t Stochastic Neighborhood Embedding 4
- 1.1.6. Prediction of Students' Academic Performance 4
- 1.1.7. Data Mining in the Higher Education Industry 5

1.2. Problem Definition 6
- 1.2.1. Technical Problem Definition 6
- 1.2.2. Business Problem Definition 7

1.3. Thesis Contribution 7

.1.4 Thesis Outline 8

CHAPTER (2): RELATED WORKS 9

2.1. Prediction of Students' Performance 9

2.2. t-SNE Applications 12

CHAPTER (3): THEORETICAL BACKGROUND 21

3.1. Principal Component Analysis (PCA) 22
- 3.1.1. Summary of the approach [44] 22

المنارة للاستشارات

www.manaraa.com

.3.2 IsoMap 23
3.2.1. Summary of the Approach 24
.3.3 Locally Linear Embedding (LLE) 25
.3.4 t-Stochastic Neighbor Embedding (t-SNE) 25

CHAPTER (4): PROPOSED SOLUTION 28

.4.1 Phase 1 - Business Understanding 28
.4.1.1 Business Problem Statement 29
.4.1.2 Data Mining Problem Statement 29
.4.2 Phase 2 - Data Understanding 29
4.2.1. Identifying Relevant Data Elements for Collection 29
4.2.2. Data Quality Assessment and Challenges 34
4.3. Phase 3 – Data Preparation 35
4.4. Phase 4 – Modeling 36
4.5. Phase 5 – Model Evaluation 37
4.6. Deployment 39

CHAPTER (5): EXPERIMENTS 40

5.1............Benchmark Dataset: UCI-Predicting Portuguese Secondary School Student Performance [57] 40
5.1.1. Dataset 40
5.1.2. Data Preprocessing 41
5.1.3. Dimensionality Reduction using t-SNE 42
5.1.4. Comparison of Classification Results 44
5.1.5. Results Discussion 45
5.2. AUC Dataset (All Attributes) – Exploring visualization capabilities of t-SNE 46
5.2.1. Dataset 46
5.2.2. Dimensionality Reduction using t-SNE 46
5.2.3. Comparison of classification results before and after applying t-SNE 52
5.2.4. Results Discussion 53
5.3. AUC Dataset (Numeric Attributes) – Comparing t-SNE to PCA 54
5.3.1. Dataset 54
5.3.2. Visualizing data using PCA and t-SNE 54
5.3.3. Comparison of Classification Results 54

CHAPTER (6): CONCLUSION AND FUTURE WORK 56

6.1. Conclusion 56
6.2. Future Work 57

REFERENCES 59

المنارة للاستشارات

www.manaraa.com

LIST OF FIGURES

Fig. 2.1 Flowchart of Facial Expression Recognition including t-SNE [32] 13

Fig. 2.2 Visualization of Bacteria DNA sequences using PCA, GTM, and t-SNE [35] 14

Fig. 2.3 Visualization obtained using t-SNE on the musical loops database [36] 15

Fig. 2.4 Visualization of Chowdary dataset using t-SNE on the left and PCA on the right [37] 16

Fig. 2.5 Visualization of su dataset using t-SNE on the left and PCA on the right [37] 17

Fig. 2.6 Visualization of the MNIST dataset using IsoMap as presented in supplemental material to [15] ... 18

Fig. 2.7 Visualization of the MNIST dataset using LLE as presented in supplemental material to [15] 19

Fig. 2.8 Visualization of the MNIST dataset using t-SNE as presented in supplemental material to [15]..... 19

Fig. 3.1 The Swiss Roll dataset 24

Fig. 4.1 Flowchart for our proposed solution showing t-SNE as a pre-classification step 39

Fig. 5.1 2D Mapping of the UCI Students’ Performance Dataset 43

Fig. 5.2 3D Mapping of the UCI Students’ Performance Dataset 44

Fig. 5.3 t-SNE 2D Mapping of AUC student performance dataset 47

Fig. 5.4 t-SNE 3D mapping of AUC student performance dataset 48

Fig. 5.5 t-SNE 2D mapping of AUC dataset. Achievement scholarship status used for color coding. 49

Fig. 5.6 t-SNE 2D mapping of AUC dataset. Gender used for color coding 49

Fig. 5.7 t-SNE 2D mapping of AUC dataset. Students' majors used for color coding.. 50

Fig. 5.8 t-SNE 2D mapping of AUC dataset. Second Year GPA used for color coding 50

Fig. 5.9 PCA 2D Mapping of numerical AUC students’ performance dataset 55

Fig. 5.10 t-SNE 2D Mapping of numerical AUC students' performance dataset 55

Fig. 5.11 PCA 3D Mapping of numerical AUC students' performance dataset 55

Fig. 5.12 t-SNE 3D Mapping of numerical AUC students' performance dataset 55

www.manaraa.com

LIST OF TABLES

Table 4.1 AUC Dataset: Demographic Attributes 31

Table 4.2 AUC Dataset: Admissions Attributes 31

Table 4.3 AUC Dataset: General Academic Attributes 32

Table 4.4 AUC Dataset: Student Engagement Indicators 33

Table 4.5 AUC Dataset: Financial Assistance Attributes 33

Table 4.6 AUC Dataset: Academic Information 33

Table 5.1 Attributes of UCI Portuguese secondary school students’ performance [57] 41

Table 5.2 Comparison of the classification accuracy on the UCI dataset 44

Table 5.3 Comparison of the classification accuracy on AUC full dataset 52

Table 5.4 Comparison of the classification results using Cohen’s Kappa on AUC full dataset 52

Table 5.5 Comparison of classification accuracy of PCA And t-SNE Mappings of AUC numerical dataset. 56

المنارة للاستشارات

www.manaraa.com

LIST OF ABBREVIATIONS

IBM	International Business Machines
ACT	American College Testing
ANN	Artificial Neural Network
AUC	The American University in Cairo
BBN	Bayesian Belief Network
C4.5	An algorithm used to generate a type of decision trees that is an improvement over ID3
CART	Classification and Regression Trees
CRISP-DM	Cross-Industry Standard Process for Data Mining
DNA	Deoxyribonucleic acid
FER	Facial Expression Recognition
GTM	Generative topographic mapping
GPA	Grade Point Average
ID3	Iterative Dichotomiser 3. An algorithm used to generate a type of decision trees.
KNN	k Nearest Neighbor
LE	Laplacian Eigenmaps
GDA	General Discriminant Analysis
LDA	Linear Discriminant Analysis
LLE	Locally Linear Embedding
LMS	Learning Management System
MDS	Multi-dimensional Scaling
NN	Neural Network
PCA	Principal Component Analysis
RF	Random Forest
RMSE	Root Mean Square Error
RPART	Recursive Partitioning Trees
SAT	Scholastic Assessment Test
SNE	Stochastic Neighbor Embedding
SVM	Support Vector Machine
t-SNE	Student-t Stochastic Neighbor Embedding
WEKA	Waikato Environment for Knowledge Analysis

المنارة للاستشارات
www.manaraa.com

CHAPTER (1): INTRODUCTION

1.1. **Background**

Data visualization techniques are useful tools for data analysis. Visualizing high dimensional data is a challenge that lower dimensional embedding techniques has successfully overcome. IsoMap, Locally Linear Embedding and t-SNE are examples of such embedding techniques. In this research, we focus on t-SNE as a dimensionality reduction and visualization technique and use it as a pre-classification step to classify students according to their academic performance. The rest of this section elaborates on this idea.

1.1.1. High Dimensional Dataset

High-dimensional datasets are becoming more common with our increasing ability to gather more data about every problem and our quick advancement in improving the infrastructure that can handle vast amounts of data. In a study conducted by IBM in 2012, IBM estimated that around 2.5 quintillion bytes of data are created every day. They also added that 90 percent of the data in the world today were just produced within the two years 2010 and 2011 [1]. This provides a great potential for better understanding the world and making smarter informed-decisions. However, dealing with these vast amounts of data introduces a tough challenge to the organizations. Institutions are required to analyze and make sense of very high-dimensional datasets in order to enhance their customer experience or make cost-effective decisions and find an edge in a very competitive market. The real challenge is always uncovering the underlying structure of data, the clusters within the data, the normal pattern and the outliers. This gets more complicated as the data size

المنارة للاستشارات
www.manaraa.com

and dimensionality increase. For example, if you are trying to solve a classification problem, it would be very helpful if you were able to tell whether your classes are linearly or non-linearly separable, and accordingly decide on the most suitable algorithms for classification and the possible parameters space that you would want to try, this becomes more difficult as the data dimensionality grows.

1.1.2. Data Mining and Business Success

Researchers conducted a lot of research on machine learning and data mining as means of generating insights from large datasets and businesses have benefited a lot from these researches. Data mining and analysis are now among the main pillars of business success and survival. Netflix took over Blockbuster, the movie rental retail company that ruled the business for decades. Netflix analyzed and understood customer's behavior and introduced a new business model of movie rental through their online-streaming model which proved to be more convenient to its customers. Netflix leverages data analysis and predictive analytics to make smart movie suggestions to users based on their past movie ratings thus increasing customers' retention rates. In 2010, Blockbuster went bankrupt whereas in 2014, Netflix was estimated to be worth of $28 billion [2].

Amazon.com is another success story of employing advanced data mining techniques to improve customers' experience. Amazon.com, estimated to be worth of $292 billion and ranked among the top 15 innovative companies in 2016 by Forbes [3], analyzes your purchasing data, your click-stream and your demographics information to find the hidden patterns in users purchasing behavior. This extensive data analysis informs Amazon on how to offer you suggestions for more items that you are likely to consider buying.

المنارة للاستشارات

www.manaraa.com

1.1.3. The curse of Dimensionality and Dimensionality Reduction Techniques

Data Mining and Machine learning can solve very complicated problems but they have always been plagued by the curse of dimensionality [4]. Hence, dimensionality reduction became a very important pre-processing step for an effective machine learning or data mining model. Feature selection [5], [6], [7], [8] and feature extraction [9], [10] are two common approaches for dimensionality reduction. We will address the difference between the two techniques in chapter II: Theoretical Background.

1.1.4. Data Visualization

Data visualization is another method used by researchers and scientists to help in uncovering the hidden patterns and structure of high dimensional data. One might think that data visualization is a relatively new research area, but in fact it has roots that extend back to ancient times [11]. The field was revived in the last quarter of the twentieth century with so many developments that are so varied, across wide range of disciplines and applications.

Data visualization has proved to be an effective data analysis technique in several applications such as facial expression classification and visualization images [12], diagnostic of assembly line performance in smart factories [13], malware analysis [14].

Data visualization combined with data clustering and classification and dimensionality reduction techniques can be employed to generate lower-dimensional mapping of high dimensional datasets which reveal useful information about the data structure, similarities and differences.

المنارة للاستشارات
www.manaraa.com

1.1.5. t-SNE: Student-t Stochastic Neighborhood Embedding

In this research, we focus on using t-SNE [15], a data visualization algorithm that tries to find an "interesting" 2D or 3D mapping of high dimensional data preserving much of the local structure of the high-dimensional data very well, while also revealing global structure such as the presence of clusters at several scales. We will try to prove our hypothesis that this lower-dimensional mapping can enhance the performance of the data classification algorithms. While we try to prove our hypothesis, we will be addressing a specific application; prediction of students' academic performance.

1.1.6. Prediction of Students' Academic Performance

Student's academic performance is one of the factors contributing to overall student's success which is now a hot topic of research in the industry of higher education specially with the increasing economic pressures and competition between universities. Student's academic performance can also inform higher education institutions on the probability of his/ her attrition and help us identify students at risk of dropping out. The wasted time and paid tuition are huge losses to the student. Nevertheless, the institutional cost of a dropping-out student can also be significant. This cost can be estimated as the cost of unrealized revenue from anticipated tuition and fees plus the cost of harm to the institutional reputation in case of an unsatisfied dropping-out student, not-to-mention the effect of that on institutional ranking and accreditations as a result of drops in retention and graduation rates.

While data mining has helped the success of many businesses in different industries, it has only gained recent popularity in the educational field [16]. Universities collect large volumes of data about their students but they do not usually make the best use out of it. That makes them data-rich, however information-poor and raises the need for data mining

المنارة للاستشارات

www.manaraa.com

models that can generate insights from this wealth of data. This paved the way for the evolution of the Educational Data Mining, a new field that utilizes machine learning and data mining to solve educational problems [17].

1.1.7. Data Mining in the Higher Education Industry

Enrollment Management is a huge concern for higher education institutions. Enrollment Management starts as early in the student journey as students' recruitment. Universities allocate huge budgets for costly school visits, college fairs and hosting special events on campus to reach out to the largest population of potential applicants. Data Mining and predictive analytics have helped universities in identifying those potential applicants with maximum likelihood of becoming actual admits. Universities, such as Baylor University, raised their number of new applications from 15,000 to 26,000 in Fall 2005 after using data mining to develop a model of the different student profiles who are more likely to be admitted and concentrated their recruitment efforts to target those students [18].

Other challenges facing higher education institutions is course planning of enrolled students. Shatnawi et al [19] used association rule mining to help both students and advisors in selecting and prioritizing courses. The system uses a priori association rule mining to find association between courses that have been registered by students in many previous semesters to recommend a set of courses for the student to register sorted according to the rule confidence value.

Hegazy et al [20] proposed a framework that uses classification and clustering to recommend a specific major or track for a student. The framework included attributes related to students' academic level before college enrollment, students' major of interest and his/ her grade in the first year. The researchers compared the performance of several

www.manaraa.com

classification techniques and their experiment showed that the decision tree algorithm C4.5 is the best classifier that yields the highest F-measure value when predicting the department of major. They also used K-means clustering to cluster students' based on their department of major.

Identifying students who are academically at risk early on during their educational journey is another application that allows institutions to intervene before it is too late. This intervention can be in the form of one-on-one academic advisory sessions that can help students make better academic decisions. An institution can also offer remedial courses or suggest a different learning track. Students who are struggling due to psychological issues may be advised for counseling sessions. An institution that is proactively considering student success will automatically realize enhancement in its retention and completion rates and consequently ranking and reputation. In this study, we showcase how machine learning combined with data visualization can solve this specific problem. The definition of academically at-risk students differs from one university to another [21]. For the purpose of this research, we define students at risk in terms of their overall academic performance. More specifically and for simplicity, we define students at risk as those whose GPA falls below 2.0.

1.2. **Problem Definition**

1.2.1. Technical Problem Definition

This research aims at studying whether an effective data visualization technique such as t-SNE can guarantee minimal information loss and thus be used as a data preprocessing step to data classification or clustering. That is given a dataset D of n

للاستشارات المنارة
www.manaraa.com

dimensions and its 2D and 3D embedding using t-SNE denoted by M2 and M3 respectively, we would like to prove that using M2 and M3 which re-arrange the data points such that the true clusters within the data are more emphasized may result in enhancing a classifier X performance. The importance of such finding may help in solving the dilemma of explaining machine learning outputs to business users through an easy to understand visualization.

