CCT College Dublin

MSc in Data Analytics

Capstone Project

Analysing a dataset using both traditional data analytics methods and causal inference methods for comparison and evaluation

Milo Moran sbs23081

Supervisor: James Garza

23rd February 2024

CCT College Dublin

Assessment Cover Page

To be provided separately as a word doc for students to include with every submission

Module Titles:	Capstone Project
Assessment Title:	Thesis
Lecturer's Name:	James Garza
Student's Full Name:	Milo Moran
Student Numbers:	sb23081
Assessment Due Date:	23/02/2024
Date of Submission:	23/02/2024

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Acknowledgements

I would foremost like to thank James Garza for his supervision, support, and feedback throughout my project. I would also like to thank Ewan Mullane and Oran O'Connor for their feedback and advice on my report. I would like to thank Oscar Moran and Eoghan O'Callaghan for facilitating much of the work I have done on the project. Further big thanks to my class and everyone in our group that pitched in with suggestions, questions, and support in relation to the submission of the project. Finally thanks to my parents Helen O'Shea and Timothy Moran for their constant support and efforts to motivate and encourage me.

Contents

List of Figures

List of Tables

List of Abbreviations

1HE/OHE = One Hot Encoding
ANOVA = Analysis of Variance
ATE = Average Treatment Effect

ATT = AVerage effect of Treatment on the Treated

CATE = Conditional Average Treatment Effect

CIM = Causal Inference Method

CV = Cross Validation

DAG = Directed Acyclic Graph

DTC = Decision Tree Classifier

KNN = K Nearest Neighbours

ML = Machine Learning

PCA = Principal Component Analysis

RBF = Radial Basis Function

RCM = Rubin Causal Model

RTC = Randomised Controlled Trial

SCM = Structural Causal Model

Std Dev = Standard Deviation

SVM = Support Vector Machine

TDIDT = Top Down Induction of Decision Trees

XGBoost = Extreme Gradient Boosting

1 Introduction & Background

In this section, the Research Question and the Research Objectives will be introduced, contextualised, and justified. This section will also contain a brief theoretical outline for the relevant models and concepts used in the project.

1.1 Research Question

In investigating outcomes, typically traditional data analytics does not attempt to offer any answers to causal questions. Hernán et al.(2019) defines the data analysis tasks as description and prediction, which are the traditional tools of data science and analytics at all levels, but also counterfactual prediction, which is required for causal inference applications, and is not only rarer among data practitioners but was also explicitly avoided by mainstream statistics for the better part of a century (examples of each can be seen in Figure 1). From a learners perspective, analysing a dataset relating to student performance outcomes, can traditional methods of description and prediction provide a base of information to help ask and answer causal questions? And does the addition of these Causal Inference Methods (CIMs) to an analysis provide invaluable information about the topic?

Table 1—Examples of Tasks Conducted by Data Scientists Working with Electronic Health Records

	Data Science Task		
	Description	**Prediction**	**Causal inference**
Example of scientific question	How can women aged 60–80 years with stroke history be partitioned in classes defined by their characteristics?	What is the probability of having a stroke next year for women with certain characteristics?	Will starting a statin reduce, on average, the risk of stroke in women with certain characteristics?
Data	• Eligibility criteria • Features (symptoms, clinical parameters …)	• Eligibility criteria • Output (diagnosis of stroke over the next year) • Inputs (age, blood pressure, history of stroke, diabetes at baseline)	• Eligibility criteria • Outcome (diagnosis of stroke over the next year) • Treatment (initiation of statins at baseline) • Confounders • Effect modifiers (optional)
Examples of analytics	Cluster analysis …	Regression Decision trees Random forests Support vector machines Neural networks …	Regression Matching Inverse probability weighting G-formula G-estimation Instrumental variable estimation …

Figure 1: Examples of the tasks of Data Science according to Hernán et al.

1.2 Relevance

While attempts are commonly made to predict dependent variables using datasets like the one selected, efforts are usually focused on accuracy and generalisability of models created (Yilmaz & Sekeroglu 2019), and offer limited information about what actions could be taken in order to affect outcomes. The dataset has been selected as it has been recently donated to the UC Irvine Machine Learning Repository, and represents the kind of dataset that commonly sees extensive application of classifiers and machine learning techniques on websites like Kaggle. Such data analytics projects undertaken by the users of the websites are regularly extremely comprehensive and yet causal inference is only sometimes included. Success at achieving the research objectives will add support to the practice of causal inference within Data Analytics, and may promote it. Specifically to this dataset, understanding causal effects on student performance is desirable, and determining if there are aspects of a students academic, or personal life that could be changed resulting in improved performance would be valuable information, even if there might be difficulties in applying this information in terms of policy(von Hippel & Wagner 2018). However, determining causality for something as complex as student performance requires to any significant degree of validity requires consistently accurate assumptions and significant domain knowledge, and likely access to data that is not included in this dataset. As such, the specific causal findings of this research are assumed to have extremely limited validity.

1.3 Contribution

The novelty of the project is related to the production of a data science project that is designed to answer causal questions about effects of features on the outcome, as well as the creation of a model to determine how the features combine to classify the target for this dataset.

This project will combine an analysis of a dataset and creation of prediction models with an attempt to answer causal questions about the relationships between the "treatments" and the target outcome of student performance. Because the effect, aka the student grades, are known, and the causes are unknown, the causal questions for this project will fall under under the "Explanation" quadrant as differentiated from other question according to the Table in Figure 2. For the questions a selection of the non target features will be selected as potential treatment variables, and treatment effects will be measured for the se-

lected potential treatments, with the hopes of determining which if any features have causal relationships with the target variable.

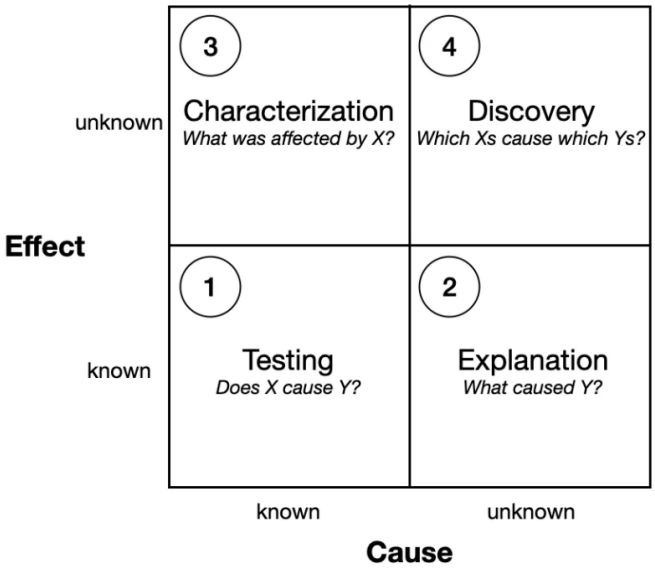


Figure 2: Types of Causal Questions, taken from Taylor(2022)

1.4 Objectives

1.4.1 Primary Objective

The primary objective of this project is to analyse the features of the dataset and their relationships with the student performance outcomes through a combination of "traditional" data analytics and Causal Inference Methods, and evaluate the advantages of adding the causal analysis, attempting to demonstrate a situation where the addition of causal analysis deepens overall analysis significantly.

1.4.2 Secondary Objectives

The following secondary objectives will be pursued and will support the primary objective:

It will be determined through experiment if there are any methods for "feature suggestion" that seem particularly suited for causal explanation. Feature

suggestion methods will be defined as methods that identify features which have strong relationships with the target. This will involve the generation of metrics to describe these relationships, and will be generated in two different ways. Initially statistical tests will be used to analyse individual features for correlation with the target independently of the other features. Then ML models will be trained and tuned in order to generate feature importance scores that accounts for the interactions between features in classifying the target. Finally, statistical testing will be used to test whether any of the "feature suggestion" methods are preferable to the others.

To support interpretation of the relationships between the independent features and the target, feature engineering methods such as sensitivity analysis, encoding, and binning will be utilised and evaluated. Feature management decisions will continuously be evaluated throughout the analysis so as to ensure interpretability of the contribution of various features on the student performance outcomes, to potentially create recommendations for further research into understanding the drivers of student success.

1.5 ML Model Background

Three supervised machine learning models are used in the project for the purpose of classifying the dependent variable, in this case the grade of the student, using all the other independent variables relating to student details. According to the No Free Lunch theorem, there is no singular best machine learning algorithm for all possible target problems (Wolpert 1996). There is no way to know for certain in advance which would be best, so to truly find the absolute most appropriate algorithm, all would need to be applied with extensive hyperparameter tuning. This generally isn't feasible however, therefore selection of models needs to be chosen appropriately based on the knowledge available. For this project, The DTC, the SVM, and the XGBoost algorithms have been chosen.

1.5.1 Decision Tree Classifier

The decision tree algorithm is an easily interpretable algorithm that is commonly used for classification. The "Tree" structure of a decision tree comes from the shape of the visualisation created to represent it. Various decision rules are defined at nodes, stemming from the root node that encompasses the entire data. Next the rules are applied to split the data on each child node iteratively until some stopping criterion is met, whereupon each data point in the node is

classified and the node becomes a leaf. In the default Scikit-learn (Pedregosa

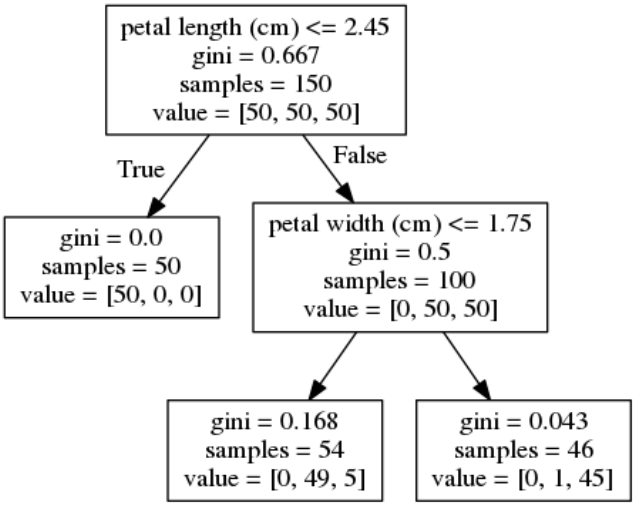


Figure 3: Example Decision Tree Classifier demonstrating use of Gini impurity for the popular "Iris" Dataset

et al. 2011) implementation of the algorithm for classification, the Decision Tree evaluates candidate splits of the data for each feature using the Gini impurity measure to find the best split for a given node. By maximising Gini impurity at the parent node, the impurity of the child nodes is minimised. This process is repeated at each node until either a minimum tree depth is hit or some minimum number of samples in the node has been reached. As this algorithm is based on choosing the locally optimal split, it is an example of a greedy algorithm.

1.5.2 Support Vector Machine

Support Vector Machines attempt to classify the dataset accurately by defining decision boundaries between classes, known as hyperplanes, using "support vectors". The support vectors for a hyperplane are the data points of the training set within the space that influence the hyperplane. The algorithm iteratively optimises the hyperplanes to maximise the perpendicular distance between the support vectors and the hyperplanes, as demonstrated for two dimensions in the linear plane in Figure 4, while minimising classification error. In implementing the model, the C parameter controls the tradeoff between those two functions, with larger C being more strict on misclassifications, typically decreasing the number of support vectors. The parameter Gamma defines the distance at

which support vectors are taken into account. At lower Gamma, points further away from the decision boundary will be considered, whereas at higher gamma only the closest points will be included in calculation. Kernel functions allow for the mapping of the data into a higher dimensional space where the data may become linearly separable. The hyperparameter is then defined within the higher dimensional space. Common Kernel functions include the Radial Basis Function (RBF) and polynomial kernels

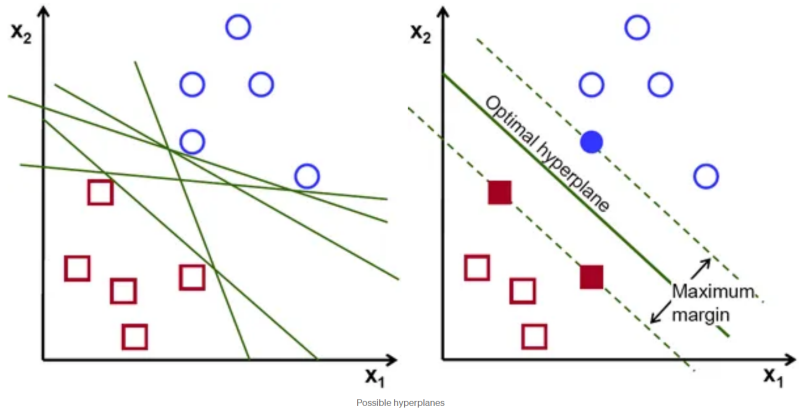


Figure 4: Visualisation of (L) potential hyperplanes , and (R) a resulting optimal hyperplane with maximised margins based on a number of surrounding support vectors (filled in) (Gandhi 2018)

1.5.3 Extreme Gradient Boosting

Boosting is an example of an ensemble method, whereby weak models are aggregated together in order to produce a much stronger model. For a boosting algorithm this process is done sequentially, with each new model being trained in an attempt to improve upon its predecessor, as opposed to bagging, where the weak models are all trained in parallel. To improve between iterations in the boosting algorithm, the weights of previously misclassified data are increased for each new model, allowing for weaknesses to be addressed.

Extreme Gradient Boosting, or XGBoost, is based upon an ensemble of the decision trees. Gradient in the name refers to the gradient based optimisation that is performed improve the optimisation function, which is a combination of the training loss and an additional regularisation term to reduce overfitting (Rodruiguez 2023). At the individual tree level the first order (gradient) and

second order (Hessian) derivatives of the loss function are calculated in order to define the Gain equation which is then minimised for each node of the tree.

