# EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR PSYCHOSIS PROGNOSIS PREDICTION

MAARTEN ROEST

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

1274468

COMMITTEE

dr. Seyed Mostafa Kia
dr. Marijn van Wingerden

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

June 30th, 2023

WORD COUNT

8769

ACKNOWLEDGMENTS

I would like to take this opportunity to express my gratitude to those who have supported me throughout the process of completing this thesis and acknowledge their invaluable contributions to this work. Foremost, I extend my deepest appreciation to Dr. Seyed Mostafa Kia, whose support, valuable insights, and willingness to assist me have been instrumental in the successful completion of this research. His profound knowledge and expertise in the field of Data Science have greatly contributed to my growth as a proficient Data Scientist. Furthermore, I would like to express my gratitude to Dr. Marijn van Wingerden for devoting time in reading and reviewing my thesis.

Maarten Roest
Amsterdam, June 2023

# EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR PSYCHOSIS PROGNOSIS PREDICTION

MAARTEN ROEST

## Abstract

*Extensive research has been conducted on the prediction of psychosis prognosis and treatment outcomes, yielding valuable insights into the use of machine learning models for accurate prediction. Multiple studies have demonstrated the potential of these models in effectively forecasting the prognosis and outcomes associated with psychosis, highlighting their promising application in clinical settings. This thesis aims to investigate the predictive value of various features in psychosis prognosis prediction. The study distinguishes itself by utilizing a random forest classifier with static feature inputs and employing random forest feature importances, permutation feature importance, and SHAP to determine feature importance scores. This empowers clinicians and researchers to prioritize the most influential factors for targeted interventions. The OPTiMiSE dataset is utilized, which includes information on clients with schizophrenia, schizophreniform, or schizoaffective disorder who participated in a three-phase switching study on antipsychotic medication. The main findings indicate that Positive and Negative Syndrom Scale scores, a clinical assessment tool that measures the severity of positive and negative symptoms in individuals with schizophrenia, at baseline are the most influential factors for psychosis prognosis prediction. This knowledge enables healthcare professionals to identify high-risk patients and provide timely interventions for enhanced outcomes. Furthermore, a novel approach has emerged for determining the extent to which individual features need to be modified at a local scale in order to influence prediction outcomes, providing insights into the direction and magnitude of feature modifications necessary to influence the classifier's outcomes.*

## SOURCE/CODE/ETHICS/TECHNOLOGY STATEMENT

The dataset used in this thesis was collected by Kahn et al. (2018). It is important to note that this project did not involve the collection of data from human participants or animals. Prior to their participation, all subjects provided informed consent. To access the data, the author has signed a non-disclosure agreement. The original data and code used in this thesis are the property of their respective owner, who retains ownership rights both during and after the completion of this research. The author of this thesis acknowledges and confirms that they do not possess any legal claim to this data or code. Consequently, the complete code utilized in this thesis is not accessible to the public. All figures belong to the author. In my academic writing process, I have employed a thesaurus tool from "Quillbot" (2023) to enhance the variety, richness of my vocabulary, and alert me for grammar and spelling errors.

# 1 INTRODUCTION

## 1.1 *Project Description*

Psychosis refers to a mental state where an individual experiences a disconnection from reality. It is a symptom of several mental disorders, including schizophrenia and schizophreniform disorder. The treatment for schizophrenia and schizophreniform disorder typically involves a combination of medication, therapy, and support. Currently, there is no single best treatment for schizophrenia and schizophreniform disorder as treatment is highly individualized and can vary depending on the severity of symptoms and personal circumstances. As a result, the prognosis for clients with schizophrenia and schizophreniform disorder is generally unpredictable (Patel et al., 2014).

Machine learning (ML) has emerged as a promising tool for improving health treatment outcomes and optimizing clinical decision-making. To assist medical professionals, ML algorithms should incorporate a degree of explainability, enabling human experts to trace the decisions made and exercise their own judgement (Adlung et al., 2021). The main objective of this paper is to thoroughly examine and emphasize the significant features that play a crucial role in predicting the prognosis of psychosis. This investigation is carried out by employing various techniques for feature importance assessment, as well as carrying out a novel counterfactual explanation method. These post-hoc interpretability methods are used in the field of Explainable Artificial Intelligence (XAI).

## 1.2 *Motivation*

This section outlines the societal and scientific relevance of the stated problem. Firstly, the societal relevance will be addressed, followed by a discussion of proposed solutions for existing gaps in the scientific literature.

### 1.2.1 *Societal relevance*

The global prognosis for several mental disorders, such as schizophrenia and schizophreniform disorder, is deemed unacceptably poor due to inadequate resourcing of mental health care. Timely intervention and treatment during the first phase of psychosis assume critical importance for improving the long-term outcomes of clients. The primary objectives of first-phase psychosis treatment include symptom reduction, improvement in functioning, and prevention of relapse (D MCGORRY, 2002). Enhanced comprehension of the critical factors that significantly impact the prognosis of clients has the potential to improve both the accuracy and expediency of psychosis prognosis prediction. Identifying the critical factors that significantly influence the prognosis allows healthcare providers to develop targeted interventions that are more effective in treating the condition, leading to a reduction in treatment costs (Ricciardi et al., 2008). According to Fusar-Poli et al. (2017) the identification of risk factors and indicators to facilitate a more personalized and effective treatment plan for clients leads to improved overall mental health, enhancing the quality of life for both clients and their families.

### 1.2.2 *Scientific relevance*

There are several ongoing investigations aimed at optimizing the psychosis prognosis prediction. ML algorithms can facilitate the development of personalized treatment plans for client by analyzing their medical history, genetic makeup, and other relevant factors. Although promising, the European General Data Protection Regulation (GDPR) and other regulations are making it more difficult to utilize complex ML techniques since retraceability of decisions is a requirement (European Commission, 2016). Therefore, post-hoc interpretability techniques have recently made substantial progress. These XAI techniques aim to approximate complex ML techniques using simpler, interpretable models that can be examined to explain the behavior of these ML techniques (Gunning et al., 2019). The objective of this study is to employ various ML methodologies to evaluate and contrast their efficacy in predicting the prognosis of first phase psychosis. The top-performing algorithm will undergo an XAI approach to elucidate the rationale behind its decisions, thereby rendering the findings comprehensible to human. The first step involves the measurement of feature importance scores. Then, a thoroughly researched technique known as a counterfactual explanation method will be employed. This technique can facilitate the analysis of client data by generating hypothetical scenarios based on alterations to the input data. Additionally, counterfactual explanations can be used to identify the most influential factors in a client's data, identifying potential risk factors or indicators of disease progression (Verma et al., 2020). The objective is to contribute to enhancing the interpretability of an algorithm's decision-making process. Specifically, the aim is to determine the extent to which different features predominantly contribute to decision. Additionally, the research aims to ascertain the extent to which certain features need to change in order to invert the target variable by generating counterfactual explanations. In order to achieve this, this research uses the Diverse Counterfactual Explanations (DICE) method. DICE is introduced by Mothilal et al. (2020). The DICE method generates counterfactual explanations by optimizing diversity while ensuring similarity to the original input. In other words, the method aims to generate counterfactual explanations that differ from the original input yet remain plausible and share as much as possible similarities with it.

