



Article

Predicting Co-Occurring Mental Health and Substance Use Disorders in Women: An Automated Machine Learning Approach

Nirmal Acharya ^{1,*} , Padmaja Kar ², Mustafa Ally ³ and Jeffrey Soar ³ ¹ Australian International Institute of Higher Education, Brisbane, QLD 4000, Australia² St Vincent's Care Services, Mitchelton, QLD 4053, Australia³ School of Business, University of Southern Queensland, Toowoomba, QLD 4350, Australia; mustafa.ally@unisq.edu.au (M.A.); jeffrey.soar@usq.edu.au (J.S.)

* Correspondence: nirmal.acharya@aiihe.edu.au

Abstract: Significant clinical overlap exists between mental health and substance use disorders, especially among women. The purpose of this research is to leverage an AutoML (Automated Machine Learning) interface to predict and distinguish co-occurring mental health (MH) and substance use disorders (SUD) among women. By employing various modeling algorithms for binary classification, including Random Forest, Gradient Boosted Trees, XGBoost, Extra Trees, SGD, Deep Neural Network, Single-Layer Perceptron, K Nearest Neighbors (grid), and a super learning model (constructed by combining the predictions of a Random Forest model and an XGBoost model), the research aims to provide healthcare practitioners with a powerful tool for earlier identification, intervention, and personalised support for women at risk. The present research presents a machine learning (ML) methodology for more accurately predicting the co-occurrence of mental health (MH) and substance use disorders (SUD) in women, utilising the Treatment Episode Data Set Admissions (TEDS-A) from the year 2020 (n = 497,175). A super learning model was constructed by combining the predictions of a Random Forest model and an XGBoost model. The model demonstrated promising predictive performance in predicting co-occurring MH and SUD in women with an AUC = 0.817, Accuracy = 0.751, Precision = 0.743, Recall = 0.926 and F1 Score = 0.825. The use of accurate prediction models can substantially facilitate the prompt identification and implementation of intervention strategies.

Keywords: mental health; substance use disorder; machine learning; AutoMLcheck for
updates

Citation: Acharya, N.; Kar, P.; Ally, M.; Soar, J. Predicting Co-Occurring Mental Health and Substance Use Disorders in Women: An Automated Machine Learning Approach. *Appl. Sci.* **2024**, *14*, 1630. <https://doi.org/10.3390/app14041630>

Academic Editors: Chien-Hung Yeh, Wenbin Shi, Xiaojuan Ban, Men-Tzung Lo and Shenghong He

Received: 8 January 2024

Revised: 15 February 2024

Accepted: 17 February 2024

Published: 18 February 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

An association between co-occurring substance use disorders (SUDs) and various mental health disorders is linked to substantial levels of sickness, death, and impairment [1]. Twenty-five percent of patients seeking medical care have at least one mental or behavioural issue; however, these conditions frequently remain undetected and untreated [2]. Substance addiction affects both genders, although there is evidence to suggest that women may face a more rapid progression toward addiction, encounter greater difficulties in sustaining abstinence, and have a higher susceptibility to relapse compared to men [3]. Women tend to resort to substance consumption as a response to negative emotions [4,5], and prior research has also revealed the distinctive mental health dimensions experienced by women who have substance-related issues [6]. These dimensions include higher levels of depression, traumatic stress, and borderline features in comparison to men [7]. The implications of these interconnected issues have broader consequences, as substance use disorders (SUDs) have been linked to increased risks of suicide and aggressiveness [7,8]. Women grappling with co-occurring disorders often navigate a multitude of hurdles, spanning familial conflicts, depression, educational barriers, economic hardships, past trauma, physical health concerns, reproductive health complications, infertility, early onset of menopause, and complications during pregnancy, breastfeeding, childbirth, unemployment, and more, highlighting the multifaceted nature of their challenges [9,10].

Machine learning (ML) has emerged as a promising tool for understanding and addressing these challenges. Previous studies have explored its application in identifying predictors for suicide, treatment success, and more. Acion, et al. [11] aimed to investigate disparities in substance use disorder treatment completion in the U.S. using 2017–2019 data from TEDS-D by SAMHSA. Employing a two-stage virtual twins model (random forest + decision tree), the research identified factors influencing completion probability (e.g., race/ethnicity, income source), revealing that those without co-occurring mental health conditions, with job-related income, and white non-Hispanics are more likely to complete treatment. Miranda, et al. [12] employed deep learning and natural language processing to develop DeepBiomarker2 that accurately predicts alcohol and substance use disorder risk in post-traumatic stress disorder patients and identifies medications and social determinants of health parameters that may reduce this risk. Adams, et al. [13] performed a study in Denmark that focused on individuals with substance use disorders (SUDs) and their elevated suicide risk. Using machine learning, the analysis identified key predictors for suicide in men and women with SUDs, highlighting specific factors such as antidepressant use, poisoning diagnoses, age, and comorbid psychiatric disorders. The findings suggest that individuals with prior incidents of poisoning and mental health disorders, especially women, are at increased risk of suicide among those with substance use disorders in Denmark. Aishwarya, et al. [14] investigated the use of machine learning, including AutoML and ensemble classifiers, to predict potential cardiovascular diseases by analysing real-time IoT-based healthcare data, highlighting improved accuracy and efficiency in data analytics for healthcare devices. Kundu, et al. [15] explored the application of machine learning (ML) in investigating mental health and substance use concerns within the LGBTQ2S+ population. Examining 11 recent studies, the findings suggested ML as a promising tool. A lack of studies evaluating substance use treatments in women with severe mental illness who differ in their needs and capacity has been noted [16], there are opportunities to explore the potential application to research in this field of Automated Machine Learning (AutoML) interfaces.

The current research utilises data from the Treatment Episode Data Set Admissions (TEDS-A) for the year 2020 and utilises an AutoML interface to predict co-occurring mental health and substance use disorders among women. The rationale behind leveraging AutoML stems from its growing significance in healthcare analysis [17–19], particularly within the domain of mental health and substance use disorders [20–22], where it often leads to enhanced precision and accuracy [23].