While many hyperparameters are shared with standard decision tree models, others are more specific to the XGB instance: max depth of tree; gamma controlling stopping on leaves; lambda applying L2 regularisation and alpha applying l1 regularisation; learning rate defining step size for the gradient descent function; min child weight defining the minimum value of the hessian for which a split will occur; subsample defining how much of the training set, as a split, to sample for each tree iteration, and colsample bytree which selects how much of the dataset, as features, to use for each tree; and finally n estimators describing the total number of trees to use.

1.6 Causal Inference Background

Determining causality between a treatment and an effect is typically done via Randomised Controlled Trials (RTCs), as in complex systems there are usually significant confounds, both observed and unobserved, that must be accounted for. However running RCTs at scale is difficult, expensive, time consuming, and occasionally unethical. Therefore as an alternative, the potential to use observational data for causal inference has gained popularity. Two frameworks have developed to solve these problems, the Structural Causal Model (SCM), and the potential outcome framework. (Yao et al. 2021)

The SCM involves the creation of causal graphs to describe the causal mechanisms associated with a system, supporting modelling of the causality of the system via a set of structural equations. The potential outcome framework aims to estimate all potential outcomes for a treatment in the situation where only one of the potential outcomes has actually occurred, and then calculate a treatment effect. There are several variations of the treatment effect which are commonly used, that is the Average Treatment Effect (ATE), Conditional Average Treatment Effect (CATE), and Average effect of Treatment on the Treated (ATT).

In this section, the problem statement of the project has been defined and analysed. Relevant theory has been included related to the DTC, SVM, and XGBoost ML models, as well as an introduction to the frameworks of causal inference. Further discussion of how this theory relates to the project, and why specific models and libraries were selected can be found in the Methodology

Section In the next section, literature related to the project and the topic of causal inference will be investigated to further contextualise the project from a theoretical perspective.

2 Literature Review

In this section, an overview of some of the areas of causal inference will be outlined, referencing and evaluating the related literature.

Causal inference describes the analytical task of inferring causal relationships and answering causal questions based on observational data. There are two main frameworks for performing these analyses that have been championed by practitioners, and a host of methods and libraries for applying the methods. As causal effects are rarely known for certain, thorough evaluation is difficult, but several researchers have suggested novel solutions.

2.1 Motivations and Causal Questions

An article byNarula (2022) opens with the following quote which succinctly outlines an issue with the current state of data science and analytics that this project hopes to address:

> "We often talk about correlation vs causation in theory, but while implementing Data Science solutions towards solving business problems, not much influence is given to validating causation amongst independent and dependent features."

The article also mentions a key issue in ML models in that when patterns of data change due to significant disruptive or unforeseen events, the models tend to fail. Contrary to just learning patterns, causal inference is focused around learning treatment effects, and if estimated appropriately, these effects don't change when patterns in the overarching data shift for whatever reason. Much of the novelty of this project comes from its attempts to demonstrate the power of causal inference methods and their viability for pointing towards answers to causal questions in all data based quantitative research. The importance of an uptake of these methods is argued inHernán et al. (2019), and is also referenced in Martinaitis (2023) and Taylor (2022). He et al. is a recently published textbook that highlights the increasing demand and uptake in the use of causal inference in the field of medical research. Hernán et al. (2019) outlines

the historical context whereby attempts to infer causality from observational data have been suppressed by mainstream statistics, and claims that there is an essential opportunity to redefine data analysis to naturally accommodate a science - wide framework for causal inference from observational data. The article classifies the tasks of data analytics, as in Figure 1, and emphasises the differences between prediction and causal inference. It claims that while the most traditional data scientists (statisticians, epidemiologists, economists and political scientists) have typically attempted to tackle much more complex questions, and as a result focus on narrower causal questions, relative newcomers to data science, such as computer scientists, often focus on closed systems where the rules can be defined more easily. As such the latter group tends to approach causal questions less cautiously, which has the potential to be problematic in certain complex fields, like health. Finally the article makes recommendations about the teaching of data science, and suggests the dividing of learning efforts equally into the three tasks of description, prediction, and causal analysis. This project will attempt to bring the tasks together in an interdisciplinary integration demonstrating all three. The article describes an incredibly important and pertinent issue and gives recommendations with strong justification, and if adopted will have a very significant impact on the way data science and causal inference are taught to perspective data scientists.

Taylor (2022) explains the set of causal questions (Figure 2), emphasising that they can be applied for far more than experimental circumstances. It explains that within the "hypothesis space" of potential cause-effect relationships, there are a lot of unknown relationships because the variables involved either haven't been considered or are too difficult to measure. The article then goes on to describe a query engine developed by Motif analytics which facilitates exploratory causal analysis, to assist analysts in discovering and investigating potential causal relationships. This is quite a useful idea, and in this project, various traditional data analytics methods will be used in an attempt at a similar exploratory analysis. The article provides a theoretical example but no further qualitative or quantitative evidence of any results or successes with the engine.

In a preface to the textbook Real-World Evidence in Medical Product Development (He et al. 2023), the authors describe the recent increase in the usage of real world data and evidence, and the changing role of RCTs. It claims that the driving force behind this uptake of observational data is that it increases efficiency of drug development and results in time and cost savings. The outlined

goal of the book is to give a balanced and comprehensive end to end coverage of the use of real world evidence in drug development. The third section of the book contains several chapters dedicated to causal inference frameworks and methodologies as they relate to research using real world evidence. The existence of the book as well as the preface demonstrates that with the exponential rise in data in the recent years of the information age, there is increasing desirability to measure causal effects from observational data, and there have been increasing efforts to make this more accessible to practitioners and increase its uptake.

2.2 Frameworks and applications for Causal Inference

There exists two primary frameworks or models for causal inference. There is the Structural Causal Model (SCM) associated with Judea Pearl, who has written the book *Causality* (Pearl 2009), where the model is discussed at length. There is also the Rubin Causal Model(RCM), or potential outcome framework, primarily associated with Donald Rubin (described comprehensively in Imbens & Rubin (2015)).

Pearls book *Causality* is largely a continuation of his earlier work. It describes the SCM and outlines the ladder of causal queries: association, describing correlations; intervention, where the effects of deliberate actions are predicted; and counterfactuals, where a model of a counterfactual situation is created. The Structural Causal Model is mathematically defined as a set of equations that describe the relationships between exogenous, external variables influencing the system, and endogenous, internal variables that are influenced by either exogenous variables or other endogenous ones. These relationships are initially captured by a causal graph, which represents the model of the causal equations qualitatively without any quantitative component, and then analyses are carried out to determine the values of variables within the equations. The library in use for causal estimation in this project (Sharma & Kiciman 2020) involves the creation of a causal graph prior to identifying routes for esimation, before generating estimations.

The paper by Rubin (1974) is the seminal work on the potential outcome framework, with over 11,000 citations as of February 2024 according to Google Scholar. The premise of the paper is that the use of carefully controlled nonrandomised data is a reasonable second choice in the case where randomised data is not available for the estimation of causal effects. This assertion has paved the

way for all modern estimations of causal effects under the RCM across a broad range of fields and has massive influence on the topics of statistics and data science. In another much later article (Rubin 2019), Rubin describes the history and development of causal inference and the potential outcome framework where the ideas and principles that lead to the RCM as it is now known are discussed and contextualised, explaining the contributions of Neyman, Fisher, and others to the formalisation of the framework. The article describes the two essential messages of Rubin's foundational 1974 paper that was so influential as follows:

> "The comparison of potential outcomes on a common set of units define causal estimands in all settings, not just in randomized experiments."

and

> "A Posited assignment mechanism is needed for causal inference using standard statistical inference tools."

These two assertions were both fundemental breakthroughs in justifying the use of observational data to infer causality, setting the stage for an entire domain of statistics and analytics.

Ibeling & Icard (2023) compares the two frameworks, attempting to represent RCMs by corresponding SCMs. It finds that all RCMs emerge from other RCMs that are representable as SCMs and goes some way to connect the two theoretically, which is a valuable contribution to a unification, or at the very least effective translation between the two models. The ability to combine the two methods together is an important underpinning of the DoWhy end to end library for causal inference.

'A Survey on Causal Inference'(Yao et al. 2021) is a review of the potential outcome framework of causal inference. In Table 3, the paper lists 4 tool boxes designed to facillitate the application of causal inference methods, some in Python -DoWhy(Sharma & Kiciman 2020), CausalML(Chen et al. 2020) and EconML(Microsoft Research 2019), and one in R - causalToolbox(Künzel et al. 2019). Sharma & Kiciman (2020) describes the DoWhy library, and explains all the features it has to enable for end to end causal inference. The library builds on both the SCM and RCM to provide a 4 step framework for causal inference: Modelling the data, identifying if there is an opportunity to measure the desired effect according to the model, estimating the effect of the model

using a causal inference method, and refuting the obtained estimate through robustness checks. The library also has compatibility to use estimation methods from CausalML and EconML. The library is singular in accounting for the modelling, identification, and refutation steps of the causal inference process in comparison with other libraries that only enable the estimation step. One feature of the library is that causal graphs need not complete, allowing for use of partial graphs with DoWhy accounting for remaining variables itself. The documentation for DoWhy also provides a significant number of examples for the purposes of learning. All these features serve to make it an invaluable tool for analysts in performing end to end causal inference.

Yao et al. (2021) divides causal inference methods based on whether they are independent of the assumptions of the potential outcome framework or not. In Section 4, it describes the workarounds that researchers have used to infer causality in scenarios where the assumptions or parts thereof do not hold. This is important for this research project, as when dealing real world projects, it is unlikely that all three assumptions of the framework will be met, and being able to use causal inference methods in these situations is invaluable. When discussing future directions in Section 7, the paper notes the need for more research into cases in which the assumptions of the causal model can be relaxed, as practical settings frequently do not allow all the assumptions to be met and current assumption independent methods may not always be appropriate. This outlines a potential hurdle for this research, as all assumptions of the model may not be met, and it may not be possible to compensate for that when applying the causal inference model.

2.3 Evaluation of Causal inference models

There are often significantly differing ATEs for the same effect when estimated using different causal inference methods(Figure 5). as a result the evaluation and refutation of CIMs are quite important. There are three general approaches to validation of causal methods: Firstly the face validity test, whereby the result is compared with the intuition of an expert in the domain. Second the placebo test, where the nature of the data allows for the separation into placebo and treated groups based on either time or selection within the sample. Thirdly, synthetic data is used for testing methods to measure a known created treatment effect. Schuler et al. (2017) proposes Synth Validation, whereby simulations are used to test and select causal inference methods for use in given scenarios,

Table 1. Lalonde's NSW Sample ATE Estimated using few commonly used causal effect estimation methods.

Estimators	**ATE Estimate**	**Std. Dev.**
Difference of Means	886.30	277.37
Double Machine Learning	370.94	394.68
Causal BART	818.79	184.46
Propensity Score Matching	1079.13	158.59

Figure 5: An example of different causal methods resulting in drastically different ATEs for the same dataset, from Table 1 of Parikh et al.(2022)

allowing the most appropriate method to be used for that scenario. The paper notes that previous research has been inconsistent and has failed to find any one-size-fits-all methods that tests methods against hand crafted benchmark data. The results of the experiments performed to compare synth validation with individual methods showed that synth outperformed them all, with Synth validation using constrained gradient boosted-trees performing the best. This paper claims to have submitted the first formal approach to evaluating CIMs in a way that is tailored to a given dataset, and shows that different methods are more suited to different datasets compared to others, likely consistent with the NFL theorem. These are quite important findings as they provide a valuable tool to researchers and practitioners for CIM selection.

Parikh et al. (2022) presents *Credence*, which is a deep generative based framework for the validation of causal systems, allowing for the generation of synthetic data with user specified causal and confounding effects for the evaluation of different causal effect estimation methods. The paper uses Credence to perform experiments and finds that for purely synthetic data, it can reproduce rankings of individual model performance compared to an "oracle" with access to the true DGP. The paper also finds that for real data case studies where observational causal inference was performed alongside experiment, it can reproduce ATE results using the observational data that are similar to those found through experiment. These are convincing findings and show promise for using Credence to benchmark causal models against each other.

In summary, the field of causal inference is becoming increasingly relevant, and there are many different ways in which people try to make less wrong estimates of causal effects. In the next section, the methodology for carrying

out the research of the project will be outlined in detail.

3 Methodology

In this section, the method and framework of the research project will be outlined, with an emphasis on justification of the approaches chosen for primary research, feature management, analysis and causal inference. The section will also include a description of the data, and an evaluation of the potential ethical risks associated with it.

3.1 The Data

The dataset used for this project describes survey results from a survey of students in higher education at Near East University in Cyprus and was collected from the Faculty of Engineering and Faculty of Educational Sciences students in 2019. It includes information about personal details, family details, education habits, and performance outcomes. The dataset was sourced from the UC Irvine Machine learning repository. The dataset was first introduced in the paper by Yilmaz & Sekeroglu (2019), in which the authors experimented with the data and used Radial Basis Function Neural Network to achieve accuracy of 70-88%. The dataset is licensed under the Creative Commons Attribution 4.0 International (available at Appendix A), which allows for the sharing and adaptation of this dataset for any purpose once it is appropriately credited. The dataset contains a total of 33 features and 145 observations. All the data are numerically encoded as values ranging from 0-9. There is no missing data. The first feature of the dataset is an identifier for the observations, each one relating to a unique StudentID. This feature has no relevance for analysis and is dropped immediately. The answers to the questions in Table 1 correspond to features 1-30 of the dataset. The next feature, COURSE ID, assigns a number to the course the student was undertaking for each of the 9 courses covered. The last feature is GRADE, and represents the grade band that the students final grade fell into. This feature is discussed in detail in Section 3.3.1.

