UNIVERSITY OF MINNESOTA

PUL-thru:

a high-precision PU-Learner-based EWS for student depression

Nicole Sullivan

Advisor: Dr. Erich Kummerfeld

Committee: Dr. Sisi Ma, Dr. Dan Knights

December 2, 2023

Disclaimer

Views discussed herein are those of the author's, and do not represent thoughts or strategies of the local partnering public school district. This work was performed to fulfill the requirements of the University of Minnesota Data Science Master's program capstone.

Artifact availability

The data used in this project are not publicly available; however, for transparency on the implementation of logic described in this paper, some of the source code for the project has been made available at `https://github.com/sullivannicole/PUL-thru`.

Contact

Computer Science & Engineering
200 Union St. SE
Minneapolis, MN 554555

Contact: sull1120@umn.edu

Changelog

v1.0	2023-10-02	Initial draft.
v1.1	2023-12-02	Report disseminated to committee.

Table of Contents

Abstract Suicide is the second leading cause of death for adolescents in the US [2], and the rate of suicide among Minnesota teens currently outpaces the national rate [2]. Undiagnosed or untreated depression is one of the leading causes of suicide; however, depression in young people often goes undiagnosed [26]. Moreover, there's an oft-overlooked racial component to suicides in the state: suicide rates amongst Native American/Indian adolescents in Minnesota is triple that of any other racial/ethnic group, and suicide rates amongst young Black Minnesotans are increasing [2]. To provide treatment and prevent suicide, though, detecting depression and doing so early is critical [22] [35]. Current work in predicting student depression has either relied too heavily on custom data collection, or failed to attain usable precision [20] [10] [13]. Therefore the objectives of this project were two-fold: develop an early warning system (EWS) for student depression that (1) doesn't require any custom data collection and (2) achieves high precision and F1. To meet the first requirement, we limited features in our EWS to only those that could be engineered from administrative school data already collected as a matter of other regulatory or functional requirements. Towards our second end, we applied a machine learning approach called Positive and Unlabeled Learning (PU Learning); using the resulting framework, which we've dubbed PUL-thru, we were able to achieve a **maximum precision of 1** (average: 0.74), a **maximum F1 of 0.81** (average: 0.66), and a **maximum AUPRC of 0.93** (average: 0.71) across all back-test sets, identifying **94** students that would go on to have a depression diagnosis 6 months in advance of their depression diagnosis, with **only 36 false positives across all 5 years back-tested**. That means that, were the district to deploy PUL-thru, an average of 13 students would receive mental health outreach each semester, with *74%* of those students (on average) actually needing those services (according to our ground truth). To help the district develop meaningful, individualized interventions for students whom our EWS predicts will receive a depression diagnosis, we constructed a causal model, and found the strongest deterrent to a depression diagnosis was high academic performance in the previous semester, while an increase in excused absences conferred a slight elevation in risk of depression diagnosis.

1 Introduction

On Fri, Jun 11, 2021, just a few days after his 17th birthday, Jonas Wagner, a student at a local Minnesota high school, lost his battle with suicidal depression [25], sending a "tidal wave" [6] of grief through the community at the loss. Classmates remembered Wagner as a "really nice guy" [6], one who was actively involved in jazz and marching bands [6], an avid swimmer, waterskiier and wakeboarder [29] and "affectionate with family and friends" [29]. Less than two years later, Aaron Husmann, a teen at the same high school would lose his life to the same illness [15], eliciting similar reactions from classmates, friends and family [23].

Wagner and Husmann's deaths are troubling symptoms of a larger mental health crisis amongst adolescents in the US, with suicide rates amongst 10-24 year-olds increasing by a staggering 57% from 2007 to 2018 [36]. The decline in mental health of US adolescents has been so dramatic that in 2021 it even evoked an Advisory from the U.S. Surgeon General [36].[1].Minnesota, especially, has cause for urgency in the fight against teen suicide: the rate of adolescent suicide in the state of Minnesota has been higher than the national rate for years, and "among communities where suicide is prevalent, the risk of suicide among adolescents can increase by as much as 4 times" [2]. Additionally, even more concerning are the ethnic and racial disparities amongst Minnesota teen suicides: American Indian and Alaskan Native youths in Minnesota die by suicide at higher rates than other races/ethnicities in the state, and suicide deaths amongst Black/African American Minnesotan adolescents have been on the rise [2].

[1] A U.S. Surgeon Advisory is a statement "reserved for significant public health challenges that need the nation's immediate awareness and action". [36]

Combating this crisis, however, requires early detection and diagnosis of suicide precursors like depression [38] [22] [35], which can be difficult,

even for primary care physicians utilizing standardized assessments [18]. In fact, failure to detect depression in young people is common [26], with just half of all individuals who die by suicide ever receiving a formal mental health diagnosis [5], though autopsy studies indicate that the rates should be much higher [38]. Sometimes, even family and close friends are in the dark: parents of another local Minnesota high schooler who committed suicide in 2012 said, in hindsight, their son "hid his depression well" [17].

Though highly-minute patterns generally elude human perception, there are many examples where machine learning models have successfully learned patterns too nuanced for humans to discern [40] [12]. In an attempt to port this success to predicting student depression, a plethora of ML-based EWSes for youth depression have been developed; however, past attempts have either (1) relied too heavily on bespoke data collection [20] [13] [31] that, while innovative, would be either too burdensome or expensive for many districts to maintain long-term [20] or too invasive to student's privacy (such as mining social media posts [13] or data from wearables [20] [31]); or (2) failed to report or achieve a system with sufficient precision and F1 to be usable in the real-world [10] [13].

Therefore, our objectives in this project were two-fold: (1) develop an EWS from only existing, administrative data that is already collected by the district for other purposes and (2) ensure the EWS achieves both decent recall and precision, which would minimize time-burden on and alarm fatigue experienced by school faculty and staff [14]. In this work, we train a deployment-worthy EWS for student depression, which we've dubbed PUL-thru as an homage to its underlying technique, on school administrative data. We intend this system as a decision support tool to aid school faculty and staff in targeting mental health services to students who need them most, when they need them most.

Novel contributions. Novel contributions of this work, therefore, are:

1. Limiting input to our EWS to only those datapoints that are collected in the regular course of school administration **and**
2. Achieving high precision relying only on the aforementioned features

To our knowledge, this is the first-ever high-precision, back-tested EWS for junior-high and high-school student depression that doesn't require *any* special data collection. Other research such as [13], and even our own early efforts, can attain high recall and AUROC, but produce far too many false positives to be truly usable in the real-world. PUL-thru is unique in that it attains remarkable precision without sacrificing recall, thereby representing a framework that's real-world-ready.

2 Background

2.1 Predictive modeling

Machine learning and predictive modeling are powerful tools in data-driven decision-making. At the heart of many predictive models is a mathematical function that maps input features to output predictions [9]. In predictive modeling, the primary goal is to achieve maximum predictive performance and minimum error.

To minimize error, a function must be specified that quantifies error, called the loss (or cost) function [9]. A general representation can be given as:

$$\text{minimize } L(\theta) = \frac{1}{N}\sum_{i=1}^{N}(Y_i - f(X_i;\theta))^2 \tag{1}$$

where $L(\theta)$ is the loss function, Y_i represents the observed outcome, $f(X_i;\theta)$ is the model's prediction for the i-th instance, θ are the model parameters, and N is the number of data points. For a given label or target variable Y_i, many different suitable loss functions may exist; therefore, the practitioner must determine the most costly type of error for the translational objective at hand.

To determine minima of the given loss function, optimization techniques are used to vary model parameters. Gradient descent, perhaps one of the most popular and effective methods for optimization, iteratively adjusts the model parameters in the direction of steepest descent of the loss function to find the optimal parameter values. This process continues until a convergence criterion is met [28].

2.1.1 Time-to-event modeling

Time-to-event modeling, a statistical approach widely employed in survival analysis, focuses on predicting the time until a particular event of interest occurs [19]. One prominent method in this domain is the Cox Proportional Hazards (Cox-PH) model, developed by Sir David R. Cox in 1972 [19]. The Cox-PH model has been extensively utilized to investigate the relationship between covariates and the hazard function, which represents the instantaneous risk of experiencing the event at any given time. The Cox Proportional Hazards model assumes that the hazard ratio between any two individuals remains constant over time, allowing for the examination of covariate effects on survival without specifying the baseline hazard function [19]. Mathematically, the model can be expressed as:

$$h(t, Z) = h_0(t) \cdot \exp(\beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_p Z_p)$$

where $h(t, Z)$ is the hazard function at time t for an individual with covariates $Z_1, Z_2, \ldots, Z_p$, $h_0(t)$ is the baseline hazard function, and $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients associated with the covariates. This model provides valuable insights into the impact of various factors on the timing of events, making it a fundamental tool in survival analysis.

2.1.2 Positive-Unlabeled Learning

Positive-Unlabeled (PU) learning is a specialized machine learning paradigm designed to address scenarios where the majority of available data is unlabeled, and only a small portion is positively labeled. This approach seemed particularly well-suited to our objective as we assumed that there may exist in our dataset students with undiagnosed or unreported depression - that is, there are likely positive-class observations "hidden" amongst our 0-labeled class. Traditional supervised learning methods struggle in such imbalanced settings, as they assume a balanced distribution of positive and negative samples. PU learning aims to mitigate these challenges by leveraging a set of positively labeled examples alongside a larger pool of unlabeled data.

Problem Formulation. Let X be the input space, and Y the binary output space with labels 0 (unlabeled) and 1 (positive). Let $D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ represent the training dataset, where $x_i \in X$ and $y_i \in Y$. However, the labels are incomplete, with only a subset $P \subset D$ being positively labeled. The goal of PU learning is to train a classifier that can accurately predict positive instances not only in P but also in the unlabeled set $U = D \setminus P$.

To formulate this problem mathematically, let $f : X \to [0, 1]$ be the decision function of the classifier. The classifier assigns a probability score to each instance, indicating the likelihood of it being positive. The ob-

jective is to find a decision boundary that maximizes the probability of correctly classifying the positively labeled instances while minimizing the misclassification of the unlabeled ones. This can be expressed as the following optimization problem:

$$\underset{f}{\text{maximize}} \quad \sum_{(x_i,y_i)\in P} \log(f(x_i)) + \sum_{(x_j,y_j)\in U} \log(1 - f(x_j)), \tag{2}$$

where the first term encourages high probabilities for positive instances, and the second term penalizes high probabilities for unlabeled instances.