The opportunity for AutoML arises from the need to provide a more user-friendly method for anyone to generate and implement machine learning, offering a more intuitive approach for creating and deploying models with minimal reliance on coding or complex ML infrastructure [24]. Given the limited financial resources allocated to clinical coding and the high wages of data scientists [25], it is imperative to identify a cost-effective approach that enables healthcare organisations to leverage machine learning capabilities without incurring substantial expenses. Several AutoML platforms are currently available. Certain platforms are open source whereas others are commercial. Many prominent organisations in the field of artificial intelligence, including Microsoft Azure, Google, Amazon, H₂O.ai, Dataiku, and RapidMiner, have undertaken the development and dissemination of advanced systems, such as the publicly accessible Cloud AutoML [26]. Platforms such as Dataiku exemplify this shift, providing a graphical interface empowering users to fine-tune computational settings effortlessly, enhancing accessibility. Instead of being tethered to specific algorithms or coding languages, researchers gain the flexibility to explore diverse methods within a unified space, encompassing languages such as Python, R, and more, fostering the full spectrum of ML tools. Within the AutoML framework, users leverage existing algorithms and ML frameworks. The process begins with inputting data onto the platform. Users can then opt to employ a specific method or request algorithm suggestions. Once chosen, an algorithm is set up to facilitate training, seamlessly leading into the automated testing phase. This yields immediate access to ML insights, including model predictions

and performance metrics, enabling researchers to employ validated models for forecasting or analysing various phenomena. Typically, AutoML workflows initiate with basic ML algorithms known for their simplicity, user-friendliness, and rigorously evaluated models such as *k*-nearest neighbors and decision trees. As the analysis demands more intricate scrutiny, more complex alternatives like boosted trees or deep learning (e.g., XGBoost) come into play for analysis and evaluation. Diverse intricate models can be crafted, often formed as ensembles—a fusion of basic models leveraging the strengths of each component while mitigating individual weaknesses. The synergy within an ensemble of algorithms aims to enhance overall predictive power and model robustness. ML techniques, particularly AutoML have led to improved granularity and accuracy in various studies [11,22,27]. This research capitalises on the power of AutoML to automate the process of model selection, hyperparameter tuning, and feature engineering, streamlining the analytical process and enhancing the predictive accuracy of the models. The super learning (SL) model has the potential to distinguish women with co-occurring disorders from those without. The super learning algorithm is a supervised learning method that uses a loss-based approach to choose the best combination of prediction algorithms [28]. The method achieves asymptotic performance comparable to the optimal weighted combination of the basic learners, making it a highly effective strategy for addressing various issues using the same technique; it can reduce the probability of over-fitting during the training process, employing a modified version of cross-validation [29,30]. The area under the curve (AUC) value of 0.817 achieved by the super learning model attests to its efficacy in capturing intricate patterns within the data, underscoring its potential as a robust diagnostic tool.

This study serves as an illustration of advanced statistical methods and machine learning techniques harnessed through Dataiku, an AutoML interface, in a real-world healthcare setting. It showcases the platform's ability to automate essential operations such as selecting models, optimising hyperparameters, and engineering features. The super learning model, which combines Random Forest and XGBoost, has superior performance compared to separate algorithms. It serves as a diagnostic tool for early detection of co-occurring disorders in women. The study emphasises the potential advantages of these powerful predictive models. By leveraging these technological advancements, we aim to bridge the gap between data-driven innovation and clinical practice. SUDs are often inadequately addressed in women [31,32]. The insight of the study holds the potential to develop a tool for early identification of mental health and SUDs in women. The findings also offer valuable insights that can inform future research and collaborations with policymakers, medical associations, and patient advocacy groups to develop guidelines for responsible integration and optimise the model's potential advantages while ensuring patient well-being and privacy protection.

The structure of the paper is as follows. The next section presents the materials and methods, covering the description of the dataset, the machine learning models utilised, and the statistical method employed. This is followed by the Section 3, where the findings of the study are presented. The Section 4 then follows, which discusses the application, limitations, and future prospects of the study. Finally, the paper concludes with the implications, recommendations for further research, and conclusions.

2. Materials and Methods

2.1. Dataset

This study used publicly available Treatment Episode Data Set Admissions (TEDS-A) 2020 [33], maintained by the Center for Behavioral Health Statistics and Quality (CBHSQ) of the Substance Abuse and Mental Health Services Administration (SAMHSA) to illustrate the machine learning approach to predict co-occurring mental and substance use disorders in women. TEDS, encompassing the Admissions Data Set (TEDS-A) and the Discharges Data Set (TEDS-D), is a notable representation of a substantial administrative dataset that may captivate addiction researchers in practical situations [34,35]. TEDS provides comprehensive statistics regarding admissions and discharges from substance use disorder

treatment programs across participating states. However, the analysis for the year 2020 had to exclude Oregon, North Dakota, Idaho, and Washington due to inadequate data reporting. Notably, some states contribute data that document multiple admissions for the same individual, shaping statistical analyses to accurately portray admissions rather than individual clients [36]. The dependent variable in this study was co-occurring mental health and substance use disorder which is coded as PSYPROB (1 = Yes, 2 = No) in the dataset. As we focused on women, we extracted the records where the client's biological sex was female ($n = 497,175$). We then conducted data pre-processing which consisted of three steps. First, we conducted listwise deletion for records with missing values at the dependent variable and thirty-seven relevant predictors that include PSYPROB, STFIPS, SERVICES, PREG, IDU, EMPLOY, EDUC, ETHNIC, LIVARAG, BARBFLG, MARFLG, DSMCRIT, AGE, MARSTAT, RACE, PSOURCE, AMPHFLG, ALCDRUG, STIMFLG, MTHAMFLG, ALCFLG, SEDHPFLG, INHFLG, OTCFLG, PCPFLG, HALLFLG, OPSYNFLG, BENZFLG, TRNQFLG, METHFLG, COKEFLG, HERFLG, OTHERFLG, METHUSE, FRSTUSE1, SUB1, SUB2, SUB3, NOPRIOR. Records with incomplete data in any of the predictors, the outcome, or characteristics used for defining inclusion in the study were excluded from the analysis. Second, outlier detection was performed using the `analyse` function in Dataiku at the dependent variable and all relevant predictors. The outliers were handled by performing listwise deletion, leaving us a final analytic sample ($n = 132,128$). Finally, each feature was processed using target encoding, in which its original value was substituted with a numerical value derived from the target values. Within the dataset, several features exhibit different units and scales. This trend could result in certain features having a more significant influence on the learning algorithm compared to others, thus potentially introducing bias. To tackle this issue, we employed the min-max normalisation technique to standardise all the features, consequently ensuring that they are within a consistent range, typically ranging from 0 to 1 [37]. This ensures that all features contribute equally to the model. The provided sample was then randomly split into two sets: a training set comprising 80% of the sample ($n = 105,760$), and a test set consisting of the remaining 20% ($n = 26,368$).