The data is based on personal data for these students, but attempts have been made to anonymise the data. According to the Full Guidance Note on Anonymisation and Pseudonymisation from the Data Protection Commission

Personal questions	Family questions	Educational questions
Age	Mothers'Education	Weekly study hours
Sex	Fathers'Education	Reading (non-scientific)
High School Type	Number of Brother/Sister	Reading (scientific)
Scholarship Type	Parents'Relationship	Attendance to Seminar/Conference
Additional Job	Mothers'Job	Effect of Projects and Activities
Sports/Arts	Fathers'Job	Attendance to Lectures
Relationship		Taking notes
Salary		Writing/Listening
Transportation		Effect of in-class Discussions
Accommodation		Effect of Flip Classroom
		GPA of Last semester
		Expected CGPA at graduation

Table 1: Summary of questionnaire, adapted from Table 1 in Yilmaz & Sekeroglu (2019)

(2022), personal data that has been irreversibly anonymised ceases to be 'personal data' or require compliance to Data Protection law. To determine that this data is suitably anonymised, identifiability of the subjects must be ruled out. The data set contains no information about unique identifiers relating to the students such as names, phone numbers, student numbers, birth years/days, or addresses. There is no information that allows for any student to be singled out. Numerical data such as age, salary, and grades are binned which makes identification more difficult. A combination of the Course ID factor along with the information given related about the courses in the original paper may be enough to identify the course of some of the students, but the courses feature here doesn't correspond directly to those mentioned in the original paper, and doing so would involve difficult. Some responses to the personal questions regarding family status may increase risk of linkage between values to identify students, but there is no other public data available related to the family status of students with which to corroborate this information, and re-identification in this manner would not be likely. As a result of all these factors, it appears that it is not reasonably likely for the identification of any of the subjects and given the nature of the topic, identification attempts are also unlikely. Therefore the data can be considered to be fully anonymised, and will not need to be treated as personal data for the purposes of this report, minimising ethical considerations related to data privacy.

3.2 Primary Research

Experimental research has been chosen as the primary research methodology for this project. An experiment is performed to test the hypothesis "Are there traditional data analytics methods that could be preferentially suited to Causal Explanation when combined with Causal inference methods?"

To test this hypothesis, statistical testing is used to test for correlation between the dependent and independent variables, and 2 ML algorithms are used to make classification models. Both are used to produce feature importance metrics/ pseudo feature importance metrics. The most important features as per these metrics are measured for causal effects on the target variable. The ATE measurements are then compared to determine if there are significant differences between the methods.

Establishing causality for this experiment relies on several conditions being met. There is concomitant variation between the different methods for suggesting features and the range of ATEs produced by each group of features. There is a definite temporal sequence of the events whereby the methods were chosen, applied, features were selected based on each of them, and then CIMs were applied to produce ATE's. The experiment has minimal theoretical support; there is no literature that has been found suggesting that any of the traditional data analytics methods should serve as a basis for causal explanation. On the other hand, while features showing correlational relationships with the target variable - either independently of other features (as in the statistical testing) or when combined together with the others (as in the model feature selection metrics) - don't imply a significant causal effect, it is likely that any features that have a causal relationship with the target would also be correlated with it. In terms of non-spurious association, the only factors that could influence the results are the incorrect application of the methods, and aside from that there is no additional factor that should influence the outcomes aside from the choice of method.

Internally, for the experiment to be valid the results of both feature selection from the models and ATE generation based on the CIMS must both also be valid. The design of this experiment gives no guarantees of external validity, and attempts to find such would require repetition of the experiment in different situations.

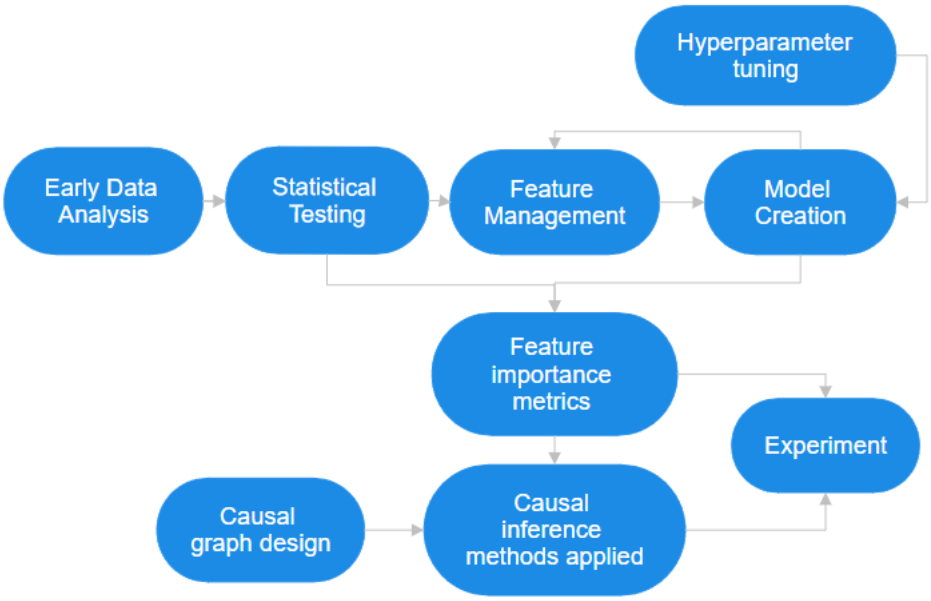


Figure 6: High level diagram of the project framework

3.3 Approach

In order to achieve the objectives of the research project, feature importance metrics must be generated, and ATEs considering various different features as treatment on the target must be estimated using causal inference. To generate the ATEs for various features, Causal inference will need to be applied to potentially causal features. A causal graph must be designed and generated for the data for the application of a CIM using the DoWhy library. Statistical tests will be performed, and and ML models will be trained and tuned to create feature importance metrics as potentially causal features for causal inference. Feature management will need to be performed in a manner that maintains and in some cases increases interpretability for the data, in order to ensure appropriate treatment of the data by the models. The statistical testing will be used to justify the feature management decisions. Early Data Analysis will be used to understand the data and serve as the basis for the justification and evaluation of all other decisions.

3.3.1 Target Feature: Student Grade

In order to better understand and more appropriately treat the target variable, some further interpretation is needed. Turkish grade conversions were used to accompany the categories with percentage ranges between 0 and 100 as visible in Table 2 below. To determine the spread of the variables a histogram of the variable was included (Figure 7). It shows quite an uneven distribution, with a DD grade being most frequent, and FF and CB grades being particularly underrepresented relative to the other categories.

Category	Grade	Percentage Range
7	AA	90.00 - 100.00
6	BA	85.00 - 89.00
5	BB	80.00 - 84.00
4	CB	75.00 - 79.00
3	CC	70.00 - 74.00
2	DC	60.00 - 69.00
1	DD	50.00 - 59.00
0	FF	0.00 - 49.00

Table 2: Corresponding Categories, Grades and Percentage Ranges

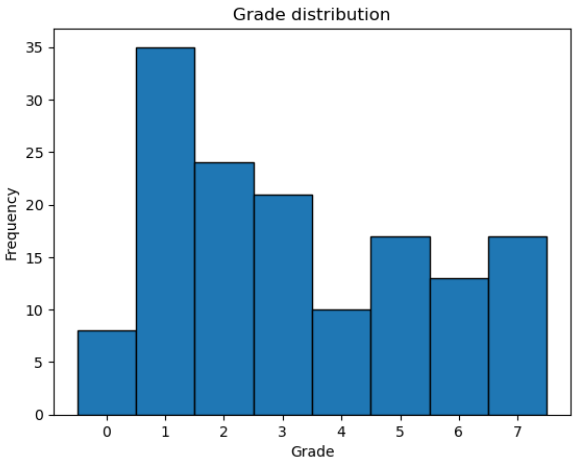


Figure 7: Histogram of the GRADE student feature

There is a question of linearity involved with the treatment of this variable. The encoded 0-7 scale is not representative of the size of the differences between different grades on the 0-100 scale at all. However this isn't wholly inappropri-

ate; the distribution amongst the different grades isn't linear, and the difficulty of increasing from one grade to the next is presumably non linear as well. For design of further research, it might be preferable to use deciles or quartiles of student of grades as bands to predict between.

One major issue with the full 8 category Grade column is highlighted by the initial decision tree algorithm runs. The amount of total points for classification assuming a 25% testing split of 36/145 observations across 8 categories is very low, with on average just over 4 occurrences per category for testing, and often significantly less. Such low amounts make any predictions extremely sensitive to noise, and hinders predicting to any significant degree of accuracy. This issues was further compounded by the uneven distribution of the categories. One choice to help counteract this effect was to increase testing split sizes for the rest of the analysis up to 0.3, providing slightly more values in each category for testing.

This issue is the primary motivator for the re-binning of the target variable from 4 bins into 8. Another motivator is that this balances the distribution of the dataset significantly vs the 8 category distribution, as can be seen in Figure 8. The new percentage bands for the 4 bins can be seen in Table 3. As a goal of this analysis is to ascertain whether certain interventions increase or decrease grades, re-binning the values like this improves the interpretability, as while data and granularity are reduced, the overall ability to predict positive or negative effects on the grade is maintained.

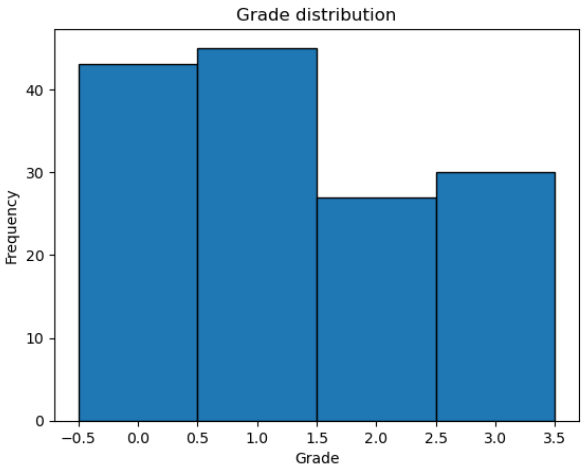


Figure 8: Histogram of the new grade feature created through rebinning

Category	Percentage Range
3	85.00 - 100.00
2	75.00 - 84.00
1	60.00 - 74.00
0	0.00 - 59.00

Table 3: Categories and Percentage Ranges after re-binning

3.3.2 Dimensionality and sensitivity analysis

With such a high ratio of features to observations in the dataset, over 1:5 in the initial set, prior to the addition of even more via the One Hot Encoding of non-ordinal categorical features, there is a definite concern that the Curse of Dimensionality will effect model results. As a result, efforts are made in several stages to justify the removal of features, on the assumption that doing so will reduce the negative effects of dimensionality. Further, to evaluate the trade off between the datasets that contain more and less features, models created using both subsets referred to as small, and large, are compared against each other throughout the analysis.

Models that are largely based on distance metrics, like KNN classifiers, are particularly susceptible to the Curse of Dimensionality, as given enough dimensions, distances between points tend to zero, making the points difficult to distinguish and degrading the performance of the model. (Karanam 2021). SVM models are also susceptible to issues with high dimensionality in the form of "data piling" (Marron et al. 2007). An SVM model has been chosen here with the expectation that a difference in accuracy will be noticed between the models based on the smaller and larger datasets in the case that dimensionality of the data is sufficiently high to affect results.

3.3.3 Feature management

As feature importance is intended to be used going forward, it is important that feature management not include any kind of irreversible feature transformation or reduction methods such as Principle Component Analysis(PCA). While the use of such methods would reduce dimensionality and collinearity by combining heavily correlated features, they would be a large hindrance to the interpretability of the individual features. Various qualities of the features were initially investigated to motivate treatment going forward. One important step

was statistical testing for correlation between features. As the target GRADE feature is in the form of numeric categoricals representing binning of a continuous feature, it was decided to consider the feature as both numerical and categorical for the purposes of statistical testing. ANOVA tests, where the target is taken to be continuous, and Chi-Square tests, where the target was taken to be categorical, were both performed for each feature. The tests generated P values and tested the hypothesis that each feature was correlated with the target dependent variable to a 95% confidence level. The tests were accompanied with box plots and group bar plots respectively for the purposes of visualisation. As the independent variables were all treated as categorical by the tests, further encoding of the values was at this point not necessary. The results and interpretation of these results led to the creation of two subsets of the data. One removing many of the features, referred to as the small dataset, and the other removing significantly less features and referred to the larger dataset.

Certain columns were non ordinal categorical, and would need to be one hot encoded for the purposes of applying ML models. As 1HE greatly increases the number of features in a dataset, this was taken as an opportunity to limit increasing the dimensionality of the dataset by condensing categories together where appropriate. Categories were merged for three of the non ordinal categorical features, 9,10 and 15 (representing transportation methods, accommodation types, and mothers occupation respectively, combining things like "Bike" and "Other" into "Bike and Other", ensuring continued interpretability) on the basis that for some of the categories, there were not enough occurrences (¡3) for a model to appropriately learn a pattern for use in prediction going forward. Consolidating categories like this for features not already marked for 1HE may have proven beneficial, but it was only performed here in the case of the double utility in limiting dimensionality. Two of the binary categories were changed to yes(1)/no(0) categories to improve interpretation by models and readers. Finally, the non ordinal categorical features were One Hot Encoded to enable independent treatment by the prediction models.

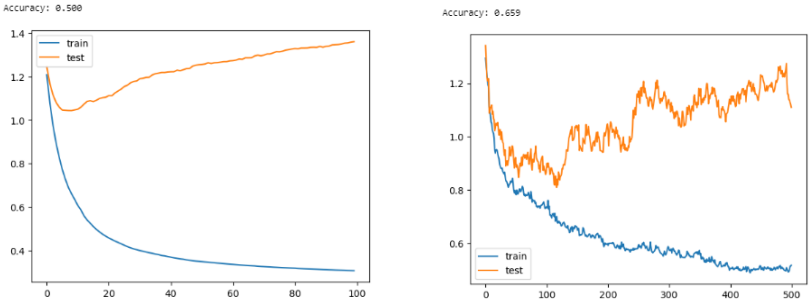
3.3.4 Predictions

Initially, a Decision Tree classifier(DTC) was used to set benchmarks for further models. Models were run with both size datasets, and compared against the full 8 category grade feature as well as the re-binned 4 category grade feature. Heatmaps and cross validated accuracy scores were used in combination to eval-

uate model performance visually and quantitatively. Feature importance scores were produced. As these models were primarily investigative, and as there are relatively few parameters involved with DTC models, hyperparameter tuning was not performed.