## 1.3 *Research Questions*

As became apparent, it is important to identify which features have the greatest impact on treatment success or failure through mapping and analysis. This research aims to gain a deeper understanding of the features that lead ML techniques to make specific prognosis predictions for psychosis. Additionally, this research aims to investigate the extent to which changing specific features can result in a different outcome in the ML model's prognosis prediction. This leads to the following main research question of this paper:

- **"Which features have the highest predictive value on psychosis prognosis prediction?"** "

In order to address the research question, a series of steps must be undertaken. The objective is to generate counterfactual explanations to investigate the factors that contribute to a ML model's psychosis prognosis prediction. To accomplish this, different classification models will be applied to the psychosis prognosis dataset, and their performances will be evaluated and compared. To assess and compare the performances of the different ML techniques, repeated stratified hold-out will be employed. This method randomly splits the dataset in a training set and test set while ensuring that the class distribution in the target variable is maintained. This method is being used due to the limited size of the dataset, which comprises 481 observations, and the unbalanced distribution of the target variable. Evaluation metrics used include Area Under the Curve (AUC) and Area Under the Precision-recall curve (AUC-PR), which have been addressed extensively (see section 3). To give answer to this, the first sub-question is

- **"What are the comparative performance results of Random Forest and Support Vector Machine models in predicting psychosis prognosis?"** "

The subsequent step involves identifying the features that significantly contribute to the decision-making process of the ML model. In order to achieve this objective, the feature importance scores derived from the ML algorithm that achieves the highest performance on the evaluation metrics will be employed. In this regard, three distinct methods will be employed to assess feature importance and determine the extent of their impact on the model's predictions. The scores are normalized so they sum up to 1.0, and higher scores indicate that a feature is more important in making the model's decision. These scores are calculated over 50 iterations and visualized in box plot charts per feature to compare the importance distribution over multiple iterations. The objective of this comparative analysis of feature importance rankings obtained from distinct methodologies is to answer the second sub-question:

- **"How consistent are the results obtained from different feature importance methods in determining the most significant features for predicting psychosis prognosis"** "

Finally, the DiCE method will be applied to generate counterfactual explanations. The main research question will be addressed through a comparative analysis of three distinct methods for obtaining feature importance scores and generating counterfactual explanations. This investigation aims to examine the differences and similarities among these methods, shedding light on their effectiveness in providing insights into the underlying factors influencing the model's decision-making process. By evaluating and comparing the results obtained from these approaches, valuable insights will be gained towards answering the primary research question.

## 1.4 *Results*

The main results of the study are as follows. Firstly, the analysis identified 24 features that demonstrated the highest predictive value for psychosis prognosis prediction, providing valuable insights into the underlying factors influencing the prognosis. Secondly, the study introduced a novel approach using a strip plot for visualizing counterfactual explanations, which showed promise as a valuable tool for understanding the impact of feature changes on the prediction outcomes. Additionally, the study compared two ML algorithms and found that one algorithm exhibited slightly superior performance across diverse evaluation metrics, suggesting its potential effectiveness in psychosis prognosis prediction. However, it is worth noting that a statistical test did not demonstrate significant superiority of one algorithm over the other, indicating the need for further investigation and comparison in future studies.

### 1.5 *Thesis outline*

This research paper follows a structured approach. Section 2 provides the theoretical background, while Section 3 outlines the methodologies used. Section 4 details the experimental setup, and Section 5 presents and analyzes the results. Section 6 discusses the findings and suggests future research directions. Finally, Section 7 offers a concise summary, encapsulating the key findings.

## 2 RELATED WORK

In this chapter, psychosis prognosis prediction strategies are briefly discussed. Thereafter, used algorithms to predict psychosis prognosis are explained. Followed by an introduction to XAI and counterfactual explanation methods.

### 2.1 *Predictive models and features for psychosis prognosis*

Over the past few years, ML has emerged as a promising approach for predicting treatment outcomes in psychosis research. A review by Del Fabro et al. (2023) examined the use of ML in predicting antipsychotic treatment outcomes in clients with schizophrenia at various stages, utilizing neuroimaging, neurophysiological, genetic, and clinical features. Multiple studies have demonstrated the potential of ML models based on sociodemographic and clinical features for effective prediction.

According to a study by Podichetty et al. (2021), which included 639 clients with psychotic disorders, PANSS scores at baseline were found to be the most predictive features. The study compared the performance of different ML classifiers (Support Vector Machine (SVM), Logistic Regression (LR), Naive Bayes, and Random Forest (RF)), and the RF algorithm demonstrated the best ability to classify clients' responses after 6 months of treatment. It achieved an AUC of 0.7 on the test set.

Li et al. (2021) compared different ML approaches (LR, stochastic gradient descendent, gradient boosting decision tree, extreme gradient boosting, and RF) to predict social functioning improvement for clients with schizophrenia treated with Second-Generation Antipsychotics (SGA). The RF algorithm achieved the highest AUC value of 0.86. In comparison with the study by Podichetty et al., Li et al. discovered that not only PANSS total scores at baseline is the most important feature, but also the use of mood stabilizers and social functioning. A study conducted by Koutsouleris et al. (2016), used socio-demographical and clinical features to predict the treatment outcome in first-episode psychosis (FEP) clients. The study employed a ML analysis using pre-treatment clinical information, specifically psychosocial, sociodemographic, and psychometric variables, as features to predict functional outcomes after treatment with First-Generation Antipsychotics or SGA. Various ML algorithms, including (non)linear SVM, decision trees, and logistic regression, were evaluated in a comparative study. The results indicated that the highest balanced accuracy of 71.7 percent was achieved by the nonlinear SVM model. Psychosocial features, rather than symptom data, were found to be the most valuable predictors.

Different studies on ML for psychosis prognosis prediction have identified various algorithms that achieve the highest prediction scores. Therefore, this research will consider multiple ML algorithms and compare their performances on the dataset.

## 2.2 *Feature importance scores for machine learning methods*

In the realm of health-related applications, feature importance scores hold significant importance in comprehending the pivotal factors that contribute to precise predictions and informed decision-making within ML models. The recognition of essential features is critical in acquiring insights into the fundamental relationships between health-related predictors and the target variable, thereby leading to enhanced diagnoses, treatments, and overall client care (Bracher-Smith et al., 2022). Demircioğlu (2022) conducted a comprehensive evaluation of 29 feature importance algorithms using 10 classifiers. Their study aimed to assess the stability of various methods and measure the pairwise similarity between them. The findings revealed that simpler methods exhibited greater stability compared to more complex ones.

This research study centers around the investigation of three feature importance methods: RF feature importances (Pedregosa et al., 2011), permutation feature importance (Breiman, 2001), and SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017b). Among these methods, SHAP is regarded as the most complex, providing a comprehensive and nuanced understanding of feature contributions. On the other hand, permutation feature importance and RF feature importances are comparatively simpler and computationally efficient. These simpler methods have been demonstrated to yield reliable and interpretable feature importance scores (Demircioğlu, 2022).

## 2.3 *Explainable Artificial Intelligence for healthcare*

As mentioned in section *2.2.2 Scientific relevance*, ML algorithms should incorporate a degree of explainability to assist medical professionals. XAI has emerged as a trend in Artificial Intelligence (AI), which emphasizes the explainability of traditional AI models by utilizing the decision-making and prediction outputs of the models. The incorporation of the explainability factor has provided new opportunities to the black-box models and instilled confidence in healthcare stakeholders in interpreting ML and deep learning models (Adadi & Berrada, 2020). In contrast to models that are transparent and able to display the most relevant features for their output, deep neural networks do not intrinsically provide a display of such feature importance. Although interpretation of a model based on feature importance may be possible at global level, it is still challenging to estimate how the model will behave for an individual example (Dey et al., 2022). A substantial amount of research has recently been concentrated on postulating several explainability approaches allowing professionals to explore global and local explanations and comprehend why trained ML models make particular predictions. Generating explanations for the output of a given algorithm in terms of reasonable yet non-occurring alternatives can be improved by suggesting modifications to the input that can lead to a different decision by the algorithm. These are referred to as counterfactual explanations (Stepin et al., 2021).