As the data utilised in this study were sourced from publicly available information without any subject identification, the research design and methodology were determined to be exempt from ethics review.

2.2. Statistical Methods

Multivariate analysis was performed using Dataiku v12 [38], an integrated coding-free platform for data science, machine learning, and analytics [24]. The modelling algorithms applied for binary classification modelling for the prediction of the probability of co-occurring mental health and substance use disorders in women were Random Forest, Gradient Boosted Trees, XGBoost, Extra Trees, SGD, Deep Neural Network, Single-Layer Perceptron, K Nearest Neighbors (grid) and a super learning model (constructed by combining the predictions of a Random Forest model and an XGBoost model) [39] (see Figure 1). In the discipline of predictive modelling, conventional techniques such as linear or logistic regression have historically been used. However, the advancement of machine learning has introduced Random Forests (RF) and XGBoost as robust alternatives in the field of health sciences [40–42]. The rationale behind incorporating a Random Forest model and an XGBoost model into a super learning framework is in their capacity to overcome the limitations of traditional regression approaches [11,43–45]. Random Forest, with its collection of decision trees, offers resistance against overfitting and excels in capturing intricate, non-linear relationships within data. Meanwhile, XGBoost utilises gradient boosting to repeatedly improve predictive accuracy by combining weak learners and tackling obstacles posed by heterogeneous data. The objective of this integrated strategy is to capitalise on the advantages of both algorithms, promoting a more robust and precise predictive model.

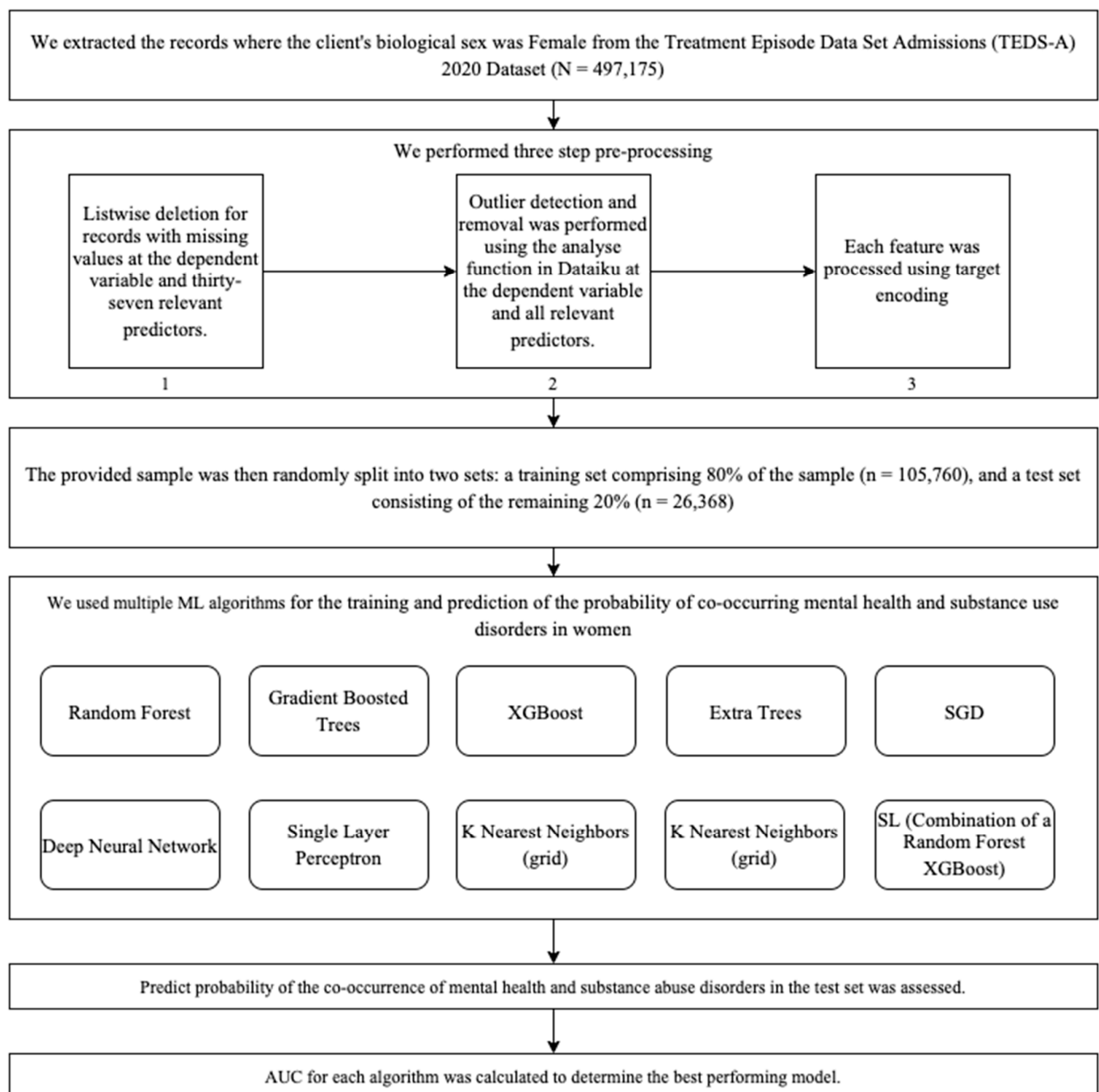


Figure 1. Analytic workflow.

The optimal analytic approach for forecasting the co-occurrence of mental health and substance abuse disorders was determined to be the model that maximises the Area Under the Curve (AUC) [46]. The AUC is a useful metric for evaluating prediction accuracy. It represents the likelihood that a randomly selected successful patient will be ranked higher than a randomly selected unsuccessful patient by any of the algorithms. The AUC (area under the curve) metric measures the performance of a prediction model, with values ranging from 0 to 1. AUC = 1 indicates a perfect forecast, while AUC = 0.5 suggests that the prediction is no better than chance.