Approaches in Positive-Unlabeled Learning. Several approaches have been proposed to tackle the PU learning problem. One common strategy is to estimate the probability distribution of the unlabeled data and adjust the classifier's decision boundary accordingly. Let $p_U(x)$ be the estimated probability density function for unlabeled instances. The optimization problem can then be reformulated as:

$$\underset{f}{\text{maximize}} \quad \sum_{(x_i,y_i)\in P} \log(f(x_i)) + \lambda \sum_{(x_j,y_j)\in U} \frac{p_U(x_j)}{1 - f(x_j)}, \tag{3}$$

where λ is a regularization parameter controlling the influence of the unlabeled instances on the objective function.

2.2 Causal inference

Causal inference is the process of understanding the causal relationships between variables in a complex system. It plays a pivotal role in various scientific disciplines, including epidemiology, social sciences, and machine learning. The objective of causal inference is to go beyond correlation and determine whether one variable directly influences another. While discussing causal inference at length is not within the purview of this paper, to aid the reader, we provide a short overview of several fundamental concepts and challenges in the field, including the Fundamental Problem of Causal Inference, interventions, counterfactuals, and the causal DAG. We refer the reader to Judea Pearl's seminal work, Causality: Models, Reasoning and Inference [33], for further details.

2.2.1 Fundamental Problem of Causal Inference

The inability to observe both the potential outcomes for the same unit simultaneously is referred to as **the fundamental problem of causal inference** (FPCI). This problem arises from the fact that the outcome can only be be observed under the actual treatment (e.g., Y_i when $X_i = 1$), and therefore what *would have* occurred under a different treatment is unknown. The fundamental problem can be expressed as:

$$Y_i = X_i \cdot Y_i(1) + (1 - X_i) \cdot Y_i(0) \tag{4}$$

In this equation, Y_i represents the observed outcome for unit i, X_i is the treatment indicator, and $Y_i(1)$ and $Y_i(0)$ are the potential outcomes under treatment and no treatment, respectively. FPCI has stymied scientists and statisticians for millennia; only in the past two or three decades have formalized techniques that combine computer science and mathematics emerged that allow for researchers to draw conclusions from observational data [32].

2.2.2 Causal DAG and the Markov blanket

Causal Directed Acyclic Graphs (DAGs) provide a powerful framework for modeling and analyzing causal relationships among variables. An essential concept in understanding the conditional independence rela-

tionships within a DAG is the Markov blanket [33]. For a random variable represented by a node X in a DAG, its Markov blanket, denoted as $MB(X)$, consists of all the nodes that shield X from the rest of the graph. Mathematically, the Markov blanket of X is defined as:

$$MB(X) = \{Y \in Parents(X) \cup Children(X) \cup Parents(Y) \mid Y \in Children(X)\} \tag{5}$$

where $Parents(X)$ represents the set of parent nodes of X, $Children(X)$ is the set of children nodes of X, and the notation $\{Y \in S \mid P(Y)\}$ indicates the set of elements in set S that satisfy the condition P(Y). The Markov blanket of a node captures all the variables that need to be considered to conditionally infer the value of X, given the structure of the DAG, ensuring that X becomes conditionally independent of all other nodes in the graph when the Markov blanket is known [32]. This concept is fundamental in probabilistic graphical models for modeling and analyzing causal relationships. Other concepts that are fundamental to interpreting causal DAGs include **colliders** and **confounders**.

2.2.3 Colliders and confounders

Colliders are variables in a causal network where two or more arrows (causal pathways) converge [32]. They have the potential to create spurious associations between their parent variables when conditioning on the collider [33]. The statistical relationship between non-colliders X and Y may be influenced by a collider Z. Mathematically, the collider bias can be expressed as:

$$P(X, Y|Z) \neq P(X|Z) \cdot P(Y|Z) \tag{6}$$

Confounding variables are third variables that affect both the cause and the effect, leading to incorrect conclusions about causality if not appropriately controlled [33]. Confounder adjustment is crucial for accurate causal inference. The relationship between variables X and Y adjusted for the confounder Z can be represented as:

$$P(Y|do(X), Z) = P(Y|X, Z) \tag{7}$$

2.2.4 Interventions and counterfactuals

Causal inference often involves studying the effects of interventions or actions. We denote an intervention on variable X as $do(X)$, which represents setting X to a specific value regardless of its previous causal history. The causal effect of X on Y after an intervention can be defined as:

$$P(Y|do(X)) \tag{8}$$

It should be emphasized that:

$$P(Y|do(X)) \neq P(Y|X) \tag{9}$$

that is, that the (observed) probability of the outcome Y given X is *not* equal to the interventional probability of the outcome given X; it's for this very reason that causal inference and causal AI are such powerful tools for exploring data [33].

Conversely, counterfactuals describe what would have happened if a different action had been taken [33]. For a unit i, the counterfactual outcome of variable Y under intervention $do(X_i)$ can be expressed as:

$$Y_i(0) \text{ and } Y_i(1) \tag{10}$$

Here, $Y_i(0)$ represents the outcome for unit i if X_i were set to 0, and $Y_i(1)$ represents the outcome if X_i were set to 1 [33].

3 Methods

Data for this project were obtained through a partnership between the University of Minnesota Institute for Health Informatics and a local Minnesota public school district. For model experimentation and hyperparameter tuning for the binary classifier ML models, we utilized the H2O AutoML framework, a mature open-source software that is discussed in more detail in Section 3.3. We also experimented with employing survival analysis techniques with time-varying covariates for prediction, detailed in Section 3.4. For examining causal mechanisms, we utilized Tetrad 7.1.0-0, which is described in further detail in Section 3.6.

3.1 Data transfer and access

Data were stored in a secure computing environment within a University of Minnesota data center. Through an agreement with a local Minnesota public school, anonymized student and faculty data were pulled manually from the school's databases and uploaded to the secure computing environment via an encrypted connection. No student data were transferred or stored outside of the secure computing environment.

3.2 Feature engineering and pre-processing

3.2.1 Cohort inclusion/exclusion criteria

We imposed relatively few restrictions on the training cohort, in order to make the EWS as widely applicable in the prediction/inference stage as possible. Pre-requisities for inclusion included grades at the time of prediction, and enrollment at one of the district's junior high or high schools[2]. Specifically, students needed to be in grades 7-12, inclusive (regardless of age). Students were also required to have at least one grade recorded in the prior 6 months, as this information was a significant contributor to model performance. Therefore, newly enrolled students for whom grade information was not ported over from their previous school would be excluded from predictions until their second semester in the district (at which point they would have grades recorded in the previous 6 months). If this model were to be deployed within the school district, this present a conundrum of treating certain students inequitably; we therefore would recommend that the district offer the mental health interventions to all new students during their first 6 months in the district. Intuitively, this is sensible as well: new students often lack a social support network scholastically and hence likely would benefit from services that address this anyways [4].

[2] That is, co-op and home-school students who were enrolled in the district but not attending either the junior high or high schools were excluded.

3.2.2 Constructing the outcome variable

The outcome in this project was a depression diagnosis, either in the 1-6 months (for the H2O models, Cox-PH model and PU learner) or 1-12 months following the date of prediction (for the Panorama models). We explored both the former, shorter forecast window, and the latter, two-step forecast window as we surmised either could be of use to the district: the shorter time window could be useful in that there is a natural bi-yearly cadence to the U.S. school year - even schools that utilize quarters or trimesters have a longer break around the Christmas holiday that becomes a hinge-point in the year. Meanwhile, the 1-12 months prediction could be used to issue two alerts, giving faculty and staff greater lead time in which to conduct their outreach.

Conditions reported that constituted a depression diagnosis in this project were: "Depression", "Depressive disorder", "Dysthymia/neurotic depression" and "Dysthmic [sic] Disorder". The dataset containing conditions for each student contained a start date and end date for each condition recorded; however, the end date was NULL for all depressive diagnoses. While depression is a chronic, often persistent disease, the fact that no resolution date was ever recorded for any student called into question whether this resolution date was actively being recorded for chronic comorbidities. It is more likely it is only recorded for acute conditions where a resolution date is more easily identified, and where it is also of greater practical use to administrators. Therefore, instead of inferring that no resolution date meant that the depression persisted beyond that date without any resolution, we chose only to attempt to predict a student's first depression diagnosis; after the occurrence of this first diagnosis, no additional predictions were made (regardless of whether or not the model correctly identified if the student would have that diagnosis). Importantly, this means that we were precluded from using a historical depression diagnosis for prediction of future diagnoses, which makes PUL-thru that much more valuable – instead of relying almost solely on historical diagnoses as some other models are trained to do [37] [27], we give PUL-thru the difficult job of identifying an extremely rare outcome for subjects with yet no past history of that outcome.

3.2.3 Constructing the feature set

Features created fell into one of the following 7 categories: attendance, incidents, comorbidities, demographics, students' socioeconomics, academic performance, and household data. A comprehensive litany of the features created is shown in Table 12 in Appendix A.

Incidents. Incident data contained information on both the violation itself (e.g. dramamongering, cheating, theft), as well as the resolution or consequence assigned for that particular incident. Since severity of an incident could vary greatly within one category (e.g. theft of an office supply such as a stapler vs. theft of another student's laptop), we chose to engineer features from the resolution rather than the incident type. Resolutions fell on a spectrum of less severe (detention, in-school suspension) to highly severe (police-involvement, expulsion). It should be noted that a pattern of repeated same-severity violations (such as theft of office supplies once a week) could incur increasingly severe resolutions over time, leading to an underlying trend in the data. We hypothesized that a pattern of behavior might actually correlate with an underlying mental health dis-

order such as depression, so this trend was intentionally preserved during the course of feature engineering. Specifically, data up to 24 months prior to the prediction window were originally created, in windows of 6 months (e.g. 1-6 months prior, 6-12 months prior, 12-18 months prior, and 18-24 months prior). However, in the course of experimentation, features with information that was less recent than the 6-month window preceding the date of prediction were found to add more noise than signal and ultimately were discarded in subsequent experiments.

Comorbidities. Strong links between other psychiatric disorders and depression have been established [30], as well as between depression and inflammatory diseases, autoimmune disorders, substance-related disorders, and reports of self-harm [7][3] A03 A04. Since data for this project were observational administrative data, comorbidity feature engineering focused only on diagnoses in the students' health data that have an already-established link in literature, rather than trying to surface possibly novel connections between depression and other conditions. Six comorbidity categories were therefore constructed, based on reports of: psychiatric disorders (other than depression), stress- or depression-related disorders, inflammatory diseases, autoimmune diseases, substance-related conditions, or self-harm. A comprehensive list of the conditions that fell into each comorbidity category is given in Table 11 in Appendix A.