### 2.3.1 *Counterfactual explanation method*

There are varying methods for generating counterfactual explanations. A Counterfactual Conditional Heterogeneous Variational AutoEncoder (CCHVAE), introduced by Pawelczyk et al. (2020), uses a conditional hetero-encoder to learn a representation of the input data and generates counterfactual examples by modifying the learned representation. In contrast, Diverse Counterfactual Explanations (DICE) is a post-hoc and model-agnostic method that generates diverse sets of counterfactual explanations that can be used with any type of ML model (Mothilal et al., 2020). Another method, Feasible and Actionable Counterfactual Explanations (FACE) generates counterfactual explanations by first identifying the most important features for a given prediction (Poyiadzi et al., 2020). Selecting an appropriate counterfactual explanation method is a crucial component for generating meaningful counterfactual explanations (Pawelczyk et al., 2021).

### 2.3.2 *Benchmarking studies*

de Oliveira and Martens (2021) conducted a framework and benchmarking study for counterfactual methods on tabular data. They compared 10 counterfactual explanation methods on 22 tabular datasets. These tabular datasets are divided into three types of data: numerical, categorical, and mixed. The novel evaluation framework consists of nine relevant metrics to evaluate counterfactual results. The results show that no single counterfactual method scored best on all evaluation metrics. The dataset employed in this thesis comprises both numerical and categorical data, thereby qualifying it as a mixed dataset. In case of mixed datasets, de Oliveira and Martens verified a statistical tie between DICE and SE. Based on the coverage evaluation metric, DICE was found to be capable of generating the most counterfactual explanations (89.6 percent).

The Counterfactual And Recourse LibrAry (CARLA) framework, conducted by Pawelczyk et al. (2021), presents a novel approach for benchmarking and evaluating counterfactual methods. This study distinguishing counterfactual methods into three distinct categories - independence-based, dependence-based, and causality-based - by emphasizing the significance of features and the advisory element. CARLA uses slightly different metrics compared to de Oliveira and Martens. CARLA employs a metric that measures the degree of support for counterfactual explanations based on the positively classified instances within the data. It also measures redundancy to identify cases where multiple counterfactual explanations provide overlapping information. Costs, constraint violation, success rate, and average time are intersecting metrics for both the CARLA framework and the framework and benchmarking study for counterfactual methods on tabular data. The findings from the CARLA framework indicate that, for both independence-based and dependence-based methods, no single algorithm demonstrated superior performance compared to its competitors on all six metrics.

The selection of the counterfactual explanation algorithm in this thesis is based on the findings from two benchmarking studies conducted by de Oliveira and Martens and Pawelczyk et al. Although the DICE counterfactual explanation method did not outperform other methods in all metrics, it demonstrated favorable performance overall.

### 2.3.3 *Diverse Counterfactual Explanations*

This thesis will utilize the DICE method for generating counterfactual explanations for psychosis prognosis prediction. Firstly, DICE is a model-agnostic counterfactual explainer (Mothilal et al., 2020). Irrespective of the performances of a given ML algorithm, DICE can be employed to generate counterfactual explanations. The method can be leveraged to provide explanations for the best-performing algorithm that is being compared, in order to address the first sub-question. A comprehensive explanation of the DICE method's functionality and workings can be found in Section 4 of this paper.

# 3 METHOD

This section begins by providing a detailed description of the target variable. Following that, the section proceeds to provide a comprehensive description of the ML methods that have been adopted for the study. This section will not go into the specifics of every method, as RF and SVM are considered common knowledge in ML. Methods for extracting feature importance scores will be described. The full pipeline can be found in the experimental setup.

## 3.1 *The Positive and Negative Syndrome Scale (PANSS)*

The Positive and Negative Syndrome Scale is a clinical tool used in psychiatry to assess symptom severity of clients with schizophrenia and other psychotic disorders. The scale, which comprises 30 items rated on a 7-point Likert scale ranging from 1(absent) to 7(extreme, was first introduced by Kay et al. (1987). The PANSS is divided into three subscales: positive symptoms (7 items), negative symptoms (7 items), and general psychopathology (16 items). In order to assess an individual's symptoms using the PANSS, a clinician or researcher conducts an interview based on these 30 items. The overall severity of symptoms is calculated by summing the ratings for each item, resulting in a total score, with higher scores indicating greater symptom severity.

This thesis focuses on symptomatic remission, defined by Andreasen et al. (2005) and used as a binary target variable. According to Andreasen's criteria, symptomatic remission was defined as the presence of no more than mild symptoms (with a maximum rating of 3) on eight specific PANSS items (P1, P2, P3, N1, N4, N6, G5, and G9), which do not interfere with daily life functioning. However, unlike the Andreasen criteria, the minimum duration of symptom severity of 6 months was not applied.

## 3.2 *Classification Algorithms*

As outlined in the literature review section, it has been established that RF and SVM models are effective models for predicting psychosis prognosis.

In this study, these algorithms will be utilized to develop predictive models for psychosis prognosis, and their performance will be evaluated and compared.

### 3.2.1 *Random Forest*

RF was first introduced by Breiman (2001), and is a supervised learning algorithm. The principles of RF aim to produce a collection of diverse and accurate decision trees that are less prone to overfitting than a single decision tree. In RF algorithm, a technique called bootstrap is used to randomly sample the training data with replacement, generating multiple datasets. Each dataset is used to train a decision tree, and the predictions of these trees are combined to produce the final prediction. RF is widely used in social science for several reasons (Schonlau & Zou, 2020). Besides their robustness, RF provide feature importance scores, which can help researchers understand the relative importance of features. RF are flexible and can be used with a wide range of data types, including categorical and continuous variables, making them suitable for many types of social science research (Schonlau & Zou, 2020).

### 3.2.2 *Support Vector Machine*

Widely utilized for binary classification problems, SVM is a supervised learning algorithm introduced by Cortes and Vapnik (1995). SVM aim to find the hyperplane that best separates the data points of different classes, while maximizing the margin between them. SVM can handle non-linearly separable data. The soft margin and regularization help to avoid overfitting and improve generalization performance.

## 3.3 *Feature importance scores*

The present study aims to investigate the feature importance of ML algorithms. Several methods are available to obtain feature importance scores for ML algorithms, including built-in feature importance methods provided by some algorithms and external libraries that are compatible with multiple algorithms. The methods employed in this research are discussed in this subsection.

### 3.3.1 *Feature importances for Random Forest*

Feature importances is an attribute from Pedregosa et al. (2011), available in ML algorithms including RF. It provides a measure of importance of each feature in the algorithm's prediction. The score is computed by measuring the total reduction in impurity, typically measured using either the Gini impurity or entropy, that is achieved by splitting on that feature across all decision trees in the forest. Features that contribute more to the reduction in impurity are assigned higher importance scores, while features that do not contribute much or at all are assigned lower scores or zero. This attribute enables easy comparison of the relative importance of each feature in the model and helps identify the most influential features for the given dataset. The formula for feature importance scores is:

$$\text{importance}j = \frac{\sum \text{trees} w_{\text{tree}} \times (\text{impurityparent} - \text{impurityleft} - \text{impurityright})}{\sum \text{trees} w_{\text{tree}}} \tag{1}$$

Here, $W$ is the weight of a tree in the forest. *Impurityparent* is the impurity of the parent node before the split. *Impurityleft* and *impurityright* are the impurities of the left and right child nodes after the split. The summation is done over all the decision trees and the scores are normalized by dividing the score by the sum of all feature importance scores.

