Implementing Machine Learning in the Workplace of a Further Education Provider

Contents

# [Introduction](#_Introduction)

# [Predicting Student Retention](#_Predicting_Student_Retention)

## [Support Vector Machine Model](#_Support Vector_Machine_Model)

## [Decision Tree Model](#_Decision_Tree_Model)

# [Estimating Student Grades/Test Scores](#_Estimating_Student_Grades/Test)

## [Simple Linear Regression Model](#_￼Simple_Linear_Regression)

## [Multiple Linear Regression Model](#_Multiple_Linear_Regression)

# [Concluding Remarks](#_Concluding_Remarks)

# [References](#_References)

# Introduction

The purpose of this report is to explore various ways in which machine learning (ML) can enhance the data analysis processes within a further education (FE) provider's workplace. It will delve into two business use cases involving different datasets and discuss the application of statistical models to address the respective business problems. Additionally, given the increasing adoption of machine learning within the education sector (Korkmaz and Correia, 2019), this report will investigate the current use of statistical models in educational settings. By doing so, it aims to deepen the understanding of how these approaches can be applied effectively within the context of a further education provider.

# Predicting Student Retention

Student retention is a pivotal metric for educational institutions, serving not only as an indicator of institutional performance but also as a determinant for government funding allocations (Education & Skills Funding Agency, 2024). Therefore, by incorporating a machine learning model to predict student retention, educational providers can proactively identify at-risk students, enabling data-driven decisions to address their needs and ultimately improve overall retention rates.

In terms of the data that could be used to facilitate this model, the target variable would be a learner’s completion status. The features would consist of various attributes related to a student’s completion status. Given there are numerous measures within education data potentially relevant to student completion, it would be best to utilise the strongest contributors. With this in mind, attendance, previous education level, and delivery mode can be included, since these are particularly relevant to student retention (Zaman, et al., 2023). An explanation of these attributes, alongside mock sample data (table 1), is provided below:

* Attendance – Represents the number of lessons attended by a student as a percentage of their total assigned lessons.
* Previous Education Level – Indicates the highest level of education a learner achieved prior to the currently enrolled course.
* Delivery Mode – The format of how the learning is delivered.

Table 1, Mock sample of target and features for student retention data

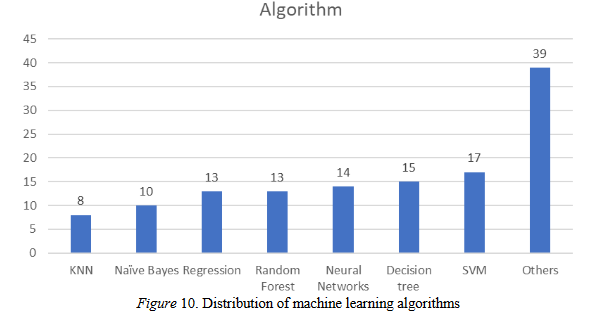
|  |  |  |  |
| --- | --- | --- | --- |
| **Attendance** | **Previous Education Level** | **Delivery Mode** | **Completion Status** |
| 88% | A-level | In-Person | Achieved |
| 93% | Higher apprenticeship | Remote | Achieved |
| 64% | GCSE | In-Person | Withdrawn |
| 90% | GCSE | Blended Learning | Achieved |

It should be noted that evidence suggests demographic factors, such as sex or race, can also influence student retention (Zaman, et al., 2023). However, these factors have been purposefully omitted from consideration in this business case to adhere to the Data Ethics Framework (Central Digital & Data Office, 2020). Including such data could inadvertently lead to discriminatory practices, implying that certain groups may or may not complete their course based on these characteristics.

## Support Vector Machine Model

A classification model is required for predicting student retention, as the intention is to classify a learner’s completion status. With numerous classification models available, it is essential to select the most appropriate one for this business case. Support Vector Machine (SVM) was chosen as one of the models to consider due to its effectiveness in handling classification tasks. Moreover, SVM is commonly utilised in educational contexts, as depicted in figure 1 below, which illustrates the frequency of ML model usage for education-based business problems.

Figure 1, Distribution of machine learning algorithms (Luan and Tsai, 2021)



Furthermore, SVM has previously demonstrated successful applications in predicting student retention based on historical data (Shah, et al., 2021), further justifying its application within a further education environment.

In order to implement an SVM model within the workplace of an FE provider, the machine learning process will need to be followed (Shah, et al., 2021). This process typically involves establishing the business case (as established earlier), acquisition and preprocessing the data, constructing and training the model, evaluating its performance, and finally deployment.