Academic performance. Since our desire was to provide depression risk scores for students as young as 13, this required joining junior high and high school grades, which are assigned on different scales in most cases in the state, including the school district from which our data were sourced. Additionally, the school converted to a new grading standard for the junior high during the time period of data available. To standardize across all of these scales, we created 5 different grading-related features - the percent of courses taken in the previous semester where:

- An above-average (relative to that grading scale) final grade was achieved (grade bucket **1**)
- A "Pass" was achieved (grade bucket **2**)
- A failing grade (relative to that grading scale) was assigned (grade bucket **4**)
- There was a withdrawal or no-credit grade assigned (grade bucket **5**)
- All other remaining grades (grade bucket **3**)

In subsequent plots and results, rather than a descriptor, the "grade bucket" listed in parentheses above after the feature is used to reference these features.

We also experimented with converting all non-A-F grades to an A-F grading scheme and calculating the percentage of total courses where a specific grade was attained (A-grades, B-grades, C-grades, etc.) in the previous semester; these features were used in the Cox-PH model, Panorama models, and in PU learner over the aforementioned 5 grade buckets.

3.2.4 Cross-validation and evaluation

To prevent overfitting, the dataset created for training the H2O binary classifiers (discussed in Section 3.3) was split longitudinally; several different

splits were explored, as well as including differing amounts of older data (i.e., using data starting with the year 2005 vs. excluding older data). To validate the trained H2O models, a specific validation set was selected to avoid data leakage; results of this are detailed further in Section 2.3. Ultimately, the dataset used to train the contender classifier models contained data from the 2011-2016, while 2017 constituted the validation set and 2018 and 2019 the test.

Back-testing. For the Cox-PH model and PU learner, we utilized back-testing, a popular form of longitudinal cross-validation (unfortunately, not available in H2O). In back-testing, the training set is composed of all data available before a specific time point, while the back-test set is simply the next time point in the dataset. Evaluation is performed by "rolling" the training set and test set forward 1 timepoint for each "fold" to be tested, until there are no future timepoints left in the dataset. A visualization of this strategy is shown in Figure 1. Due to time constraints, we selected hyperparameters for algorithms *a priori* and did not perform a grid search *a la* hyperopt. For that reason, we can consider each back-test a separate "test" set, rather than a validation set.

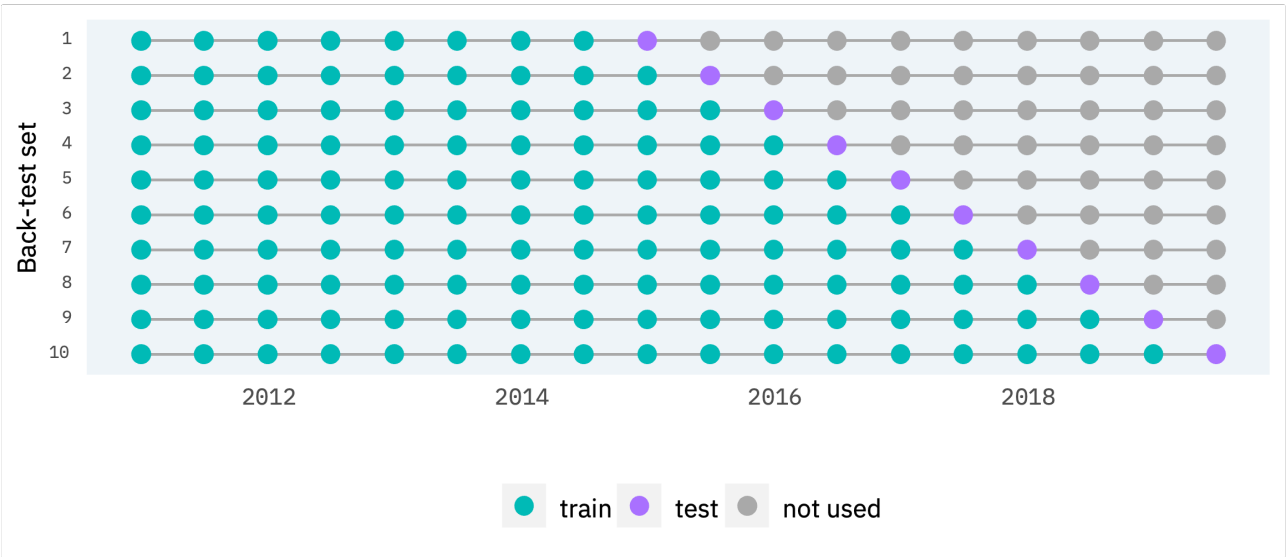

Figure 1: The back-testing strategy utilized for the Cox-PH model and PU learner.

As we were solely interested in a one-step forecasts for these models (one-step being 6 months), we did not evaluate the model's ability to forecast depression diagnoses occurring beyond 6 months in the future. Evaluation metrics were averaged across all back-tests, but we do also give the results of each individual back-test set for the PU learner in Section 4.2.

3.2.5 Feature selection

We conducted experiments both with and without performing filter feature selection beforehand. For filter feature selection, we used Pearson's correlation and pairwise comparisons on the training set only to determine features that explained the most variance in the outcome variable. More complex feature selection methods were eschewed in favor of a routine that could easily be explained to and discussed with our district stakeholders.

3.2.6 Panorama data

After exhausting most pre-processing and model hyperparameter combinations possible with the above feature engineering approach, we considered inclusion of a set of data that was only available for student cohorts enrolled 2020-2022: student responses from the Panorama Social-Emotional Learning Comprehensive assessment (SELComp) (described in more detail in the following paragraph). Given that the assessment wasn't administered prior to 2020 and is administered only on an annual basis, our feature engineering was constrained somewhat. Therefore, in model experiments utilizing the Panorama data, we departed slightly from the feature engineering framework described previously: in these experiments, we used data from the 2020-2021 school year to create predictors, and data from the full 2021-2022 school year to define the outcome variable (a depression diagnosis). In this scenario, we make only one prediction for the next 1-12 month period. We recognize then, that comparisons between Panorama models and the other models aren't apples-to-apples, but provide evaluation results regardless, as preliminary indications of the usefulness of these additional features. We leave further validation of these findings, when more data are available, to future work.

The Panorama SELComp assessment. The Panorama SELComp assessment is designed to evaluate a student's socioemotional progression along several dimensions [21]; educators and administrators can then use results to target competency development efforts for students. Prompts include questions like "How confident are you that you can complete all the work that is assigned in your classes?", "How easily do you give up?" and "How well did you get along with students who are different from you?". To determine the likeliest source of signal in the dataset (if any), we again ran pairwise tests of correlation (using Pearson's correlation coefficient) on the training set, this time between the outcome variable and each field in the assessment. For simplicity, we focused only on the two highest-correlated fields (negative and positive).

In the Panorama assessment dataset, we found the most highly correlated prompt in the negative direction (i.e. the lower the response score, the higher the risk in depression) was "How confident are you that you can complete all the work assigned in your classes?" (r = -0.077), while the most highly correlated prompt in the positive direction was "How often were you polite to other students?" (r = 0.029).

For the prompt "How confident are you that you can complete all the work assigned in your classes?" (which had the greater absolute value of the two correlations), we performed a two-sided t-test to determine if there was a statistically significant difference between the mean responses of the two groups (no depression diagnosis in the next 12 months vs. depression diagnosis in the next 12 months). Formally:

H_0: $\mu_{dx=0} = \mu_{dx=1}$
H_a: $\mu_{dx=0} \neq \mu_{dx=1}$
Significance level: $\alpha = 0.05$

We found that t = 3.308 (df = 17.327, p = 0.004073), with a mean value of 3.5 for the group with no depression diagnosis vs. 2.7 for the group with a depression diagnosis. The response distributions are visualized in Figure 2; note that while responses were discrete, we visualize them as

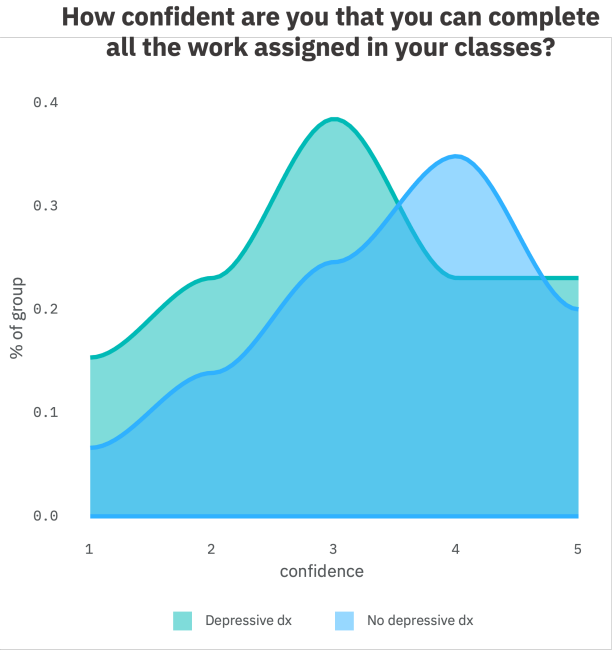

Figure 2: Student responses from the training set for a question from the Panorama SELComp assessment. Higher numbers indicate higher confidence.

continuous distributions to highlight the differences between the two.

We then trained a vanilla logistic regression model and additionally ran several H2o autoML experiments on a dataset composed of 75% of the original data, with a 25% hold-out for testing. For comparison to other models, we selected one H2O model (the Stacked Ensemble) and the logistic regression as contenders; these are designated "Pan-SE" and "Pan-LR" in the Results section.

3.3 H2O AutoML

In recent years, a number of auto-machine learning frameworks have been developed both to democratize machine learning and to abstract away mundanities like manual hyperparameter tuning and model selection, allowing data scientists to more quickly iterate with different features and pre-processing pipelines [16]. Popular autoML frameworks include AutoGluon from Amazon and auto-sklearn from the creators of the Python library sci-kit learn. In this work, for training binary classifiers, we utilized an open-source framework known as H2O autoML; the framework is relatively easy to configure and includes support in both R and Python.

12 autoML experiments were run in total; the settings for each run are shown in Table 1. In general, balancing classes proved more effective than regularization alone and unsurprisingly, truncating the training set by excluding the oldest data available provided better predictions than runs training models on all the data available.

From our 12 total H2O experiments, we selected 7 runs from two different experiments (8 and 10 in Table 1) with the best performance on the

Table 1: AutoML experiments and settings.