### 3.3.2 *Permutation feature importance*

Breiman (2001) introduced permutation importance. Permutation importance is a technique that assesses the importance of each feature in a RF model by randomly permuting the values of that feature and observing the resulting reduction in the model's performance. The extent of the reduction in performance after the permutation indicates the degree to which the model relies on that particular feature. When shuffling a feature has a negligible effect on the model's performance, it implies that the feature is not significant for the model's predictions. Conversely, when the model's performance decreases significantly, it indicates that the feature is critical for the model's predictions. Pedregosa et al. (2011), introduced a library to calculate feature importance for different classifiers. This formula can be used both for RF and SVM.

### 3.3.3 *SHapley Additive exPlanations*

The concept of Shapley Additive Explanations (SHAP) was introduced by Lundberg and Lee (2017b). SHAP is a technique that is grounded on the concept of Shapley values derived from cooperative game theory. It offers a mechanism to justly allocate the total payout among the players in a coalition. The SHAP approach generates explanations that are precise at the local level, i.e., they explain the prediction for a specific instance, while also being consistent at the global level, i.e., they explain the model as a whole.

Although methods such as feature importance scores and permutation feature importance provide explanations at the global level, SHAP has the added capability to explain models at both the local and global level. However, one common limitation of these methods is their inability to provide information on the extent to which features should be changed in order to elicit a different prediction from the ML algorithm.

### 3.4 *Diverse Counterfactual Explanations*

The input of the DICE method is a trained ML model. DICE is capable of generating a diverse set of counterfactual examples to increase the likelihood of an example being feasible. The dataset used in this research contains a large number of features, which could lead to a high number of features being modified. Therefore, DICE combines diversity and feasibility. Diversity is captures by building on determinantal point processes. From an intuitive perspective, counterfactual examples that are in close proximity to the original input can be particularly useful to a user. In the DICE methodology, proximity between the original input and its counterfactual (CF) example is quantified using a distance metric. This distance is defined as the negative vector distance between the features of the original input and the CF example. One common distance metric used is the Manhattan-distance first coined in 1951 by Pólya et al., which can be optionally weighted by a user-provided custom weight for each feature. Based on the diversity and proximity, DICE considers a combined loss function for all generated counterfactual examples. The loss function uses a gradient descent method for optimization. Lastly, the generated counterfactual examples reflect the underlying causal relationship between the input features and the model's predictions. DICE uses a causal graph-based model to identify the relevant features and their causal relationships and generates counterfactual examples that modify the relevant features while keeping the rest of the features fixed.

## 4 EXPERIMENTAL SETUP

This section will provide a formal description of the dataset used for the experiment, along with descriptive statistics. Additionally, the preprocessing of the data will be thoroughly explained, including the transformation of the data and the handling of missing values. A detailed outline of the experimental procedure will also be provided, including tools and packages that are used for the experimental setup. Evaluation criteria will be discussed, along with the robustness of the models.

## 4.1 *Dataset description*

This research will utilize the OPTiMiSE dataset, collected by Kahn et al. (2018). The dataset comprises information on clients who participated in a three-phase switching study to examine the relevance of switching antipsychotic medication for clients with schizophrenia, schizophreniform, or schizoaffective disorder. The experimental study was conducted between May 26, 2011, and May 15, 2016. This thesis used data from the first phase of the study. During this phase, clients were treated for 4 weeks with Amisulpride. Total clients who started this experimental study were 495. This research used data from participant who still participated the study after phase 1, which are 376 clients. As mentioned, our target variable is symptomatic remission based on the Positive and Negative Syndrome Scale. The target variable is unbalanced as shown in Figure 1.

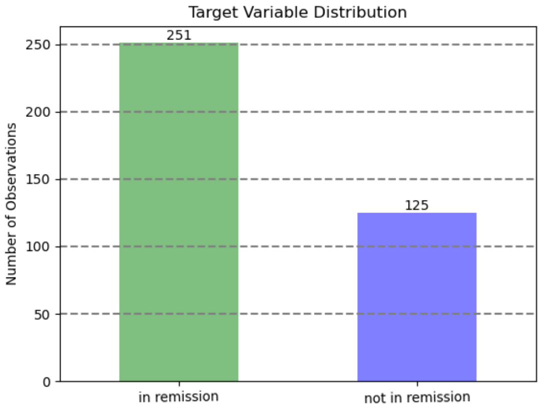


Figure 1: Target Variable Distribution

The dataset consists of categorical, continuous and binary features. All the data collected in this study has been stored and presented in a tabular format. The dataset is composed of both dynamic and static features. This research uses only static data. These features are categorized in 12 categories. These categories are shown in Table 1, with the corresponding number of features within the categories. The data for training the RF and SVM is obtained during visits 1 and 2 of the clients involved. The PANSS score, as target variable, is obtained during visit 5.

## 4.2 *Pre-processing*

The PANSS variables are extracted from the dynamic features at visit 5 to compute the output label: PANSS recovery. After extracting the features,

| Categories | Num. Features |
|---|---|
| Demographics | 16 |
| Diagnosis | 7 |
| Lifestyle | 7 |
| Somatic | 11 |
| Treatment | 1 |
| MINI | 67 |
| Cytokines | 34 |
| PANNS | 30 |
| PSP | 5 |
| CGI | 1 |
| CDSS | 9 |
| SWN | 20 |

Table 1: Categories and Number of Features from dataset

missing values are handled. SimpleImputer from Pedregosa et al. (2011) is chosen as the preferred method for handling missing values due to its simplicity, speed, suitability for non-linear relationships, compatibility with large datasets, and transparent imputation process (Graham et al., 2013).

The features are split into two categories: categorical and continuous features. This categorization is performed in order to facilitate the imputation of missing values. In the data preprocessing step, numerical features are imputed using the SimpleImputer class with the strategy set to "median". This strategy replaces missing values in numerical features with the median value of the available data. On the other hand, categorical features are imputed using the SimpleImputer class with the strategy set to "most frequent". This strategy replaces missing values in categorical features with the most frequent value observed in the available data. By utilizing these imputation strategies, missing values in both numerical and categorical features are effectively handled to ensure a complete dataset for subsequent analysis.

The next step in the data preprocessing pipeline is feature transformation. This step involves transforming categorical features using one-hot encoding and continuous features using a standard scaler. These feature transformation techniques are implemented in the scikit-learn library (Pedregosa et al., 2011). One-hot encoding converts categorical features into numerical representations, while standard scaling ensures comparability of continuous features by scaling them to have zero mean and unit variance. These transformations enable effective utilization of categorical information and equal contribution of continuous features during model training and prediction (Kusiak, 2001).

## 4.3 *Experimental procedure*

The experimental procedure involves several sequential steps. Initially, the dataset is divided into training and testing subsets. The ML algorithms are then trained on the training data to build predictive models and evaluated on the test data. Following the training phase, three methods for determining feature importances are implemented and evaluated. These methods assess the significance of individual features in the prediction process. Finally, a counterfactual explanation method is applied to generate alternative instances and understand the causal relationships between input features and model predictions.