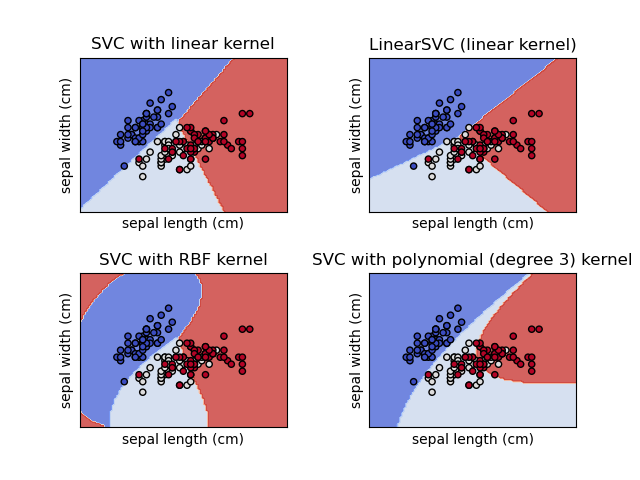
To acquire the necessary data, the feature and target attributes must be retrieved from the business’ enterprise application server, where this data is hosted. This can be accomplished by querying and extracting the relevant data using Python or SQL. Preliminary filtering may be necessary at this stage, such as removing test students or irrelevant records, to ensure the dataset's integrity before proceeding to subsequent steps.

The data can then be imported into a Jupyter Notebook for exploratory data analysis (EDA), preprocessing, and cleaning. During EDA, various aspects of the data, such as its shape and distribution, can be analysed using Pandas ‘describe’ method (Pandas, 2024). Additionally, visualisations like boxplots and correlation matrices can be generated to understand the distribution and relationships within the data. Standard data cleaning practices, such as handling missing values and outliers, can then be applied based on insights gained from the EDA process.

For preprocessing, categorical attributes must be transformed into numerical representations suitable for SVM modelling (Hsu, Chang, & Lin, 2003). Techniques like category encoding, such as one-hot encoding, can be employed for this purpose (McGinnis, 2022). Furthermore, numerical attributes may need to be scaled to ensure uniformity in their ranges, which is important for SVM performance (Hsu, Chang, & Lin, 2003). For instance, ‘StandardScaler’ from the Scikit-learn library provides a convenient method for standardising the data (Scikit-learn developers, 2024c).

Following data preparation, SVM modelling can commence using the prepared student retention data. SVM operates by finding the optimal hyperplane that best separates the data into distinct groups based on the target variable. This hyperplane aims to maximise the margin, which is the distance between the hyperplane and the nearest data points from each group (Hsu, Chang, & Lin, 2003). Thus, with SVM modelling, the objective should be to seek the widest margin. For this business case then, the goal is to train the model on past completion statuses, using the ideal hyperplane to predict the future completion statuses of current students accurately.

Various types of SVM models are available for classification, each with distinct strengths, limitations, and functionalities. For instance, figure 2 below depicts the varying results between using linear, radial basis function (RBF), and polynomial kernels for support vector classifiers (SVC) (Scikit-learn developers, 2024b).

Figure 2, SVC kernels performance differences (Scikit-learn developers, 2024b) 

In addition to selecting which kernel, different parameters can be tuned to optimise the SVM model's performance. For instance, adjusting the 'C' parameter can control the regularisation strength of the model (Scikit-learn developers, 2024f). Furthermore, SVM supports both binary and multiple class classification tasks (Scikit-learn developers, 2024b). Given the range of these available options, it will be necessary to experiment with different parameters and kernel types to identify the most appropriate SVM model for predicting students’ completion statuses.

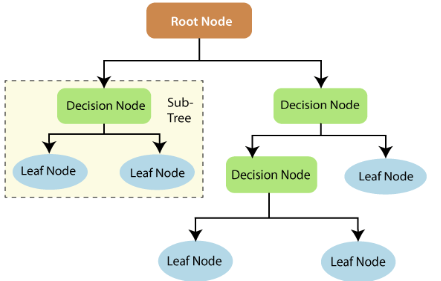
Lastly, model evaluation will be performed to assess the performance of the SVM model and make necessary adjustments based on the results. One effective technique for this is k-fold cross-validation, which involves splitting the dataset into k folds. Each fold is used as a test set while the remaining k-1 folds are used for training. (Kharb and Singh, 2021).

Metrics such as accuracy, precision, recall, and F1-score can then be calculated from the cross-validation results to evaluate the model's effectiveness. Additionally, visualisations like confusion matrices can be generated to provide a graphical representation of the model's performance, highlighting the number of correct and incorrect predictions across different classes.

## Decision Tree Model

Another viable ML model for predicting student retention is the decision tree model. This algorithm works by recursively partitioning the data into groups based on feature values, creating a tree structure where each node represents a decision based on a feature, as exemplified in figure 3 below. This process continues until the data can be classified into the target variable's categories (Charbuty and Abdulazeez, 2021). For predicting student retention, a decision tree could effectively categorise students into their most likely completion status.