Next, a SVM classifier model was used, selected as it could potentially produce evidence of the effect of dimensionality on the predictions. It was performed for both sized datasets, again using cross validated accuracy along with heatmaps to evaluate results. To investigate repeated patterns in the heatmaps, a heatmap for an alternative split of the data was also taken. While coefficient scores similar to feature importance scores do exist for SVM classifiers, they are only applicable for linear kernel SVMs, as other kernels transform the data before creating the model, as described here (BartozKP 2014). Hypereparameter tuning was performed based on recommendations for C and Gamma from (geeksforgeeks 2023) but with the inclusion of several extra kernel options, as the optimal feature space for a model many features was expected to be unpredictable. As there was only 100 possible values with 5 cross validating fits, the GridSearchCV algorithm was used, as it would excecute in an appropriately short amount of time (20sec) and is easily controllable in terms of its values.

Finally, an XGBClassifier method was used. Distributed (Deep) Machine Learning Community (2024) has compiled a list of many of the occasions where XGBoost models have placed first or second in machine learning challenges. Such a widely successful model is a promising choice in this situation as it clearly has very broad applications. As XGBClassifier models typically are tuned by changing many parameters using large grids or random search spaces, but these are too time consuming and computationally expensive for the scope of this project, so standard grid or random search algorithm methods were ruled out. Instead two different approaches combining several methods were used for hyperparameter tuning. Initially, on the smaller dataset, the method of learning curve adjustment outlined by Brownlee (2021) was used to adjust several important parameters. Based on manual manipulation and inspection of the learning curves, an apparent local minimum between 4 hyperparameters was found as can be seen in Figure 9(R). These were then selected as the basis for a gridsearchCV, with the remainder of the parameter grid selected based on suggestions from several tutorials. (Navas & Liaw 2022)(Toth 2024) The gridsearch included over 5000 fits, and was time intensive, but significantly less so than it would have had the learning curves not been investigated beforehand. For the larger set of the data, in an attempt to address some of the issues with the



(a) No tuning

(b) n_estimators=500, eta=0.7, subsample=0.3, colsample_bytree=0.4

Figure 9: untuned and tuned learning curves and associated accuracy generated for the XGBClassifier model using the small dataset

previous approach, a different method was selected: Bayesian hyperparameter optimisation. One advantage of Bayesian optimisation is that it can be explicitly scaled by running for more and more iterations, depending on time available, and it can approach closer and closer to global minima.

The Bayesian optimisation algorithm initially chooses random values from within the search space, but then proceeds to use one of two mechanisms to choose the next set of parameters by evaluating the previous ones. It uses exploitation, to select the points with the highest uncertainty, or exploration, to select a point from the region with the current best results. This informed hyperparameter selection allows for more efficient tuning, and typically finds better hyperparameters in less time compared to other tuning methods. From the tuned XGBClassifier models, feature importance plots were produced using an inbuilt method, and these were based on an f score which measures the frequency of the feature being split in the classification tree.

3.3.5 Causal inference

To implement Causal inference, the DoWhy library was used, as it provides access to a comprehensive framework for causal inference with a range of possible models, via its 4 steps of "model", "identify", "estimate", and "refute". One of the first steps of causal inference is the creation of a causal graph. A causal graph was created based on assumptions about interactions between the most important features and can be seen in full in Figure 10. For simplicity,

this graph was not created with the same rigour as would be appropriate for a true investigation into what causes Student Grades. Doing so would require acquiring much more domain knowledge to more accurately model the causal interactions between variables, and is outside the scope of this project. Several features were assumed to be independent of others (Gender, Age, Mothers Education). It was assumed that other variables caused the two unobserved variables Student Interest and Available Time, which in turn led to Weekly Study, Scientific Reading, and Taking Notes. It was assumed that all factors were partially causal to the students grade either directly or, in the case of Partner, indirectly. All the assumptions in the graph can be viewed in Appendix A: "Causal Graph Assumptions."

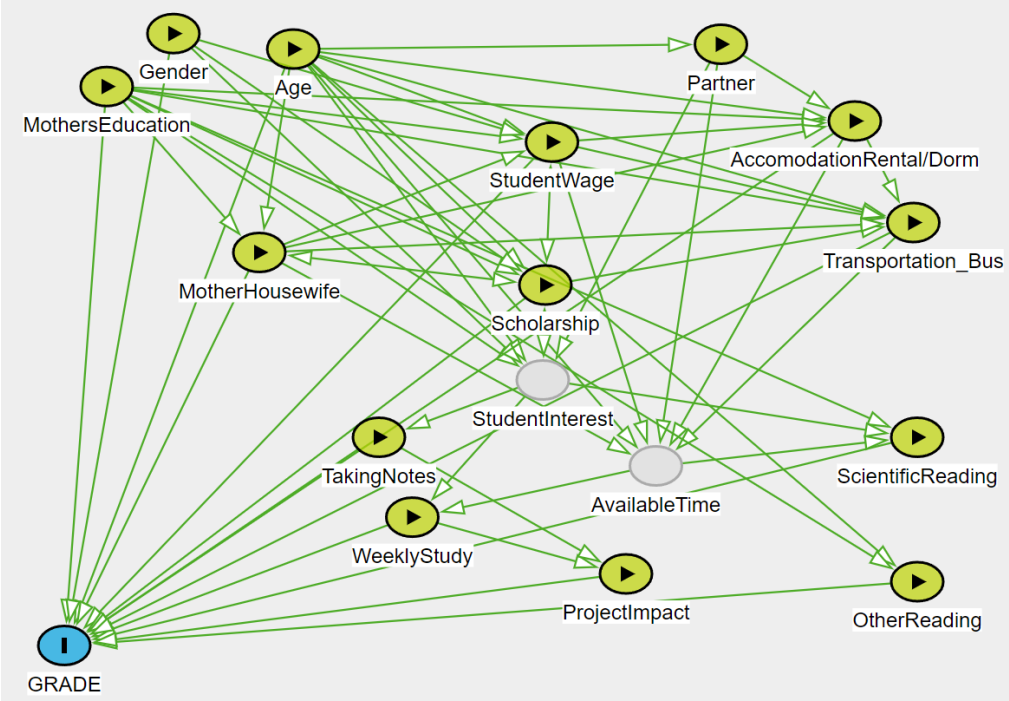


Figure 10: Directed Acyclic Graph for the Data

Using the graph, causal relationships were modeled, identified and estimated between the target and features highlighted through the use of feature importance methods. As no candidate instrumental variables were included in the causal graph, and there is no support in the DoWhy library for using propensity score methods with nonbinary treatments, the linear regression method was

the only one appropriate for estimation for the majority of the features. For the "Partner" feature however, the propensity score methods can be used.

Over the course of this section, the data, and the approaches to solving the various problems identified in the research objectives have been introduced and justified. In the following section, the results of the analyses outlined in this section will be presented and interpreted

4 Findings

In this section, the findings of the analyses and experiment will be described and presented graphically and visually. In the cases where specific results are used to further justify subsequent decisions for feature management or tuning or other things, those justifications will be considered here also. As such there is some overlap in the justification of decisions that make up the analysis between this section and the section on the 'Approach'(3.3).

4.1 Feature Exploration Findings

Looking at the histograms of the data in Figure 11, it's clear that a lot of features are unevenly distributed. From the ML point of view, difficulty can arise when lack of total observations of a category means that any rules learned by the model are more likely to be based on outlier observations and not generalise well to unseen data. As a result of this, some features are noted as highly unbalanced, and interpretation of models must account for potential bias of the model against the minority classes. For several of the features that undergo 1HE, minority categories have been combined, which significantly reduces total class imbalance, but this treatment was not applied to all important variables. Of the variables eventually selected for ML, 18 (Non-scientific reading), 19 (Scientific reading), and 21 (project impact) especially should all be cautiously interpreted.

From the ANOVA tests, only 3 features were accepted as being correlated with the target variable to any statistical significance: 2(Sex), 21(Project Impact), and COURSE ID. From the chi-square test, 3 more: 1(Age), 4(Scholarship), 11(Mothers Education) were found to be significant on top of those from the ANOVA test. 2 visualisations were produced, grouped bar plots, treating the target as categorical, and boxplots (Figure12) treating the target as continuous. The grouped plots are not particularly valuable in comparison to

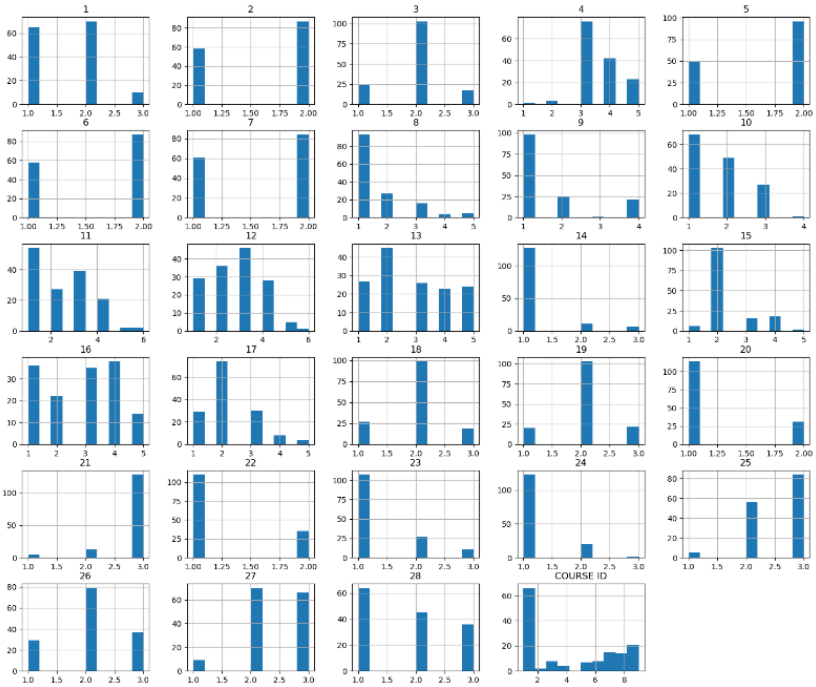
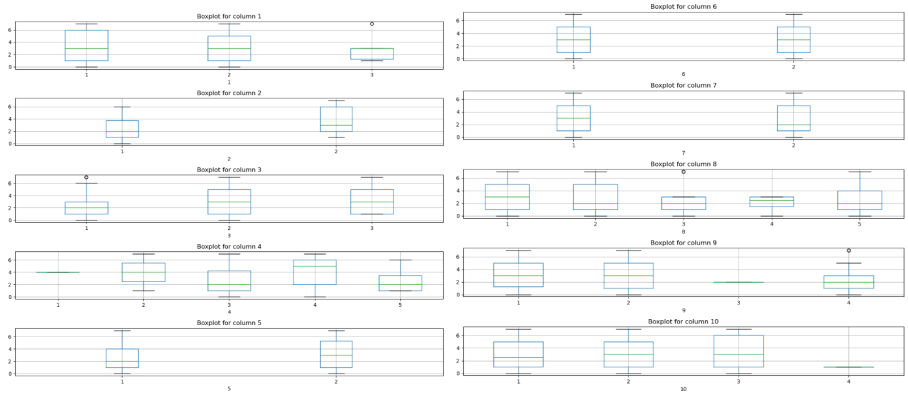


Figure 11: Frequency Distributions for all Features

the boxplots as they don't allow for any overall effect of the categories on the grade on the whole to be observed. The boxplots show that a vast majority of the categories are not very well correlated with the grade, and are distributed somewhat evenly between the categories, which fits with the results from the testing. The plots however do isolate some features, and more importantly some categories, as being better correlated with the target. Course ID especially has uneven plots, though it cannot be appropriately interpreted, and has some plots correlating with high grades whereas some skew towards lower ones. "High" and "Low" here have been defined differently, as the distribution of the grade feature itself is skewed, and both have been selected to capture a smallest categories with more than 1 or 2 different associated grades, as those are assumed to be too under-observed to accurately interpret. For example in feature 1, the third category, (1)Age: 26+, visibly correlates with lower grades. Negative correlations, defined as having the plot max ≤ 4 can also be seen for (8)Total salary: USD 271-340 and USD 341-410, (11)Mothers education: MSc, (15)Mothers employment: Self-employed, (17)Weekly Study hours: 11-20

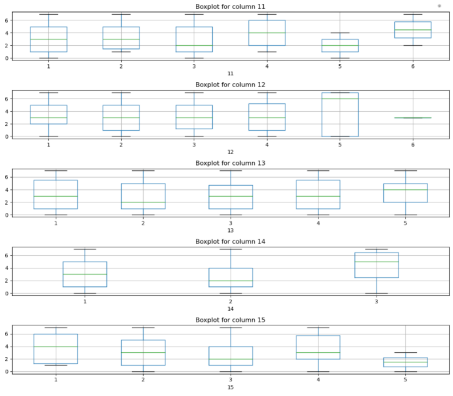
hours and More than 20 hours, (21)Project impact:Negative, and (27)Discussion improves interest: Never. Positive correlations, here defined as median ≥ 5, can be seen for (4)Scholarship: 75% (11)Mothers education: PhD, (12)Fathers education: MSc, and (24)Preparation for mid term exams: Never.