1.2.2. Business Problem Definition

Since data analytics is relatively new to the higher education industry, it is difficult to educate the community on the importance of its adoption and convince them of the accuracy of its results. In this research, we decided to use data visualization to increase human involvement in ways other than simply being a primitive input device that feeds in data as input to the data mining algorithm. Human beings are much more efficient in interpreting complex situations than any computer [22]. In order to achieve this goal, data must be mapped to a lower dimensional scale (2D or 3D) so that we can get a feel of how the data objects are arranged in the data space and accordingly use our brain powers to identify the hidden patterns by merely looking at a simple scatter-plot of the data. In this study, we will be developing a framework for a students' performance prediction model, researching students' attributes most informative about predicting their performance, trying different machine learning techniques and reporting on their performance and using visualization of the dataset structure to reason about the differences in the different algorithms performance.

1.3. Thesis Contribution

Our contribution can be summarized as follows:

المنارة للاستشارات
www.manaraa.com

- We were able to prove that using t-SNE as a dimensionality reduction technique on two and three dimensions was able to boost NN and KNN performance on a benchmarking dataset as well as on our AUC students' performance dataset. Reducing dimensionality to 2D and 3D using t-SNE boosted the classification accuracy in comparison to reducing dimensionality using PCA as well as using the raw full-dimensional dataset.
- We collected and mined data on student performance at AUC in order to predict students' third year GPA at AUC given their first two-years information with an accuracy of 72% ± 0.7 using RF on the raw full-dimensional numerical and categorical attributes dataset and an accuracy of 71% ± 0.6 by including numerical variables and reducing dimensionality using t-SNE to 3D.

1.4. Thesis Outline

In chapter II, we discuss related works focusing on the business problem of students' academic performance and the use of t-SNE in different applications that showed potential to solve our problem here. In chapter 3, we demonstrate necessary theoretical background and provide a brief analysis of the different dimensionality reduction and data mining algorithms then we present our approach to solving the problem in chapter 4. In chapter 5, we show our experiments. Finally, we discuss the results in chapter 6 and we provide a conclusion and plan for future work in the last chapter.

للاستشارات المنارة
www.manaraa.com

CHAPTER (2): RELATED WORKS

2.1. **Prediction of Students' Performance**

The work referenced as [23] was able to predict university student graduation performance based on high school scores and grades in first and second year courses without having to check any socioeconomic or demographic information. They worked on a dataset of four academic cohorts comprising 347 students and grouped them into five classes (A,B,C,D,E) based on their Graduation GPA. They used RapidMiner for exploration, statistical analysis and applying the different types of classifiers; Decision Trees, Naive Bayesian, Nearest Neighbor and Neural Networks and reported highest performance prediction accuracy of 83.65% by the Naive Bayesian classifier. They also confirmed the finding in [24] that students' grades in courses closer to graduation are better predictors of their performance than pre-university or first-year courses.

In the research conducted by Rusk et al. [25], the authors were able to identify the factors that serve as good indicators of whether an engineering student will drop out or fail the program. They focused their study on electrical and computer engineering students only and used their first academic year information only to predict the student classification as either **S**atisfactory for those who completed their degree without ever being put on probation, **P**robation for those who completed their degree but have been placed on probation at least once during their study journey and **F**ailure for those who were required to withdraw from the Engineering program. They worked on a dataset of 72 records and using principal components analysis as a dimensionality reduction technique, they found out the top three courses contributing to the first PCA which contributes by 86% to the total

www.manaraa.com

variance in the dataset. They concluded that those three foundational courses: MATH110, 133 and ENGR120 are the most informative predictors of engineering students' academic performance which confirmed a long-term belief at their institution that was never scientifically proven. Their future plans were to apply the same methodology to find out other problematic courses in later years of study that affect student performance and are probable reasons for students to drop-out.

Kozlick et al. presented a framework [26] that uses Bayesian Belief Network (BBN) to predict the performance of students early in their academic careers and advise them into their best-fit majors. The dataset was for 400 students from 4 different departments and the courses taken by these students are spread over 22 semesters (7 years). Their work showed that BBNs can easily model a curriculum graph and can be used to predict the future progress of a student.

Advising on students' admissions decision is also a huge concern for higher education experts. In [27], authors mined prospective students' data to advise on their admissions decision. They built an artificial neural network (ANN) that takes students' scores in standardized exams, including university entrance exams, high school graduation exams, high school location, type of high school (public or private), gender and elapsed time between high school graduation and applying for a university degree as input and predicted student's first-year average score. They used a real dataset of 653 enrolled students at University of Transport Technology, divided into (60%)training and (40%)testing. They were able to achieve a RMSE of 0.4819 using ANN.

Yadav et al [28] used decision trees to predict student's performance, probable drop-outs, students who need special attention based on past student's performance. They used

المنارة للاستشارات
www.manaraa.com

a real dataset for 48 enrolled students in Purchanval University in the sessions from 2008 to 2011. For every student, his previous semester grades, average of current semester grades, student's performance in the seminar class, assignments submission, student's attendance and completion of lab work are used to predict his overall performance. They compared the performance of three of the most frequently used decision tree algorithms (ID3, C4.5 and CART). They used the WEKA knowledge explorer to conduct their experiments which showed that CART is the best algorithm for the classification of students' data with 56.25% accuracy.

Al Sarem et al [29] used the difference between registered and earned credit hours to build a decision tree model based on the algorithm C4.5 to generate a set of rules that would help the academic advisors identify the students at risk. Lauria et al [30] had a different finding when comparing multiple classifiers. Specifically, when comparing logistic regression, SVMs and C4.5 decision trees to predict whether a student is at academic risk based on demographical and course enrollment data. Their experiments proved that logistic regression and SVMs outperform the decision tree C4.5 to perform this classification. Romero et al [31] presented a review of the state of the art educational data mining techniques and a survey of the different applications.

المنارة للاستشارات
www.manaraa.com

2.2. **t-SNE Applications**

In our study, we wanted to make sure that our dataset can be visually explored by our stakeholders and therefore, we are going to use t-SNE algorithm to find a 2D or 3D mapping of our dataset. t-SNE has succeeded to provide appealing visual representations for various datasets.

Xue et al. [32] used t-SNE as a data dimensionality technique as a pre-step to employ AdaboostM2 as a multi-classifier for facial expression classification. The experiments were done on the Japanese Female Facial Expression database. Their experiments showed that using t-SNE as a pre-processing step as shown in Fig. 2.1 enhances both SVM and AdaBoostM2 performance in comparison to PCA, LDA, LLE and SNE. Their best reported classification accuracy was 94.5% achieved by combining t-SNE and AdaBoostM2.

A study on the relationships between Wikipedia articles [33] compared the effect of using two different approaches to building the feature vector of each article on t-SNE's produced mapping. Their first set of experiments used the term frequency of all words in the article after removing the stop words and the second set used only the semantic role labels to compose the feature vector of each article. They were able to show that using the second approach, related articles were mapped to well clustered data points.

المنارة للاستشارات
www.manaraa.com

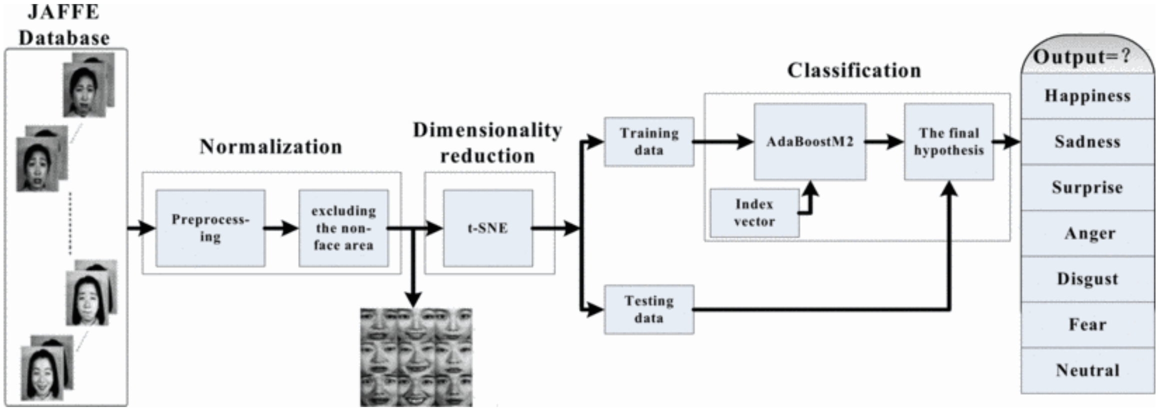

Fig. 2.1 Flowchart of proposed solution for FER including t-SNE for dimensionality Reduction [32]

Copyright © 2013, IEEE

In [34], Oord et al. the authors used t-SNE to provide a mapping for the Million Song Dataset onto 2D. They provided a visualization of the predicted usage patterns. The visualization showed clear separation of songs based on their genres.

Mokbel et al. [35] were able to detect clusters in the context of metagenomics using t-SNE and proposed a variation of the algorithm which they called “kernel-t-SNE” as a fast parametric counterpart based on t-SNE. They worked on a manually curated dataset of DNA sequences for 21 bacteria of different species. Fig. 2.2 compares the 2D mapping obtained for the dataset using PCA, generative topographic mapping (GTM) and t-SNE. PCA could not provide a mapping of clear cluster separation except for one cluster, whereas the non-linear visualization techniques; GTM and t-SNE showed better separation of the clusters within the data. The cluster structure is less clear by GTM with some overlaps between the different clusters but t-SNE succeeded to provide clear separation of all clusters.

المنارة للاستشارات
www.manaraa.com

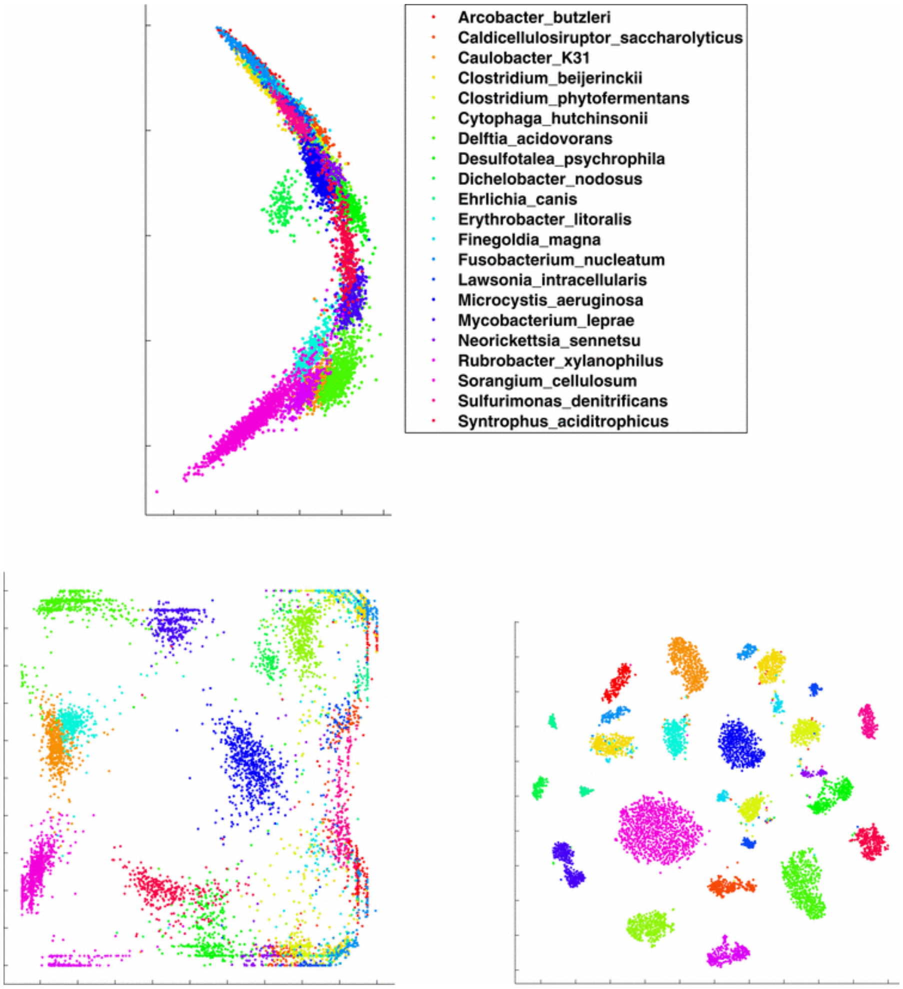

Fig. 2.2 Visualization of the data set according to PCA (top), GTM (bottom left), and t-SNE (bottom right). Coloring is done according to the real classification of the data-point. [35]

Copyright © 2013, IEEE

Another application of t-SNE in music/sound classification [36] compared the performance of t-SNE to PCA and IsoMap on two real audio use cases: musical instrument loops used in music production and sound effects used in sound editing. Their experiments proved that t-SNE presents the best visualization. Fig. 2.3 shows the t-SNE mapping obtained for the musical loop database. The clear separation of sounds of the different instruments proves the clustering power of t-SNE. They also proved that t-SNE can improve classifier performance if used as a preprocessing step when the amount of labeled data is low. They reported enhancement in the performance of the KNN classifier

www.manaraa.com

as a result of reducing the dimensionality only to 2D using t-SNE over the original dataset as well as the mappings by PCA and IsoMap.