Figure 3, Decision Tree (Charbuty and Abdulazeez, 2021)



Although decision tree models may not always be as accurate as SVM (Zaman, et al., 2023), they are generally easier to construct and more easily interpretable, even for non-technical stakeholders (Charbuty and Abdulazeez, 2021). Therefore, it would be advantageous to develop both models to compare their results and determine which one offers the most benefits for the business case in terms of accuracy, interpretability, and other relevant factors.

Similar to developing an SVM model, creating a decision tree model requires following the same machine learning process (Shah, et al., 2021), albeit with differences in data preprocessing and model tuning steps. While categorical data may still need to be converted to numerical format during the data preparation stage, normalisation and scaling are not necessary (Scikit-learn developers, 2024a), unlike with SVM or other models.

The ‘DecisionTreeClassifier’ from Scikit-learn developers (2024d) can be used for creating a student retention decision tree classifier, having various parameters for model tuning. For example, ‘criterion’ determines the method, such as Gini impurity, used for evaluating the effectiveness of each split. Additionally, ‘max\_depth’ can limit the depth of the decision tree to manage model complexity. Experimenting with these configuration options in model evaluation is important for producing an effective decision tree model and addressing some of the disadvantages associated with them such as overfitting and overly complex trees (Scikit-learn developers, 2024a).

# Estimating Student Grades/Test Scores

Another significant metric for FE providers is the grades/test scores achieved by their students. This measure not only reflects the effectiveness of the teaching but is also a key factor used by Ofsted (2024) for inspecting the quality of schools in the UK. Consequently, FE providers could benefit greatly from predicting learner outcomes in advance to proactively identify students who may be at risk of underperforming, enabling them to provide additional support and improve academic performance.

When creating a model for predicting student outcomes, the target variable would be the student grade or test score achieved. Historically, previous or mock exam scores have often been utilised as feature attributes for predicting learner outcomes (Korchi, et al., 2023). Additionally, a student’s previous level of education and their current proficiency in English and Mathematics have also been used to estimate student results (Dagdagui, 2022). Hence, these variables can be employed as the feature variables for this business case. A mock sample table of these attributes is provided in table 2 below:

Table 2, Mock sample of target and features for learner outcomes data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Previous or Mock Score/Grade** | **Level of Education at Previous School** | **English Level** | **Mathematics Level** | **Achieved Score/Grade** |
| 78% | GCSE | 9 | 8 | 82% |
| Pass | A-Level | 4 | 2 | Pass |
| Fail | Diploma | 4 | 5 | Pass |
| 55% | A-Level | 0 | 1 | Fail |

## Simple Linear Regression Model

Potentially the most straightforward model for predicting student outcomes would be a simple linear regression model. This model finds the best-fitting correlation line between a single feature variable and the target using ordinary least squares (Scikit-learn developers, 2024e). For this business case, the ‘Previous or Mock Score/Grade’ can be utilised, as it is the most prominent feature for estimating learner results. This model is also chosen due to its demonstrated success in predicting student outcomes in educational environments previously (Korchi, et al., 2023; Dagdagui, 2022).

To construct a linear regression model, the machine learning process should be followed (Shah, et al., 2021), as outlined with the aforementioned models. As with most other statistical models, the inputted attributes are required to be numeric (Scikit-learn developers, 2024e). Therefore, during the data cleansing phase, non-numeric values will need to be converted, for example, ‘Fail’ to 0. Although it isn’t strictly necessary to standardise or normalise data for linear regression, in some cases it can be beneficial, so this should also be tested when creating the student outcomes model to optimise the model performance.

During model evaluation, the previously mentioned techniques can also be employed for evaluating this linear regression model, such as k-fold cross-validation (Kharb and Singh, 2021), and confusion matrices. Additionally, metrics used to score the performance of linear regression models include mean squared error, root mean squared error, mean absolute error, and R-squared (Korchi, et al., 2023). For instance, R-squared can assess the strength of the relationship between the previous or mock score/grade and the current score/grade achieved.

## Multiple Linear Regression Model

Multiple linear regression operates similarly to simple linear regression but incorporates two or more feature variables. Instead of finding the best-fitting correlation line, it identifies the best-fitting hyperplane that correlates the features with the target variable, again using ordinary least squares (Scikit-learn developers, 2024e). The process for creating a multiple linear regression model for estimating student results will be essentially the same as with a simple linear regression model. However, all the feature columns from Table 2 can be included, and these will need to be accounted for during model tuning and evaluation.

# Concluding Remarks

This report has presented two distinct business cases for a further education provider where incorporating statistical models could effectively generate business value. Among these, predicting student retention may offer the greatest value, since it involves numerous factors to forecast a learner’s completion status, which can be more complex than estimating students’ test scores.