As described in Section 3.3.2, one intention of the statistical testing was to justify the removal of features from the data, thus reducing dimensionality. While this analysis was intended to only remove the least correlated features to contribute to predictions and causal inference, only 6 features passed the tests at this p value. Removing all other features from the prediction would be removing a lot of information from the dataset, and while correlation is low for the rejected variables, the trade off of reducing dimensionality at the expense of ability for models to learn interactions from subtle interactions between features is not worthwhile. Selection of the threshold value alpha is an important step in statistical testing, especially because it is an arbitrary value. Changing the value of α after already seeing the results of the tests in order to change whether the null hypothesis is accepted or rejected and the statistical significance of the data being tested is known as p-hacking, and is a major issue in research whereby researchers can manipulate results such that they are more in line with what the researcher expected to see. Changing the alpha value in the case of this analysis is possible, and would lead to more features being found to correlate with the target, but it could be misleading and potentially unethical to do so. Increasing the risk of false positives for the purposes of choosing features in this scenario however is not a major issue, as these statistical tests are being used for supportive purposes as opposed to conclusive ones. The solution for this is the creation of two datasets: the first one with the strict, original statistical test threshold of p¡ 0.05 referred to as the small dataset, and one with a much broader inclusion threshold of p ¡ 0.2. The results of predictions using the two datasets are compared to evaluate the differences and test the value of reduced dimensionality vs the inclusion of low correlation features. The larger dataset based around the 0.2 α includes the following features: 1(Age), 2(Sex), 4(Scholarship), 7(Partner), 8(Salary), 9(Transportation), 10(Accommodation), 11(Mothers' education), 15(Mothers' occupation), 17(Weekly Study), 18(Non-scientific Reading), 19(Scientific Reading), 21(Project Impact), 23(Midterm Study), and 25(Note Taking).

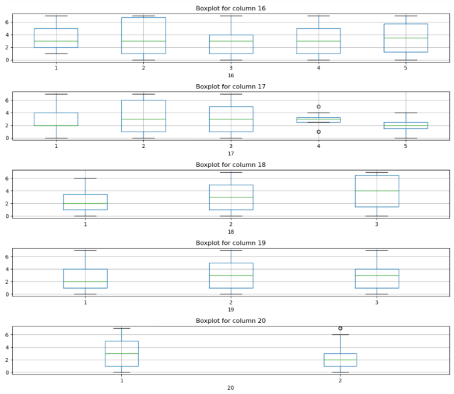


(a) 1-5

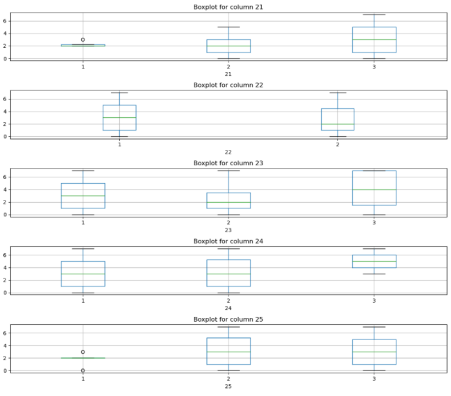
(b) 6-10



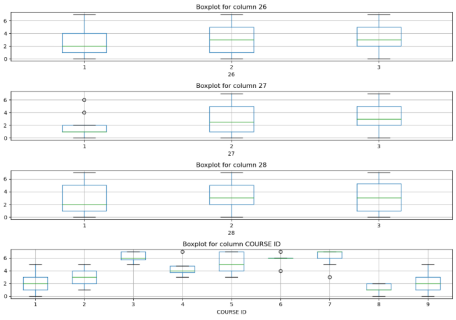
(c) 11-15



(d) 16-20



(e) 21-25



(f) 26+

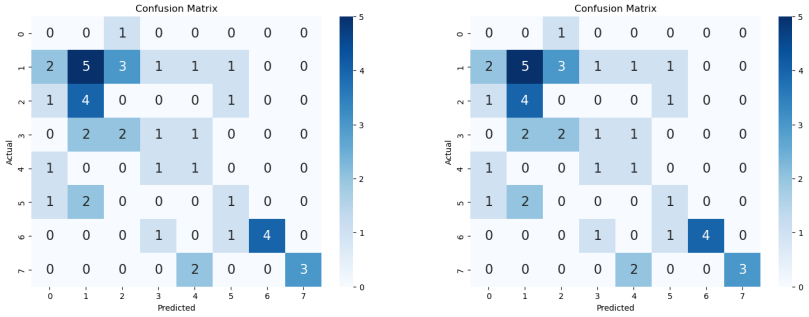
Figure 12: Boxplots for features 1-27 and Course ID

4.2 Prediction Results

4.2.1 Decision Tree Classifier

	Small Dataset		Large Dataset	
CV Accuracy	Mean	StDev	Mean	StDev
8 Grade bins (k=10)	0.303	0.141	0.275	0.078
4 Grade bins (k=10)	0.413	0.108	0.406	0.137

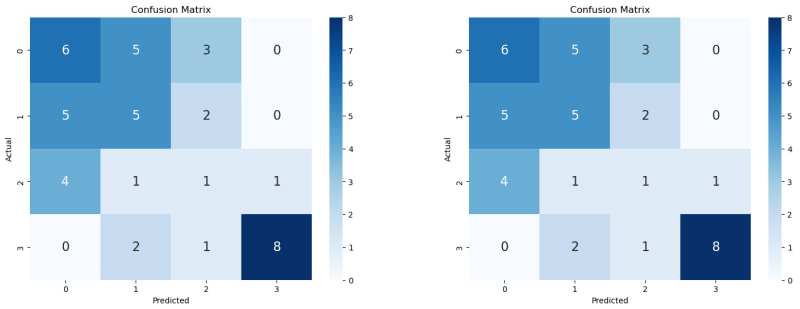
Table 4: Cross validated accuracies for Decision Tree Classifier model



(a) Small Dataset (b) Large Dataset

Figure 13: Heatmaps from DTC predicting 8 bin grades

The goal of implementing the DTC models was primarily investigative. Table 4 shows the Cross validated mean and standard deviation accuracies for the model in 4 domains. Contrasting the accuracies of the 8 bin and 4 bin results shows a marked increase. This is an expected result, based on there being less possible opportunities for mis-classification, and the overall simplification of the target variable. As a result of this, and because the trend of whether a grade is lower or higher due to the independent variables is more important than the explicit category, the 4 bin grade category is used for all further predictions. Differences between the mean CV accuracy between the two sizes of dataset are quite small, and considering the relative size of the standard deviation and the general sensitivity to the train/test split, the differences between the two sets of predictions can be taken as insignificant. This implies that there is little additional accuracy gained as a result of the inclusion of more features in the larger dataset with this model.



(a) Small Dataset (b) Large Dataset

Figure 14: Heatmaps from DTC predicting 4 bin grades

Feature importance of the (encoded) features created from the DTC classifier can be found in Figure 15. This particular metric for feature importance gives each feature a score denoting its contribution to the total predictions, with the combined scores adding to 1. For the small dataset model, features 4(Scholarship Type) and 11(Mothers' Education) were two of the most contributing to the predictions, while in the big dataset model, 17(Weekly Study Hours), 18(Non- Scientific Reading) are some of the highest, followed closely by again 11, and 7(Relationship). The Course ID feature dominates both models quite significantly however as mentioned before, the Course ID feature has been encoded for anonymisation purposes and will not be of much value for interpretation or causal inference, though it can be said that students from different courses received significantly different grades on average.

1	: 0.068
2	: 0.068
4	: 0.219
11	: 0.253
21	: 0.059
Course ID	: 0.334

1	: 0.020
2	: 0.048
4	: 0.027
7	: 0.074
8	: 0.045
11	: 0.088
17	: 0.136
18	: 0.104
19	: 0.029
21	: 0.050
Course ID	: 0.202
9 cat _1	: 0.054
10 cat _2	: 0.031
15 cat _1	: 0.018
15 cat _2	: 0.028
23 cat _1	: 0.031
23 cat _2	: 0.016

Figure 15: Feature importances extracted from the DTC classifier model for the small and large datasets (including 1HE)

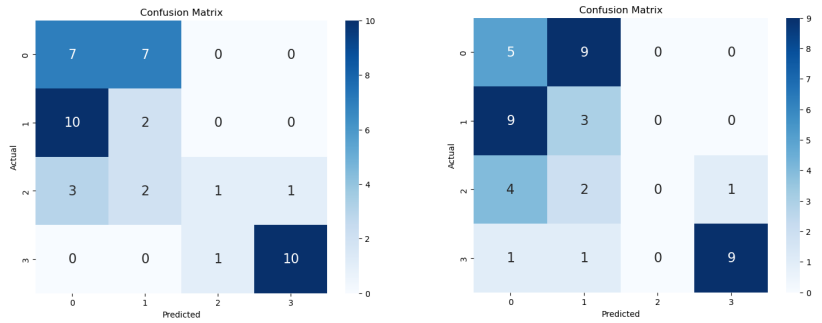
4.2.2 Suport Vector Machine

	Small Dataset		Large Dataset	
CV Accuracy	Mean	StDev	Mean	StDev
Default parameters (k=10)	0.331	0.119	0.344	0.180
Hyperparameters tuned (k=5)	0.446	0.073	0.504	0.093

Table 5: Cross validated accuracies for SVM Classifier model

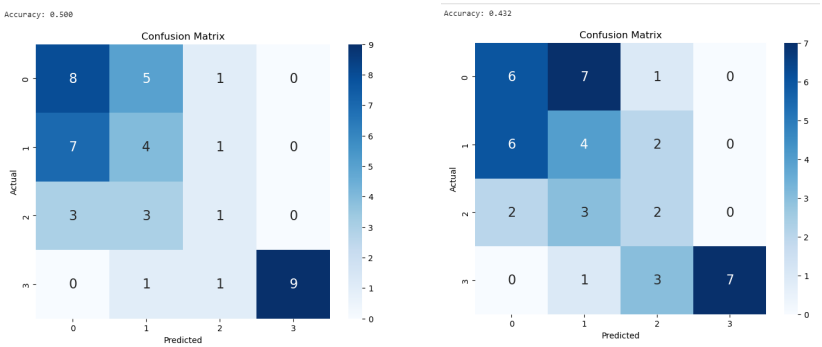
One goal of using the SVM model was to demonstrate the value of hypertuning models through comparison with untuned ones. As can be seen from Table 5, the differences in accuracy are quite significant, around 34% & 46% increases in accuracy for the small and large dataset models respectively. Compared to the DTC models, the tuned SVM is notably more accurate, though the difference is not particularly large and the DTC models are completely untuned. Another observation is that the hyperparameter tuned models have significantly smaller standard deviations of accuracy. Through finding the most optimal parameters in the grid, the model has found parameters that produce the least poor results, leading here to a tighter clustering of accuracies. It isn't clear whether

increasing k in this case would reduce or increase the stdev. There is not a significant difference between the error patterns of these classifications compared to either the DTC classifiers, or with each other. The only one that could potentially be significant is that the untuned models underpredicts category 2 more significantly than the others.



(a) Small Dataset (b) Large Dataset

Figure 16: Heatmaps from SVM default parameters

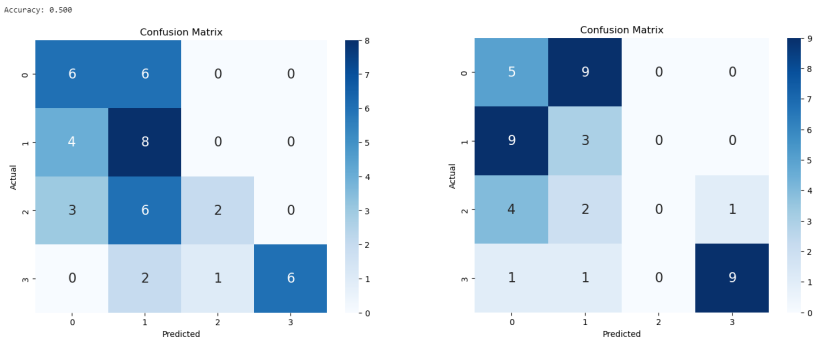


(a) Small Dataset (b) Large Dataset

Figure 17: Heatmaps from SVM after Hypertuning. Accuracies for specific split included

For the hypertuned models, there is a moderately significant difference between the accuracies using the small vs the large dataset.

To evaluate the dependence on the training/testing split being used, the



(a) Alternative split (b) Default split

Figure 18: Heatmap of an alternate split vs the default split used for the SVM with the Large Dataset (no tuning)

separate split of the data was compared with the arbitrarily chosen default split in Figure 18. While the alternate split seems to have a bias towards predicting category 1 as opposed to category 0, and seems to be slightly less accurate for category 3, it still follow the same broad pattern, in that it underpredicts category 2 and frequently misclassifies categories 0 and 1. As a result it can be assumed that these aspects of the heatmaps are not just a function of split.

4.2.3 XGBoost Classifier

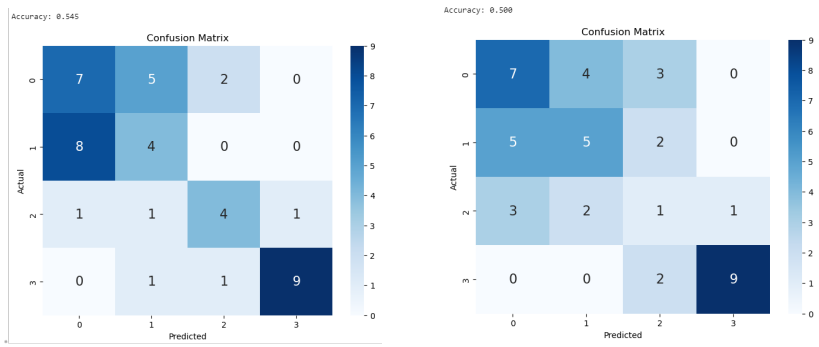
With just the default parameters, and increase in accuracy vs the previous two type of untuned models can already be seen. Note that for the visualisation of larger dataset on the used split in Figure 19, the majority of category 2 observations are predicted correctly, better than for any other model. XGBoost models have large numbers of parameters and hypertuning such models can be quite computationally/time intensive. As a result two strategies for hyperparameter tuning have been used. For the model on the small dataset, creating a heatmap based on the hyperparameters seen on the right in Figure 9 generated by the manual learning curves , Figure 20 was generated. This is the highest accuracy obtained by any model on the chosen split, and as can be seen in the heatmap it correctly predicts a majority of the data points. However, the intention was to combine the learning curve tuning with gridsearchCV in order to produce a cross validated accuracy for the model, and it turns out that this was not successful at all in significantly increasing the accuracy, with a very similar cross validated accuracy to the default model.