Dataiku's ML diagnostics feature was enabled to conduct comprehensive checks on the dataset, modelling parameters, training speed, overfitting, leakage, model checks, ML assertions, and abnormal predictions. A Bayesian search strategy was employed to

optimise the hyperparameters of the machine learning models. The search was guided by a probabilistic model that intelligently selected hyperparameter combinations for evaluation. The goal was to find the best-performing set of hyperparameters for the model's task. The search process was limited to exploring five different combinations of hyperparameters. This approach allowed for an efficient and systematic exploration of the hyperparameter space, leading to improved model performance. A super learning model was constructed by combining the predictions of a Random Forest model and an XGBoost model using the "average" method. Each model was trained independently on the training data to capture distinct patterns and relationships. During prediction, the outputs of both models were averaged for each data point, resulting in a final prediction for the super learning model. This approach leverages the strengths of both Random Forest and XGBoost, providing a potentially more robust and accurate prediction by blending the insights from these two diverse algorithms.

3. Results

Dataiku automatically ranks the best-performing interpretable model based on the set performance metric (AUC in this case). The characteristics of the sample are presented in Table 1. It presents an overview of a cohort and only includes gender as well as the top 10 predictors of PSYPROB that were selected as the most essential based on the Shapley values, to keep it concise. The cohort predominantly resides in states such as New York, Colorado, and Illinois. Notably, around 30% had no prior treatment episodes, reflecting a significant proportion seeking treatment for the first time, while diverse referral sources: individuals, legal systems, and community referrals highlight the multifaceted pathways to treatment. In terms of race, the majority of the individuals classified themselves as Black or African American (73.9%), and a range of substance use patterns emerged, encompassing various substances across primary, secondary, and tertiary categories. The significant unemployment rate of 52.3% among the cohort highlights the possible socioeconomic factors at play. The substances encompass alcohol, cocaine/crack, marijuana/hashish, prescription opiates/synthetics, methamphetamine/speed, and various other substances. The diagnostic data indicated a significant occurrence of opioid dependence, with a prevalence rate of 33.8%. Additionally, around 26.3% of individuals received medication-assisted opioid therapy.

Table 1. Baseline characteristics of the cohort (n = 132,128).

Factor	Value	Number (%)
GENDER	Female	132,128 (100)
State (STFIPS)	New York	35,662 (27)
	Colorado	11,247 (8.5)
	Illinois	9351 (7.1)
	Michigan	9161 (6.9)
	North Carolina	8410 (6.4)
	New Jersey	7520 (5.7)
	Indiana	6806 (5.2)
	Connecticut	5502 (4.2)
	Kentucky	4954 (3.7)
	Tennessee	4052 (3.1)
Missouri	3593 (2.7)	
Pennsylvania	3236 (2.4)	
Ohio	2771 (2.1)	
Other	19,863 (15)	

Table 1. Cont.

Factor	Value	Number (%)
Previous substance use treatment episodes (NOPRIOR)	No prior treatment episodes	39,112 (29.6)
	One prior treatment episode	28,334 (21.4)
	Five or more prior treatment episodes	24,478 (18.5)
	Two prior treatment episodes	19,212 (14.5)
	Three prior treatment episodes	13,101 (9.9)
	Four prior treatment episodes	7891 (6)
Type of treatment service/setting (SERVICES)	Ambulatory, non-intensive outpatient	70,284 (53.2)
	Rehab/residential, short-term (30 days or fewer)	21,110 (16)
	Ambulatory, intensive outpatient	15,802 (12)
	Detox, 24-h, free-standing residential	13,945 (10.6)
	Other	10,983 (8.4)
Referral source (PSOURCE)	Individual (includes self-referral)	62,232 (47.1)
	Court/criminal justice referral/DUI/DWI	29,059 (22)
	Other community referral	15,135 (11.5)
	Alcohol/drug use care provider	14,176 (10.7)
	Other	11,535 (9.7)
Race (RACE)	Black or African American	97,677 (73.9)
	Asian or Pacific Islander	20,931 (15.8)
	White	8788 (6.7)
	Other	4732 (3.5)
Substance use (secondary) (SUB2)	None	48,513 (36.7)
	Cocaine/crack	19,126 (14.5)
	Marijuana/hashish	18,044 (13.7)
	Methamphetamine/speed	11,659 (8.8)
	Alcohol	10,974 (8.3)
	Other opiates and synthetics	7012 (5.3)
	Other	10,246 (12.7)
Substance use (tertiary) (SUB3)	None	92,205 (69.8)
	Marijuana/hashish	10,689 (8.1)
	Alcohol	6454 (4.9)
	Cocaine/crack	5465 (4.1)
	Methamphetamine/speed	3373 (2.6)
	Other	13,942 (10.5)
Employment (EMPLOY)	Unemployed	69,048 (52.3)
	Not in labour force	51,075 (38.7)
	Part-time	12,005 (9.1)
Diagnostic and Statistical Manual of Mental Disorders diagnosis (DSMCRIT)	Opioid dependence	44,699 (33.8)
	Alcohol dependence	22,077 (16.7)
	Other substance dependence	17,766 (13.4)
	Other mental health condition	14,540 (11)
	Cannabis dependence	6567 (5)
	Cocaine dependence	6196 (4.7)
	Other	20,283 (15.4)
Medication-assisted opioid therapy (METHUSE)	Yes	34,807 (26.3)
	No	97,321 (73.7)

Table 2 shows the performance matrices for each algorithm applied for binary classification in the test set (N = 26,368). The primary evaluation criterion in this study was the AUC. All AUC values were between 0.631 and 0.817. This range signifies the probability that any of the algorithms would correctly rank a randomly selected woman with co-occurring mental and substance use disorders higher than one without such disorders. As hypothesised, the super learning model showed the largest AUC of 0.817, demonstrating robust predictive capability. The performance of the super learning model is closely

followed by XGBoost with an AUC of 0.809. Several ensemble techniques such as Random Forest (AUC = 0.807) and Extra Trees (AUC = 0.803) showed significant discriminatory ability, closely following the top-performing algorithm. Meanwhile, conventional methods such as Gradient-Boosted Trees (AUC = 0.799) and Single-Layer Perceptron (AUC = 0.776) demonstrated comparable but slightly lower AUC scores. Nevertheless, the utilisation of K Nearest Neighbors (grid) resulted in a relatively reduced prediction accuracy, as indicated by an AUC of 0.670. Interestingly, models employing Deep Neural Network architecture exhibited the least satisfactory performance among the investigated algorithms, achieving an AUC of 0.631. These findings highlight the superiority of the proposed ensemble-based approach in achieving higher AUC values and therefore more successful binary classification in this experimental environment.