#	Min. run	Feat. engineering	Feat. selection	Yrs.	Balancing	Stopping metric
1	5	One-hot encoded	F	2006-2019	T	AUROC
2	5	One-hot encoded	F	2006-2019	F	AUROC
3	5	One-hot encoded	F	2006-2019	F	AUPRC
4	25	One-hot encoded	F	2006-2019	T	AUROC
5	5	One-hot encoded	F	2011-2019	F	AUROC
6	5	One-hot encoded	F	2011-2019	T	AUROC
7	20	One-hot encoded	F	2011-2019	F	AUPRC
8	20	OHE, feats >= 6m prior	F	2011-2019	T	AUROC
9	20	OHE, feats >= 6m prior	F	2006-2019	T	AUROC
10	20	One-hot encoded	T	2011-2019	T	AUROC
11	20	Continuous	F	2011-2019	T	AUROC
12	20	Discretized, 4 quantiles	F	2011-2019	T	AUROC

test set and varying advantages to serve as contender models. Experiment information is given in Table 2. Two models, GBM-41 and GBM-22, contained all features engineered; the other five models were trained using only the top 15 features determined from the feature selection process. ANN-2 and ANN-3 have 2 and 3 hidden layers, respectively, each with 20 units, 30% dropout and a reLu activation function. ANN-2 was trained for a total of 39 epochs, while ANN-3 was trained for a total of 35 epochs, both with a mini batch size of 1 and an adaptive learning rate.

Table 2: H2O contender models.

Experiment	Model	Model type
8	GBM-41	Gradient boosted machine
8	GBM-22	Gradient boosted machine
10	ANN-2	Artificial neural network
10	ANN-3	Artificial neural network
10	GBM-46	Gradient boosted machine
10	GBM-54	Gradient boosted machine
10	StackedEnsemble	H2O Meta-learner

3.4 Cox proportional hazards model

While we originally formatted our data to allow for training a binary classifier, after repeated failures to increase precision with different binary classifiers, we hypothesized time-to-event modeling might be more suitable, given the temporal nature of our outcome variable and possible censoring of students due to moves or disenrollment; we therefore also trained a Cox-PH model using the survival package in R and utilized rolling origin back-testing to evaluate its performance over time. The event of interest

was a depression diagnosis, and we trained the model to predict a student's risk of this event, given covariate effects in the previous 6 months.

3.5 PU learner

After persistently poor precision despite extensive experimentation with binary classifiers and time-to-event modeling, we implemented a Positive-Unlabeled learner, which surpassed previous models on almost every performance metric and therefore ultimately became our algorithm-of-choice. We utilized the pulearn package from Python to train an Elkanoto Positive-Unlabeled learner without weighting (hold-out ratio = 0.1)[3], combined with an scikit-learn support vector machine (hyperparameters: C = 10, γ = 0.4, and an RBF kernel).

[3] This hold-out is distinct from the back-test set; each back-test set is a separate set used *only* for evaluation.

3.6 Tetrad

For causal modeling in this project, we used the free Java-based software Tetrad. The overall routine used to generate causal parameters and effect sizes is visualized in Figure 3. To determine the causal graph structure underlying our data, we used the BOSS (Better Order Score Search) algorithm [24].[4] We did not alter the default settings for BOSS within Tetrad – that is, we utilized a recursion depth = 4, a penalty discount = 2, and lambda = 1 (Chickering). Cutoff for p-values was 0.01.

[4] "Enable Experimental" has to be selected to use the BOSS algo in Tetrad 7.1.0-0.

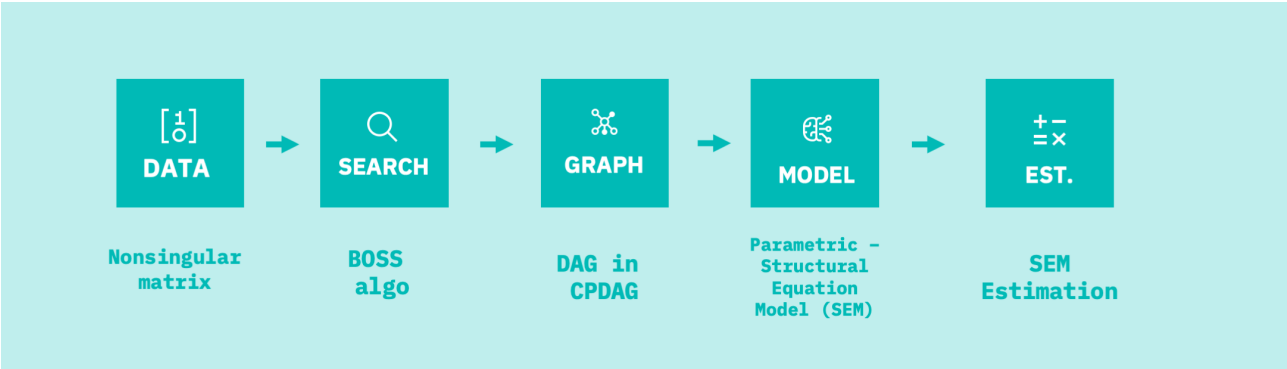

Figure 3: Workflow in Tetrad.

Because causal discovery requires the assumption that there are no unmeasured confounders, we did not perform feature selection prior to running the search algorithm and instead included all of the 59 engineered features. After running BOSS, we selected the DAG in the Conditional Probability DAG (CPDAG) and fit a Structural Equation Model (SEM) to the generated graph. Finally, we instantiated the given SEM to output estimates of each of the parameters given in Section 4.3.2.

4 Results

4.1 Summary statistics

In total, the cohort was composed of 13,609 distinct students across the entire 2011-2019 period (spanning both train and test). We report demographic information and comorbidities in Table 3. We note that a significant portion of students (2,484) were missing reported gender (mainly in data from less-recent years); this missingness was implicitly encoded when the gender feature was one-hot encoded into 2 separate features (a zero for both features indicating missing gender). For race/ethnicity, we used the federal designations reported, which fall into seven categories: African American/Black, American Indian/Native Alaskan, Asian, Hispanic, multi-racial, Native Hawaiian/Pacific Islander, white. Due to low N, we combined students identifying as American Indian/Native Alaskan, Native Hawaiian/Pacific Islander, or multi-racial into one group, designated "other/multi race/ethn" in Table 3. African-American/Black and Asian students experienced the lowest reported rates of depression diagnosis (1.3% and 1.7%, respectively, of each group); all other race/ethnicity groups had a rate of reported depression of approximately 3.0%. Amongst comorbidities examined, about 5.5% of students with a psychiatric disorder diagnosis also had a depression diagnosis at some point following; similarly, 5.4% of students with a stress- or depression-related diagnosis (defined in Table 11 in Appendix A) had a depression diagnosis at some point in the time after.

4.2 Predictive modeling

We give a brief definition of each of the metrics used to evaluate models in the subsequent section; equations for each of the discussed metrics is given in Table 4. General results of evaluation of models on their respective test sets are given in Tables 5 and 7. The results for each back-test set used with the PU learner are given in Table 6; the mean across all back-tests is italicized in the final line of the table. For binary discrimination metrics (F1, recall, precision, logloss, accuracy, confusion matrices), the threshold that maximized F1 along the precision-recall curve was selected to convert continuous model outputs to binary predictions.

Metric definitions. *Confusion matrix.* A model confusion matrix gives counts of model predictions, categorized into the following 4 groups (for a binary classifier): true negatives (TNs), false positives (FPs), false negatives (FNs), and true positives (TPs). True negatives and true positives are correct classifications of the 0 and 1 classes, respectively, while false negatives and false positives are incorrect classifications of the 1 and 0 classes, respectively.

Recall, precision and F1. Recall (also called sensitivity, or true positive rate) is a measure of the model's ability to correctly "catch" all the positive observations; for example, in our case, a high recall would mean that most or all the students that went on to have a depression diagnosis in the next 1-6 months were classified as "1" by our model. Notably, recall does not take into account false positives, so theoretically a model could classify all observations as positive and achieve a recall of 1.00. Precision, on the

Table 3: Summary statistics for the cohort.

Group	No depression dx	Depression dx	Total
Gender			
female	5,165 (96.6%)	180 (3.37%)	5,345
male	5,626 (97.5%)	145 (2.51%)	5,771
unknown	2,484 (99.6%)	9 (0.36%)	2,493
Race/ethnicity			
African American/Black	2,904 (98.7%)	39 (1.3%)	2,943
Asian	709 (98.3%)	12 (1.7%)	721
Hispanic	940 (97.0%)	29 (3.0%)	969
other/multi race/ethn	479 (97.0%)	15 (3.0%)	494
white	8,114 (97.1%)	239 (2.9%)	8,353
unknown	129 (100%)	0	129
Language spoken at home			
English	10,246 (97.7%)	245 (2.3%)	10,491
Spanish	606 (97.6%)	15 (2.4%)	621
other	2,423 (97.0%)	74 (3.0%)	2,497
Comorbidities			
Psychiatric disorder dx[5]	463 (94.5%)	27 (5.5%)	490
Stress- or depression-related	123 (94.6%)	7 (5.4%)	130

[5]Other than depression

other hand, measures the proportion of predicted positives that the model actually got right. Emphasizing precision penalizes the production of false positives by a model; however, "missing" a high number of positive observations will have no affect on precision. In this work, that means a model could miss any number of students that go on to have a depression diagnosis and still attain high levels of precision, so long as it does not emit very many false positives. Thus, often a balance of both recall and precision is desired - enter the F1 metric. F1 serves as a balanced measure of both recall and precision, as both recall and precision must be high to obtain a high F1 value; high recall but low precision will lead to lower F1 values, and vice versa. Recall, precision, and F1 all takes values from 0 to 1 (inclusive), with 1 being the best possible value.

AUROC and AUPRC. Area under the receiver-operator curve, also known as AUROC or simply ROC, assesses a model's ability to discriminate between positive and negative instances across various classification thresholds, graphically represented by the ROC curve. AUROC quantifies the trade-off between true positive rate (TPR) and false positive rate (FPR). Notably, precision has no impact on a model's AUROC or graphical ROC curve. On the other hand, area under the precision-recall curve (AUPRC) focuses on the precision-recall trade-off, providing a comprehensive measure of a model's performance in scenarios where imbalanced class distribution exists. The precision-recall curve plots precision against recall, emphasizing the model's capability to correctly classify positive instances while minimizing false positives, similar to F1. Both curves serve as maps that indicate the variability of the TPR-FPR and precision-recall values when using different thresholds, and which can then be used to select the threshold that maximizes either ROC or F1.

Both AUROC and AUPRC metrics also take values between 0 and 1 (inclusive, 1 being the best).

Demographic parity. Demographic parity is a concept in the context of fairness in machine learning and statistics. It is often used to evaluate whether the predictions or outcomes of a model are consistent across different demographic groups. The goal is to ensure that the model's performance is fair and unbiased w.r.t. sensitive features, such as race, ethnicity, or gender.

Consider a binary classification task with dataset D with sensitive attribute S (e.g., gender) and a binary outcome variable Y (e.g., prediction or actual outcome). Demographic parity is achieved if the following condition holds:

$$P(Y = 1|\hat{Y} = 1, S = 0) = P(Y = 1|\hat{Y} = 1, S = 1)$$

Here, $\hat{Y}$ represents the model's predicted outcome. This equation ensures that the rate of true positive predictions is consistent across different demographic groups. Demographic parity takes values between 0 and 1, inclusive, with *0* indicating perfect parity.