### 4.3.1 *Data splitting procedure*

The initial step in the process is to split the dataset into input and output components. The input comprises all the static features, while the output corresponds to the target column. Given the description of the dataset, it is evident that the target variable exhibits a significant class imbalance. Additionally, the dataset itself is relatively small, with a size of N=376 observations. In order to address these challenges, a stratified hold-out method is employed to effectively partition the training and test data. This approach ensures that the model's performance can be accurately assessed in terms of its generalization capabilities on unseen data. The model undergoes training using the stratified k-fold cross-validation technique on training data.

### 4.3.2 *Model comparison*

To compare the performances of SVM and RF models, we stored the models in a vector. For the RF model, we utilized the RandomForestClassifier from scikit-learn library with default parameters. On the other hand, for the SVM model, we employed the RBF (Radial Basis Function) kernel, which is a non-linear kernel. Cross-validation scores are computed through a process that involves training the model on different subsets of the training data. This process is repeated for a total of 20 iterations. During each iteration, the training data is shuffled using different random states, ensuring variability in the subsets used for training. The model is trained on the training subset and then evaluated on the separated test set, which is excluded from the training process. This iterative approach allows for a more robust assessment of the model's performance across different data configurations.

### 4.3.3 *Feature importance scores*

After the model comparison, the feature importance analysis will be conducted for the highest scoring model. This involves creating the model using the same configuration as in the model comparison section, as well as applying stratified hold-out cross-validation. The number of iterations for the analysis will be set to 50. For the permutation test and the SHAP method, the absolute feature importance values will be extracted to ensure a robust comparison. Box-plot charts will be created for each method, with the feature importance values sorted in ascending order based on their median importance values. This visualization allows for a comprehensive comparison of the feature importance across all methods.

### 4.3.4 *Counterfactual explanations*

Before applying the diverse counterfactual explanation method, we need to distinguish between categorical and continuous features. To do so, we utilize the list that was defined during the pre-processing phase. The model we use for this purpose is the same model with the same parameters that were employed in the model comparison part. We utilize the entire dataset and specify the continuous features within the DICE data object. The backend we utilize is 'sklearn.' To generate counterfactual explanations using DICE, a method needs to be specified. In this case, the method used is set to "random," which randomly samples features during the counterfactual generation process. The random method explores different regions of the feature space by selecting features in a random manner to generate diverse counterfactual examples. This output file will be compared with the original input data to examine the differences introduced by the counterfactual explanations. The comparison allows us to analyze the impact of the generated counterfactuals on the predicted outcomes and understand the changes made to the input data.

## 4.4 *Evaluation criteria*

### 4.4.1 *Evaluating Machine Learning models*

To evaluate the performance of the RF and SVM models, two metrics will be employed: Area Under the Curve (AUC) and Area Under the Precision/Recall curve (AUPRC). These models do not depend on choosing a threshold for classification, which is advantageous because it eliminates the subjectivity involved in threshold selection. This ensures that the evaluation is objective and consistent across different scenarios. Moreover, the metrics of AUC and AUPRC offer robust quantitative measures that capture various aspects of the classification performance. By employing these metrics, a comprehensive evaluation of the models' effectiveness can be achieved, providing valuable insights into their discriminative capabilities and overall performance.

### 4.4.2 *Evaluating feature importance scores*

The three methods will be assessed based on two distinct metrics. The first metric employed is the Pairwise Feature Overlap test, wherein the Jaccard similarity coefficient is applied to the top 30 influential features derived from each importance method. The second metric is a correlation matrix for the feature importances. A correlation matrix will be constructed to analyze the relationship between feature importances. This matrix will provide insights into how the importance scores of different features are correlated with each other.

### 4.4.3 *Evaluating counterfactual explanations*

In order to assess the quality of the generated counterfactual explanations, a comparison will be made between the transformed instances and the original instances to examine the extent of their variation. The visualization of these feature changes will be presented using a strip plot. The Y-axis will represent the features, while the X-axis will indicate the variation scale. Each dot on the plot will symbolize a feature for a specific client at a given index. By employing this approach, it becomes feasible to observe the extent to which features need to vary in order to alter the outcomes at a local level. This analysis provides insights into the relationship between feature variations and their impact on the desired outcomes. Considering the computational demands associated with the DICE model, the features selected for variation will be limited to those that exhibit overlap among the feature importance scores obtained from the three different methods.

## 4.5 *Algorithms and Software*

The programming language used for this experiment is Python 3.8.8 (Van Rossum & Drake Jr, 1995). For data manipulation the following libraries were utilized; Numpy (Oliphant et al., 2006) and Pandas (pandas development team, 2020). For data transformation, ML algorithms, splitting data, and feature importance scores the scikit-library is mainly used (Pedregosa et al., 2011). For shapley additive explanations, the SHAP library is utilized (Lundberg & Lee, 2017a). For generating counterfactual explanations, the DICEML package is utilized (Mothilal et al., 2020). For visualizations Seaborn (Waskom, 2021) and Matplotlib (Hunter, 2007) were utilized.

### 4.6 *Overview*

The process pipeline is depicted in Figure 2.