Between the two proposed models for predicting student completion statuses, SVM and decision trees, it is not immediately clear which would be the superior choice without comprehensive testing. However, SVM might be preferred due to its potential for higher accuracy. Regardless, both models should be experimented with to determine which offers the most accuracy and delivers the most business benefits.

# References

Central Digital & Data Office (2020) *Data Ethics Framework* [Online]. Available at: <https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework-2020> (Accessed 18th May 2024)

Charbuty, B. and Abdulazeez, A. (2021) ‘Classification based on decision tree algorithm for machine learning’ *Journal of Applied Science and Technology Trends*, *2*(01) [online]. Available at: <https://doi.org/10.38094/jastt20165> (Accessed 19th May 2024)

Dagdagui, R. T. (2022). ‘Predicting Students’ Academic Performance Using Regression Analysis’ *American Journal of Educational Research*, 10(11) [online]. Available at: <https://pubs.sciepub.com/education/10/11/2> (Accessed 23rd May 2025)

Education & Skills Funding Agency (2024) *Funding guidance for young people 2023 to 2024 rates and formula* [Online]. Available at: <https://www.gov.uk/government/publications/funding-rates-and-formula/funding-guidance-for-young-people-2023-to-2024-rates-and-formula> (Accessed 16th May 2024)

Hsu, C.W., Chang, C.C., and Lin, C.J. (2003) ‘A practical guide to support vector classification’ [online]. Available at: <https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf> (Accessed 18th May 2024)

Kharb, L. and Singh, P. (2021) ‘Role of machine learning in modern education and teaching’, *Impact of AI Technologies on Teaching, Learning, and Research in Higher Education*, IGI Global [online]. Available at: <https://doi.org/10.4018/978-1-7998-4763-2.ch006> (Accessed 18th May 2024)

Korchi, A., et al. (2023) ‘Machine Learning and Deep Learning-Based Students’ Grade Prediction’, *Operations Research Forum*, 4(87) [online]. Available at <https://doi.org/10.1007/s43069-023-00267-8> (Accessed 23rd May 2025)

Korkmaz, C. and Correia, A.P. (2019) ‘A review of research on machine learning in educational technology’, *Educational Media International*, 56(3) [online]. Available at: <https://doi-org.ezproxy.neu.edu/10.1080/09523987.2019.1669875> (Accessed 16th May 2024)

Luan, H. and Tsai, C.C. (2021) ‘A review of using machine learning approaches for precision education’ *Educational Technology & Society*, 24(1) [online]. Available at: <https://www.jstor.org/stable/26977871> (Accessed 16th May 2024)

McGinnis, W. (2022) *Category Encoders* [Online]. Available at: <https://contrib.scikit-learn.org/category_encoders/> (Accessed 18th May 2024)

Ofsted (2024) *School inspection handbook* [Online]. Available at: <https://www.gov.uk/government/publications/school-inspection-handbook-eif/school-inspection-handbook-for-september-2023#evaluating-the-quality-of-education-part-2> (Accessed 23rd May 2024)

Pandas (2024) *pandas.DataFrame.describe* [Online]. Available at: <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.describe.html#pandas-dataframe-describe> (Accessed 18th May 2024)

Shah, D., et al. (2021) ‘Exploiting the capabilities of blockchain and machine learning in education’, *Augmented Human Research*, 6 [online]. Available at: <https://doi.org/10.1007/s41133-020-00039-7> (Accessed 16th May 2024)

Scikit-learn developers (2024a) *1.10. Decision Trees* [Online]. Available at: <https://scikit-learn.org/stable/modules/tree.html> (Accessed 23rd May 2024)

Scikit-learn developers (2024b) *1.4. Support Vector Machines* [Online]. Available at: <https://scikit-learn.org/stable/modules/svm.html> (Accessed 18th May 2024)

Scikit-learn developers (2024c) *6.3. Preprocessing data* [Online]. Available at: <https://scikit-learn.org/stable/modules/preprocessing.html> (Accessed 18th May 2024)

Scikit-learn developers (2024d) *DecisionTreeClassifier* [Online]. Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier> (Accessed 23rd May 2024)

Scikit-learn developers (2024e) *LinearRegression* [Online]. Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html#sklearn.linear_model.LinearRegression> (Accessed 23rd May 2024)

Scikit-learn developers (2024f) *sklearn.svm.SVC* [Online]. Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC> (Accessed 18th May 2024)

Zaman, B., et al. (2023) ‘Modeling education impact: a machine learning-based approach for improving the quality of school education’, *Journal of Computers in Education* [online]. Available at: <https://doi-org.ezproxy.neu.edu/10.1007/s40692-023-00297-5> (Accessed 16th May 2024)