Table 2. Performance matrices for each algorithm applied for binary classification in the test set (N = 26,368).

Model	AUC	Accuracy	Precision	Recall	F1 Score
Random Forest	0.807	0.742	0.734	0.923	0.818
Gradient Boosted Trees	0.799	0.733	0.726	0.921	0.812
XGBoost	0.809	0.745	0.739	0.918	0.819
Extra Trees	0.803	0.738	0.733	0.916	0.814
SGD	0.778	0.725	0.728	0.898	0.804
Deep Neural Network	0.631	0.628	0.628	1	0.771
Single Layer Perceptron	0.776	0.721	0.718	0.916	0.805
K Nearest Neighbors (grid)	0.670	0.661	0.655	0.971	0.782
Super Learning	0.817	0.751	0.743	0.926	0.825

Table 3 provides the mean and standard deviation (SD) for each metric across the different models. Based on these findings, the “super learning” model emerged as the better option, as it consistently demonstrated high performance across various parameters with low variability.

Table 3. Mean and standard deviation of key metrics across the evaluated models.

Metric	Mean	Standard Deviation (SD)
AUC	0.772	0.059
Accuracy	0.713	0.037
Precision	0.711	0.037
Recall	0.935	0.032
F1 Score	0.804	0.020

Figure 2 provides a visual representation of the AUC in the test set (n = 26,368) for the super learning model. The super learning model exhibited the highest AUC among all models, boasting an AUC of 0.817, a score that is typically considered a strong performance for prediction models. Other matrices for the super learning model include accuracy (0.751), precision (0.743), recall (0.926), and F1 score (0.825). This outcome underscores the model’s strong ability to distinguish women with co-occurring mental and substance use disorders from those without.

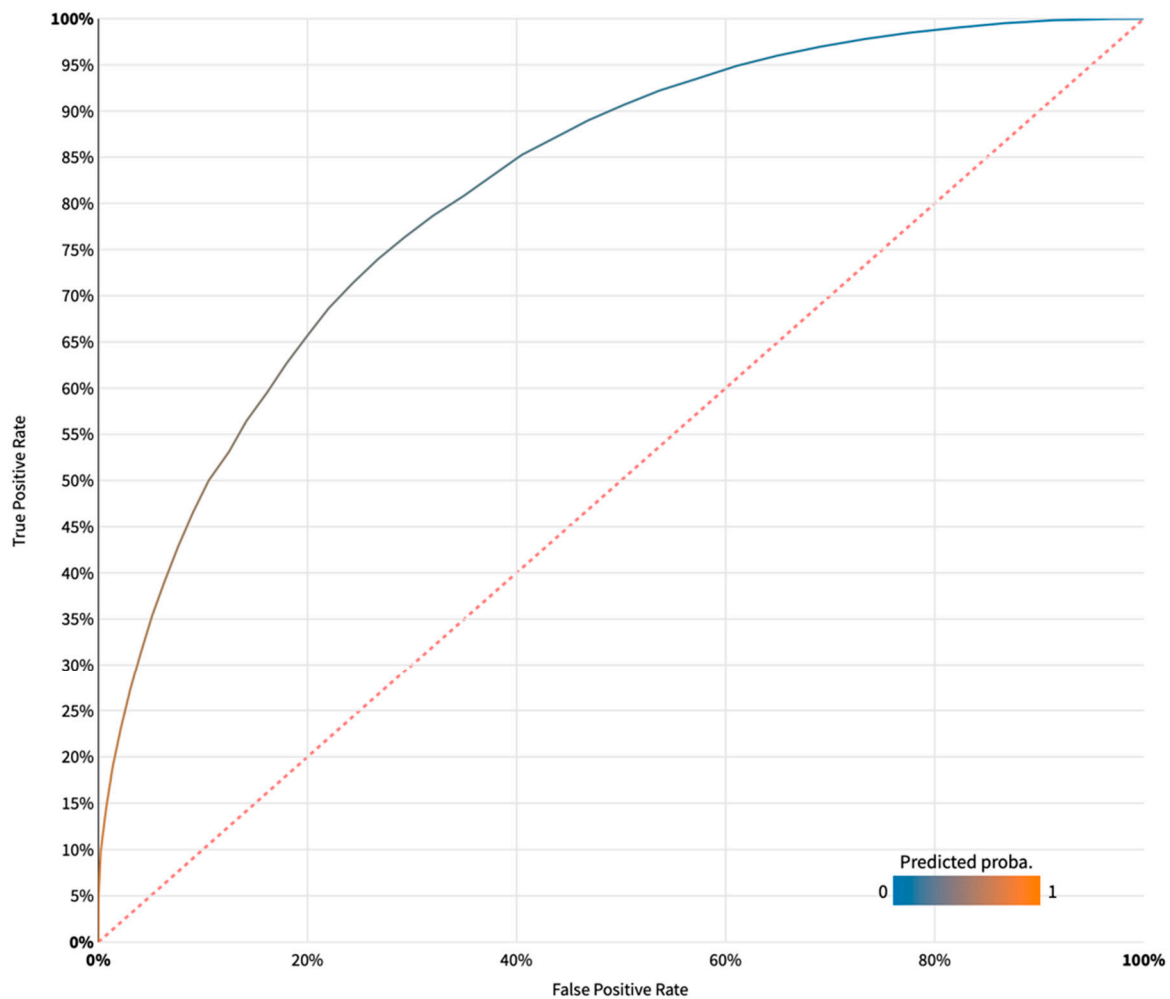


Figure 2. Distribution of the performance metric (AUC) of the super learning model.

4. Discussion

This section offers a discussion on the application, limitations, and future prospects of the research findings. It highlights the practical implications, potential challenges, and opportunities for further progress in the field of co-occurring mental health and substance use disorders among women.