For the bayesian hyperparameter methods on the other hand, there was a noticeable difference. It is also important to mention that increasing the number of iterations for the bayesian model notably increased the maximum CV accuracy of the model, for n=10, best score = 0.448, for n = 100, best score = 0.483, and for n=200, best score = 0.497. There were issues trying to run the model for n=200 and it was very time consuming relative to the project, so further increasing the accuracy via running the model with higher n wasn't pursued.

	Small Dataset		Large Dataset	
CV Accuracy	Mean	StDev	Mean	StDev
Default parameters (k=15)	0.426	0.167	0.419	0.176
Learning Curve and GridSearchCV tuned (k=5)	0.432	0.164	-	-
Bayesian Optimiser tuned, n=200 (k=5)	-	-	0.497	-

Table 6: Cross validated accuracies for XGBClassifier model

The feature importance plots produced for the XGB model using the large dataset(Figure 21) and the smaller dataset(Figure 22) contain only the features included for their prediction. The plot with less features shows greater magnitudes for the f score, and amongst the shared features there are notable



(a) Large (b) Small

Figure 19: Heatmap of the untuned XGB classifier models

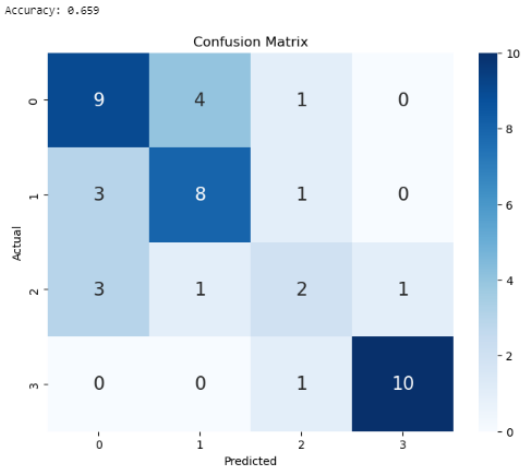


Figure 20: (Highest Accuracy) Heatmap based on learning curve tuning method alone

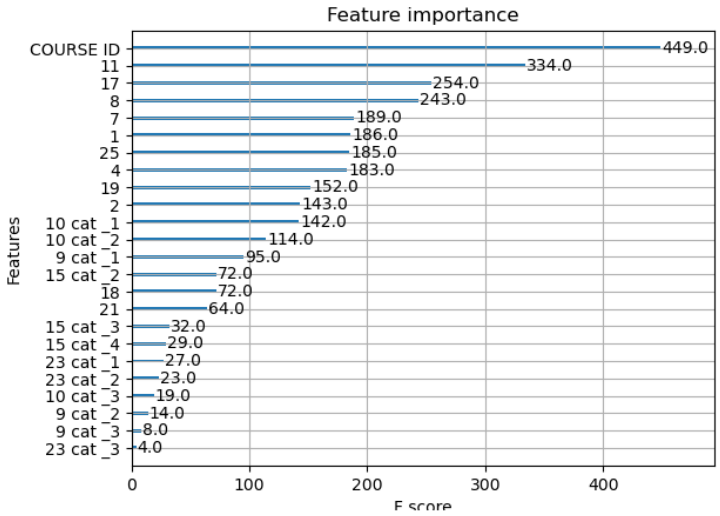


Figure 21: Feature importances from the hypertuned XGBClassifier on the larger dataset

differences in the order of features. For the purposes of modelling causality, the causal graph was created based on the most important features from the plot based on the larger dataset. A minimum of 50 F score is used as a threshold for feature inclusion. Features from the original that were not included in the creation of the ML models in the first place are also excluded. It is assumed that the excluded features don't have sufficient effect on the target variable to warrant inclusion. This assumption is very unlikely to be correct, however it serves an important purpose in allowing the causality of the scenario to be modeled without excessive detail in the causal graph. The incorrectness of this assumption does render the resulting causal inference itself significantly less valid, but the assumption is stated so the causal inference model can be evaluated on that basis.

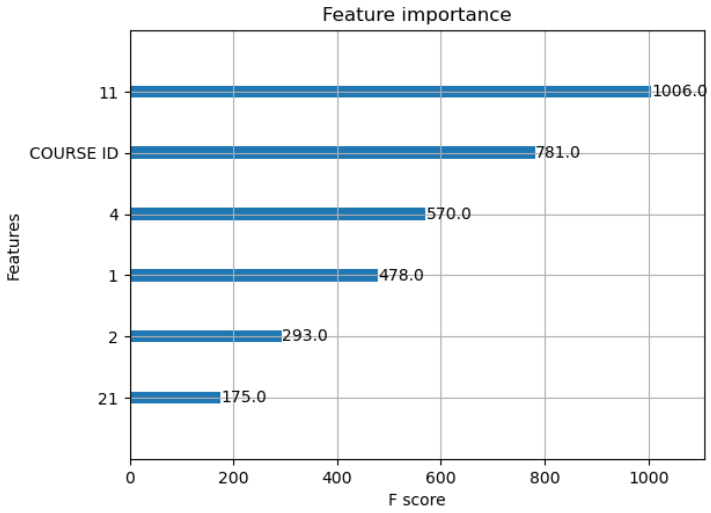


Figure 22: Feature importances from the hypertuned XGBClassifier on the smaller dataset

4.3 Causal Inference

For one of the features used, WeeklyStudy, it was tested whether using the 8 bin or 4 bin Grade as target made any difference to the outcome of the causal inference. It was found that there was no difference when predicting the ATE using the 8 bin or 4 bin grade feature as the target. This result represents the total magnitude of the grade being cut in half, but the distances between the bins doubling proportionately. Another one of the features, Partner, is in the form of a binary treatment, and so propensity score methods can also be applied to it. 7 shows the results of applying two different propensity scores as well as the linear regression estimation methods. The effects are all small, but differ significantly relative to their size, as well as their direction. Some variance between ATE for different methods is not uncommon, as described in Parikh et al. (2022) and shown in Figure 5. However especially given the the low absolute magnitude of effect as well as the direction of causality being flipped in the ATE for the stratification method, there is cause to doubt the validity of these results.

Across all potential "target" features, ATEs were measured using the Backdoor Linear Regression estimation method and can be seen in 8. Positive and negative effects can be seen ranging in absolute magnitude from 0.018 to 0.672.

	Propensity Score	
Linear Regression	Weighting	Stratification
0.018	0.063	-0.008

Table 7: ATE for the "Partner" feature comparing linear regression, propensity score weighting and stratification

Treatment variable	Category #	Average Treatment Effect
Mothers education	11	0.082
Weekly study	17	-0.076
Student wage	8	-0.161
Scholarship	4	0.079
Non-Scientific reading	18	0.383
Partner	7	0.018
Gender	2	0.672
Project Impact	21	0.383
Age	1	-0.273

Table 8: ATE for various features using the linear regression causal inference method

4.4 Experiment

To determine which traditional data analytics method was most useful as a basis for causal explanation, mean and Std Dev of absolute treatment effects for the top 4 or 5 "highest importance" features as treatments will be compared based on the Chi Square test, the DTC, and the XGBClassifier. The feature importance methods of the models trained on the large dataset will be used, as the features included in the smaller dataset are indistinguishable compared to the Chi Square test. The Course ID feature will again not be included as it isn't interpretable as discussed in Section 4.2.1. The Chi Square test highlighted features 1, 2, 4, 11, and 21. The DTC highlighted 7, 11, 17, and 18. The XGBClassifier highlighted 1, 7, 8, 11, and 17. There is some amount of overlap between the three. As the distributions are not assumed to be normalised and have different lengths, a Kruskal-Wallis test is applied to the data. The test returns a p value of 0.302 which is above a default alpha of 0.05 and thus the null hypothesis, that the is no statistically significant difference between the medians, cannot be rejected. This is not sufficient evidence to say that any one method for suggesting treatment variables outperformed the others.

Over the course of this section, the results and findings for the Feature Exploration, the DTC, SVM, and XGBClassifier models, the causal inference, and the experiment have all been presented, and interpreted. In the following section, the findings of the project will be discussed and evaluated with respect to the research objectives.

5 Discussion

In this section, the results from the traditional data analytics methods as well as the causal inference methods will be discussed and further interpreted, and evaluations of the results will be performed with an emphasis on how the results achieve the research objectives(or not).

5.1 Traditional Data Analytics Methods

A major component of evaluating traditional data analytics methods is evaluating the methods used to describe the data, such as the generation of box plots. Some of the relationships noticed in the box plots (figure 12) would be expected intuitively, like never taking notes and project work having a negative impacts on the success of a student, or parents being better educated, but some effects are quite counterintuitive and bring the interpretations into question: Do Mothers with Masters degrees reduce student grades? Do students who never prepared for their midterms succeed while those who studied the most did worse? While these effects may not be significant when combined with analysis of the entire dataset as a whole, another possible reason for these correlations is underobserved categories, with outlier values that don't represent the population distribution influencing the plots. Of these categories with visible effects, Age), Total Salary(For the 341-410 category), Mothers Education(for both categories), Fathers education, Weekly Study Hours(for both categories), Project Impact, and Preparation for midterm exams, are all particularly underobserved, with the frequency plots putting the observation of their specific categories as $\leq$ 10. As a result, it can be said that the majority of correlated categories detected in the box plots are underobserved. This isn't entirely unexpected, as rarer categories associated with edge of distribution events are likely to be related to edge of distribution grades, but the possibility that these values are caused by outliers that are not generalisable to the population remains. The creation of these box plots as a form of descriptive data analytics is quite valuable pre-

cisely because it shows that most of the features involved do not independently correlate with the target feature, and in combination with the frequency plots, the categories that do appear to correlate with the target can be appropriately interpreted as possibly being correlated with the target and possibly being a function of underobservation and bias of the sample. All of this combines to give quite a solid interpretation of the independent feature relationships from a descriptive perspective. However evaluation of interactions with the target feature independently are inherently limited, as they can't account for combined effects, especially in this case for observations with so many different features. ML classifiers on the other hand are in theory much better equipped to deal with the problem.

For the statistical testing, despite two types of tests being performed along with their corresponding visualisation, only the results of the Chi-square test were particularly useful. This is in part because the set of features that were found to correlate to the target in the ANOVA tests was a subset of the features in the Chi square test and provided no new information, but also because the treatment of the target feature as numeric made much more sense: knowing if a change for a feature correlated with them getting between 75% and 85% is much less valuable than knowing that the change in the feature resulted in a quantifiable increase in the student score. By measuring the association between the features and the grade, the test creates a quasi feature importance metric which effectively describes the independent relationships. In the case where there was significant validity to the ATEs and consequentially the experiment, the power of this method as a feature importance metric could be compared to the other methods, but this is unfortunately not possible.

It was intended for the SVM method to be used to test for the effects of dimensionality, as it has been known to affect SVM models, hypothesising that models created with the larger data set might suffer in performance due to the Curse of Dimensionality. This was not in fact the case, as the larger dataset, with almost 4x the number of features, outperformed the smaller dataset by $\sim$13%. This strongly implies that the Curse of Dimensionality has not come into effect here. The model predicts slightly better when using the extended dataset. It is unclear whether this means that there are no negative effects present from dimensionality, or if the accuracy of second model is better despite effects from dimensionality, but based on the results for the XGB model where similar increase in accuracy was seen for the model trained on the larger dataset and there is no expectation of an influence of dimensionality, the former seems

more likely.

While the decision to run the analysis separately for both large and small datasets was in part motivated by determining the effects of dimensionality on the dataset, it also enabled a significant and valuable sensitivity analysis of the data. In terms of predictions, as can be seen in Table 9, untuned models saw almost no differences between the large and small datasets, in two cases favouring the smaller dataset over the larger one by a very small margin. For the tuned models there is a preference for the models trained on the larger dataset, albeit still quite small, ∼10-15% of the overall accuracy. An increase for the tuned model is expected, as hypertuned models are inherently more complex and will pick up on more subtle patterns in the data, but it is still true that having 4x the data complexity for the second dataset didn't improve predictions by a significant margin.

Model	Change for untuned model	Change for tuned model
DTC	-0.007	N/A
SVM	0.013	0.058
XGB	-0.007	0.065

Table 9: Differences in CV accuracy between the large and small datasets for all three models

A significant result across all the predictions is that in most cases the standard deviations of the mean accuracy are frequently between 1/2 and 1/3 of the mean accuracy itself. This reveals that the models are quite sensitive to the split of the training and testing data, and implies that these models will likely generalise poorly. This is typical of predictions made with datasets that are lower in data-points, as low numbers of observations for given categories will limit the ability of the models to learn and facilitate overfitting. These results are expected given the size and nature of the dataset and emphasise the importance of cross validation in evaluating accuracy for classifications based on smaller datasets.