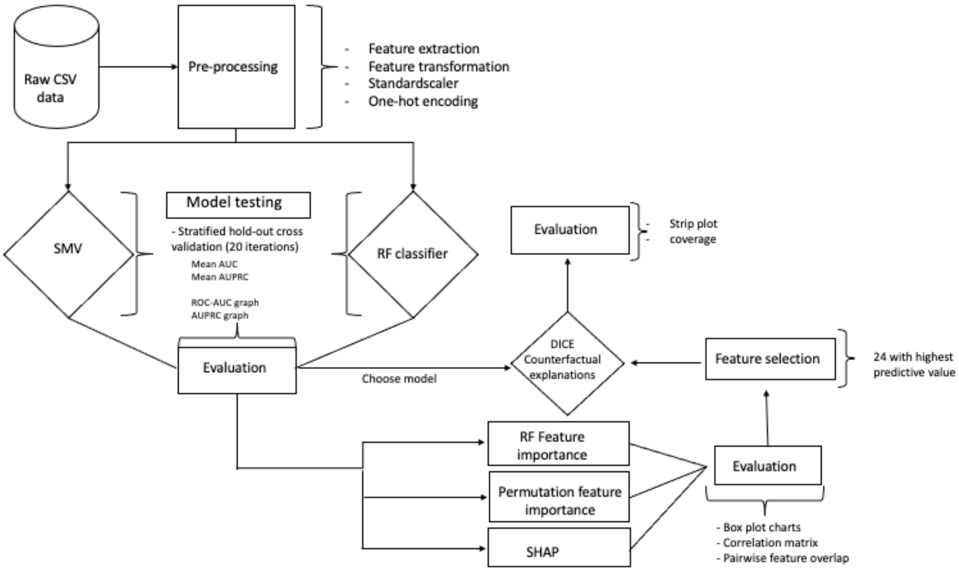


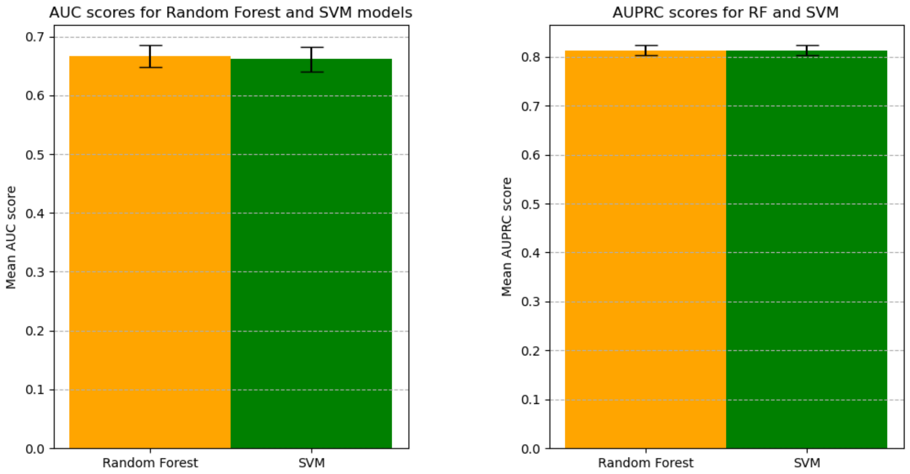
Figure 2: Experiment pipeline

## 5 RESULTS

This section is structured into three parts to present the results of the experiment. The first part is the performance comparison of the RF an SVM algorithms. The second part is the comparison of the feature importance scores methods. The last part contains the counterfactual explanations results.

### 5.1 *Model results*

In this subsection the results of the RF and SVM models are evaluated. First the mean AUC and AUPRC scores are compared over 20 iterations. These results are visualized in Figure 3. The RF model exhibits a mean AUC score of 0.667 with a standard deviation of 0.019. On the other hand, the SVM model has a mean AUC score of 0.662 with a standard deviation of 0.021. The results of the two-sided paired t-test indicate a t-statistic of 0.09666 and a p-value of 0.92401. Based on these findings, the difference in mean AUC scores is not statistically significant. The comparison of the mean AUPRC scores indicate that the RF model exhibits a higher score, with a mean AUPRC of 0.809 with a standard deviation 0.010. The SVM has a mean AUPRC of 0.806 with a standard deviation of 0.011. The conducted two-sided paired t-test indicates that there is no statistically significant difference in the performances of the models, as evidenced by a t-statistic of 1.7494 and a p-value of 0.0963.

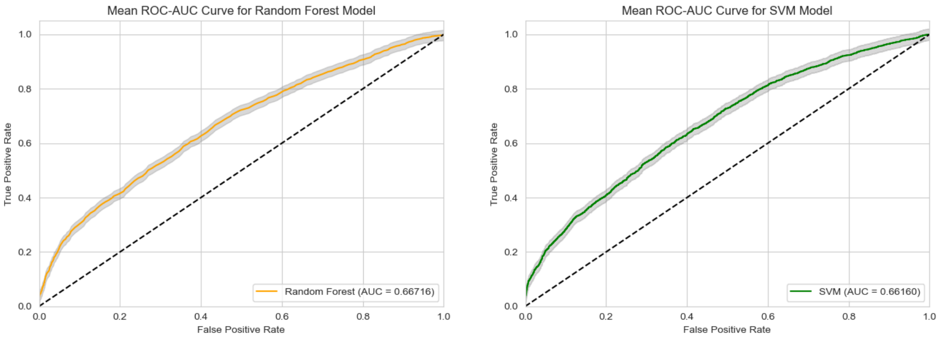


(a) AUC scores for RF and SVM

(b) AUPRC scores for RF and SVM

Figure 3: Comparison of AUC and AUPRC scores for Random Forest and SVM models

A graphical representation of the relationship between the true positive rate (TPR) and false positive rate (FPR) can be found in Figure 4, visualized in a ROC-AUC curve. The curve, shown for both models, exhibits a steeper rise in the beginning than towards the end, indicating that the models achieve higher TPR values at low FPR values, suggesting good early classification performance. Additionally, both ROC-AUC curves do not exhibit a flat section, implying that the models are capable of differentiating between positive and negative instances.

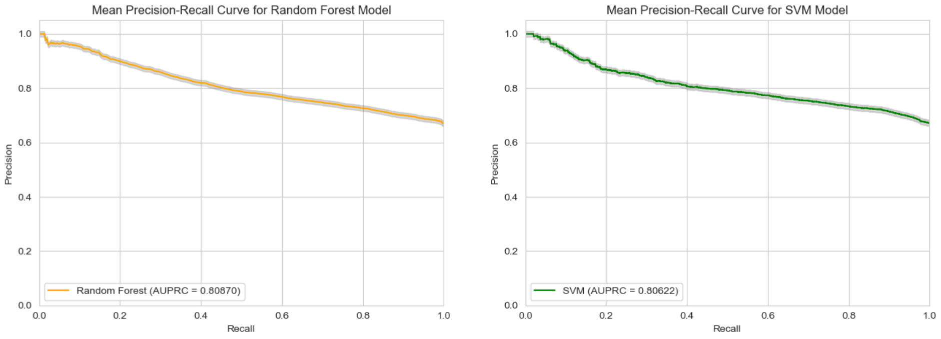


(a) ROC-AUC curve for RF

(b) ROC-AUC curve for SVM

Figure 4: Comparison of ROC-AUC curves for Random Forest and SVM models

The AUPRC curves are depicted in Figure 5. Both the curves for the RF and SVM models exhibit similar characteristics. When employing a high threshold, the proximity of precision to 1.0 implies that the model demonstrates a highly cautious nature in predicting positive instances. Nevertheless, the low recall value signifies a substantial number of actual positive instances that remain undetected. This behavior can be attributed to the model's inclination towards minimizing false positives, potentially leading to an elevated count of false negatives. Conversely, adopting a low threshold yields a recall of 1.0, indicating the model's ability to identify all positive instances, encompassing all true positives. However, the precision hovering around 0.7 suggests a noteworthy presence of false positives. This implies that the model becomes more inclusive in its predictions but at the expense of an increased number of false positives.



(a) AUPRC curve for RF

(b) AUPRC curve for SVM

Figure 5: Comparison of AUPRC curves for Random Forest and SVM models

## 5.2 *Feature importance scores results*

In this section, box plot charts are presented to visualize and compare feature importance scores obtained from three different methods. The feature importance scores are derived using the RF classifier, which was selected as the model for training based on the results of the model comparison phase. For each model, the 30 most influential features are shown in Figures 6, 7, 8.