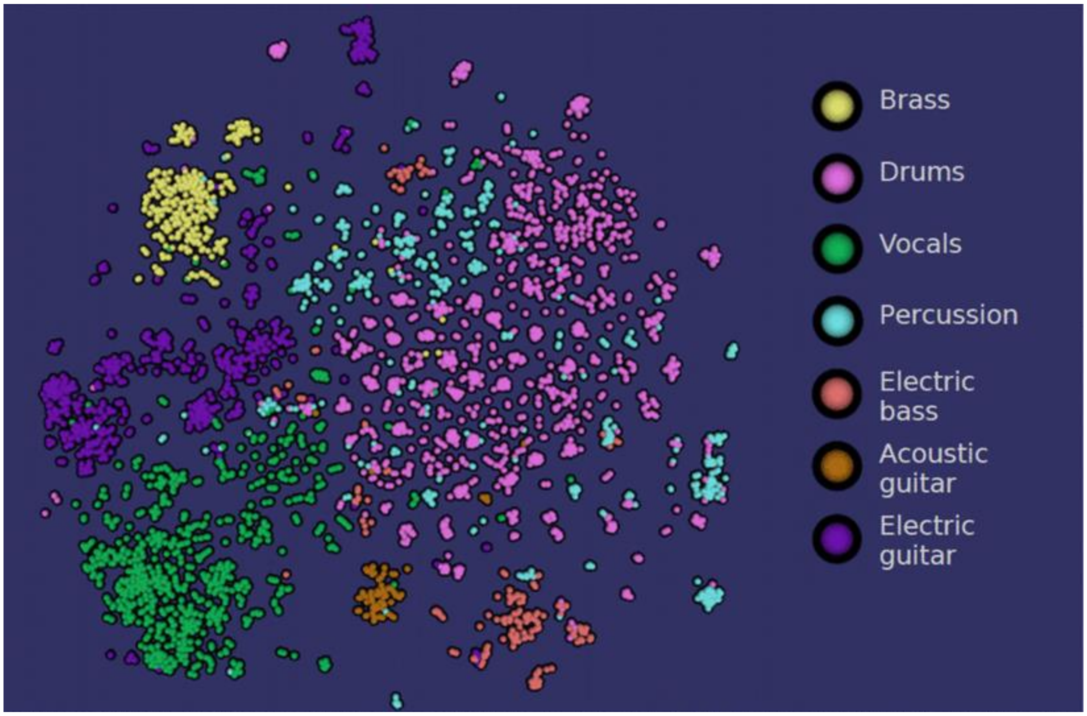

Fig. 2.3 Visualization obtained using t-SNE on the musical loops database [36]

Copyright © 2013, IEEE

A comparative study [37] that studied the impact of seven dimensionality reduction techniques on the quality of the partitions produced by cluster analysis of micro-array data. The different techniques under study were: PCA, IsoMap, Locally Linear Embedding (LLE), Laplacian Eigenmaps (LE), t-SNE and General Discriminant Analysis (GDA). They also tested the consistency of their results by applying the different techniques to five different micro-array datasets. Their first finding was that applying any dimensionality reduction improved the clustering result over that obtained on the original dataset. The study concluded that t-SNE and Laplacian eigenmaps outperform PCA, kernel PCA, IsoMap and the Locally linear embedding techniques. In Fig. 2.4 and Fig. 2.5, we show

www.manaraa.com

their provided visualizations to two of the datasets using t-SNE (The technique that achieved best overall results) and PCA (The most widely applied dimensionality reduction technique) and showed how t-SNE provides better separation of the different clusters within the data.

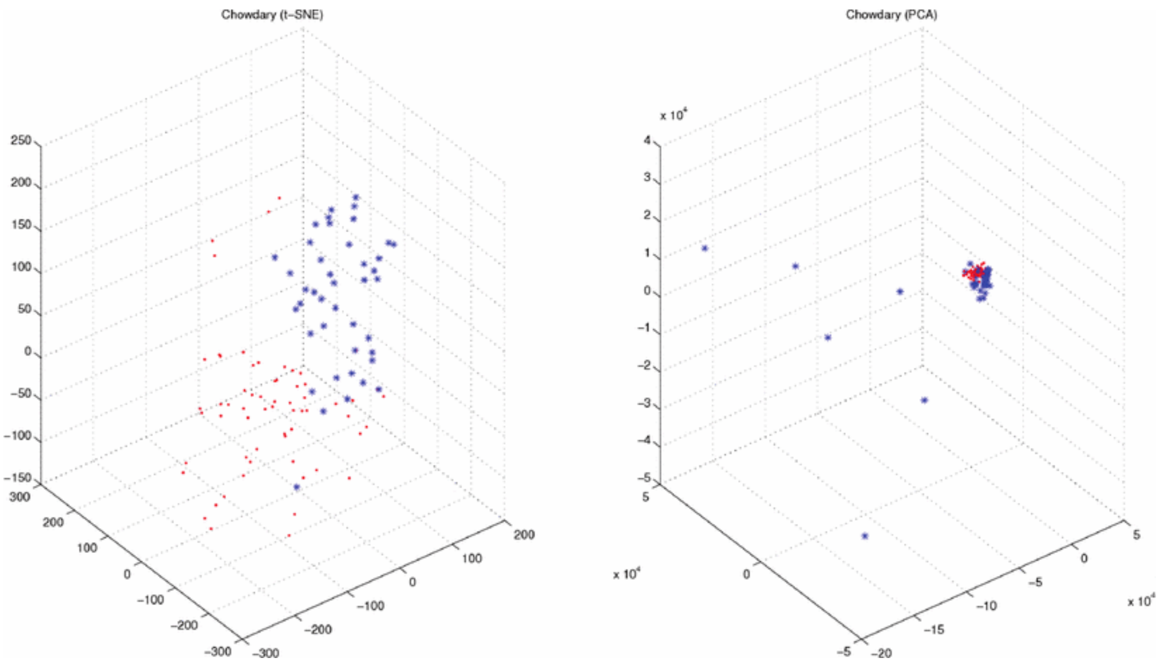

Fig. 2.4 Visualization of Chowdary dataset using t-SNE on the left and PCA on the right [37]

Copyright © 2011, IEEE

www.manaraa.com

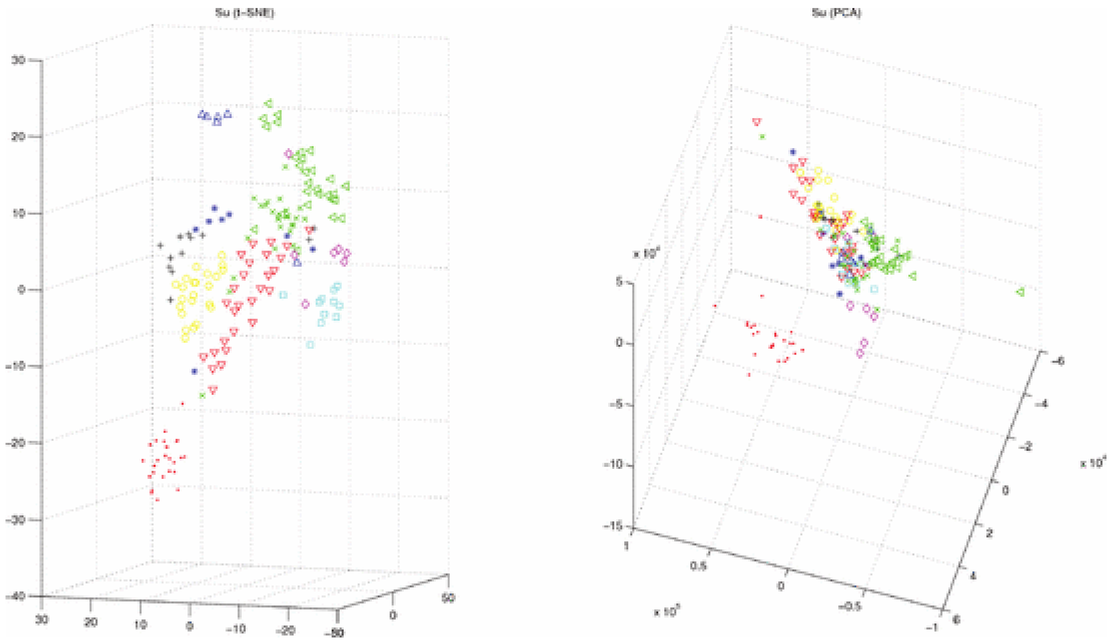

Fig. 2.5 Visualization of su dataset using t-SNE on the left and PCA on the right [37]

Copyright © 2011, IEEE

The work referenced in [38] identified faults in system elements and abnormal behaviors in circuits of the heating system of one of the buildings of University of Leon when the used t-SNE to reduce the dimensionality of the daily heating consumption data collected from the different sensors which enabled them to do proper analysis and identify the abnormal hidden patterns.

In supplemental material to [15], the author compared the visualizations of t-SNE to those of IsoMap and Locally Linear Embedding on 6000 images from the MNIST dataset [39]. The MNIST dataset has 28x28 pixels-images of handwritten digits and proved that t-SNE generates a mapping that is most suitable for visual exploration of the underlying structure of the dataset. Fig. 2.6 and Fig. 2.7 show the mapping generated by IsoMap and Locally Linear embedding respectively which fail to clearly separate the images of the

للاستشارات المنارة
www.manaraa.com

same digits into separate clusters. On the other hand, in Fig. 2.8 t-SNE mapped the images of the digits in well-separated 10 clusters each representing images of same digit.

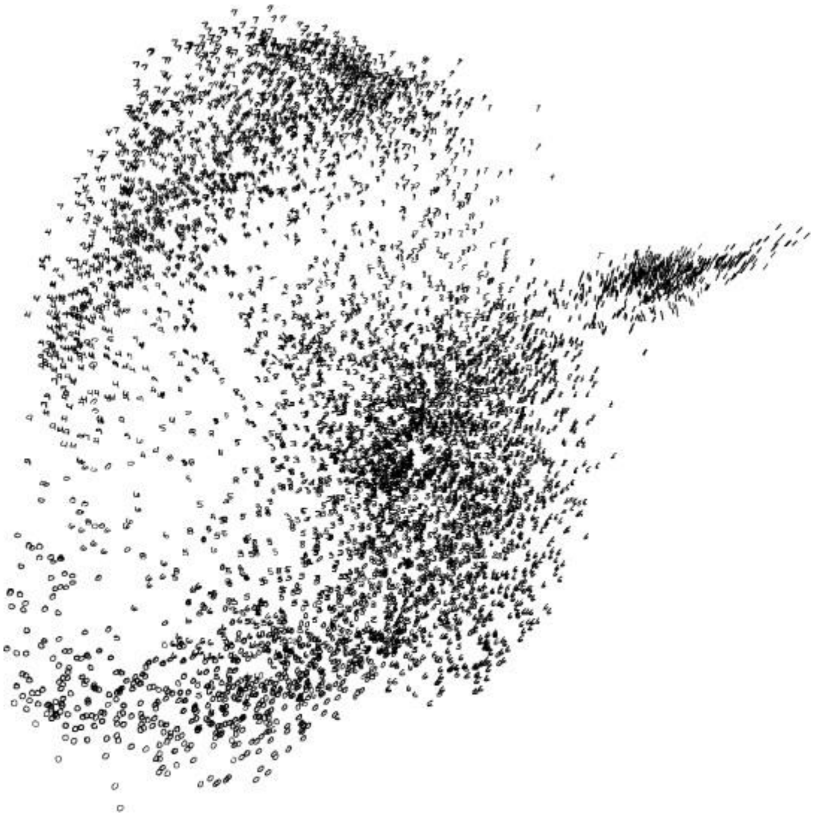

Fig. 2.6 Visualization of the MNIST dataset using IsoMap as presented in supplemental material to [15]

www.manaraa.com

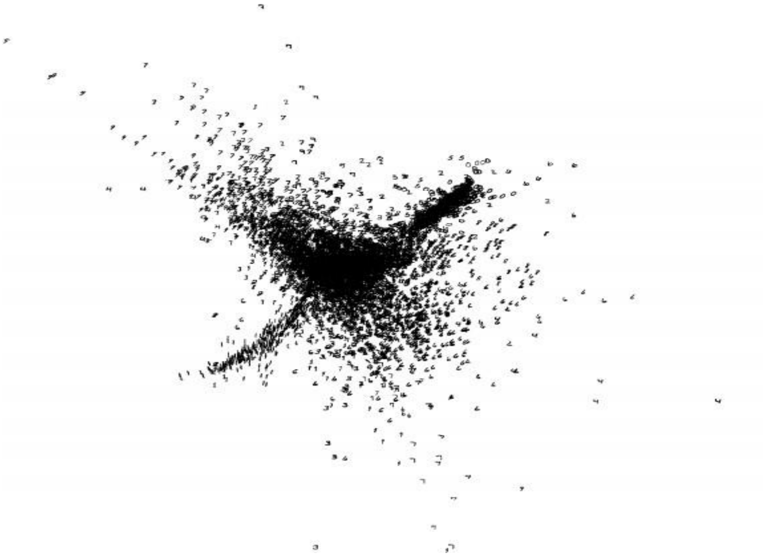

Fig. 2.7 Visualization of the MNIST dataset using LLE as presented in supplemental material to [15]

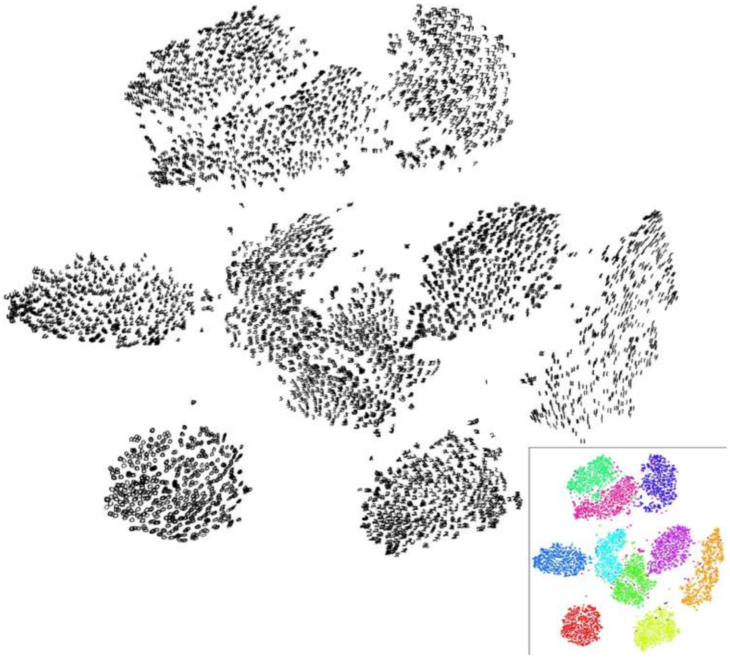

Fig. 2.8 Visualization of the MNIST dataset using tSNE as presented in supplemental material to [15]

المنارة للاستشارات
www.manaraa.com

t-SNE succeeded and outperformed other visualization techniques to provide visually appealing mapping of different datasets for a wide variety of applications. The fact that it preserves and uncovers the real structure of the dataset motivated us to test its applicability in solving the problem of predicting students' performance. We believe that students who perform similarly also possess similar characteristics and those who differ in their performance have different ones. Therefore, by using t-SNE, these clusters can be revealed and further studied for common characteristics.

المنارة للاستشارات
www.manaraa.com

CHAPTER (3): THEORETICAL BACKGROUND

The curse of dimensionality has long been known as a real impediment to the success of machine learning and data mining techniques. The term has first been coined by Richard Bellman in 1957 when considering problems in dynamic programming [40], [41]. Bellman argued that as dimensionality increases, the data becomes sparse. Accordingly, one would want to increase the number of observations in order to provide the machine learning algorithm with good space coverage and enable a satisfactory learning process. However, in the real world we are usually tied to a limited dataset size. This could be due to difficulty in collecting more data or limited time or processing resources to handle more data. Hence, dimensionality reduction became an inevitable step for an effective and efficient learning process.