Analyzing the heatmaps generally, shows that while the 4th(3) and highest grade category is fairly well predicted, the 3rd(2) is misclassified quite significantly. According to Table 3, the third bin(Category 2) contains the least integer percentages, at 10 compared to 15, 16, and 60 for the other bins. The assumption of an uneven grade distribution seems appropriate but it may not be entirely true and result in this effect. The distribution of the 4 bin grade fea-

ture as can be seen in Figure 8, where it can be seen that the third bin contains the least values, though not to the same degree that the total percentage points included at the bin varies from others. The smaller total number of points for the bin leads to less potential patterns and more outliers for models to learn, especially in cases where training testing splits reinforce this. However this effect is not seen at all for the best predicted 4th bin, which is also underrepresented in the distribution (though not by as much), and other mechanisms likely explain this discrepancy. For the first two bins, there is a lot of misclassification between them. All these patterns in the heatmaps generally hold for all models. The heatmaps provide quite a valuable insight into how prediction is happening and enables for additional evaluation of the prediction.

Using cross validation as a mechanism to ensure some level of generalisability for the models, predictive power was not found to be high for any of the models, with CV accuracy ranging from 0.33 to 0.5 for all the models built using the 4 column grade variable. Because the prediction is between 4 categories as opposed to 2, there are more opportunities for the model to incorrectly classify data. While a 2 category classifier operating purely on chance would expect to see 0.5 accuracy, for a classifier with 4 categories that expected value is 0.25, meaning that the highest accuracy models produced here are significantly more accurate at prediction than chance. From the perspective of prediction as a traditional data science task, the models at best can correctly predict what grade range a students grades range will fall into about half of the time. These models could have some mild practical uses, but they would not be able to encourage any actions or interventions to take for a student to increase their grade, merely predict them in advance.

For the SVM models, significant increases in accuracy were seen with the addition of hyperparameter tuning (Table 5). This corresponds to a significant increase in complexity of the model but also generalisability. One reservation about this is that these improvements may just be a function of lower number of k fold splits in the data, and the closer groupings of accuracy according to the Std Dev are also a function of the split. But while higher k results in more total chances to produce a model with outlier accuracy, there is no reason to assume that the splits chosen under high k would necessarily result a less balanced distribution of accuracy overall than those selected under high k, as the increase in probability of outlier splits is proportional to the increase in near-to-mean splits. Further manipulation of k to determine if this holds true could prove insightful.

The XGBC model with the smaler dataset saw a minimal increase in accuracy for hyperparameter tuning with the combination of the learning curve method alongside GridSearchCV. This result, especially in comparison with the accuracy achieved from learning curve tuning without CV, implies that the additional accuracy gained by manual tuning was not generaliseable, and was a result of overfitting. Further, while high variance in maximum accuracy cross validated scores, implying that the hyperparameter combination underfitted some models and overfitted others, was typical behaviour, the inability of the gridsearchCV to find any combination of hyperparameters beyond the ones set with manual tuning points to those hyperparameters being particularly not generaliseable, with a significant majority of the k fold splits underfitting compared to the one used to tune the learning curve model for every iteration of the grid search. From a time consideration perspective, the computation for this method was much shorter than the bayesian tuning even for lower n, but considering the time investment to manually test hyperparameters for the learning curve and the negligible increase in accuracy this improvement on time was not a good trade off. This interpretation of the results would imply that the combination manual and gridsearch tuning method is neither robust nor effective. However, in circumstances where computation costs for any searching algorithms were especially high, and for data that wasn't as prone to overfitting as this, it is possible that this method for tuning would be more comparable in value to the Bayesian tuning The model based on the larger dataset saw a higher increase in accuracy for the Bayesian hyperparameter tuning, with a positive relationship being found (Table 10) between n (number of iterations in the tuning instance) and the cross validated accuracy score for the best parameters. There appears to be some kind of diminishing returns for accuracy with regards to n, and it would be valuable to create a plot of the relationship, however this would be computationally quite intensive to create, especially for exponentially higher n where the highest accuracies would be found. Computational intensity however also increases proportionally to n, and is the cause for this project not attempting to achieve higher accuracy for n ¿ 200.

This scaling power of the Bayesian hyperparameter tuning combined with its mechanisms of exploration and exploitation to progressively make better "educated guesses" gives it a lot of potential in comparison to other hyperparameter tuning methods. It is also vastly superior to scale the complexity of the algorithm merely by increasing one parameter rather than manually adding more granularity to a grid search instance. This project however has missed an

n	Max CV accuracy
10	0.448
100	0.483
200	0.497

Table 10: Scaling of accuracy with n for Bayesian hyperparameter tuning on the XGBClassifier with the large dataset

opportunity to compare the Bayesian tuning method directly with methods like grid search and random search from a computation and accuracy perspective.

The feature importance methods derived from the two models have some descriptive power, but it must be acknowledged that the features they determine as important differ significantly. There was minor overlap with the features from the statistical testing, which is to be expected as by capturing interactions between variables, different variables could be found to be important. The DTC model produced several features also highlighted by the XGBClassifier (7,11,17), made up of an ensemble of decision trees, but there were others that didn't (1,8,18). This is an unexpected finding as while one model is more complex than the other, they are both trained on the same data, and the differences in accuracies between the two were not very large. The finding that different models can produce similar accuracies based on modelling drastically different patterns in the data highlights the value of analysing feature importances to learn about the variables and their relationship to the target rather than simply creating a model and deploying it to predict the target variable for unseen data, which gives no insight into the variables themselves.

The binning of the grade variable has been discussed, and while it results in an interpretively unimportant loss of granularity, the increases in prediction accuracy as can be seen in Table 4 are inherently quite valuable for the rest of the project. This was an good decision, however looking at the patterns of the heatmaps suggests that there were few overall rules being learned by the models in relation to predicting the third bin, and potentially a different breakup of the grade feature might have lead to much better predictive accuracy, such as a three or two bin split. Further advantages would be expected but tradeoffs would also need to be considered and evaluated.

One of the issues with the feature management decisions taken throughout the project was their sequential nature, they were generally applied in a manner mirroring a "greedy" algorithm, in that locally optimal decisions were propa-

gated throughout the rest of the project without being returned to. One aspect is that little of the EDA that was applied prior to re-encoding of the dataset was repeated on the encoded data, which could have provided an updated correlation plot as well as more accurate grouped bar and frequency plots. Another is that the box plots and grouped bar charts were not performed for the new 4 bin grade column, although this would probably only lead to better interpretation for the grouped bar plots, reducing the granularity of the box plots that was important for analysis. Also when features were dropped for various justifications, while the specific decision was later evaluated, there was no design element to the project whereby these decisions could be reversed, and applied again with different starting conditions as a counterfactual analysis compared to the original decision. Specifically for the splitting of features into two datasets as a result of the statistical testing, it may have been beneficial to have returned to this splitting once an ML model had been selected, and to have tested a much greater variety of subsets of the data, which could have found the most optimal feature combinations for prediction. On the whole, the feature management decisions maintained interpretability over the project, as there was no loss of information as to what any of the features mean through to the final stage of the analysis.

5.2 Causal Inference

As an attempt to infer causality and to compare the causal effects for various features on the target student grade variable, the results of the project highlighted quite a lot of the limitations to the approach chosen to infer causality. Taking data created for one purpose and trying to use it for another purpose is inherently inferior to designing data collection methods with the goal of the research in mind and based on a significant amount of domain knowledge. For this project, the attempt was made to infer causality without specialist domain knowledge, relying instead largely on intuitions related to student performance and relationships between the collected variables. Also, the graph was designed based only on features deemed to be important from the prediction models, as opposed to considering the question of "What possible variables are likely to have the strongest causal effect on the target" as the starting point for the approach and coming up with features based on that. While causal inference is inherently imperfect, and all created causal models are based on assumptions that may not be correct, increasing the number of assumptions and using less

well founded assumptions combines to significantly increase the error in the measurement of treatment affects, and reduce the validity of the causal model.

In creating the causal DAG graph for this project, many assumptions were made that are not likely to be accurate. The first being that all the features were assumed to be directly or indirectly considered to be causal to the target, especially the fact that many are assumed to be both. In Huntington-Klein (2021) Chapters 6, unimportance is listed as a vitally important stage of in the creation of causal graphs, and this was not taken into account in the creation of the graph. Many relationships that are more likely to be either negligibly causal, or merely correlational, were added to the graph as causal. Previous results could have been used to infer whether there were strong enough relationships between variables and the target to imply the possibility of a significant causal relationship, but the prior results were not leveraged to this effect. While there was no guarantee of finding any significant causal relationships in the data, review of the approach taken for causal analysis points to the results from it being invalid.

It is difficult to know what kind of ATEs would have been measured if the causal graph had been modeled differently and based on more robust assumptions. The results of the causal inference may have been showed similar patterns but possibly with greater magnitude, or it could have shown completely different patterns. Taking the results at face value, which is appropriate if the former possibility is assumed, it is seen that Gender, Non-Scientific Reading, and Project Impact are the three "most causal" of the features based purely on ATE. For Gender, an increase of "1 unit" corresponding to an increase in their grade of 0.672 along the scale in Table 3 represents being male vs not being male. for non scientific reading, an increase of one position along the scale "1: None, 2: Sometimes, 3: Often" corresponds to an increase of 0.383 on the Table, while an change of 1 unit along the scale of "1: positive, 2: negative, 3: neutral" results in the same increase in grade for the "Impact of your projects/activities on your success" question. This feature was re-encoded differently for the "traditional" analysis so as to have some level of ordinality, and not doing so for the causal inference is suboptimal for interpretation, but the result can still be understood - either going from positive to negative or negative to neutral, or both, contribute to increases in the grade of a student. It is important to note that these results have very low validity and interpretations are useful to be able to do, the findings of the interpreted results should not be expected.

Assuming the project was repeated with an appropriate causal graph, it may

have been beneficial to use multiple types causal methods and compare between them. To comply with the restrictions of the DoWhy library for propensity score matching on binary treatments only, the data could have been engineered to create binary treatments, allowing for the use of a more diverse range of scores. creating something like "Mother third level education: y/n" may have been appropriate in this regard, as would several others for the other treatment variables. Alternatively, the nonbinary categorical potential treatments could be analysed using a threshold crossing approach, or discontinuity analysis, by creating several binary thresholds from the treatment feature and combining multiple causal inference instances created with them to analyse the resulting pattern.

While the experiment based on the causal effect measurements will inherit any validity issues from the causal inference, the results or potential for results can still be considered. The Kruskal-Wallis test was used as evidence to determine whether there was any stand-out method of feature importance metric generation for exploratory causal analysis as described in the first secondary objective. The test failed to provide evidence of any of the methods outperforming the others, which is unsurprising given the magnitude of the ATEs was so small, and much of the features were shared between methods. however, purely based on mean ATEs, the Chi Square test seems to produce slightly higher average magnitudes, with a mean of 0.298, compared to 0.122 and 0.140 for the XGBClassifier and DTC models respectively. While this is not a statistically significant result according to the test, it is noticeable. It may be possible that a statistical test that compares means as opposed to medians such as Welsh's ANOVA, might have provided evidence to suggest that there was a standout method of feature importance generation. If proceeding results were more valid, it might be appropriate to infer that the Chi Square test is more valuable for exploratory causal analysis than the feature importances generated by the models. The question warrants significant further analysis.

6 Conclusion

This section will conclude the project and deliver verdicts on how the objectives of the research were achieved, outlining inherent limitations with the goals themselves as well as issues that arose during the handling of the project itself and their possible solutions.

It was a consistent occurrence that when an irregularity was observed in the analysis, it was related to a feature that had a very uneven category balance. While the rationale that this could either be due to irregularities inherently being rare or outlier values being able to disproportionately influence the category, it is clear that this issue causes issues with the reliability of the results in the project. This is not an uncommon problem, and a review of solutions to deal with class imbalance (Abd Elrahman & Abraham 2013), outlines several methods as having potential to be useful: Under and over sampling; cost sensitive methods; recognition based methods, and ensemble methods. For further investigation of this data it may be prudent to select and utilise one or several of these methods.

For the predictions of the dataset, and the analysis of feature contribution to those features, some significance can be assigned to the results, especially in cases where features were found to be important based on multiple sources. Mother's education is found to be one of the most important variables for the majority of feature importance metrics, as seen from the chi square test, the DTC and XGB feature importances for both the smaller and larger datasets. Irrespective of whether correlation with the students grade is taken independently or in combination with the other features, the level of education for a students mother is robustly positively related to their own grades. From a causal analysis standpoint, as the temporal sequence of these events rules out causality from the grade to the mothers education, it is likely that either increased mothers education level has the causal effect of increasing grades, or some other confounding variable is causal to both variables, such as something like multigenerational socioeconomic status, or mothers IQ. Despite the robustness of this feature across feature importance measures, there is still the possibility of this finding doesn't represent reality and is merely an artifact dependent on outlier-from-population data within underobserved categories as seen in the frequency distributions (Figure 11).

Upon review of other features, it has found that from the point of ML modelling onwards, the Course ID feature was mistakenly included without One Hot Encoding, leading to the models treating the variable as ordinal, while there is no reason to believe ordinality in the feature itself based on what little information there is on it. While there are no resulting issues with the statistical tests as these both involve treatment of the independent features as categorical, and it was found to correlate with the dependent variable and was as such included. It was understood that the feature had no useful interpretation, and so it was ig-

nored for the purposes of maintaining interpretability across One Hot Encoding, and included in the models without appropriate transformation. This mistake is quite significant to the rest of the results, as according to the feature importance plots the Course ID feature, and the false ordinality produced by the likely random allocation of courses to different integers in the initial data aquisition, dominated the ML models. This puts quite a lot of the results based on the models and their feature importance metrics into doubt. Given its ranking in the feature importance metrics, it is likely that if the feature was either correctly encoded or omitted, the accuracy of various models would be significantly less than have been reported. Also, it is likely that the importance ranking of other features could change, as based on the premise that the feature importances for the ML models capture interactions between models, interactions between other features and the Course ID would disappear, and would not necessarily be evenly distributed among the other features. To rectify this mistake, the project could be repeated in one of two ways. As the feature categories do not have a useful interpretation, the feature could be dropped. This limits the models however in being unable to explore possible states in which the course is held constant, which could be essential for learning patterns in the data that do not hold between but only within courses. It is important to consider that the accuracy including the feature is already not high and would drop even more with the removal of its dominant feature, so in a repeat scenario maximising accuracy will be essential. Alternatively, the Course ID feature could be encoded and included for the predictions, possibly involving rebinning of courses, as their interpretation is marginally useful as is. When treated in the categorical form by both the chi square and ANOVA tests, there was a definite relationship with the target variable, and so this relationship might prove useful for classification of the target despite the lack of interpretability.