Table 4: Evaluation metrics.

Metric	Definition
TN	y = 0 $\wedge \hat{y}$ = 0
TP	y = 1 $\wedge \hat{y}$ = 1
FN	y = 1 $\wedge \hat{y}$ = 0
FP	y = 0 $\wedge \hat{y}$ = 1
recall	$\frac{TP}{TP+FN}$
precision	$\frac{TP}{TP+FP}$
F1	$\frac{2 \times precision \times recall}{precision+recall}$
AUROC	TPR by FPR
AUPRC	recall by precision

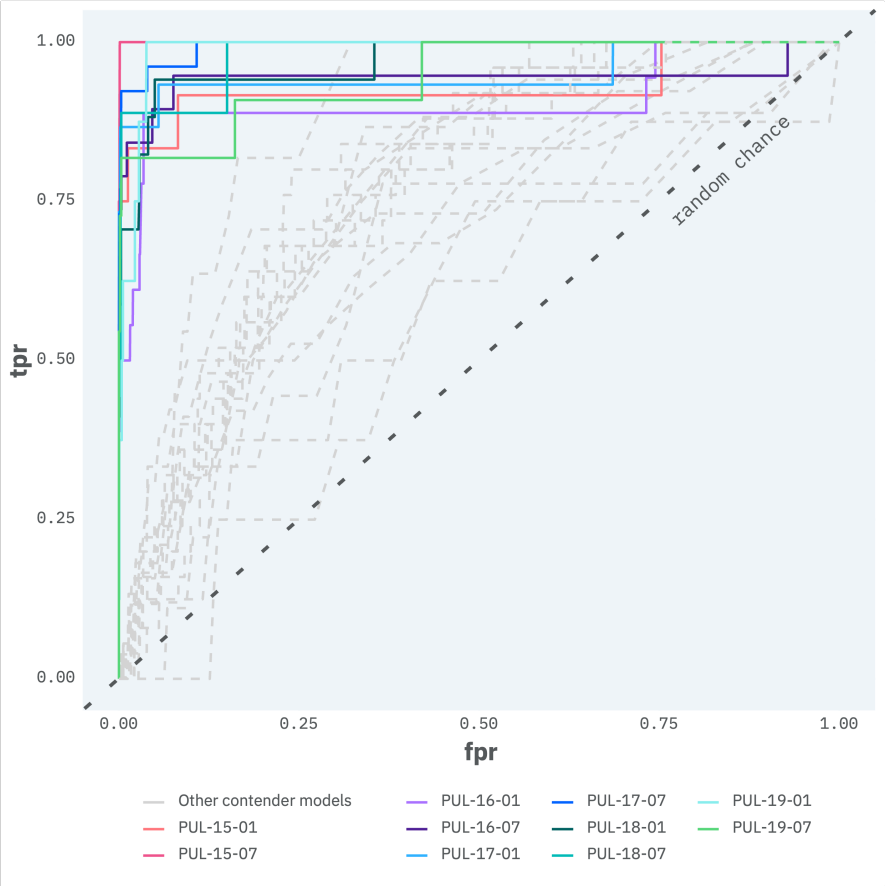

Figure 4: Receiver-operator curves on test sets for 9 of the 11 contender models (curves for the champion model in color). The curve for each back-test is shown for both the Cox-PH and PU learner models.

Table 5: Performance of contender models on a withheld test set. Top metrics are bolded.

model	AUROC	AUPRC	F1	recall	precision	accuracy
GBM-41	0.77	0.007	0.02	0.56	0.008	0.79
GBM-22	0.79	0.007	0.01	0.84	0.007	0.64
ANN-2	0.80	0.009	0.01	0.91	0.005	0.48
ANN-3	0.79	0.008	0.008	**1.00**	0.004	0.31
GBM-46	0.78	0.008	0.01	0.84	0.006	0.58
GBM-54	0.77	0.008	0.01	0.68	0.006	0.70
StackedEnsemble	0.76	0.008	0.02	0.32	0.01	0.91
Mean-Cox-PH	0.71	0.014	0.03	0.70	0.018	0.71
Pan-SE	0.75	0.041	0.16	0.60	0.06	0.92
Pan-LR	0.66	0.042	0.18	0.60	0.107	0.96
Mean-PUL	**0.961**	**0.710**	**0.661**	0.628	**0.742**	**0.997**

AUROC and AUPRC. ROC for the 7 H2O contender models as well as the 10 back-tests used for the Cox-PH model and PU learner are visualized in

Figure 4. The Panorama models contained too few positive observations in the test set to construct a reliable curve; they are, therefore, omitted. Aggregate AUROC and AUPRC values are given in Table 5. All contender models except Pan-LR achieved > 0.7 test ROC, with several H2O models narrowly missing the 0.8 mark (ANN-2 achieves 0.8 when rounded to two decimal points). Overall, the PU learner achieved the highest AUROC: **an average of 0.961**, with a minimum of 0.907 and a maximum of 1.000. The PU learner's average AUPRC was likewise highest: **0.710**, achieving an impressive 0.935 at its maximum, and just 0.408 at its *minimum* (which still represents a 10x improvement from the next-best AUPRC, and a 45x improvement from the best AUPRC of the H2O binary classifiers).

Table 6: Performance of the PU learner on each back-test set. The mean across all back-tests provided in the bottom row, in italics.

Back-test start	AUROC	F1	recall	precision	logloss	accuracy	AUPRC
2015-01-01	0.929	0.696	0.667	0.727	0.081	0.998	0.731
2015-07-01	1.000	0.770	0.625	1.000	0.043	0.998	0.935
2016-01-01	0.907	0.533	0.444	0.667	0.163	0.995	0.519
2016-07-01	0.944	0.722	0.684	0.765	0.140	0.996	0.726
2017-01-01	0.950	0.800	0.667	1.000	0.065	0.998	0.836
2017-07-01	0.993	0.807	0.885	0.742	0.156	0.996	0.877
2018-01-01	0.970	0.667	0.648	0.688	0.227	0.993	0.682
2018-07-01	0.982	0.583	0.778	0.467	0.143	0.996	0.664
2019-01-01	0.988	0.364	0.250	0.667	0.139	0.996	0.408
2019-07-01	0.947	0.667	0.636	0.700	0.101	0.997	0.723
Mean:	*0.961*	*0.661*	*0.628*	*0.742*	*0.123*	*0.997*	*0.710*

Recall, precision and F1. Table 5 shows a number of metrics evaluating models' discrimination performance on the withheld test set. Recall was highest in the H2O artificial neural network models: 1.0 and 0.91 in ANN-2 and ANN-3, respectively. Unequivocally, though, precision for all models aside from the PU learner was dismal, dragging down F1 and AUPRC with it. Poor precision has frequently plagued machine learning systems involving rare events or outcomes [1]; in our case, we hypothesized that discrimination issues might stem from a noisy outcome variable - that is, there might be students with undiagnosed or unreported depression in the "negative" set causing a cascade of false positives when using standard binary classifiers. The PU learner, originally developed for exactly such a scenario, successfully overcame this issue, attaining an average F1 of 0.661, a **33x** improvement from maximum F1 attained with the H2O binary classifiers[6]. Maximum F1 of the PU learner (0.807) was on the 2017-07-01 back-test set. Recall and precision for the PU learner were similarly high, with means of 0.628 and 0.742 respectively, and maximums of 0.885 and 1.000 respectively.

[6]We do not consider the F1s of the Panorama models due to the low total number of positives (5) available in the test set.

Confusion matrix. The confusion matrix for the 11 contender models is given in Table 7. The PU learner model was the only model to achieve metrics that demonstrated real-world usability; the full confusion matrix for the PU learner is given in Table 8. We note that the large reduction in false positives, thanks in large part to the λ regularization parameter in the PU optimization function, is responsible for the huge improvement

Table 7: Confusion matrix for contender models on a withheld test set.

model	TN	FP	FN	TP
GBM-41	6,937	1,806	11	14
GBM-22	5,571	3,172	4	21
ANN-2	4,213	4,530	2	23
ANN-3	2,665	6,078	0	25
GBM-46	5,075	3,668	4	21
GBM-54	6,140	2,603	8	17
StackedEnsemble	7,967	776	17	8
Mean-Cox-PH	1,794	712	6	10
Pan-SE	539	44	2	3
Pan-LR	559	25	2	3
Mean-PUL	2,503	4	5	10

in precision, F1, and AUPRC. Overall, these metrics give the district a fair idea of the outreach workload that could be expected from semester to semester - an average of 13 students are identified each semester as needing outreach (false positives + true positives), and on average, 74% of these students **actually** go on to have a depression diagnosis in the next 1-6 months, absolute numbers that represent a not overly-onerous outreach workload for staff to take on in 6-month windows.

Table 8: Confusion matrix for the PU Learner on back-test sets.

Back-test start	TN	FP	FN	TP
2015-01-01	3,119	3	4	8
2015-07-01	2,477	0	3	5
2016-01-01	3,079	4	10	8
2016-07-01	2,543	4	6	13
2017-01-01	2,762	0	5	10
2017-07-01	2,510	8	3	23
2018-01-01	1,728	5	6	11
2018-07-01	2,511	8	2	7
2019-01-01	1,799	1	6	2
2019-07-01	2,510	3	4	7

Feature importance. We evaluated feature importance for our champion model, the PU learner, using scikit-learn's implementation of permutation feature importance, which is defined as "the decrease in a model score when a single feature value is randomly shuffled" [11]. Per the scikit-learn documentation, "this procedure breaks the relationship between the feature and the target; thus the drop in the model score is indicative of how much the model depends on the feature" [34]. We evaluated the PU learner model trained on all data up to 2019-07-01, and repeated the permutation procedure 10 times for each feature. As estimating feature importance using permutation importance is computationally expensive[7],

[7] As an example, to run the routine described at left for a single back-test, it took 63 minutes on a 16-core Intel(R) Xeon(R) Gold 6152 CPU @ 2.10GHz.

we did not attempt to estimate feature importance for any of the models cross-trained on less data. The mean and standard deviation feature importance for the top features is shown in Figure 5.