4.1. Application

In ML, classification emerges as a vital task, involving the nuanced prediction of target classes for individual data instances [28]. Achieving optimal performance on diverse datasets necessitates the careful selection of suitable individual classifiers. The challenge lies in pinpointing the most suitable data mining or machine learning model tailored to a specific problem. To tackle this complexity, researchers often deploy an array of models to ascertain the utmost performance for a given scenario. AutoML platforms are appealing pre-packaged tools for constructing predictive models using healthcare data [47]. In this study, the AutoML interface was employed to forecast and differentiate the simultaneous co-occurrence of mental health and substance use disorders in women. Notably, AutoML consistently demonstrates enhanced precision and specificity across various research investigations. By harnessing the capabilities of AutoML, the study automated critical tasks encompassing model selection, hyperparameter optimisation, and feature engineering. This streamlined approach simplifies the analytical pipeline and substantiates an elevation in the accuracy of prediction models, marking a promising stride within computational health research.

A significant novelty of our study represents the statistical analysis using an AutoML interface to predict co-occurring mental and substance use disorders in women. This research serves as a practical demonstration of the presented statistical methods utilising an AutoML interface within a real-world context, offering valuable insights into predictive analytics in health-related domains.

The research finding showcasing the performance of the super learning model in distinguishing co-occurring mental health and substance use disorders among women carries substantial potential for transformative impact. The super learning model comprised two base learners: a Random Forest model and an XGBoost model, and it outperformed the individual base learners. The super learning model's accuracy in identification would enable early identification of women at risk of co-occurring mental health and substance use disorders. Women who receive substance use treatment that is tailored to their gender experience a longer duration of stay in treatment and have a higher probability of maintaining abstinence after completing treatment [1]. The emergence of more accurate and timely diagnosis has significant consequences for the development of improved treatment techniques, aimed at reducing the complications, illness, and death associated with these disorders [26]. Efficient resource allocation would be facilitated as the model's precision allows healthcare providers to focus on those at elevated risk, ensuring that support and treatment resources are channelled where they can yield the most significant benefits.

4.2. Limitations

AutoML platforms expedite the process of developing machine learning pipelines, and the models they produce can be used as initial frameworks for constructing predictive models. It is crucial to approach the integration of such a model with caution when determining the best output levels based on the research topic, considering ethical considerations, data security, and ongoing clinical supervision to ensure that its application aligns with the envisioned positive impact. The study acknowledges the fear of stigma preventing women from seeking substance abuse treatment, leading to a lower likelihood of them pursuing help compared to men [10], highlighting a potential limitation in real-world application. Although the super learning model demonstrated enhanced accuracy in differentiating co-occurring mental health and substance use disorders in women, it is important to acknowledge its limitations, including the possibility of false positives and false negatives. While the potential benefits of early identification and intervention on the health of women are promising, it may require time for these effects to become evident. Further research and validation are necessary to validate and measure these possible long-term benefits.

4.3. Future Prospects

The utilisation of AutoML platforms, as demonstrated by the study's implementation of Dataiku, signifies a progression towards enhancing patient outcomes in the medical domain, particularly in efficiently analysing and examining large collections of patient data. The implementation of streamlined methodologies and enhanced diagnostic procedures enhances the efficiency of healthcare operations, while potentially preserving resources. This could reduce the need for physical infrastructure such as storage rooms, as data administration and utilisation become more optimised. The research results regarding the performance of the super learning model have the potential to bring about significant changes, especially in the early detection of women who are at risk of experiencing both mental health and substance use issues simultaneously. Early diagnosis enables proactive intervention, customised support, and trauma-informed treatments [48], which address the fear of social disapproval and encourage more compassionate approaches to women's mental well-being. The research highlights that the model's accuracy in allocating resources effectively could help overcome the obstacles related to stigma, leading to improved health outcomes for both impacted women and their families and communities. Ultimately, the precise forecasts generated by the model have the potential to accelerate progress in

research methodologies and shape policy choices in the field of co-occurring mental health and substance use problems.

5. Conclusions

This study investigated the potential of AutoML for predicting co-occurring mental health and substance use disorders among women using TEDS-A data for 2020. Employing advanced statistical and machine learning techniques through Dataiku's AutoML interface, a super learning model achieved a high AUC of 0.817, demonstrating robust predictive capability. These findings highlight the promise of AutoML in healthcare, particularly the super learning model's potential as a diagnostic tool for early identification of co-occurring disorders in women. Future research should focus on disseminating knowledge about AutoML's advantages and ethical considerations in healthcare integration. Collaboration with policymakers, medical associations, and patient advocacy groups is crucial for establishing guidelines on responsible implementation, data privacy, and continuous performance monitoring. This holistic approach ensures maximising the model's benefits while adhering to the highest ethical standards, safeguarding patient well-being and privacy.

Author Contributions: Conceptualisation, N.A. and P.K.; methodology, N.A. and P.K.; software, N.A. and P.K.; validation, N.A., P.K., M.A. and J.S.; formal analysis, N.A. and P.K.; writing—original draft preparation, N.A. and P.K.; writing—review and editing, N.A., P.K., M.A. and J.S.; supervision, M.A. and J.S.; project administration, P.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. The data can be found here: <https://www.datafiles.samhsa.gov/dataset/treatment-episode-data-set-admissions-2020-teds-2020-ds0001> (accessed on 2 June 2023).