Other than the mistreatment of the Course ID feature and the failure to re-encode the Project impact variable for the causal inference, the feature management for the project has been successful at maintaining interperetability. The features referenced in results from the experiment, the causal inference, and the feature importance can all be interpreted fully by looking at the transformation steps involved and comparing them all to the key in Appendix B. The SVM model not having any options for feature importance under the transformed RBF kernel was one unanticipated issue, but there were still two models with options to produce the metrics. Given the two mistakes in feature treatment however, it may have been beneficial to have created incremental versions of the

key in the Appendix in order to capture and keep a record of changes to feature interpretation as the changes were applied.

Based on the limitations for the estimation of causal effect outlined in the discussion, the validity of the measured treatment effects should not be taken at face value, and no conclusions about causal interactions between variables included in the dataset and the grade of students should be made based on the results. Ethical concerns relating to implications of the causal findings are as such minimal, and are more related to ensuring that readers do not assume these findings to be valid.

While the experiment has been performed based on the results obtained, it is clear that with the lack of validity of the causal analysis as well as the misused Course ID feature, the results of the experiment must have similarly issues. The intention was to find methods of isolating features that would be more likely to produce a variable that is not just correlated but causal, but this isn't necessarily possible, and if seemingly valid results for the question were found, it is likely the feature isolation methods would not generalise to outperforming other methods on other datasets. To thoroughly test this hypothesis, the use of a synthetic causal dataset generation and testing framework, such as *Credence* (Parikh et al. 2022), might be used in combination with other methods to test a large range of potential feature suggestion or isolation against a very large number of simulated datasets with known causal features.

This project has unfortunately not demonstrated a situation where the addition of causal analysis has significantly added value to the overall analysis of the data, and it cannot demonstrate the benefits of adding causal analysis to a project. However, it does thoroughly document the limitations, difficulties, and pitfalls that are involved in the attempt, which makes it a useful resource for others to reference in undertaking to combine the traditional data science/analytics tasks of description and prediction with the less common but potentially more valuable task of causal inference.

References

Abd Elrahman, S. M. & Abraham, A. (2013), 'A review of class imbalance problem', *Journal of Network and Innovative Computing* **1**(2013), 332–340.

BartozKP (2014), 'How to obtain features' weights'.
URL: *https://stackoverflow.com/a/21260848*

Brownlee, J. (2021), 'Tune xgboost performance with learning curves'.
URL: *https://machinelearningmastery.com/tune-xgboost-performance-with-learning-curves/*

Chen, H., Harinen, T., Lee, J.-Y., Yung, M. & Zhao, Z. (2020), 'Causalml: Python package for causal machine learning', *arXiv preprint arXiv:2002.11631* .

Data Protection Commission (2022), 'Anonymisation and pseudonymisation:full guidance note'.
URL: *https://www.dataprotection.ie/en/dpc-guidance/anonymisation-and-pseudonymisation*

Distributed (Deep) Machine Learning Community (2024), 'Awesome xgboost'.
URL: *https://github.com/dmlc/xgboost/tree/master/demoawesome-xgboost*

Gandhi, R. (2018), 'Support vector machine — introduction to machine learning algorithms'.
URL: *https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47*

geeksforgeeks (2023), 'Svm hyperparameter tuning using gridsearchcv — ml'.
URL: *https://www.geeksforgeeks.org/svm-hyperparameter-tuning-using-gridsearchcv-ml/*

He, W., Fang, Y. & Wang, H. (2023), *Real-World Evidence in Medical Product Development*, Springer Nature.

Hernán, M. A., Hsu, J. & Healy, B. (2019), 'A second chance to get causal inference right: A classification of data science tasks', *CHANCE* **32**(1), 42–49.
URL: *https://doi.org/10.1080/09332480.2019.1579578*

Huntington-Klein, N. (2021), *The Effect: An Introduction to Research Design and Causality*, 1st edn, Chapman and Hall/CRC.
URL: *https://doi.org/10.1201/9781003226055*

Ibeling, D. & Icard, T. (2023), 'Comparing causal frameworks: Potential outcomes, structural models, graphs, and abstractions'.

IBM (nodate), 'Boosting'.
URL: *https://www.ibm.com/topics/boosting?mhsrc=ibmsearch$_a$mhq = boosting*

Imbens, G. W. & Rubin, D. B. (2015), *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press.

Karanam, S. (2021), 'Curse of dimensionality - a "curse" to machine learning'.
URL: *https://towardsdatascience.com/curse-of-dimensionality-a-curse-to-machine-learning-c122ee33bfeb*

Künzel, S. R., Walter, S. J. & Sekhon, J. S. (2019), 'Causaltoolbox—estimator stability for heterogeneous treatment effects', *Observational Studies* **5**(2), 105–117.

Marron, J. S., Todd, M. J. & Ahn, J. (2007), 'Distance-weighted discrimination', *Journal of the American Statistical Association* **102**(480), 1267–1271.

Martinaitis, D. (2023), 'Causal inference as a blind spot of data scientists'.
URL: *https://dzidas.com/ml/2023/10/15/blind-spot-ds/*

Microsoft Research (2019), 'Econml: A python package for ml-based heterogeneous treatment effects estimation'.

Narula, A. (2022), 'Causal analysis utilizing causalml', Available at `https://medium.com/@publiciscommerce/causal-analysis-utilizing-causalml-bfabd8015860` (29/12/2023).

Navas, J. & Liaw, R. (2022), 'Guide to xgboost hyperparameter tuning'.
URL: *https://www.anyscale.com/blog/how-to-tune-hyperparameters-on-xgboost*

Parikh, H., Varjao, C., Xu, L. & Tchetgen, E. T. (2022), 'Validating causal inference methods'.
URL: *https://proceedings.mlr.press/v162/parikh22a.html*

Pearl, J. (2009), *Causality*, Cambridge university press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011), 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research* **12**, 2825–2830.

Rodruiguez, C. (2023), 'The notorious xgboost'.
URL: *https://towardsdatascience.com/the-notorious-xgboost-c7f7adc4c183*

Rubin, D. B. (1974), 'Estimating causal effects of treatments in randomized and nonrandomized studies.', *Journal of educational Psychology* **66**(5), 688.

Rubin, D. B. (2019), 'Essential concepts of causal inference: a remarkable history and an intriguing future', *Biostatistics & Epidemiology* **3**(1), 140–155.
URL: *https://doi.org/10.1080/24709360.2019.1670513*

Schuler, A., Jung, K., Tibshirani, R., Hastie, T. & Shah, N. (2017), 'Synthvalidation: Selecting the best causal inference method for a given dataset', *arXiv preprint arXiv:1711.00083* .

Sharma, A. & Kiciman, E. (2020), 'Dowhy: An end-to-end library for causal inference', *arXiv preprint arXiv:2011.04216* .

Taylor, S. J. (2022), 'Bringing more causality to analytics'.
URL: *https://motifanalytics.medium.com/bringing-more-causality-to-analytics-d378108bb15*

Toth, D. J. (2024), 'Binary classification: Xgboost hyperparameter tuning scenarios by non-exhaustive grid search and cross-validation'.
URL: *https://towardsdatascience.com/binary-classification-xgboost-hyperparameter-tuning-scenarios-by-non-exhaustive-grid-search-and-c261f4ce098d*

von Hippel, P. & Wagner, C. (2018), 'Does a successful randomized experiment lead to successful policy? project challenge and what happened in tennessee after project star', *Project Challenge and What Happened in Tennessee After Project STAR (March 31, 2018)* .

Wolpert, D. H. (1996), 'The lack of a priori distinctions between learning algorithms', *Neural computation* **8**(7), 1341–1390.

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J. & Zhang, A. (2021), 'A survey on causal inference', *ACM Transactions on Knowledge Discovery from Data* **15**, 1–46.

Yilmaz, N. & Sekeroglu, B. (2019), Student performance classification using artificial intelligence techniques, *in* 'International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions', Springer, pp. 596–603.

A Dataset License

https://creativecommons.org/licenses/by/4.0/legalcode

B Feature and category information (prior to encoding/transformation)

Student ID
1- Student Age (1: 18-21, 2: 22-25, 3: above 26)
2- Sex (1: female, 2: male)
3- Graduated high-school type: (1: private, 2: state, 3: other)
4- Scholarship type: (1: None, 2: 25%, 3: 50%, 4: 75%, 5: Full)
5- Additional work: (1: Yes, 2: No)
6- Regular artistic or sports activity: (1: Yes, 2: No)
7- Do you have a partner: (1: Yes, 2: No)
8- Total salary if available (1: USD 135-200, 2: USD 201-270, 3: USD 271-340, 4: USD 341-410, 5: above 410)
9- Transportation to the university: (1: Bus, 2: Private car/taxi, 3: bicycle, 4: Other)
10- Accommodation type in Cyprus: (1: rental, 2: dormitory, 3: with family, 4: Other)
11- Mothersâ€™ education: (1: primary school, 2: secondary school, 3: high school, 4: university, 5: MSc., 6: Ph.D.)
12- Fathersâ€™ education: (1: primary school, 2: secondary school, 3: high school, 4: university, 5: MSc., 6: Ph.D.)
13- Number of sisters/brothers (if available): (1: 1, 2:, 2, 3: 3, 4: 4, 5: 5 or above)
14- Parental status: (1: married, 2: divorced, 3: died - one of them or both)
15- Mothersâ€™ occupation: (1: retired, 2: housewife, 3: government officer, 4: private sector employee, 5: self-employment, 6: other)
16- Fathersâ€™ occupation: (1: retired, 2: government officer, 3: private sector employee, 4: self-employment, 5: other)
17- Weekly study hours: (1: None, 2: <5 hours, 3: 6-10 hours, 4: 11-20 hours, 5: more than 20 hours)
18- Reading frequency (non-scientific books/journals): (1: None, 2: Sometimes, 3: Often)
19- Reading frequency (scientific books/journals): (1: None, 2: Sometimes, 3: Often)
20- Attendance to the seminars/conferences related to the department: (1: Yes, 2: No)
21- Impact of your projects/activities on your success: (1: positive, 2: negative, 3: neutral)

22- Attendance to classes (1: always, 2: sometimes, 3: never)
23- Preparation to midterm exams 1: (1: alone, 2: with friends, 3: not applicable)
24- Preparation to midterm exams 2: (1: closest date to the exam,
    2: regularly during the semester, 3: never)
25- Taking notes in classes: (1: never, 2: sometimes, 3: always)
26- Listening in classes: (1: never, 2: sometimes, 3: always)
27- Discussion improves my interest and success in the course: (1: never, 2: sometimes,
    3: always)
28- Flip-classroom: (1: not useful, 2: useful, 3: not applicable)
29- Cumulative grade point average in the last semester (/4.00): (1: <2.00, 2: 2.00-2.49,
    3: 2.50-2.99, 4: 3.00-3.49, 5: above 3.49)
30- Expected Cumulative grade point average in the graduation (/4.00): (1: <2.00,
    2: 2.00-2.49, 3: 2.50-2.99, 4: 3.00-3.49, 5: above 3.49)
31- Course ID
32- OUTPUT Grade (0: Fail, 1: DD, 2: DC, 3: CC, 4: CB, 5: BB, 6: BA, 7: AA)

C Causal Graph Assumptions

"AccomodationRental/Dorm" -> AvailableTime
"AccomodationRental/Dorm" -> GRADE
"AccomodationRental/Dorm" -> Transportation_Bus
Age -> "AccomodationRental/Dorm"
Age -> AvailableTime
Age -> GRADE
Age -> MotherHousewife
Age -> OtherReading
Age -> Partner
Age -> Scholarship
Age -> StudentInterest
Age -> StudentWage
Age -> Transportation_Bus
AvailableTime -> ScientificReading
AvailableTime -> WeeklyStudy
Gender -> GRADE
Gender -> Scholarship
Gender -> StudentInterest

```
Gender -> StudentWage
MotherHousewife -> "AccomodationRental/Dorm"
MotherHousewife -> AvailableTime
MotherHousewife -> GRADE
MotherHousewife -> StudentWage
MotherHousewife -> Transportation_Bus
MotherHousewife <-> Scholarship
MothersEducation -> "AccomodationRental/Dorm"
MothersEducation -> GRADE
MothersEducation -> MotherHousewife
MothersEducation -> OtherReading
MothersEducation -> Scholarship
MothersEducation -> ScientificReading
MothersEducation -> StudentInterest
MothersEducation -> StudentWage
MothersEducation -> Transportation_Bus
OtherReading -> GRADE
Partner -> "AccomodationRental/Dorm"
Partner -> AvailableTime
Partner -> StudentInterest
ProjectImpact -> GRADE
Scholarship -> GRADE
Scholarship -> Transportation_Bus
Scholarship <-> StudentInterest
Scholarship <-> StudentWage
ScientificReading -> GRADE
StudentInterest -> ScientificReading
StudentInterest -> TakingNotes
StudentInterest -> WeeklyStudy
StudentWage -> "AccomodationRental/Dorm"
StudentWage -> AvailableTime
StudentWage -> GRADE
StudentWage -> Transportation_Bus
TakingNotes -> GRADE
TakingNotes -> ProjectImpact
Transportation_Bus -> AvailableTime
Transportation_Bus -> GRADE
```

```
WeeklyStudy -> GRADE
WeeklyStudy -> ProjectImpact
```