Dimensionality reduction techniques are classified into two types: Feature selection [5], [6], [7], [8] and feature extraction [9], [10]. Both techniques use different approaches to reduce dimensionality while maximizing some objective function. Feature selection methods try to find a small subset of the original set of features while feature extraction techniques apply transformations on the original dataset to come up with a completely new set of features on lower dimensional space. While feature selection has the advantage of retaining meaningful features, being restricted to the same number of features, feature extraction techniques may guarantee smaller value of information loss. This can be due to the fact that it combines important information from several real features in smaller number of artificial features. In this work, we focus on feature extraction techniques and we refer to them merely as dimensionality reduction techniques for simplicity. Among the most popular feature extraction techniques are: PCA, LLE, IsoMap and t-SNE.

المنارة للاستشارات

www.manaraa.com

3.1. Principal Component Analysis (PCA)

Principal Component Analysis is one of the earliest dimensionality reduction techniques [42]. Yet, it remains one of the most popular and commonly used in several applications [43]. It provides linear projection of high dimensional data such that the variance of the projected data is maximized. Unfortunately, PCA fails to provide good-enough visualizations for datasets with non-linear manifolds[1].

PCA tries to project the data of n-dimensionality onto a lower sub-space taking into consideration the minimum loss of information by trying to find the principal components of the data set and maximizing the variance between data points such that each data point tells exactly how far it is from the main trend-lines of the data. The PCA tries to find new axes to plot the data in such a way that this transformation expresses the pattern between the data rather than mere plotting the data points in terms of the x and y axes (or more axes if the data is not in two dimensions). Doing so, it emphasizes variation and brings out the strong patterns in the dataset. However, the new projected co-ordinates do not actually mean something physical, they are combinations of the principal components of the data.

3.1.1. Summary of the approach [44]

Step 1: Subtract the mean from your data points: Subtract the mean of each dimension from each value for that dimension. If you have a two dimensional dataset, you would transform it such that x becomes $x-\bar{x}$ and y becomes $y-\bar{y}$ Where $\bar{x}$ is the mean of the x values for all the data points and $\bar{y}$ is the mean of the y values for all the data points. This produces a new dataset whose mean is Zero, *AdjustedData*.

Step 2: Calculate the covariance matrix

Step 3: Calculate the eigenvectors and eigenvalues of the covariance matrix

[1] A manifold is a topological space that locally resembles Euclidean space near each point. [59]

www.manaraa.com

Since the covariance matrix is a square matrix, we can calculate its eigenvectors and eigenvalues. Eigenvectors are particularly important because they inform us about the pattern in the data. Accordingly, by extracting the eigenvectors of our covariance matrix, we are also extracting the lines that characterize the data.

Step 4: Choose your eigen-vectors and compute your feature vector

Each eigenvector is accompanied with an eigenvalue that expresses its importance in terms of the amount of variance that it can capture. If we aim at dimensionality reduction, then we would only consider eigenvectors that have significant eigenvalues. This action results in information loss but if the eigenvalues of the discarded eigenvectors are really small, we guarantee the minimum loss of data. Therefore, we order the eigenvectors in terms of their eigenvalues highest to lowest and omit the eigenvectors with the least significance. The matrix of chosen eigenvectors forms our new feature vector.

$$FeatureVector = (eig_1,eig_2,...,eig_n) \tag{1}$$

Step 5: Calculating the new dataset

Your final data is the multiplication of the transposed feature vector by the transposed adjusted data.

$$FinalData = FeatureV\ ector^T * AdjustedData^T \tag{2}$$

3.2. **IsoMap**

IsoMap tries to overcome the challenge of non-linear dimensionality reduction illustrated by [45] using the “Swiss Roll” synthetic dataset. Fig. 3.1 shows the problem as demonstrated. IsoMap tries to preserve structure by using the geodesic distance to measure pair-wise similarity rather than the Euclidean distance.

المنارة للاستشارات

www.manaraa.com

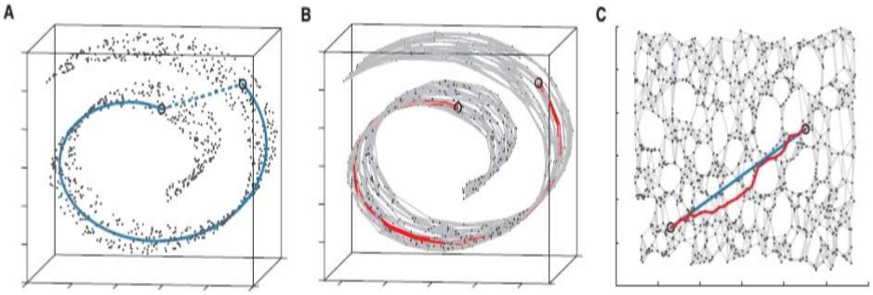

Fig. 3.1 The Swiss Roll dataset

From *[45]*. Reprinted with permission from AAAS.

The diagram illustrates how IsoMap utilizes geodesic distance to find the distance between two points on a non-linear manifold. (A) The dashed line representing the Euclidean distance between two arbitrarily chosen points (circled in blue) may not reflect their intrinsic similarity as measured by the geodesic distance on the lower dimension 2D space denoted by the solid blue line. (B) The red line segments show an approximation of the geodesic distance between the two points as achieved by traversing a constructed neighborhood graph between the two points (C) The two dimensional embedding recovered by the IsoMap preserves the shortest path distances in the neighborhood graph (overlaid). Now, straight lines in the embedding (blue) represent simpler and cleaner approximations to the true geodesic distance between the two points than do the corresponding graph paths (red).

3.2.1. Summary of the Approach

IsoMap has three main steps:

Step 1: Construct a neighborhood graph G between all data-points such that i and j are only connected if i is one of the nearest neighbors to j. The edge between i and j is set to the value of d_{ij}.

Step 2: Compute a matrix M of the shortest path between each pair of points in G.

Step 3: Apply classical multi-dimensional scaling MDS to the matrix of graph distances to construct an embedding of the data in a d-dimensional Euclidean space that best approximates the manifold's intrinsic geometry.

However, IsoMap will fail if the data points are not uniformly distributed or the dataset is of low-density. It is also known to fail if the structure of the dataset has holes [15].

المنارة للاستشارات
www.manaraa.com

3.3. **Locally Linear Embedding (LLE)**

Locally linear embedding is another dimensionality reduction technique that addressed the problem of non-linearity [46] approximately at the same time as the IsoMap. The technique assumes that the data is well sampled such that it preserves the structure of its underlying manifold. Accordingly, it expects that each data point and its k-nearest neighbors lie on or close to a linear patch of the manifold. It tries to arrange the data in a lower-dimensional map such that the relation between the point and its neighboring reconstructions is maintained, setting the weight of points that do not fall into the set of neighbors to 0. The first problem with this technique is how can one choose the value of the k nearest neighbors. Another problem is in the main assumption that the data is well sampled and sufficient so that the nearest neighbors actually fall close to each other. The algorithm's biggest failure mode is that it can map faraway inputs to nearby outputs when the manifold is under-sampled thus failing to maintain the global structure of the data.

3.4. **t-Stochastic Neighbor Embedding (t-SNE)**

t-SNE was first introduced in 2008 [15] as a data visualization technique that visualizes high-dimensional data by giving each data-point a location in a two or three dimensional map. t-SNE overcomes the overcrowding problem that its ancestor SNE suffered from. The main power of t-SNE comes from the fact that it is capable of capturing the local structure of the high-dimensional data very well, while maintaining the global structure such as the presence of clusters.

t-SNE tries to conserve the structure of the data by guaranteeing that if two points are close together in the higher dimension, they remain as close in the lower dimension. That is, if x_i is a data-point in the original D-dimensional data space denoted by R^D and we are

المنارة للاستشارات
www.manaraa.com

interested in finding a proper mapping for our data-point y_i in a two-dimensional space R^2, such that if x_i and x_j were close to each-others in the high dimensional space, y_i and y_j remain close in the lower dimensional space.

t-SNE measures how close x_j is from x_i by centering a Gaussian distribution around x_i with a given variance and normalizes it over all distances between x_i and other points falling under that Gaussian. The variance differs for every point such that points in dense areas are given smaller variance than that given for points in sparse area. Then, it assumes that the conditional probability of picking the point x_j given that x_i was picked is proportional to their similarity or in other words, inversely proportional to the distance between the two points denoted by $|x_i - x_j|$.

$$p_{j|i} = \frac{e^{(\frac{|x_i - x_j|^2}{2\delta_i{}^2})}}{\sum_{k\neq i} e^{(\frac{|x_i - x_j|^2}{2\;\delta_i{}^2})}} \qquad (3)$$

Then a joint distribution is calculated as the average of $p_{j|i}$ and $p_{i|j}$. This symmetrizes the joint probability p_{ij} and p_{ji} which guarantees that each data point makes a significant contribution to the cost function even if x_i is an outlier.

$$p_{ij} = \frac{(p_{j|i} + p_{i|j})}{2N} \qquad (4)$$

where N is the number of data points

For the lower dimensional mapping, t-SNE starts with random assignment of the data points on a lower dimensional map, hence the “Stochastic” part of the technique’s name. The distance between any two points on the lower dimensional space is calculated in the same way with just one difference that we consider the Student-t distribution with 1 degrees of freedom rather than normal distribution to make use of its heavier tails property.

للاستشارات المنارة

www.manaraa.com

$$q_{ij} = \frac{(1+|y_i - y_j|^2)^{-1}}{\sum_{k \neq m}(1+|y_k - y_m|^2)^{-1}} \qquad (5)$$

This proved to help t-SNE alleviate the crowding problem suffered by its ancestor SNE.

t-SNE tried to find a lower dimensional embedding with a probability distribution Q such that the distance between the probability distribution in the high dimensional space P and Q is minimized by minimizing the Kullback-Leiber divergence between the two probability distributions:

$$KL\ (P\ \approx Q) = \sum_{i \neq j} p_{ij}\ log\ (\frac{p_{ij}}{q_{ij}}) \qquad (6)$$

المنارة للاستشارات
www.manaraa.com

CHAPTER (4): PROPOSED SOLUTION

Our research can be categorized under data mining applications. Therefore, we followed the standard process "**CR**oss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining – **CRISP-DM**" reference model [47] to guide our research efforts. In this section, we briefly introduce our solution to the problem of students' performance prediction at AUC in light of the CRISP-DM Model.

4.1. **Phase 1 - Business Understanding**

This initial phase focuses on understanding of the business objective and formulating the data mining problem statement. This phase was fast-forwarded due to the author's professional experience in the Data Analytics and Institutional Research Department at AUC[2]. Universities are concerned about student success and providing premium support services to guide their students through their academic journey. In [48], Cuseo et al. listed student academic achievement as one of the key indicators for student success and outlined 12 potent principles of effective student support program delivery highlighting the need for an institution to be proactive taking early, preventative actions to address students' problems in an anticipatory fashion. AUC, being a no-exception, aims at guiding academically struggling students and helping them enhance their performance. This should automatically reflect on its retention and graduation rates. Our literature review covered in Chapter *2* further familiarized us with the problem and the research conducted

[2] The Office of Data Analytics and Institutional Research at AUC offers a variety of services and products to support AUC academic and administrative units, as well as outside constituencies, with planning, assessment, accreditation, research and data analysis needs.

www.manaraa.com

on how to solve it. Finally, we came up with our business problem statement and the Data mining problem definition discussed in Chapter *1* and we further elaborate on it below.

4.1.1. Business Problem Statement

We need to provide for a solution that can help the AUC administration identify students whose GPAs are more likely to fall below 2.0 in their third year of studies at AUC. The prediction will be based on two-years performance in order to leave an ample of time (1 year) for employing strategies of interventions and corrective actions to help those students enhance their performance.

4.1.2. Data Mining Problem Statement

Given a dataset of historical data on students' two-years performance at AUC, we want to classify the students based on their expected third-year GPA into seven groups; starting with the lowest achievers (GPA below 2.0) and ending with the highest achievers (GPA is 3.8 or more).

4.2. Phase 2 - Data Understanding

The second phase of the CRISP-DM Model is to provide a thorough understanding of the students' data. Initially, we started with researching what features would be relevant to our model then we proceeded to the data collection and a set of tasks that helped us get familiar with the data, identify data quality problems and discover first insights into the data.

4.2.1. Identifying Relevant Data Elements for Collection

This phase required research on our end on similar models and studies conducted in the higher education industry on factors affecting student performance. A commissioned report

المنارة للاستشارات
www.manaraa.com

for the National Symposium on Post-Secondary Student Success [49] divided factors affecting student success into:

- Student background characteristics (Gender, Ethnicity, Family Background, educational aspirations, Native Vs. International students, first generation students, Transfer students, Athletes) and pre-college experiences (Academic Intensity in High School)
- Availability of Financial Aid
- Student Engagement (Enrollment in Extra-Curricular Activities, Students living on campus, students' interaction with their peers, students' interaction with faculty, Students' perception on how supportive is their college of their academic and social needs)
- Student Satisfaction
- Structural and Organizational Characteristics of the institution and the support services provided.

A research conducted in University of Maryland [50] presented a model for student success that focuses on how campuses should integrate data across systems. They focused on integrating data elements from the Learning Management System (LMS), Student Information system and Extracurricular activities into one predictive model for student success.

Northern Arizona University [51] prepared a predictive model of student performance in gateway courses and used the following set of data elements as input to the model:

- Age, gender and ethnicity

المنارة للاستشارات
www.manaraa.com

- Cumulative earned hours and GPA
- Current semester enrolled credits and GPA
- High school GPA and rank
- ACT and SAT scores
- Socioeconomic status

We tried to map these data elements to data collected at AUC, extracted the data from the main student information system database, Banner®[52]. Data tables such as academic history, enrollment and admissions were joined to create one unique record per student. Table 4.1 to Table 4.6 show a brief description of the main data elements collected.