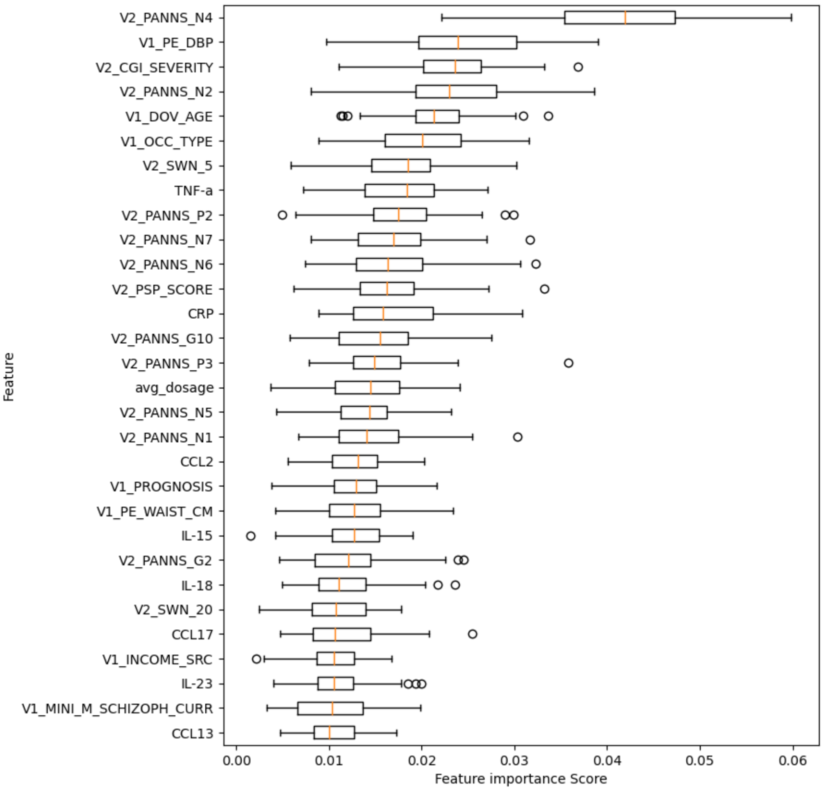


Figure 6: Random Forest feature importance

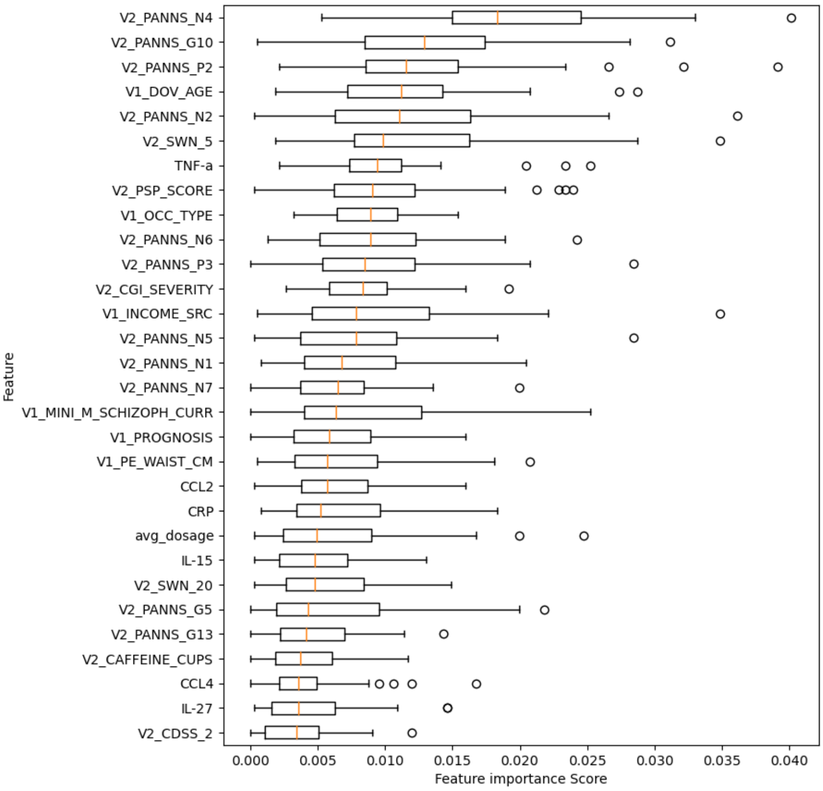


Figure 7: Permutation feature importance

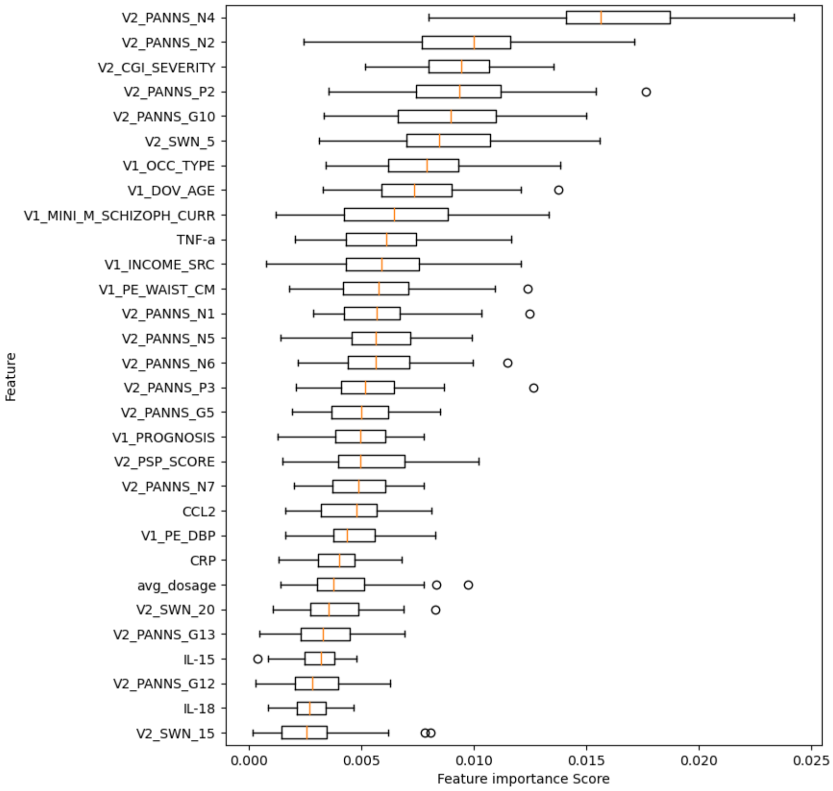


Figure 8: SHAP feature importance

By analyzing the 5 most influential features from each feature importance method, it is observed that two features consistently rank within the top 5 across all methods. Additionally, four features exhibit their significance by being present in at least 2 out of the three methods' top 5 most influential features. These notable features are presented in Table 2.

Table 2: Most Influential Features. **FI:** Feature importances. **PFI:** Permutation Feature importance. **SHAP:** SHAP

| **Feature** | **Methods** | **Description** |
|---|---|---|
| V2_PANNS_N4 | All | Passive/apathetic social withdrawal |
| V2_PANNS_N2 | All | Emotional withdrawal |
| V1_DOV_AGE | FI, PFI | Age in years begin trail |
| V2_CGI_SEVERITY | FI, SHAP | Clinical Global Impression of Severity |
| V2_PANNS_G10 | PFI, SHAP | Disorientation |
| V2_PANNS_P2 | PFI, SHAP | Conceptual disorganization |

The correlation matrix to assess the level of agreement between the feature importance scores obtained from the three different methods is illustrated in Figure 9.