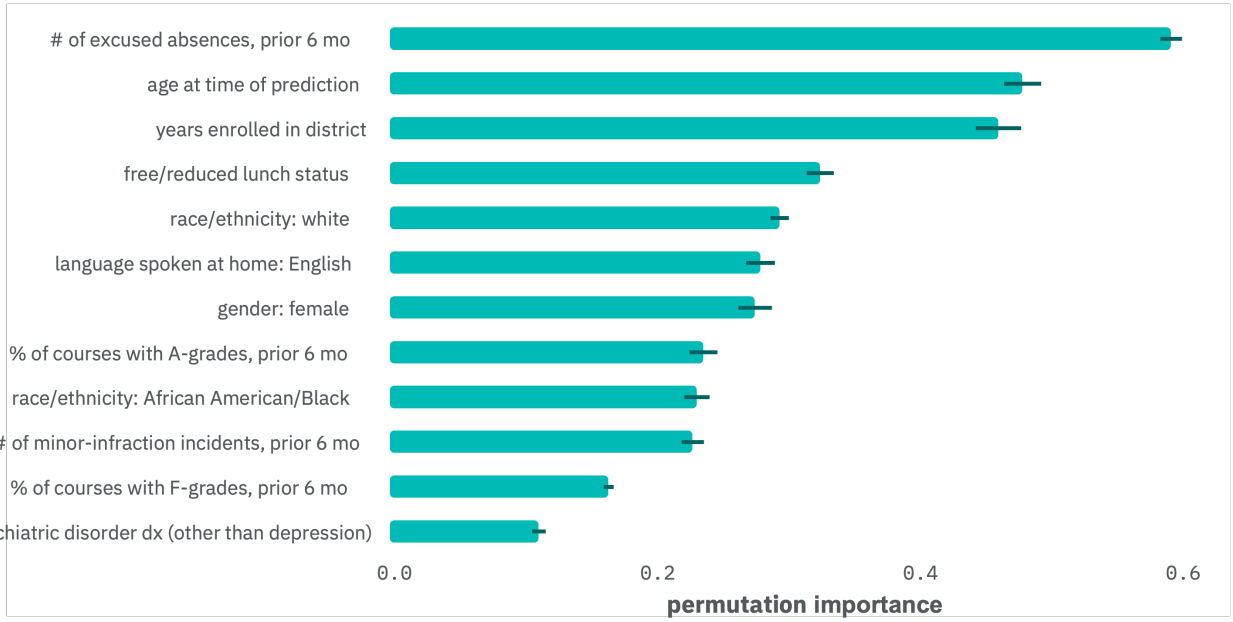

Figure 5: Permutation-based feature importance for the PU learner model. Error bars give the standard deviation of the feature importance across all 10 permutations generated.

The most important feature, with a mean permutation feature importance of 0.594, was the number of *excused* absences a student had in the 6 months prior to the prediction; the fourth most-important feature was socioeconomic-related (a student's free/reduced lunch status, as reported at the beginning of the school year), and the fifth most-important was whether or not the student was white. Unsurprisingly, other top features included age and years enrolled in the district. Significantly, it's clear that the PU learner was not overly relying on one single feature, but seemed to distribute the variance in the outcome across a number of different predictors; this is especially important in mitigating risk, in the event that a variable were to become unavailable temporarily (for example, for a period of time during the global COVID-19 pandemic, attendance tracking ceased or was conducted less robustly in many U.S. junior high and high schools).

Fairness. To ensure that no allocation harms would occur were the PU learner to be deployed, we reviewed the model's demographic parity, a measure of fairness across sensitive features, such as race/ethnicity or gender; results are reported in Table 9. For race/ethnicity, we combined individuals identifying as American Indian, Alaskan Native, Pacific Islander, Native Hawaiian, or as multi-racial/ethnic (such as Black-Hispanic or Asian-Am. Indian), as N of each group was too small to reliably assess fairness for any one of these groups on their own. We see excellent demographic parity across all sensitive features to which we had access: gender, race/ethnicity, language spoken at home, and age. We, therefore, did not need to pursue applying any mitigation techniques to improve performance among sub-groups.

Table 9: Fairness evaluation for the PU learner.

Sensitive feature	Demographic parity
Gender	0.003
Race/ethn	0.005
Language	0.006
Age	0.009

4.3 Causal inference

4.3.1 Causal DAG

The structure of the directed acyclic graph (DAG) obtained via the BOSS search is shown in Figure 6; the Markov blanket of the outcome variable (depress_dx_in_next_6m) is shown in Figure 7. We note that especially the Markov blanket for a depression diagnosis is highly sensible using domain knowledge - the nodes directly affecting a depression diagnosis are: being of African American/Black race, having above-average grades, being female, and the number of excused absences in the 6 months prior. We discuss the causal effect size and additional parameters of this DAG in the next section.

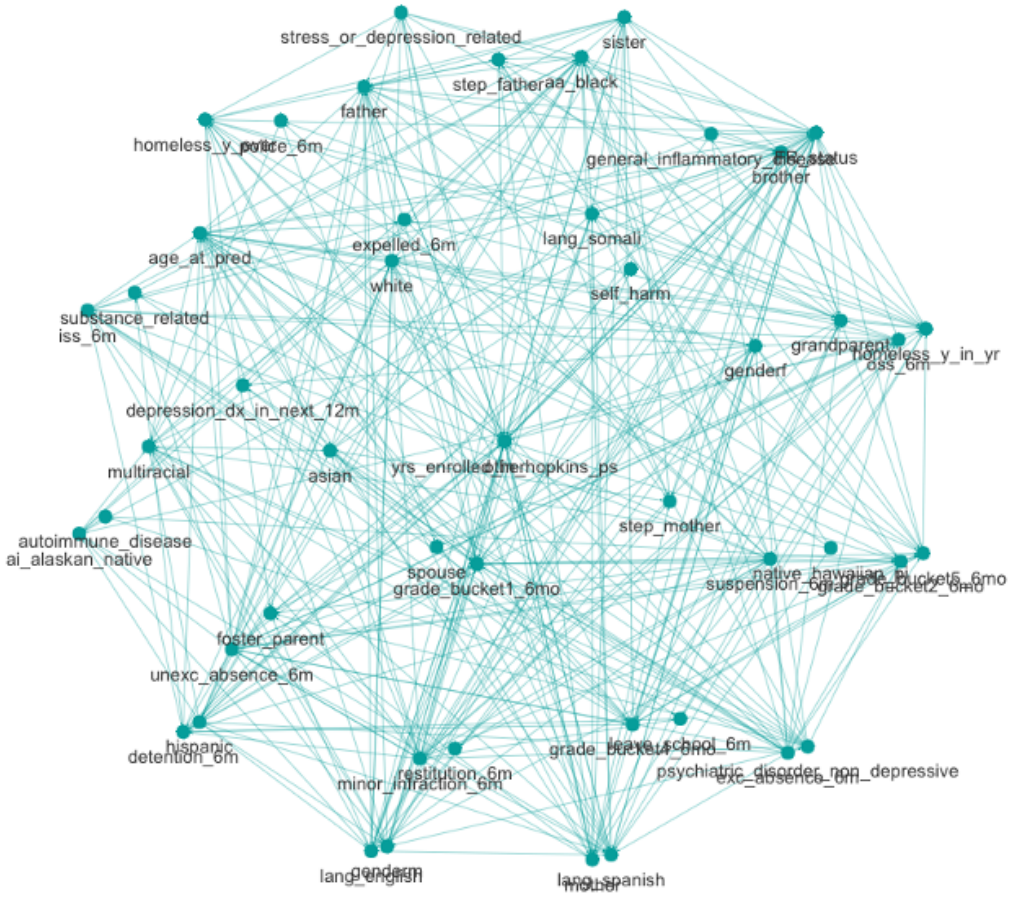

Figure 6: The full causal DAG.

4.3.2 Structural Equation Model (SEM)

Table 10, containing only the Markov blanket for the outcome variable, is provided below. Happily, all the coefficients pass the litmus test of domain knowledge: a higher percentage of courses in the prior 6 months with above average grades is protective against depression (β = -0.0133 $\pm$ 0.0010) while being a female confers slightly greater risk (β = 0.0034 $\pm$ 0.0007). Similarly, being African American/Black is also protective (at least from *reporting* a depression diagnosis to the district)(β = -0.0060, $\pm$ 0.0009), while an increase in the number of excused absences in the

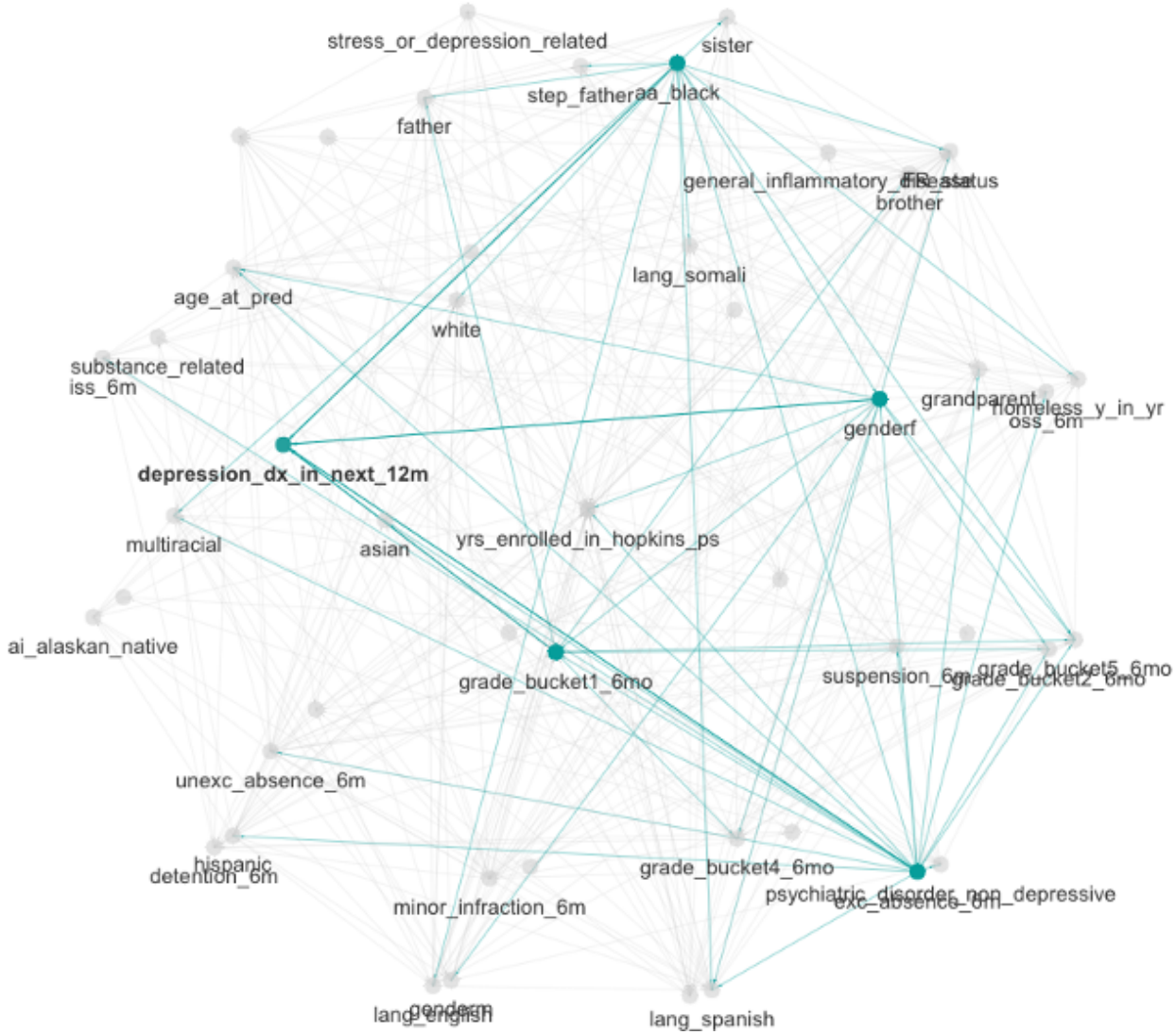

Figure 7: The causal DAG with the Markov blanket for the outcome variable highlighted.

prior 6 months resulted in increased rates of depression diagnoses in the following 1-6 month window.