Table 4.1 AUC Dataset: Demographic Attributes

	Attribute	**Attribute Description**
Demographics	GEN	Student gender (M: Male or F: Female)
	NATN	Student's birth nation (EG: Egyptians, AR: Arabs, US: Americans, OT: Others)
	LEGACY	Student's parents or siblings attended AUC (M: Mother, F: Father, S: Sibling, T:Staff, N:Not Available)

Table 4.2 AUC Dataset: Admissions Attributes

	Attribute	**Attribute Description**
Admissions	Prior School	Student's high school name (24 unique values and OT: Others)
	HS.DPLM	Student's high school certificate (AB: Abitur, IG: IGCSE/GCSE, IB: International Baccalaureate, HS: American Diploma, TS: National Certificate - Science Track, TA: National Certificate - Arts Track, FA: French Baccalaureate, OT: Others)

www.manaraa.com

	HS.Score	Numerical value for student's score in high school. Scale differs per certificate.
	ENGL.PLACE	Student's English Placement based on standardized exams (Numerical)
	FRST.CNTCT	First contact between student and AUC (WLK: Walk-in, WEB: Web Inquiry, SVA: School Visit to AUC, HSV: High School Visit, WBA: Web Application, MER: Merit Scholarship, LED: LEAD program)
	ADMS.MAJR	The major the student was admitted to (UNDU: Undeclared, AENG: Architectural Engineering, CENG: Construction Engineering, MENG: Mechanical Engineering, EENG: Electrical Engineering, PENG: Petroleum Engineering, CSCI: Computer Science, CSCE: Computer Engineering, BIOL: Biology)

Table 4.3 AUC Dataset: General Academic Attributes

	Attribute	**Attribute Description**
General Academic Attributes	UNDU.SEMS	Number of semesters before declaring a major
	PROB.SEMS	Number of semesters placed on probation
	INACTIV.SEMS	Number of semesters a student was inactive
	CHNG.MAJR	Number of times a student changed his major in his first two years
	CRSES.FAILED	Number of courses failed in the first two years
	CRSES.WITHDREW	Number of courses dropped or withdrew from in the first two years
	CRSES.INCOMP	Number of incomplete courses on the academic history of the student
	SCI.120	Student grade in the Scientific and Critical Thinking course

www.manaraa.com

Table 4.4 AUC Dataset: Student Engagement Indicators

	Attribute	**Attribute Description**
Student Engagement Indicators	WRK.STUDY	Indicator of enrolling in a work-study program
	ACTIVITIES	Indicator of enrolling in extracurricular activities in any of his first 4 semesters
	HOUSING	Indicator of whether the student is living in the AUC dorms or not
	SPRTS.SCH	Indicator of being on a Sports scholarship in any of his first 4 semesters
	CULT.SCH	Indicator of being on Cultural scholarship in any of his first 4 semesters

Table 4.5 AUC Dataset: Financial Assistance Attributes

	Attribute	**Attribute Description**
Financial Assistance	FIN.AID	Indicator of receiving financial aid in any of his first 4 semesters
	ACHIEV.SCH	Indicator of receiving achievement scholarship in any of his first 4 semesters

Table 4.6 AUC Dataset: Academic Information

	Attribute	**Attribute Description**
First Year Academic Information	FRST.YR.ERNED.HR	Total number of hours earned in the first year (Numeric)
	FRST.YR.GPA	Student’s GPA in his/her first year
Second Year Academic Information	SCND.YR.ERNED.HR	Total number of hours earned in the second year (Numeric)
	SCND.YR.GPA	Student’s GPA in his/her second year
Third Year Academic Information	THRD.YR.GPA	Student’s GPA in his/her third year

المنارة للاستشارات

www.manaraa.com

4.2.2. Data Quality Assessment and Challenges

We assessed our data based on two factors: Comprehensiveness to the model and Quality of the existent data. Comprehensiveness is whether the data covers all aspects and factors contributing to student success and that was subject to our accessibility to the data and whether AUC collects this data in the first place.

- Shortage on data related to students' perspectives and satisfaction as at the time of this study, AUC conducted surveys in anonymous fashion which hindered linking students' perspectives to their academic and demographic facts.

- Difficulty in acquiring data related to students' memberships in extra-curricular activities and sports teams. We tried to approximate this data to transactions on students' requests of extra-curricular activities transcript, however, this did not prove significant as the numbers were very low and we know that a much larger population of AUC students engage in extra-curricular activities. As for the sports information, we included indicators for students on sports scholarships. That was not an indicator of students' engagement, rather another indicator of receiving financial assistance. We hope that revisiting the model when this data is made available may result in enhancing its prediction power and uncover relationships between student engagement and academic performance.

- Difficulty in collecting data about family socioeconomic status and educational background. We hope that this data starts to be collected in the admissions application for future enhancement.
- We faced difficulty securing access to the data collected through the Learning Management System. AUC is taking its first steps in building a full-stack Business Intelligence System which will provide a 360° view of the student and will make this data readily available and accessible for future enhancements.

المنارة للاستشارات
www.manaraa.com

For data quality assessment, we believe the collected data is in good shape. Few records had some missing values and we preferred completely excluding them from our dataset over using imputation techniques. The reason for that is that we wanted to avoid adding to the uncertainty of the model by imputing the missing values.

4.3. **Phase 3 – Data Preparation**

Before feeding the data into the model, data cleansing and transformation are inevitable in order to put the data in its final format that is more suitable for the data model. Below are some examples of the data cleansing processes undertaken:

- Data Aggregations: We combined different data values into groups based on our business understanding of the problem. Example: We aggregated 47 nationalities into four values only (EG: Egyptians, AR: Arabs, US: Americans and OT: Others). We also grouped 28 high school certificates into seven main certificates (AB: Abitur, IG: IGCSE/GCSE, HS: American Diploma, FA: French Baccalaureate, IB: International Baccalaureate, TS: *Thannaweya Amma* - Science Track and TA: Thannaweya Amma - Arts Track) and the rest of the certificates into “Others”
- Categorical Attributes: A categorical attribute of n values was transformed into $n-1$ boolean dummy variables using the “Hot-one Encoding” technique.

- Numerical Attributes: Highly skewed numerical attributes such as GPA were transformed using the BoxCox Transformation function [53] in attempts to enforce data normality. The labeling class THRD.YR.GPA was converted from a continuous numerical value into a categorical variable based on business rules.

- Constructed Attributes: We calculated new variables to reflect important student performance indicators not readily available in the database such as number of

المنارة للاستشارات
www.manaraa.com

semesters before declaration, number of semesters on probation, number of courses failed, number of incomplete courses, ...etc

Missing Values: We took the decision to exclude missing values because we had a fairly large dataset with few records having missing values. We preferred to remove the records over using imputation methods to find estimates to the missing values which would add uncertainty to our model.

4.4. **Phase 4 – Modeling**

We wanted to compare the performance of different algorithms that belong different families of machine learning techniques that are well suited to solve our classification problem so we chose:

- An instance-based classifier (The lazy KNN – K-Nearest Neighbor)
- A Decision Tree algorithm (RPART – Recursive Partitioning for classification and regression trees)
- An ensemble classifier (RF – Random Forest)
- A neural network classifier (NN - feed-forward neural network with a single hidden layer)
- A kernel based support vector machine (SVM - Radial Basis Kernel)

For the UCI dataset -benchmarking set of experiments, we sticked to the ones that were used by the original authors from the list above.

We picked R for implementation because it is one of the top Analytics and Data Science tools, as per the KDnuggets software 2016 Poll [54]. We used the R package "caret" (short for classification and regression training) which contains functions to streamline the model

المنارة للاستشارات

www.manaraa.com

training process for complex regression and classification problems [55], [56] and the "Rtsne" package for t-SNE implementation.

We used the repeated cross-validation resampling technique for training and evaluating the model. The rationale behind this decision is to keep as much data for training and yet evaluate the model performance on data other than that used for training. We employ 10-fold cross-validation with 5 repeats. In other words, we split the data into 10 subsamples, train the model on 90% of the data and use the model to predict the remaining 10%. The average of the 10 model evaluation metrics becomes our estimate of the actual goodness of the model. We repeat this 5 times and average the results.

We utilized the "caret" package to evaluate, using resampling, the effect of different model parameters on the model performance and choose the optimal model across those parameters.

Fig. 4.1 shows a flowchart diagram of our proposed solution, starting with the data preparation phase, dimensionality reduction and splitting of the data into training and testing sets in a repeated cross-validation fashion, followed by the classification step which provides us with the predicted GPA interval for each student. Chapter 5 discusses our conducted experiments in details.

4.5. **Phase 5 – Model Evaluation**

In order to evaluate our model, it was necessary to test our proposed solution against previously reported benchmarking results. Therefore, we used the Portuguese Student Performance dataset publicly available on the UCI-Repository and compared our prediction accuracy to that reported by [57] in Chapter 5. The prediction accuracy achieved

المنارة للاستشارات
www.manaraa.com

on this benchmarking dataset was also useful in informing us about the industry average for similar prediction problems. We also compared using our prediction performance using the t-SNE mappings to that achieved on the full AUC dataset without dimensionality reduction. Finally, we conducted a set of experiments that compare the visualization achieved by PCA and t-SNE and the classification results on the lower dimensional mapping.

For the model performance evaluation, we used the two metrics: Accuracy (Percentage of correct classification) and Kappa.

$$Accuracy = \frac{Total\ Number\ of\ Correct\ Classifications}{Total\ Number\ of\ Cases} \tag{7}$$

$$Kappa = \frac{p_{\circ} - p_e}{1 - p_e} \tag{8}$$

where $p_{\circ}$ is the relative observed agreement among raters and p_e is the hypothetical probability that the agreement is due to chance.

Since we employed repeated cross-validation resampling, we report the average of both metrics and the relevant 95% confidence interval. We used Kappa because it is another single-value metric that takes into account the possibility of correct classification happening by chance. We provide a discussion on the results of our experiments in Chapter 5.

Also, as part of this phase, we identified other student attributes that can enhance performance prediction and summarized them in Chapter 6: Conclusion and Future Work.

المنارة للاستشارات
www.manaraa.com

4.6. **Deployment**

List of students identified to be at-risk (GPA less than 2.0) should be reported every semester to the responsible departments in order to start a persistent communication process with the students. This may include special advising and career mentoring sessions. During these sessions, students may be advised to change majors to ones that are more suitable to their interests and academic abilities. Counseling sessions may help uncover psychological or physical challenges and enable suitable and timely interventions. We also aim at integrating this model into AUC's 2017 Business Intelligence Project. This will significantly speed up the data preparation phase and will ensure systematic data transformations every semester.

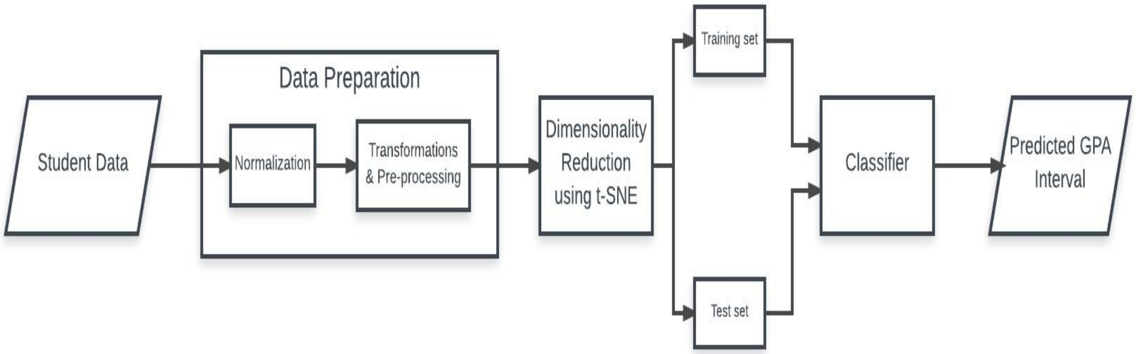

Fig. 4.1 Flowchart for our proposed solution showing t-SNE as a pre-classification step to reduce the dataset dimensionality.

المنارة للاستشارات
www.manaraa.com

CHAPTER (5): EXPERIMENTS

We conducted three sets of experiments for different purposes. The first set of experiments test the effectiveness of t-SNE as a dimensionality reduction technique against a benchmarking dataset, the second set explores t-SNE's clustering and visualization abilities on AUC dataset and the third set focuses on comparing the classification results using t-SNE and PCA as dimensionality reduction techniques on AUC dataset. Details of each dataset is provided before the set of relevant experiments.

5.1. Benchmark Dataset: UCI-Predicting Portuguese Secondary School Student Performance [57]

5.1.1. Dataset

The dataset is for real-world data from two Portuguese secondary schools. The dataset contains students' academic information (eg. grades and number of absences) as well as demographical and social information (eg. student's age, alcohol consumption, mother's education). Table 5.1 shows the dataset attributes in details. Cortez et al. [57] worked on two different datasets; one for predicting student performance in Mathematics course and another for Portuguese language course. In our work, we focus on the latter dataset. They also had three variations of the dataset; each treating the class label attribute differently. They conducted a set of regression experiments where the target variable: "Portuguese grade" is a continuous variable. They also conducted two sets of classification experiments where the target variable was set to boolean (Pass/Fail) or five-level classification. In the five-level classification experiments, the continuous grade attribute (target attribute) was binned into five bins 0-9, 10-11, 12-13, 14-15, 16-20. The dataset

المنارة للاستشارات
www.manaraa.com

contained 649 records with 32 predictors divided into 3 nominal, 12 binary and 17 numerical variables and a single nominal class label. In our work, we focus on one variation of their dataset (Five-level classification of the final Portuguese grade) and compare our solution to theirs.

Table 5.1 - Attributes of UCI Portuguese secondary school students' performance [57]

Attribute	Description
Sex	Student's sex (binary: female or male)
Age	Student's age (numeric: from 15 to 22)
School	Student's school (binary: Gabriel Pereira or Mouzinho da Silveina)
Address	Student's home address type: (binary: urban or rural)
Pstatus	Parent's cohabitation status: (binary: living together or apart)
Medu	Mother's education (numeric: from 0 to 4)
Mjob	Mother's job (nominal)
Fedu	Father's education (numeric: from 0 to 4)
Fjob	Father's job (nominal)
Guardian	Student's guardian (nominal: mother, father or other)
Famsize	Family siza (binary: <= 3 or > 3)
Famrel	Quality of family relationships (numeric: from 1- very bad to 5 – excellent)
Reason	Reason to choose this school (nominal: close to home, school reputation, course preference or other)
Traveltime	House to school travel time (numeric: 1- <15min, 2 – 15 to 30 min, 3 – 30 min to 1 hour or 4 - > 1 hour).
Studytime	Weekly study time (numeric: 1 - < 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 - > 10 hours)
Failures	Number of past class failures
Schoolsup	Extra education school support (binary: yes or no)
Famsup	Family educational school support (binary: yes or no)
Activities	Extra-curricular activities (binary: yes or no)
Paidclass	Extra paid classes (binary: yes or no)
Internet	Internet access at house (binary: yes or no)
Nursery	Attended nursery school (binary: yes or no)
Higher	Wants to take higher education (binary: yes or no)
Romantic	With a romantic relationship (binary: yes or no)
Freetime	Free time after school (numeric: from 1 – very low to 5 – very high)
Goout	Going out with friends (numeric: from 1 – very low to 5 – very high)
Walc	Weekend alcohol consumption (numeric: from 1 – very low to 5 –very high)
Dalc	Weekday alcohol consumption (numeric: from 1 – very low to 5 –very high)
Health	Current health status (numeric: from 1 – very bad to 5 – very good)
Absences	Number of school absences (numeric: from 0 to 93)
G1	First period grade (numeric: from 0 to 20)
G2	Second period grade (numeric: from 0 to 20)
G3	Final grade (numeric: from 0 to 20)

5.1.2. Data Preprocessing

We preprocessed the dataset as a preparation step for t-SNE. We created dummy variables [54] to replace the nominal variables such that for each nominal attribute with *n*

المنارة للاستشارات
www.manaraa.com

values, we create $n - 1$ binary dummy variables. For example, Medu refering to mother's education takes 5 different values, therefore we replaced it by 4 dummy variables: health, teaching, civil and home where the value for the variable health is 1 only if Medu = "health". If the 4 dummy variables are set to zero, this would mean that Medu was equal to its last value, "Other". Numerical values were also standardized by subtracting the mean and dividing by the standard deviation.