Conflicts of Interest: Author Padmaja Kar was employed by the company St Vincent's Care Services. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Louison, L.; Green, S.L.; Bunch, S.; Scheyett, A. The problems no one wants to see: Mental illness and substance abuse among women of reproductive age in North Carolina. *North Carol. Med. J.* **2009**, *70*, 454–458. [CrossRef]
2. Stewart, D.; Ashraf, I.; Munce, S. Women's mental health: A silent cause of mortality and morbidity. *Int. J. Gynecol. Obstet.* **2006**, *94*, 343–349. [CrossRef] [PubMed]
3. Kokane, S.S.; Perrotti, L.I. Sex Differences and the Role of Estradiol in Mesolimbic Reward Circuits and Vulnerability to Cocaine and Opiate Addiction. *Front. Behav. Neurosci.* **2020**, *14*, 74. [CrossRef] [PubMed]
4. McCaul, M.E.; Roach, D.; Hasin, D.S.; Weisner, C.; Chang, G.; Sinha, R. Alcohol and women: A brief overview. *Alcohol. Clin. Exp. Res.* **2019**, *43*, 774. [CrossRef]
5. Fox, H.C.; Sinha, R. Sex differences in drug-related stress-system changes: Implications for treatment in substance-abusing women. *Harv. Rev. Psychiatry* **2009**, *17*, 103–119. [CrossRef] [PubMed]
6. Prieto-Arenas, L.; Díaz, I.; Arenas, M.C. Gender differences in dual diagnoses associated with cannabis use: A review. *Brain Sci.* **2022**, *12*, 388. [CrossRef]
7. Ruiz, M.A.; Douglas, K.S.; Edens, J.F.; Nikolova, N.L.; Lilienfeld, S.O. Co-occurring mental health and substance use problems in offenders: Implications for risk assessment. *Psychol. Assess.* **2012**, *24*, 77–87. [CrossRef]
8. Forster, M.; Rogers, C.J.; Tinoco, S.; Benjamin, S.; Lust, K.; Grigsby, T.J. Adverse childhood experiences and alcohol related negative consequence among college student drinkers. *Addict. Behav.* **2023**, *136*, 107484. [CrossRef]
9. Larsen, J.L.; Johansen, K.S.; Mehlsen, M.Y. What kind of science for dual diagnosis? A pragmatic examination of the enactive approach to psychiatry. *Front. Psychol.* **2022**, *13*, 825701. [CrossRef]
10. Agerberg, S.; Schubert, N.; Overington, L.; Corace, K. Treatment barriers among individuals with co-occurring substance use and mental health problems: Examining gender differences. *J. Subst. Abus. Treat.* **2020**, *112*, 29–35. [CrossRef]
11. Acion, L.; Kelmansky, D.; van der Laan, M.; Sahker, E.; Jones, D.; Arndt, S. Use of a machine learning framework to predict substance use disorder treatment success. *PLoS ONE* **2017**, *12*, e0175383. [CrossRef]

12. Miranda, O.; Fan, P.; Qi, X.; Wang, H.; Brannock, M.D.; Kosten, T.R.; Ryan, N.D.; Kirisci, L.; Wang, L. DeepBiomarker2: Prediction of Alcohol and Substance Use Disorder Risk in Post-Traumatic Stress Disorder Patients Using Electronic Medical Records and Multiple Social Determinants of Health. *J. Pers. Med.* **2024**, *14*, 94. [[CrossRef](#)]
13. Adams, R.S.; Jiang, T.; Rosellini, A.J.; Horváth-Puhó, E.; Street, A.E.; Keyes, K.M.; Cerdá, M.; Lash, T.L.; Sørensen, H.T.; Gradus, J.L. Sex-Specific Risk Profiles for Suicide Among Persons with Substance Use Disorders in Denmark. *Addiction* **2021**, *116*, 2882–2892. [[CrossRef](#)]
14. Aishwarya, N.; Yathishan, D.; Alageswaran, R.; Manivannan, D. AutoML Based IoT Application for Heart Attack Risk Prediction. In Proceedings of the Decision Intelligence Solutions, Singapore, 2–3 March 2023; 2023; pp. 19–29.
15. Kundu, A.; Chaiton, M.; Billington, R.; Grace, D.; Fu, R.; Logie, C.; Baskerville, B.; Yager, C.; Mitsakakis, N.; Schwartz, R. Machine Learning Applications in Mental Health and Substance Use Research Among the LGBTQ2S+ Population: Scoping Review. *JMIR Med Inf.* **2021**, *9*, e28962. [[CrossRef](#)] [[PubMed](#)]
16. Johnstone, S.; Dela Cruz, G.A.; Kalb, N.; Tyagi, S.V.; Potenza, M.N.; George, T.P.; Castle, D.J. A systematic review of gender-responsive and integrated substance use disorder treatment programs for women with co-occurring disorders. *Am. J. Drug Alcohol Abus.* **2023**, *49*, 21–42. [[CrossRef](#)] [[PubMed](#)]
17. Waring, J.; Lindvall, C.; Umeton, R. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artif. Intell. Med.* **2020**, *104*, 101822. [[CrossRef](#)] [[PubMed](#)]
18. Obermeyer, Z.; Powers, B.; Vogeli, C.; Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **2019**, *366*, 447–453. [[CrossRef](#)] [[PubMed](#)]
19. Mustafa, A.; Rahimi Azghadi, M. Automated Machine Learning for Healthcare and Clinical Notes Analysis. *Computers* **2021**, *10*, 24. [[CrossRef](#)]
20. Beam, A.L.; Kohane, I.S. Big Data and Machine Learning in Health Care. *JAMA* **2018**, *319*, 1317–1318. [[CrossRef](#)] [[PubMed](#)]
21. Rajkomar, A.; Dean, J.; Kohane, I. Machine learning in medicine. *N. Engl. J. Med.* **2019**, *380*, 1347–1358. [[CrossRef](#)]
22. Tsamardinos, I.; Charonyktakis, P.; Papoutsoglou, G.; Borboudakis, G.; Lakiotaki, K.; Zenklusen, J.C.; Juhl, H.; Chatzaki, E.; Lagani, V. Just Add Data: Automated predictive modeling for knowledge discovery and feature selection. *NPJ Precis. Oncol.* **2022**, *6*, 38. [[CrossRef](#)] [[PubMed](#)]
23. Thomaidis, G.V.; Papadimitriou, K.; Michos, S.; Chartampilas, E.; Tsamardinos, I. A characteristic cerebellar biosignature for bipolar disorder, identified with fully automatic machine learning. *IBRO Neurosci. Rep.* **2023**, *15*, 77–89. [[CrossRef](#)] [[PubMed](#)]
24. Naser, M.Z. Machine learning for all! Benchmarking automated, explainable, and coding-free platforms on civil and environmental engineering problems. *J. Infrastruct. Intell. Resil.* **2023**, *2*, 100028. [[CrossRef](#)]
25. Perotte, A.; Pivovarov, R.; Natarajan, K.; Weiskopf, N.; Wood, F.; Elhadad, N. Diagnosis code assignment: Models and evaluation metrics. *J. Am. Med. Inf. Assoc.* **2014**, *21*, 231–237. [[CrossRef](#)] [[PubMed](#)]
26. Zhuhadar, L.P.; Lytras, M.D. The Application of AutoML Techniques in Diabetes Diagnosis: Current Approaches, Performance, and Future Directions. *Sustainability* **2023**, *15*, 13484. [[CrossRef](#)]
27. Barenholtz, E.; Fitzgerald, N.D.; Hahn, W.E. Machine-learning approaches to substance-abuse research: Emerging trends and their implications. *Curr. Opin. Psychiatry* **2020**, *33*, 334–342. [[CrossRef](#)]
28. Kabir, M.F.; Ludwig, S.A. Enhancing the Performance of Classification Using Super Learning. *Data-Enabled Discov. Appl.* **2019**, *3*, 5. [[CrossRef](#)]
29. Van der Laan, M.J.; Rose, S. *Targeted Learning: Causal Inference for Observational and Experimental Data*; Springer: Berlin/Heidelberg, Germany, 2011; Volume 4.
30. Laan, M.J.V.D.; Polley, E.C.; Hubbard, A.E. Super Learner. *Stat. Appl. Genet. Mol. Biol.* **2007**, *6*. [[CrossRef](#)]
31. Comartin, E.B.; Burgess-Proctor, A.; Harrison, J.; Kubiak, S. Gender, Geography, and Justice: Behavioral Health Needs and Mental Health Service Use Among Women in Rural Jails. *Crim. Justice Behav.* **2021**, *48*, 1229–1242. [[CrossRef](#)]
32. Zhao, Q.; Kong, Y.; Henderson, D.; Parrish, D. Arrest Histories and Co-Occurring Mental Health and Substance Use Disorders Among Women in the USA. *Int. J. Ment. Health Addict.* **2023**. [[CrossRef](#)]
33. SAMHSA. *Treatment Episode Data Set Admissions (TEDS-A) 2020*; SAMHSA: Rockville, MD, USA, 2023.
34. Standeven, L.R.; Scialli, A.; Chisolm, M.S.; Terplan, M. Trends in cannabis treatment admissions in adolescents/young adults: Analysis of TEDS-A 1992 to 2016. *J. Addict. Med.* **2020**, *14*, e29–e36. [[CrossRef](#)]
35. Baird, A.; Cheng, Y.; Xia, Y. Use of machine learning to examine disparities in completion of substance use disorder treatment. *PLoS ONE* **2022**, *17*, e0275054. [[CrossRef](#)]
36. Yang, J.C.; Roman-Urrestarazu, A.; Brayne, C. Differences in receipt of opioid agonist treatment and time to enter treatment for opioid use disorder among specialty addiction programs in the United States, 2014–2017. *PLoS ONE* **2019**, *14*, e0226349. [[CrossRef](#)]
37. Pozo-Luyo, C.A.; Cruz-Duarte, J.M.; Amaya, I.; Ortiz-Bayliss, J.C. Forecasting PM2.5 concentration levels using shallow machine learning models on the Monterrey Metropolitan Area in Mexico. *Atmos. Pollut. Res.* **2023**, *14*, 101898. [[CrossRef](#)]
38. Egger, R. *Software and tools. Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications*; Springer: Cham, Switzerland, 2022; pp. 547–588.
39. Tapeh, A.T.G.; Naser, M.Z. Artificial Intelligence, Machine Learning, and Deep Learning in Structural Engineering: A Scientometrics Review of Trends and Best Practices. *Arch. Comput. Methods Eng.* **2023**, *30*, 115–159. [[CrossRef](#)]