Table 10: The Markov Blanket for the outcome variable, depression in the next 1-6 months.

From	To	β	SE	T
African American/Black	depression dx	-0.0060	0.0009	-6.9839
% of courses (prior 6 mo) w/ above-avg. grade	depression dx	-0.0133	0.0010	-13.0101
female	depression dx	0.0034	0.0007	5.2244
# of excused absences in prior 6 months	depression dx	0.0008	0.0001	14.4516

4.3.3 Counterfactuals

While, admittedly, our data do not contain many predictors that the district could realistically intervene on, we present several counterfactuals as examples of how the impact of different, individualized interventions could be estimated. Given that PUL-thru only predicts 13 new students for outreach a semester, it is entirely in the realm of possibility that, for this smaller subset of students, the district (in partnership with the student's family) could collect additional qualitative and quantitative data on the student's home life, emotional health, social network, chronic and acute stressors, etc., and then use a combination of domain knowledge and causal modeling to estimate which intervention(s) are most appropriate for the identified students. For now, we simply present the given causal model as a window into what this workflow could look like in practice.

Justin. For our case study, we use the mean across all students with a depression diagnosis for most of the predictors in an effort to avoid divulging any meaningful information on any one of the students. To add dimension to the given examples, we arbitrarily perturb demographic and socioeconomic information, resulting in a fake student, who we refer to as "Justin". Justin is a Black, English-speaking 14-yo male, who is not eligible for free/reduced lunch. We ran 3 different counterfactuals for Justin: (1) we set his total excused absences in the prior semester (originally 12) to 0, (2) we increase the percent of courses in the previous semester in which he achieved an above-average grade (originally 50%) to 80%, and (3) we overwrite his original 0 suspensions with 3 suspensions. The results of these interventions are shown in 8.

While the levers explored in these counterfactuals are likely not the most ideal interventions for an impending depression diagnosis, the varying (albeit small) impacts these interventions would have on Justin's depression outcome show how valuable it is in to understand *which* intervention is optimal for the given student. For example, were the district to collect data on, say, whether or not a student has an outside tutor, they could quantify the impact on the student's school-workload and the impact of that workload on their mental health. They could then estimate the impact that outside tutoring would have on a particular student's risk of a depression diagnosis.

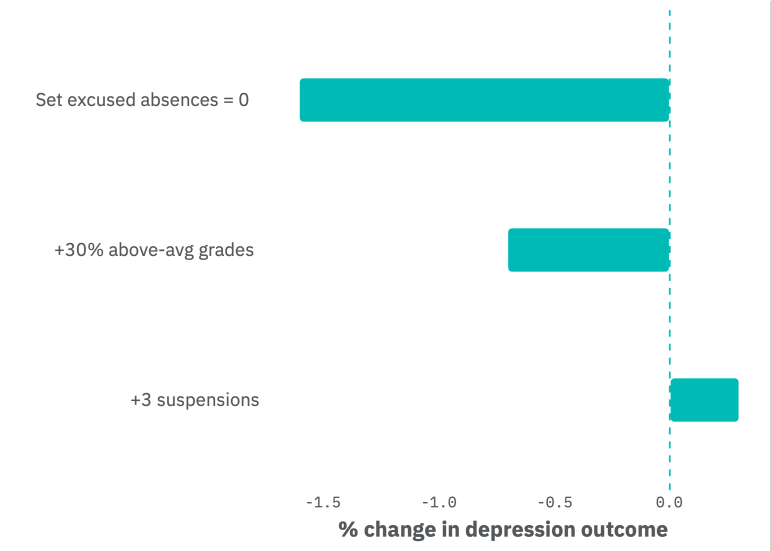

Figure 8: Impact of interventions on Justin's risk of depression dx in the next 1-6 months.

5 Discussion

To achieve our objective in designing a student-depression EWS, we started with training a number of binary classifier ML models as well as a time-to-event predictive model, with which we achieved up to 0.88 AUROC, but which was marred by unusably-low precision (max precision = 0.01); we followed this up with the incorporation of responses to a student social-emotional learning assessment and obtained a 9-fold increase in F1 and an 11-fold improvement in precision of the trained binary classifiers (maximum F1 = 0.18, max precision = 0.11). Finally, we experimented with a PU learner, which enhanced precision and F1 so markedly that the resulting model has potential for real-world deployment.

Back-testing the PU learner, which we've designated "PUL-thru", gives a decent picture of the outreach workload the district could expect from semester to semester: an average of 13 students needing mental health outreach (minimum = 3, maximum = 31), with **74%** of these students (on average) actually going on to have a depression diagnosis in the next 1-6 months. This precision is exceptional, not only given the rarity and noisiness of our outcome, but also given the failure of any other machine learning model to achieve even a fraction of this precision or, even more importantly, the resulting F1.

This leads us to note that machine learning by itself is no silver bullet - despite the advanced capabilities of deep learning and autoML platforms, there still is no substitute for simple, intelligent exploratory data analysis and reasoning. Stopping to ask the question "*Why* do we think this is happening?" often yields more fruitful lines of exploration than throwing all possible features into a workflow and attempting to brute-force one's way into superior ML configurations and results. This finding is not unique to our work, but has also been shown in several recent studies of ML, especially with tabular data [39]. In the end, ML practitioners still must intimately understand the data at hand to make successful modeling decisions.

Finally, we note that all of this work focuses on the data available to us - *reported* depression diagnoses. It is possible - indeed, highly likely - that the PU learner succeeded where more classical techniques failed *because* of the existence of unreported or undiagnosed depression in the negative-labeled observations; therefore, caution should be exercised in assuming that all false positives amongst these results are truly "false". Unfortunately, medical under-diagnosis *and* diagnosis-underreporting is pervasive, especially amongst historically marginalized communities and communities of color [8], and deploying even a well-performing framework that doesn't account for this could serve to (though unintentionally) exacerbate existing disparities. While currently there is no roadmap for the translational aspect of this work, we note that this quandary would need to be carefully considered and protocols developed before deployment could begin.

With this work, we imagine a future world where the burden to reach out is no longer on a student wrestling with depression. Instead, a faculty or staff member would tap the family and student on their proverbial shoulders proactively and check in-on how they're doing - *before* they've even reported a diagnosis to the district. We envision a community of administrators, families, staff members and students alike coming together to support one another in these efforts. Suicide prevention requires all of us, and we hope that, with PUL-thru, we've given schools a tool to fight more effectively. Importantly, we note that PUL-thru alone can't do the job - even the semester that PUL-thru exhibited the highest F1, there were three students that the system failed to identify that went on to have a depression diagnosis in the following 6 months. Ultimately, EWS or not, the responsibility is all of ours - **to listen, to care, and to respond.**

Reference List

[1] Samrachana Adhikari et al. “Revisiting performance metrics for prediction with rare outcomes”. In: *Statistical methods in medical research* 30.10 (2021), pp. 2352–2366.

[2] “Adolescent Suicide”. In: *Minnesota Department of Health reports* (2021). DOI: https://www.health.state.mn.us/docs/communities/titlev/adolescentsuic2021.pdf.

[3] Conner K.R. et al. “Substance-induced depression and independent depression in proximal risk for suicidal behavior”. In: *Journal of Studies on Alcohol and Drugs* 75 (2014), pp. 567–572. DOI: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9798469/pdf/jsad.2014.75.567.pdf.

[4] Li M. et al. “Sensitive periods of moving on mental health and academic performance among university students”. In: *Frontiers in Psychology* 10.1289 (2019). DOI: 10.3389/fpsyg.2019.01289.

[5] Yeh H. et al. “Diagnosed mental health conditions and risk of suicide mortality”. In: *Psychiatric Services* 70.9 (2019), pp. 750–757. DOI: 10.1176/appi.ps.201800346.

[6] Santos B. “Eden Prairie students spread mental health awareness in wake of death”. In: (2022). DOI: https://www.fox9.com/news/eden-prairie-students-spread-mental-health-awareness-in-wake-of-death.

[7] Leonard B.E. “The concept of depression as a dysfunction of the immune system”. In: *Curr Immunol Rev* 6.3 (2010), pp. 205–212. DOI: 10.2174/157339510791823835.

[8] Rahn Kennedy Bailey, Josephine Mokonogho, and Alok Kumar. “Racial and ethnic differences in depression: current perspectives”. In: *Neuropsychiatric disease and treatment* (2019), pp. 603–609.

[9] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. 1st ed. Springer, 2007. ISBN: 0387310738. URL: http://www.amazon.com/Pattern-Recognition-Learning-Information-Statistics/dp/0387310738%3FSubscriptionId%3D13CT5CVB80YFWJEPWS02%26tag%3Dws%26linkCode%3Dxm2%26camp%3D2025%26creative%3D165953%26creativeASIN%3D0387310738.

[10] Rachel Brathwaite et al. “Predicting the risk of future depression among school-attending adolescents in Nigeria using a model developed in Brazil.” In: *Psychiatry Research* 294 (2020), p. 113511.

[11] Leo Breiman. “Random forests”. In: *Machine learning* 45 (2001), pp. 5–32.

[12] Patrick Bryant, Gabriele Pozzati, and Arne Elofsson. “Improved prediction of protein-protein interactions using AlphaFold2”. In: *Nature communications* 13.1 (2022), p. 1265.

[13] Fidel Cacheda et al. “Early detection of depression: social network analysis and random forest techniques”. In: *Journal of medical Internet research* 21.6 (2019), e12554.

[14] Maria Cvach. “Monitor Alarm Fatigue: An Integrative Review”. In: *Biomedical Instrumentation & Technology* 46.4 (2012), pp. 268–277. DOI: 10.2345/0899-8205-46.4.268.