We used the Gower distance metric [58] which uses a different metric for each type of variables when calculating the distance between the data points in the high dimensional space. The greatest added value comes from the fact that Gower distance uses the dice coefficient metric for calculating the distance for nominal variables. It takes into consideration that if two data points do not have a specific property then that should not count as a similarity. For example, if the variable [Medu = "health"] for two students is equal to 0, then that does not indicate any similarity between them. We also assigned different weights for the variables based on the Information gained by each. Finally, we calculated a similarity matrix for the data points.

5.1.3. Dimensionality Reduction using t-SNE

We used the R package "Rtsne" to calculate the t-SNE mapping using the calculated similarity matrix then we plotted the results in 2D and 3D. The mapping was done in an unsupervised learning mode where we completely hid the final grade label during the training and only used it to color the data points on the graphs. Fig. 5.1 shows the two-dimensional mapping obtained which we believe that it has

المنارة للاستشارات
www.manaraa.com

succeeded in finding a lower dimensional mapping that uncovers the underlying data structure because it has the following two properties:

- Students belonging to the same cluster are mapped close to each-others

- Semantically distant clusters are mapped far from each-others; (for example: the highest achievers (blue) are the farthest from the lowest achievers (red)).

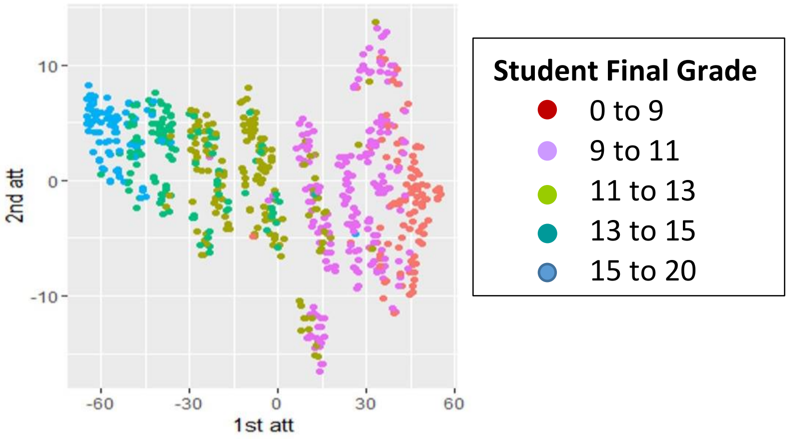

Fig. 5.1 2D Mapping of the UCI Students' Performance Dataset, the final grade was only used for coloring the data-points.

المنارة للاستشارات
www.manaraa.com

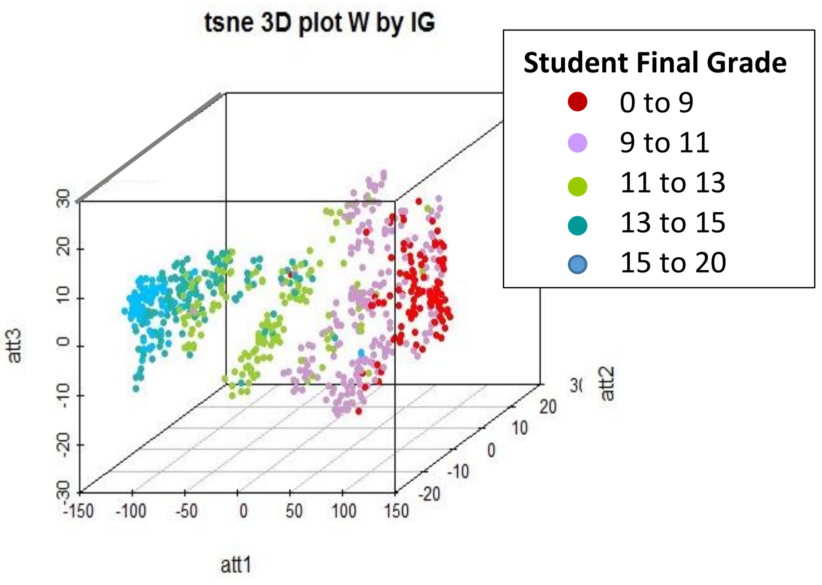

Fig. 5.2 3D Mapping of the UCI Students' Performance Dataset, the final grade was only used for coloring the data points.

Table 5.2 Comparison of the classification accuracy on the UCI dataset

	NN	RF	RPART	KNN, k = 11
t-SNE 2D Mapping	74% ± 1.5	71% ± 1.5	58% ±1.1	73% ± 1.4
t-SNE 3D Mapping	**77%** ± 1.3	73% ± 1.4	71% ± 1.2	**74%**± 1.4
Previously Reported [57]	65% ±0.9	74% ±0.2	76%±0.1	-

Similarly, Fig. 5.2 shows the three-dimensional mapping obtained which successfully maintains the same properties of the 2D mapping.

5.1.4. Comparison of Classification Results

Visually exploring the data, we were confident that if one data point was uncolored, we will be able to guess to which group it should belong based on its nearest neighbors. We were then curious to know if this would also inform a classifier about the structure of the data and hence, improve its performance. Accordingly, we trained different types of classifiers on the 2D and 3D mapping and assessed their performance with respect to the

المنارة للاستشارات
www.manaraa.com

best achieved and reported on the full-dataset by Cortez et al. [57], namely Neural Network, Random Forest and Decision Tree. Table 5.2 shows percentage of correct classifications (PCC) of the aforementioned classifiers when applied to t-SNE 2D mapping and t-SNE 3D mapping in comparison to the originally reported values on the full dataset. K-nearest neighbor (KNN) which was not considered by the original authors is considered here to represent simple classifier. We applied KNN to the full dataset and obtained an average of 40% for the percentage of correct classifications achieved on the full dataset.

5.1.5. Results Discussion

The neural network witnessed a performance improvement from 65% on the full dataset to 74% on the 2D mapping and 77% on the 3D mapping to beat the 76% which is the best achieved using the decision tree on the full dataset. Cortez et al. explained the low performance on the full-dataset dataset as a result of the high number of irrelevant attributes. This emphasizes the role of t-SNE as an effective dimensionality reduction technique. For the decision trees algorithms, there had been a decrease in the performance which can be explained as a normal information loss resulting from dimensionality reduction. The significant enhancement to the KNN classifier can be justified by the fact that the KNN is sensitive to data dimensionality when calculating distances between data points, and that t-SNE's lower dimensional mapping uncovered the local structure which made it easier to find the true nearest neighbors.

المنارة للاستشارات
www.manaraa.com

5.2. AUC Dataset (All Attributes) – Exploring visualization capabilities of t-SNE

5.2.1. Dataset

The dataset was prepared by joining different tables from the main students' information system at AUC. The dataset contains historical data for students' (3141 records) whose first semester was in between Spring 2011 and Spring 2014, excluding any previously attended semesters in the preparatory Intensive English Language Program. For example, we used information about students' first attending AUC in Spring 2011 to predict their performance as of the end of their third year, Spring 2013. The reason we could not include students whose first semester is later than Spring 2014 is that at the time of our study, we only had information about the students' grades in Spring 2016 counting as the third year from Spring 2014. For more details on the dataset and its attributes, please refer to Chapter 4.

5.2.2. Dimensionality Reduction using t-SNE

The dataset was preprocessed to handle categorical variables and remove near-zero variance attributes then it was mapped on 2D and 3D using t-SNE in an unsupervised learning mode where we completely hid the Third Year GPA label during the training and only used it to color the data points on the graphs. Fig. 5.3 shows the 2D mapping obtained. Once again, t-SNE succeeded in finding a lower dimensional mapping that uncovers the underlying data structure. The mapping maintained the same properties of the mapping obtained for the UCI dataset.

- Students were separated into three major clusters: Students maintained or gained achievement scholarship on one side and students who were never awarded an achievement scholarship on the other side. Students who lost their achievement

المنارة للاستشارات
www.manaraa.com

scholarship are very few and mapped close to those who were never awarded indicating their low performance.

- Students belonging to the same cluster are mapped close to each-others

- Semantically distant clusters are mapped far from each-others; (for example: the highest achievers (denoted by blue and purple points) are the farthest from the lowest achievers (denoted by red points)).

Similarly, Fig. 5.4 shows the three-dimensional mapping obtained which successfully maintains the same properties of the 2D mapping.

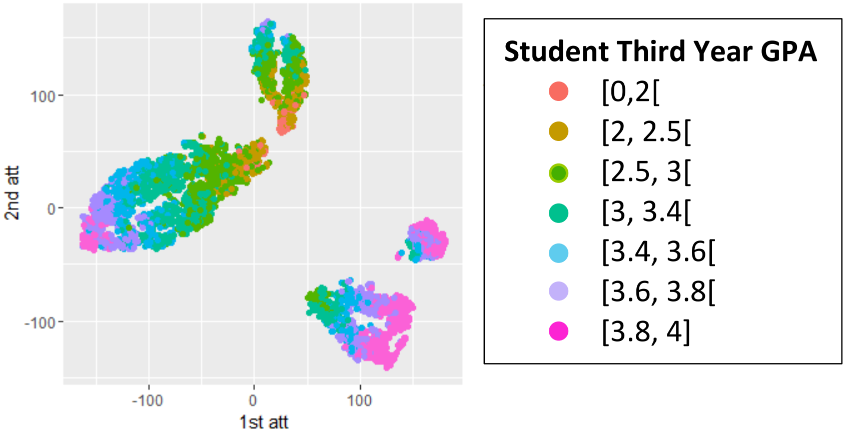

Fig. 5.3 t-SNE 2D Mapping of AUC student performance dataset

المنارة للاستشارات
www.manaraa.com

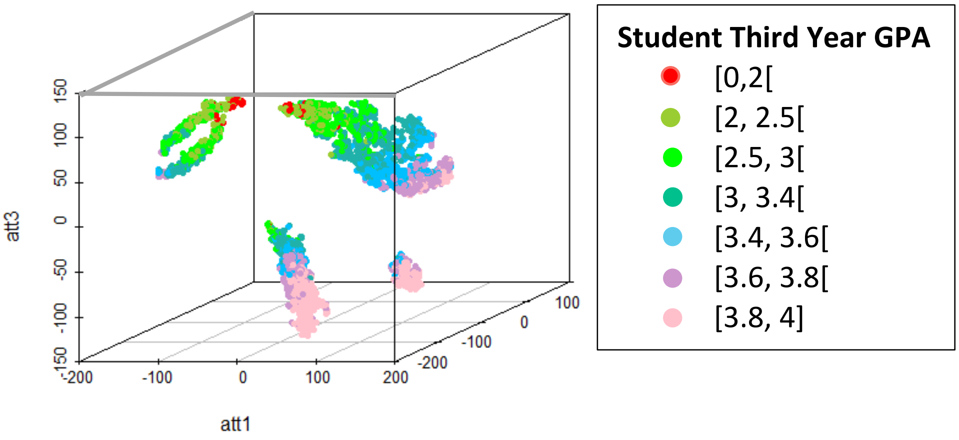

Fig. 5.4 t-SNE 3D mapping of AUC student performance dataset

المنارة للاستشارات
www.manaraa.com

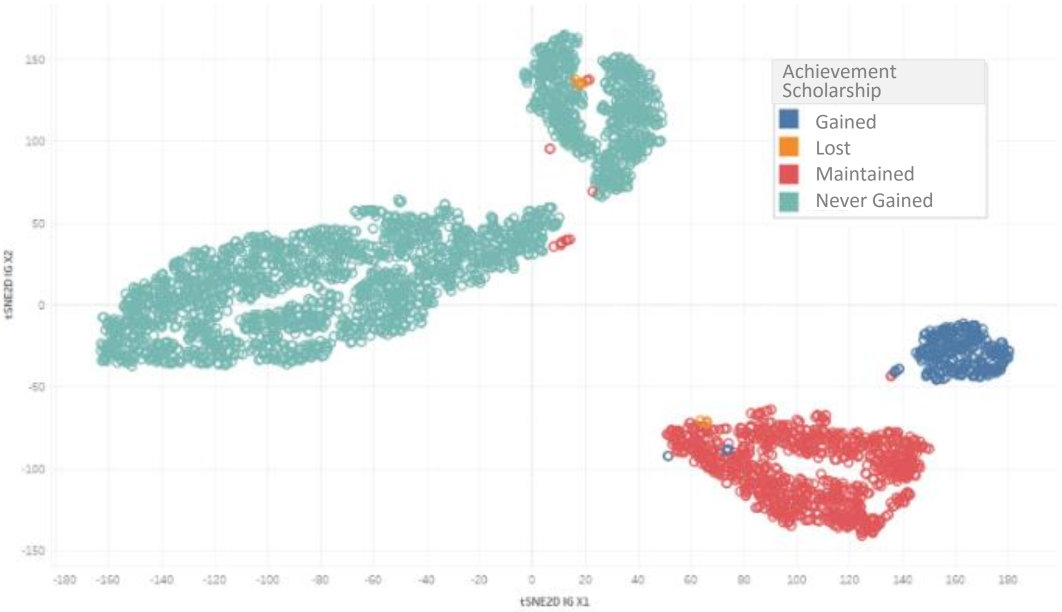

Fig. 5.5 t-SNE 2D mapping of AUC dataset. Achievement scholarship status used for color coding.