40. Sahker, E.; Acion, L.; Arndt, S. National analysis of differences among substance abuse treatment outcomes: College student and nonstudent emerging adults. *J. Am. Coll. Health* **2015**, *63*, 118–124. [[CrossRef](#)]
41. Glasheen, C.; Pemberton, M.R.; Lipari, R.; Copello, E.A.; Mattson, M.E. Binge drinking and the risk of suicidal thoughts, plans, and attempts. *Addict. Behav.* **2015**, *43*, 42–49. [[CrossRef](#)]
42. Alang, S.M. Sociodemographic disparities associated with perceived causes of unmet need for mental health care. *Psychiatr. Rehabil. J.* **2015**, *38*, 293. [[CrossRef](#)] [[PubMed](#)]
43. Huang, J.-C.; Tsai, Y.-C.; Wu, P.-Y.; Lien, Y.-H.; Chien, C.-Y.; Kuo, C.-F.; Hung, J.-F.; Chen, S.-C.; Kuo, C.-H. Predictive modeling of blood pressure during hemodialysis: A comparison of linear model, random forest, support vector regression, XGBoost, LASSO regression and ensemble method. *Comput. Methods Programs Biomed.* **2020**, *195*, 105536. [[CrossRef](#)]
44. Hong, W.; Zhou, X.; Jin, S.; Lu, Y.; Pan, J.; Lin, Q.; Yang, S.; Xu, T.; Basharat, Z.; Zippi, M. A comparison of XGBoost, random forest, and nomograph for the prediction of disease severity in patients with COVID-19 pneumonia: Implications of cytokine and immune cell profile. *Front. Cell. Infect. Microbiol.* **2022**, *12*, 819267. [[CrossRef](#)]
45. Meng, D.; Xu, J.; Zhao, J. Analysis and prediction of hand, foot and mouth disease incidence in China using Random Forest and XGBoost. *PLoS ONE* **2021**, *16*, e0261629. [[CrossRef](#)] [[PubMed](#)]
46. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [[CrossRef](#)]
47. Romero, R.A.A.; Deypalan, M.N.Y.; Mehrotra, S.; Jungao, J.T.; Sheils, N.E.; Manduchi, E.; Moore, J.H. Benchmarking AutoML frameworks for disease prediction using medical claims. *BioData Min.* **2022**, *15*, 15. [[CrossRef](#)] [[PubMed](#)]
48. Apsley, H.B.; Vest, N.; Knapp, K.S.; Santos-Lozada, A.; Gray, J.; Hard, G.; Jones, A.A. Non-engagement in substance use treatment among women with an unmet need for treatment: A latent class analysis on multidimensional barriers. *Drug Alcohol Depend.* **2023**, *242*, 109715. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.