[15] Adams E. “Aaron Husmann, Eden Prairie HS student died by suicide”. In: *SNBC13* (2023). DOI: file:///Users/nicolesullivan/Documents/Academic/2021-2023/MS_in_DS/Capstone/Hopkins_PS/literature/05%20Aaron%20Husmann,%20Eden%20Prairie%20HS%20student%20died%20by%20Suicide.html.

[16] LeDell E. and Poirier S. “H2O AutoML: scalable automatic machine learning”. In: *7th ICML Workshop on Automated Machine Learning* (2020). DOI: https://www.automl.org/wp-content/uploads/2020/07/AutoML_2020_paper_61.pdf.

[17] Mohr E. “Youths’ suicides: How to honor a life, but not risk similar deaths?” In: *Pioneer Press* (2012). DOI: https://www.twincities.com/2012/09/14/youths-suicides-how-to-honor-a-life-but-not-risk-similar-deaths/.

[18] John W Feightner and Graham Worrall. “Early detection of depression by primary care physicians.” In: *CMAJ: Canadian Medical Association Journal* 142.11 (1990), p. 1215.

[19] Lloyd D Fisher and Danyu Y Lin. “Time-dependent covariates in the Cox proportional-hazards regression model”. In: *Annual review of public health* 20.1 (1999), pp. 145–157.

[20] Eiko I Fried, Ricarda KK Proppert, and Carlotta L Rieble. “Building an early warning system for depression: rationale, objectives, and methods of the WARN-D study”. In: *Clinical Psychology in Europe* 5.3 (2023), pp. 1–25.

[21] Ma H. and Cashiola E. “Longitudinal Measurement Invariance Analysis of Panorama Student Survey in a Large, Urban School District in Texas.” In: *International Journal of Intelligent Technologies & Applied Statistics* 15 (2022).

[22] Aron Halfin. “Depression: the benefits of early and appropriate treatment”. In: *American Journal of Managed Care* 13.4 (2007), S92.

[23] Allen J. “EPHS grieves the losses of a student and a staff member”. In: *Eden Prairie Local News* (2023). DOI: https://www.eplocalnews.org/2023/04/12/ephs-grieves-the-losses-of-a-student-and-a-staff-member/.

[24] Ramsey J. “Improving accuracy of permutation DAG search using Best Order Score Search”. In: *CS-Artificial Intelligence* (2021). DOI: arxiv-2108.10141.

[25] “Jonas Michael Wagner obituary”. In: *Star Tribune* (2021). DOI: https://www.startribune.com/obituaries/detail/0000394376/.

[26] Stein K. and Fazel M. “Depression in young people often goes undetected”. In: *The Practitioner* (2015). DOI: https://www.thepractitioner.co.uk/docs/1782-May-2015/SympDepression-youngpeople1782.pdf.

[27] Kate Keenan et al. “Depression begets depression: Comparing the predictive utility of depression and anxiety symptoms to later depression”. In: *Journal of Child Psychology and Psychiatry* 50.9 (2009), pp. 1167–1175.

[28] Jiadong Lin et al. “Nesterov accelerated gradient and scale invariance for adversarial attacks”. In: *arXiv preprint arXiv:1908.06281* (2019).

[29] Wagner N. “Just one night: A local mother’s grief over loss of son to suicide”. In: *Eden Prairie Local News* (2021). DOI: https://www.eplocalnews.org/2021/06/17/just-one-night-a-local-mothers-grief-over-loss-of-son-to-suicide/.

[30] Kalin N.H. “The critical relationship between anxiety and depression”. In: *Am J Psychiatry* 177.5 (2020), pp. 365–367. DOI: https://ajp.psychiatryonline.org/doi/pdf/10.1176/appi.ajp.2020.20030305.

[31] Kennedy Opoku Asare, Aku Visuri, and Denzil ST Ferreira. “Towards early detection of depression through smartphone sensing”. In: *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 2019, pp. 1158–1161.

[32] Judea Pearl. “Causal inference”. In: *Causality: objectives and assessment* (2010), pp. 39–58.

[33] Judea Pearl. *Causality: Models, Reasoning, and Inference*. 2nd. New York, NY: Cambridge University Press, 2009. ISBN: 978-0521895606. URL: http://bayes.cs.ucla.edu/jp_home.html.

[34] “Permutation feature importance.” In: *scikit-learn documentation.* (2023). URL: https://scikit-learn.org/stable/modules/permutation_importance.html#id2.

[35] A Picardi et al. “A randomised controlled trial of the effectiveness of a program for early detection and treatment of depression in primary care”. In: *Journal of affective disorders* 198 (2016), pp. 96–101.

[36] “Protecting Youth Mental Health”. In: *U.S. Surgeon General Advisory* (2021). DOI: https://www.hhs.gov/surgeongeneral/priorities/youth-mental-health/index.html.

[37] Karen D Rudolph et al. “Why is past depression the best predictor of future depression? Stress generation as a mechanism of depression continuity in girls”. In: *Journal of Clinical Child & Adolescent Psychology* 38.4 (2009), pp. 473–485.

[38] Cash S.J.C. and Bridge J.A.B. “Epidemiology of youth suicide and suicidal behavior”. In: *Curr Opin Pediatr* 21.5 (2009), pp. 613–619. DOI: doi:10.1097/MOP.0b013e32833063e1.

[39] Ravid Shwartz-Ziv and Amitai Armon. “Tabular data: Deep learning is not all you need”. In: *Information Fusion* 81 (2022), pp. 84–90.

[40] David Silver et al. “Mastering the game of Go with deep neural networks and tree search”. In: *nature* 529.7587 (2016), pp. 484–489.

Appendices

A Additional tables

Condition classifications are shown in table 11. Table 12 contains all features that were engineered from the data; note that not every experiment used all of these features.

Table 11: Conditions that make up each comorbidity.

Condition	Comorbidity
Adjustment disorder	stress- or depression-related
Anxiety disorder	psychiatric disorder (excl. depression)
Autoimmune disease	autoimmune disease
Avoidant Restrictive Food Intake Disorder (ARFID)	stress- or depression-related
Bipolar disorder	psychiatric disorder
Body Dysmorphic disorder	stress- or depression-related
Cannabis abuse	substance-related
Chemical dependency	substance-related
Chronic Adjustment Disorder	stress- or depression-related
Dietary surveillance and counseling	stress- or depression-related
Eating disorder	stress- or depression-related
EBD	stress- or depression-related
Diabetes Type 2	inflammatory disease
Diabetes insipidus	autoimmune disease
Celiac disease	inflammatory disease
Crohn’s Disease	inflammatory disease
Encephalitis	autoimmune disease
Eosiniphilic esophagitis	inflammatory disease
Fibromyalgia syndrome	inflammatory disease
Graves disease	autoimmune disease
Grief and loss	stress- or depression-related
Guillain-Barre syndrome	autoimmune disease
Hashimoto’s disease	inflammatory disease
Inflammatory Bowel Disease (IBD)	inflammatory disease
Juvenile Rheumatoid Arthritis	inflammatory disease
Lupus Erythematosis	autoimmune disease
Mental health crisis	stress- or depression-related
Mood disorder NOS	psychiatric disorder
Myasthenia Gravis	autoimmune disease
Obsessive-Compulsive disorder	psychiatric disorder
Osgood-Schlatter Disease	inflammatory disease
PANDAS	psychiatric disorder
Panic attacks	stress- or depression-related
Personality Disorder	psychiatric disorder
Post-Traumatic Stress Disorder (PTSD)	psychiatric disorder
Psoriasis	autoimmune disease
Reactive Attachment Disorder	stress- or depression-related
Schizophrenic Disorder	psychiatric disorder
Selective Mutism	stress- or depression-related
Self-injurious behavior	self-harm
Sleep Disorder	stress- or depression-related
Ulcerative colitis	inflammatory disease

Table 12: Features engineered from raw data.

Category	Feature	Time period
attendance	# of excused absences	0-6 months prior
attendance	# of unexcused absences	0-6 months prior
attendance	# of excused tardies	0-6 months prior
attendance	# of unexcused tardies	0-6 months prior
attendance	# of absences (exc/unexc unknown)	0-6 months prior
attendance	# of tardies (exc/unexc unknown)	0-6 months prior
attendance	# of absences due to detention	0-6 months prior
attendance	# of absences due to suspension	0-6 months prior
incidents	# of restitution resolutions	0-6, 6-12, 12-18, 18-24 months prior
incidents	expulsion	0-6, 6-12, 12-18, 18-24 months prior
incidents	# w/ police involvement	0-6, 6-12, 12-18, 18-24 months prior
incidents	# of out-of-school suspensions	0-6, 6-12, 12-18, 18-24 months prior
incidents	# of in-school suspensions	0-6, 6-12, 12-18, 18-24 months prior
incidents	# resulting in leaving school	0-6, 6-12, 12-18, 18-24 months prior
incidents	# of detentions	0-6, 6-12, 12-18, 18-24 months prior
incidents	# of minor infractions	0-6, 6-12, 12-18, 18-24 months prior
comorbidities	psychiatric disorder(s) (excl. Depression)	Ever, prior to prediction
comorbidities	stress or depression-related condition	Ever, prior to prediction
comorbidities	inflammatory disease	Ever, prior to prediction
comorbidities	autoimmune disease	Ever, prior to prediction
comorbidities	substance-related	Ever, prior to prediction
comorbidities	self-harm	Ever, prior to prediction
demographics	Hispanic	As reported at enrollment
demographics	American-Indian/Alaskan Native	As reported at enrollment
demographics	Asian	As reported at enrollment
demographics	African American/Black	As reported at enrollment
demographics	Native Hawaiian/Pacific Islander	As reported at enrollment
demographics	white	As reported at enrollment
demographics	multi-racial/ethnic	As reported at enrollment
demographics	sex at birth: female	As reported at enrollment
demographics	sex at birth: male	As reported at enrollment
demographics	language spoken at home: English	As reported at enrollment
demographics	language spoken at home: Spanish	As reported at enrollment
demographics	language spoken at home: Somalian	As reported at enrollment
demographics	age	At time of prediction
demographics	years enrolled in this district	At time of prediction
socioeconomics	free/reduced lunch status	As reported for that school year
socioeconomics	homeless status	Ever, or in previous year
academic performance	grades (highest, above average, average, below average, lowest)	Prior 0-6 months
household	brother present	As reported for that school year
household	sister present	As reported for that school year
household	father present	As reported for that school year
household	mother present	As reported for that school year
household	step-father present	As reported for that school year
household	step-mother present	As reported for that school year
household	grandparent present	As reported for that school year
household	foster-parent present	As reported for that school year
household	spouse present	As reported for that school year
household	other members present	As reported for that school year