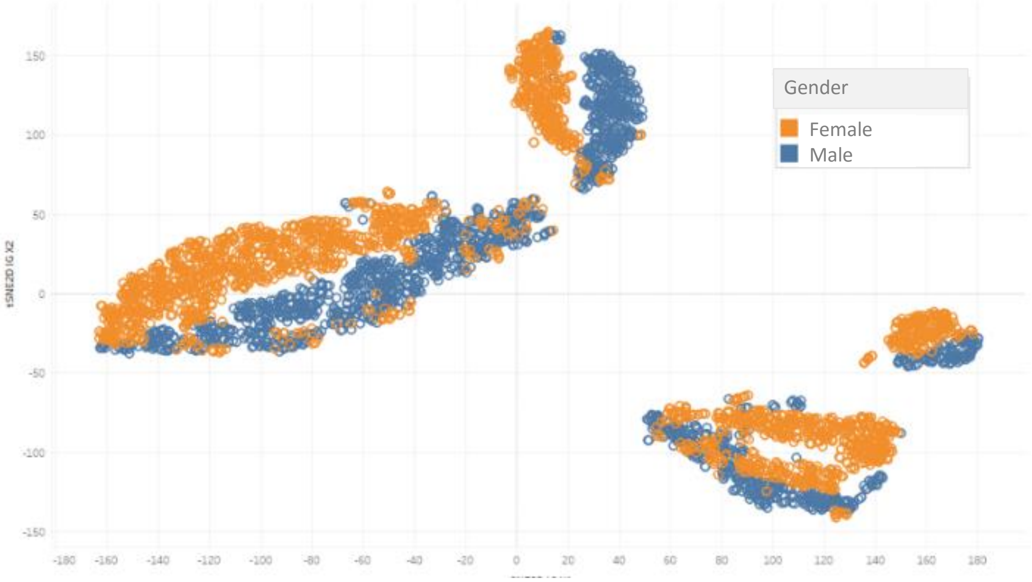

Fig. 5.6 t-SNE 2D mapping of AUC dataset. Gender used for color coding

www.manaraa.com

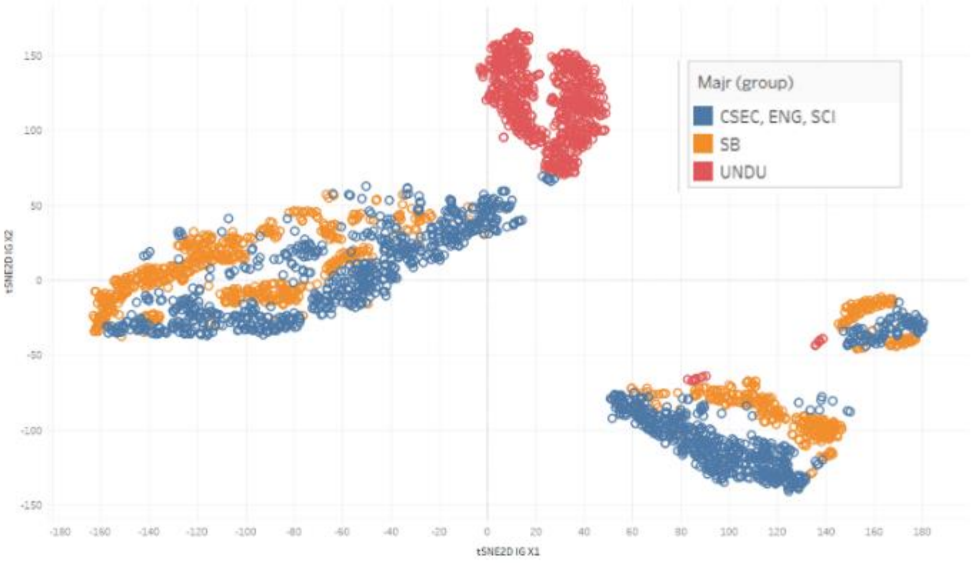

Fig. 5.7 t-SNE 2D mapping of AUC dataset. Students' majors used for color coding. Sciences and Engineering in blue, Business in orange and Undecided in red.

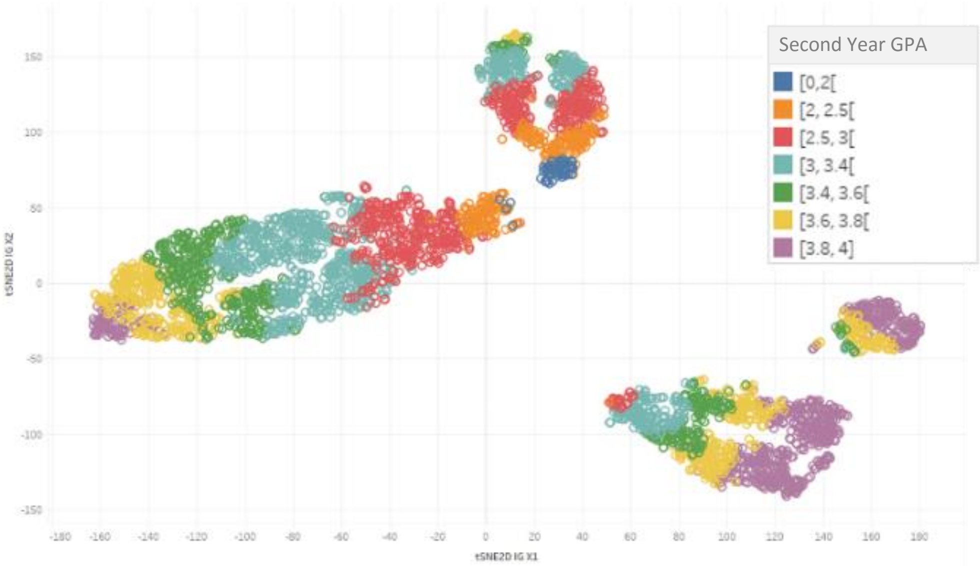

Fig. 5.8 t-SNE 2D mapping of AUC dataset. Second Year GPA used for color coding

We conducted further visual exploration of t-SNE mappings to understand the arrangement of the data points. In every plot, we used a different feature for color coding

www.manaraa.com

in order to visually check if this feature contributed to the clustering of the data points. We used Tableau for this exercise for fast and easy visual exploration.

Fig. 5.5 shows that t-SNE used the status of students' academic achievement scholarship to separate students into four main clusters; students who never received an achievement scholarship (two clusters on the top right corner), those who maintained it over the course of two years and those who gained in the second year. A few students lost it and they are mapped near those never received achievement scholarship. A clear separation between students who never received achievement and the other two main clusters indicates that students belonging to that cluster behave differently than others. This makes sense because students who receive academic achievement scholarship are academically distinguishable.

Fig. 5.6 shows that t-SNE used the student gender to further dissect the clusters into clusters of males and females and Fig. 5.7 shows major enrollment of these students in the main three schools at AUC; Sciences and Engineering, School of Business and undecided as of their second year at AUC. It is interesting to see that all undecided students who were not able to declare a major by their second year fall on the side of students who never received achievement scholarship and are even forming a separate cluster. Fig. 5.8 shows that t-SNE used the students' second year GPA to arrange them within each cluster such that students with high GPA fall close to each-others but far from those who have low GPAs.

المنارة للاستشارات

www.manaraa.com

5.2.3. Comparison of classification results before and after applying t-SNE

t-SNE mapping succeeds for the second time in boosting the performance of the lazy K-nearest neighbor from an accuracy of 35% on the full dataset to 66% which is a comparable performance to the other two non-linear classifiers. We relate that to our claim that t-SNE simplifies the problem by bringing it down to a lower dimensional space with minimal information loss. Table 5.3 shows a comparison of the performance of the different algorithms on the full dataset, the 2D mapping and the 3D mapping. Random Forest achieves highest accuracy 72% on the full-dataset. However, the second best is equally achieved by the NN and the KNN 68% on the 3D t-SNE was only 4% lower. We can see that the three algorithms performed almost equally on the 3D mapping with a slight improvement over the 2D mapping. This means that the algorithm performance became fully dependent on how good the data mapping was.

Table 5.3 Comparison of the classification accuracy on AUC full dataset

	NN	RF	RPART	SVM	KNN
t-SNE 2D Mapping	66% ± 0.7	66% ± 0.6	51% ± 0.7	54% ± 0.6	**68% ± 0.8**
t-SNE 3D Mapping	**68% ± 0.7**§	68% ± 0.6	52% ± 0.6	**57% ± 0.5**§	68% ± 0.7
Raw Dataset *No dimensionality Reduction*	62% ± 0.9	**72% ± 0.7**§	**65% ± 0.8**§	50% ± 0.5	35% ± 0.7

§ - Statistical significance under pairwise comparison with same method under different variations of the dataset

Table 5.4 Comparison of the classification results using Cohen's Kappa on AUC full dataset

	NN	RF	RPART	SVM	KNN
t-SNE 2D Mapping	58% ±0.9	58% ±0.8	40% ± 0.8	45% ± 0.7	**62% ±0.1**
t-SNE 3D Mapping	**61%** ±0.9	61% ±0.8	41% ± 0.7	**49% ± 0.8**	61% ±0.8
Raw Dataset *No dimensionality Reduction*	53% ±0.9	**66% ±0.9**	**57%** ± **1.1**	41% ± 0.6	20% ±0.9

المنارة للاستشارات
www.manaraa.com

5.2.4. Results Discussion

The enhancement in performance of the Neural Network and the KNN is aligned with our findings on the benchmarking dataset. The Random Forest still records highest performance on the full-dataset. Random Forest is an ensemble classifier that runs efficiently on large datasets which may be is the reason for their superior performance on the full-dataset. The performance of random forests is determined by two measures: strength and correlation. The strength of each individual tree in the forest. A tree with a low error rate is a strong classifier. Increasing the strength of these individual classifiers increases the overall strength of the forest. The correlation is the average correlation between the pairs of trees. Increasing the correlation between the individual trees increases the error rate of the forest. By reducing the dimensionality of our dataset to only two or three dimensions, we are reducing the strengths of our individual trees and increasing the correlation which would consequently negatively affect the overall performance of the random forest. Accordingly, the main advantage of using t-SNE mapping as a dimensionality reduction technique is to decrease the training time of the random forest at the expense of losing some accuracy (4%). For AUC purposes and because we are still introducing the culture of predictive modeling to the community with this being the first predictive model, we are willing to sacrifice some of the accuracy in our first deployment with the aim to make build the trust of our community into predictive analytics. Therefore, we found the combination of t-SNE and KNN appealing because they are easy to explain and will be more comprehended by the audience. After putting the model in effect, we will be migrating to more sophisticated models with higher accuracy and enhanced datasets.

المنارة للاستشارات
www.manaraa.com

5.3. AUC Dataset (Numeric Attributes) – Comparing t-SNE to PCA

5.3.1. Dataset

This dataset is the same dataset used in the previous set of experiments in section 5.1.1 with the exception that it has numeric attributes only; not including categorical attributes such as gender, nationality, …etc. We checked the effect of this variation on the prediction accuracy of the model. We also compared the visualizations produced by both techniques: PCA and t-SNE.

5.3.2. Visualizing data using PCA and t-SNE

Fig. 5.9 and Fig. 5.10 show the 2D and 3D PCA mapping respectively. The data points are cluttered and are highly overlapping. Fig. 5.11 and Fig. 5.12 show the 2D and 3D t-SNE mapping respectively. The area of overlap between data points belonging to different classes is smaller in comparison to the PCA mapping. This tempted us to compare the classification results.

5.3.3. Comparison of Classification Results

Table 5.5 compare the classification results for the PCA and t-SNE mappings. It is clear that t-SNE not only provided a better mapping, it also guaranteed higher prediction accuracy for all the classification models.

المنارة للاستشارات
www.manaraa.com

Student Third Year GPA

● [0,2[● [2, 2.5[● [2.5, 3[● [3, 3.4[● [3.4, 3.6[● [3.6, 3.8[● [3.8, 4]

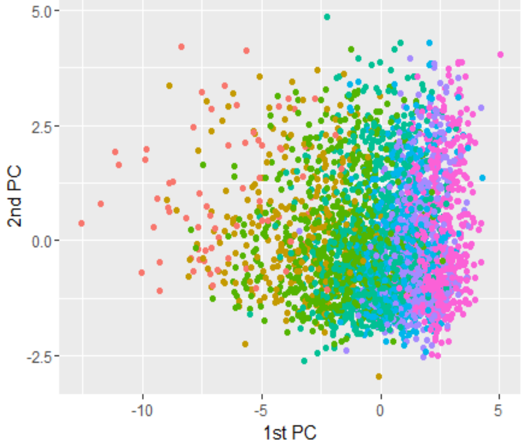

Fig. 5.9 PCA 2D Mapping of numerical AUC students' performance dataset

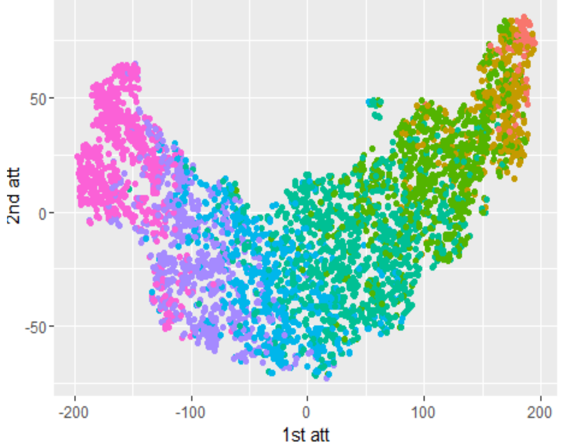

Fig. 5.10 t-SNE 2D Mapping of numerical AUC students' performance dataset

Student Third Year GPA

● [0,2[● [2, 2.5[● [2.5, 3[● [3, 3.4[● [3.4, 3.6[● [3.6, 3.8[● [3.8, 4]

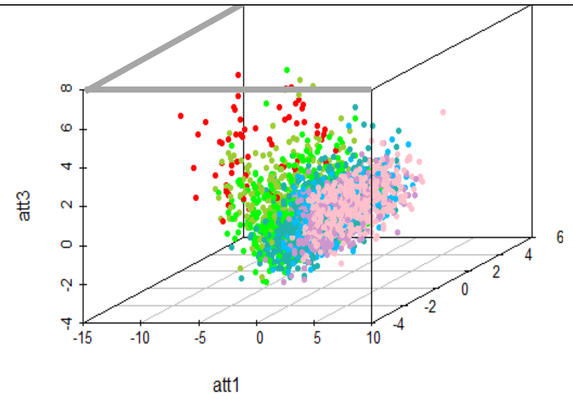

Fig. 5.11 PCA 3D Mapping of numerical AUC students' performance dataset

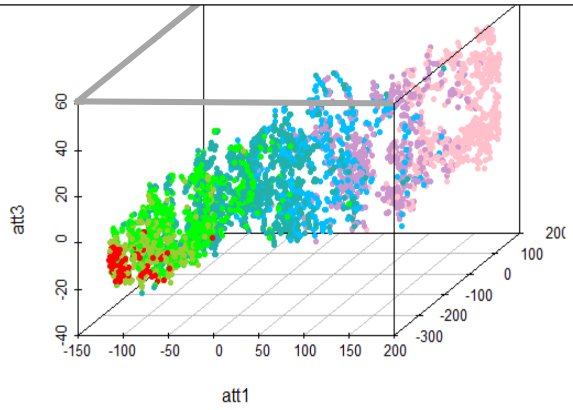

Fig. 5.12 t-SNE 3D Mapping of numerical AUC students' performance dataset

المنارة للاستشارات

www.manaraa.com

Table 5.5 - Comparison of classification accuracy of PCA And t-SNE Mappings Of AUC Numerical Dataset

	NN	**RF**	**RPART**	**SVM**	**KNN**
PCA 2DMapping	47% ± 0.8	41% ± 0.7	44% ± 0.6	41% ± 0.7	43% ± 0.6
PCA 3D Mapping	52% ± 0.6	49% ± 0.7	44% ± 0.5	44% ± 0.6	49 ± 0.6
t-SNE 2D Mapping	70% ± 0.6	67% ± 0.8	61% ± 0.6	**55% ± 0.6**	**69% ± 0.7**
t-SNE 3D Mapping	**71% ± 0.6**	**68% ± 0.7**	**63% ± 0.7**	54% ± 0.5	69% ± 0.7

CHAPTER (6): CONCLUSION AND FUTURE WORK

6.1. **Conclusion**

Predictive data modeling can help institutions become more proactive and provide better services and customer experience. In our research, we were able to utilize a recent visualization and dimensionality reduction technique (t-SNE) to leverage human cognitive power to understand the underlying structure of high dimensional data. We used t-SNE to visualize large datasets in 2D and 3D. t-SNE guarantees that data instances which are similar in the high dimensional space get mapped to nearby data points on the lower dimensional space. Accordingly, humans can easily identify similarities and differences. Another interesting finding of our study is that lower dimensional mapping simplifies the classification problem and enables a simple classifier such as KNN to compete with more sophisticated algorithms such as the NN. The simplicity of the concept of the KNN can help explain the concept of predictive analytics to a wide audience of various literacy levels.

المنارة للاستشارات
www.manaraa.com

6.2. **Future Work**

While t-SNE has great advantages, it also comes with the limitation that it requires retraining the model with every new data point. There have been some attempts to use regression techniques to train a model on predicting the projections of new data-points. We would like to extend our solution to include prediction of mappings of new data points to save the model re-training time at the beginning of every semester.

We also realized that usually, we did not get significant improvement in classifier performance when increasing the dimensionality from 2D to 3D. We would like to dig deeper into this observation and conduct more experiments to confirm or reject this hypothesis.

It is also important to try comparing PCA and t-SNE performance at dimensionalities greater than three. We would also like to compare t-SNE to other techniques such as Sammon Mapping and Fisher Discriminant Analysis.

It will also be useful if we dig deeper into the random forest performance degradation caused by the dimensionality reduction. We would also like to try other algorithms such as Logistic Regression, gradient boosting machines and deep learning techniques.

On the dimension of the students' performance prediction problem, it is worth mentioning that working on better data collection mechanisms to avail more data about students' satisfaction, students' engagement, family educational background and data about students' learning behaviors logged by the learning management system can significantly enhance our prediction model performance. It will be useful if we can test

المنارة للاستشارات
www.manaraa.com

robustness of our prediction models to missing values. Finally, we would like to solve other problems such as student recruitment, student attrition and major declaration advising using predictive modeling.

المنارة للاستشارات
www.manaraa.com

REFERENCES

[1] IBM, “IBM What Is Big Data: Bring Big Data to the Enterprise,” 2012. [Online]. Available: https://www-01.ibm.com/software/in/data/bigdata/. [Accessed: 04-Apr-2016].

[2] G. Satell, “A Look Back At Why Blockbuster Really Failed And Why It Didn’t Have To,” *Forbes,* 2014. [Online]. Available: https://www.forbes.com/sites/gregsatell/2014/09/05/a-look-back-at-why-blockbuster-really-failed-and-why-it-didnt-have-to/#8122c771d64a. [Accessed: 22-Apr-2017].

[3] “The World’s most Innovative Companies,” *Forbes,* 2016. [Online]. Available: https://www.forbes.com/innovative-companies/list/#tab:rank. [Accessed: 22-Apr-2017].

[4] A. Bernstein, “Dimensionality Reduction in Statistical Learning,” in *13th International Conference on Machine Learning and Applications*, 2014, pp. 330–335.

[5] A. L. Bluma and P. Langley, “Selection of relevant features and examples in machine,” *Artif. Intell.*, vol. 97, no. 97, pp. 245–271, 1997.

[6] R. Kohavi and H. John, “Artificial Intelligence Wrappers for feature subset selection,” vol. 97, no. 97, pp. 273–324, 2011.

[7] S. Das, “Filters , Wrappers and a Boosting-Based Hybrid for Feature Selection,” in *ICML ’01 Proceedings of the Eighteenth International Conference on Machine Learning*, 2001, pp. 74–81.

[8] L. Yu and H. Liu, “Feature Selection for High-Dimensional Data : A Fast Correlation-Based Filter Solution,” in *Proceedings of the Twentieth International Conference (ICML 2003)*, 2003.

[9] I. Rish, G. Grabarnik, G. Cecchi, F. Pereira, and G. Gordon, “Closed-Form Supervised Dimensionality Reduction with Generalized Linear Models,” in *Proceedings of the 25 th International Conference on Machine Learning*, 2008, pp. 832–839.

[10] H. Liu and H. Motodo, Eds., *Feature Extraction, Construction and Selection*, 1st ed. Springer US, 1998.

[11] M. Friendly, “A Brief History of Data Visualization,” in *Handbook of Computational Statistics: Data Visualization*, vol. III, C. Chen, W. Härdle, and A. Unwin, Eds. Heidelberg: Springer-Verlag, 2006.

[12] G. K. L. Tam, V. Kothari, and M. Chen, “An Analysis of Machine- and Human-Analytics in Classification,” *IEEE Trans. Vis. Comput. Graph.*, vol. PP, no. 99, 2016.

[13] P. Xu, H. Mei, L. Ren, and W. Chen, “ViDX : Visual Diagnostics of Assembly Line Performance in Smart Factories,” *IEEE Trans. Vis. Comput. Graph.*, vol. preprint, no. 99, 2016.

المنارة للاستشارات

www.manaraa.com

[14] M. Wagner, F. Fischer, R. Luh, A. Haberson, A. Rind, D. A. Keim, and W. Aigner, “A Survey of Visualization Systems for Malware Analysis,” in *Eurographics Conference on Visualization (EuroVis) - STARs*, 2015.

[15] L. Van Der Maaten and G. Hinton, “Visualizing Data using t-SNE,” *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

[16] Eduventures, “Predictive Analytics in Higher Education,” Boston, MA, 2013.

[17] C. Romero, S. Ventura, M. Pechenizkiy, and R. S. J. d. Baker, *Handbook of Educational Data Mining*. CRC Press, 2010.

[18] J. P. Campbell, P. B. DeBlois, and D. G. Oblinger, “Academic Analytics: A New Tool for a New Era,” *Educ. Rev.*, vol. 42, no. August 2007, pp. 40–57, 2007.

[19] R. Shatnawi, Q. Qlthebyan, B. Ghalib, and M. Al-Maolegi, “Building A Smart Academic Advising System Using Association Rule Mining,” *arXiv:1407.1807*, 2014.

[20] H. M. Nagy, W. M. Aly, and O. F. Hegazy, “An Educational Data Mining System for Advising Higher Education Students,” *Int. J. Comput. Electr. Autom. Control Inf. Eng.*, vol. 7, no. 10, pp. 622–626, 2013.

[21] H. Krase and E. Nyatepe-Coo, “Identifying and Supporting Academically At-Risk Students in Canadian Universities,” Washington, D.C, 2012.

[22] G. Grinstein and B. Thuraisingham, “Data mining and data visualization: Position paper for the second IEEE workshop on database issues for data visualization,” *Database Issues Data Vis.*, pp. 54–56, 1996.

[23] R. Asif, A. Merceron, and M. K. Pathan, “Predicting Student Academic Performance at Degree Level : A Case Study,” *Intell. Syst. Appl.*, vol. 1, no. December 2014, pp. 49–61, 2015.

[24] J. Zimmermann, K. H. Brodersen, J. Pellet, E. August, J. M. Buhmann, and E. T. H. Zurich, “Predicting graduate-level performance from undergraduate achievements,” in *Proceedings of the 4th International Conference on Educational Data Mining*, 2007, pp. 2–3.

[25] K. F. Li, D. Rusk, and F. Song, “Predicting Student Academic Performance,” in *Seventh International Conference on Complex, Intelligent, and Software Intensive Systems Predicting*, 2013, pp. 27–33.

[26] A. Slim, G. L. Heileman, J. Kozlick, and C. T. Abdallah, “Predicting Student Success Based on Prior Performance,” in *Computational Intelligence and Data Mining (CIDM)*, 2014.

المنارة للاستشارات
www.manaraa.com

[27] J. Chen, H. Hsieh, and Q. H. Do, “Predicting Student Academic Performance: A Comparison of Two Meta-Heuristic Algorithms Inspired by Cuckoo Birds for Training Neural Networks,” *Algorithms*, vol. 7, pp. 538–553, 2014.

[28] S. Yadav, B. Bharadwaj, and S. Pal, “Data mining applications: A comparative study for predicting student’s performance,” *Int. J. Innov. Technol. Creat. Eng.*, vol. 1, no. 12, pp. 13–19, 2012.

[29] M. Al-sarem, “Building a Decision Tree Model for Academic Advising Affairs Based on the Algorithm C4 . 5,” *Intenational J. Comput. Sci. issues*, vol. 12, no. 5, pp. 33–37, 2015.

[30] E. J. M. Lauría, J. D. Baron, M. Devireddy, V. Sundararaju, and S. M. Jayaprakash, “Mining academic data to improve college student retention : An open source perspective,” *Proc. Second Int. Conf. Learn. Anal. Knowl. - LAK ’12*, no. May, pp. 139–142, 2012.

[31] C. Romero and S. Ventura, “Educational Data Mining : A Review of the State of the Art,” *Trans. Syst. MAN, Cybern. C Appl. Rev.*, vol. 40, no. 6, pp. 601–618, 2010.

[32] J. Yi, X. Mao, and Y. Xue, “Facial Expression Recognition Based on t-SNE and AdaBoostM2,” in *2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing*, 2013, no. 30400002012102000, pp. 1744–1749.

[33] J. S. Sabharwal, “Visualizing Wikipedia using t-SNE,” 2008.

[34] V. Den Oord, S. Dieleman, and B. Schrauwen, “Deep content-based music recommendation,” in *NIPS’13 Proceedings of the 26th International Conference on Neural Information Processing Systems*, 2013, pp. 2643–2651.

[35] A. Gisbrecht, B. Hammer, B. Mokbel, and A. Sczyrba, “Nonlinear dimensionality reduction for cluster identification in metagenomic samples,” in *Information Visualisation (IV), 2013 17th International Conference*, 2013, pp. 174–179.

[36] T. Ravet, S. Dupont, C. Picard-Limpens, and C. Frisson, “Nonlinear Dimensionality Reduction Approaches Applied To Music and Textural Sounds,” in *IEEE International Conference on Multimedia and Expo (ICME)*, 2013.

[37] D. Ara, A. Martins, and J. Melo, “Comparative Study on Dimension Reduction Techniques for Cluster Analysis of Microarray Data,” in *Proceedings of International Joint Conference on Neural Networks*, 2011, pp. 1835–1842.

[38] M. Domínguez, S. Alonso, A. Morán, M. A. Prada, and J. J. Fuertes, “Dimensionality reduction techniques to analyze heating systems in buildings,” *Inf. Sci. (Ny).*, vol. 294, pp. 553–564, 2015.

[39] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” in *Proceedings of the IEEE*, 1998, pp. 2278–2324.

المنارة للاستشارات

www.manaraa.com

[40] R. Bellman, *Dynamic Programming*, 1st ed. Princeton, NJ, USA: Princeton University Press, 1957.

[41] R. E. Bellman, *Adaptive Control Processes: A Guided Tour*. MIT Press, 1961.

[42] I. T. Jolliffe, *Principal Component Analysis*. Springer Verlag, 1986.

[43] W. L. W. Liu and C.-I. C. C.-I. Chang, “Variants of Principal Components Analysis,” *2007 IEEE Int. Geosci. Remote Sens. Symp.*, pp. 1083–1086, 2007.

[44] L. I. Smith, “A tutorial on Principal Components Analysis Introduction,” *Statistics (Ber).*, vol. 51, p. 52, 2002.

[45] J. B. Tenenbaum, V. De Silva, and J. C. Langford, “A Global Geometric Framework for Nonlinear Dimensionality Reduction,” *Science (80-.).*, vol. 290, no. December, pp. 2319–2323, 2000.

[46] S. T. Roweis and L. K. Saul, “Nonlinear Dimensionality Reduction by Locally Linear Embedding,” vol. 290, no. December, pp. 2323–2326, 2000.

[47] P. C. Ncr, J. C. Spss, R. K. Ncr, T. K. Spss, T. R. Daimlerchrysler, C. S. Spss, and R. W. Daimlerchrysler, “Crisp-dm 1.0,” 2009.

[48] J. (Marymount C. Cuseo, “What Really Matters Defining Student Success and College Quality: Questionable,” in *22ND International Conference on the FirstYear Experience*, 2009.

[49] J. A. Buckley, B. K. Bridges, J. C. Hayek, G. D. Kuh, and J. Kinzie, “What Matters to Student Success : A Review of the Literature Spearheading a Dialog on Student Success,” 2006.

[50] J. Seuss and H. Childers, “Modeling an IT Strategy for Student Success,” 2016.

[51] R. Benford and J. Gess-newsome, “Factors affecting Student Academic Success in Gateway courses at Northern Arizona University,” Flagstaff, AZ, 2006.

[52] Ellucian Company L.P. and its affiliates, “Banner®.” Fairfax, Virginia.

[53] D. Cox and G. Box, “An Analysis of Transformations,” *J. R. Stat. Soc.*, no. ii, pp. 211–252, 1964.

[54] G. Piatetsky, “R, Python Duel As Top Analytics, Data Science software – KDnuggets 2016 Software Poll Results,” 2016.

[55] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York: Springer, 2013.

[56] M. Kuhn, “A Short Introduction to the caret Package,” 2016.

المنارة للاستشارات

www.manaraa.com

[57] P. Cortez and A. Silva, “Using Data Mining To Predict Secondary School Student Performance,” *5th Annu. Futur. Bus. Technol. Conf.*, vol. 2003, no. 2000, pp. 5–12, 2008.

[58] J. . Gower, “A General Coefficient of Similarity and Some of Its Properties,” in *Biometrics*, 2016, vol. 27, no. 4, pp. 857–871.

[59] “Manifold,” *Wikipedia*. .

المنـــارة للاستشارات

www.manaraa.com

www.manaraa.com

www.manaraa.com

www.manaraa.com