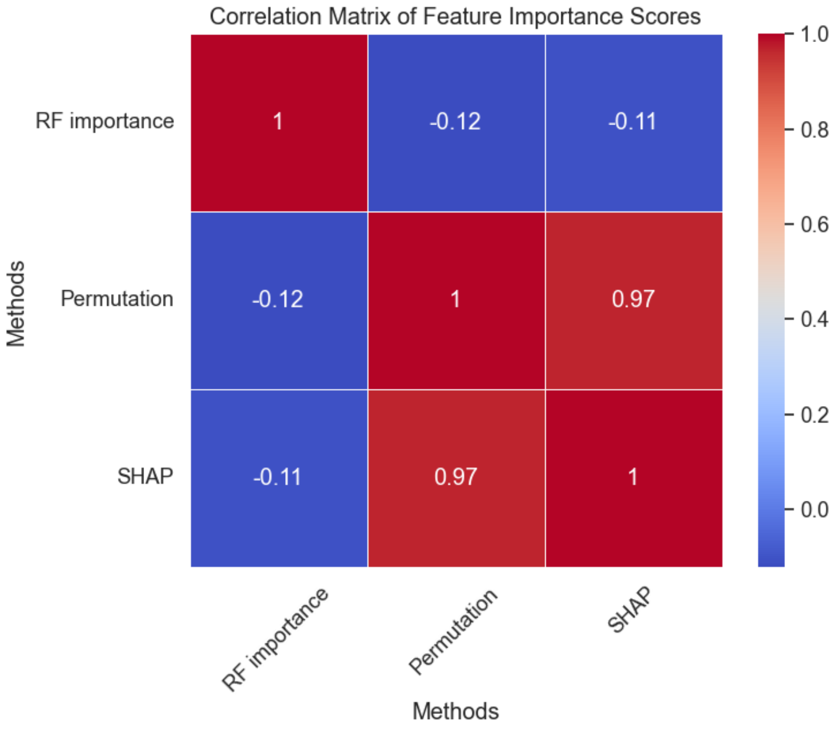


Figure 9: Correlation matrix feature importance scores

The correlation coefficient between RF importances and Permutation is -0.12, and the correlation coefficient between RF importances and SHAP is -0.11. This indicates a weak negative correlation for both comparisons. It suggests that there is a slight inverse relationship between RF importance scores and Permutation scores/ SHAP. However, the correlation is not strong, indicating that the two methods might capture different aspects of feature importance. According to Pedregosa et al. (2011), the impurity-based feature importances for RF can be misleading for high cardinality features.

The correlation coefficient between Permutation and SHAP is 0.97, indicating a strong positive correlation. This suggests that Permutation scores and SHAP scores are highly correlated, implying that they capture similar aspects of feature importance.

The Pairwise Feature Overlap test result, wherein the Jaccard similarity coefficient is applied to the top 30 influential features derived from each importance method are shown in Table 3. The pairwise overlap between RF importance and permutation is 0.714 which indicates a moderate degree of similarity between these two lists. The pairwise overlap between RF importance and SHAP, and between permutation and SHAP is both 0.765 which indicates a relatively higher degree of similarity between these lists. Among the 30 features examined, 24 of them were found to be consistent across all three methods employed in the study. These features can be found in Appendix A.

Table 3: Pairwise feature overlap

| | **RF importance** | **Permutation** | **SHAP** |
|---|---|---|---|
| **RF importance** | 1.000 | 0.714 | 0.765 |
| **Permutation** | 0.714 | 1.000 | 0.818 |
| **SHAP** | 0.765 | 0.818 | 1.000 |

Based on the results, the correlation matrix indicates a weak correlation between RF importance and Permutation, as well as between RF importance and SHAP. However, it is important to note that the Pairwise Feature Overlap test reveals a higher degree of similarity between the most influential features across these methods. Therefore, despite the weak correlation observed in the correlation matrix, the Pairwise Feature Overlap test highlights a notable overlap and similarity in the selection of important features. Upon careful examination of the box plot charts depicting feature importance scores, it becomes apparent that the RF method consistently assigns higher scores to each feature compared to the permutation method and SHAP method.

### 5.3 *Counterfactual explanations results*

It is important to note that the computational cost of running DICE increases as the number of columns in the dataset increases. The primary focus of this thesis is to identify features with high predictive value and investigate how modifying these features can potentially influence the outcome of first-phase schizophrenia and psychotic disorders. The identified features are those that overlap among the three models used to calculate feature importance scores and exhibit the highest predictive value. The remaining input remains unchanged without any modifications. The desired class is symptomatic remission. The objective of this research is to leverage the DICE method to uncover counterfactual explanations for clients who did not meet the criteria of symptomatic remission after phase 1. This entails that the method endeavors to discover counterfactual explanations for 151 clients. The DICE method successfully altered the outcomes of 87 patients out of a total of 151, resulting in a coverage rate of 0.576. The output of the DICE method is then compared with the original data, and the changes in the features are visualized in Figure 10.

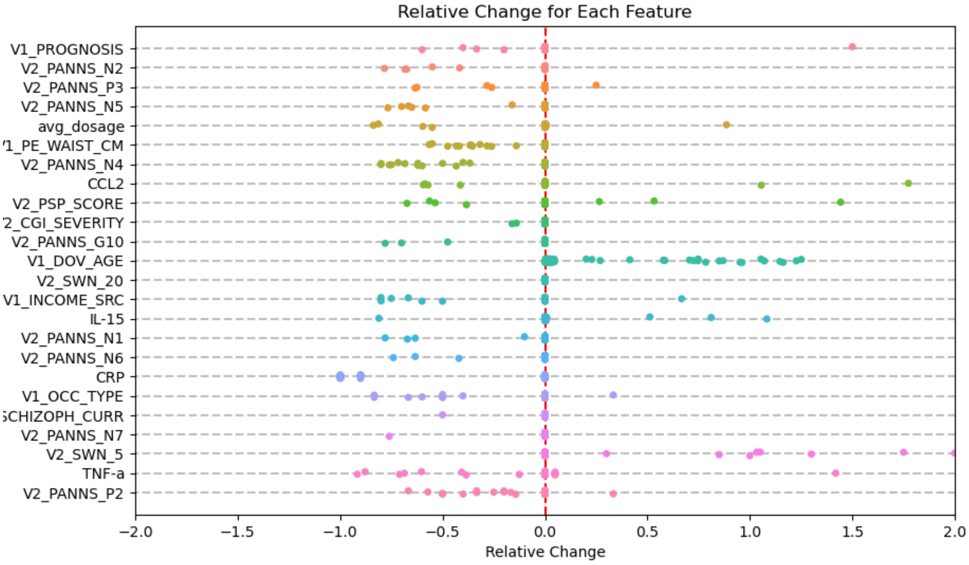


Figure 10: Strip plot of changed features

In Figure 10 all feature changes for all clients are combined. One notable finding is that for the majority of PANSS scores, a lower score at baseline suggests a higher likelihood of achieving remission at visit 5. This trend indicates that lower PANSS scores before starting the treatment are associated with favorable outcomes in the later stages of the assessment according to the DICE method. The analysis of SWN and age scores reveals an interesting pattern, suggesting that a higher score at baseline may be indicative of a different outcome for a subgroup of patients at visit 5. Although, before giving any conclusions, the result should be analyzed on individual level to see how other features have changed. Ultimately, the combination of different feature changes causes the method to alter the outcome.

## 6 DISCUSSION

This section is dedicated to discussing the findings and addressing the research questions that were posed in this study. The key findings and their implications will be thoroughly examined. Additionally, the limitations of this research will be acknowledged to provide a comprehensive understanding of the study's scope and potential constraints. Finally, based on the insights gained from this study, recommendations and suggestions for future research directions will be proposed.

## 6.1 *Goal of the study*

The objective of this study is to make a contribution towards improving the interpretability of algorithmic decision-making processes for psychosis prognosis prediction. Specifically, the study aims to investigate the degree to which various features contribute to the decision outcomes of the algorithm. Additionally, the study seeks to determine the magnitude of feature changes required to reverse or invert the algorithm's outcomes. By addressing these objectives, this research endeavors to provide valuable insights into the inner workings and interpretability of the algorithm.

## 6.2 *Research questions*

The main research question central in this study was: *Which features have the highest predictive value on psychosis prognosis prediction?*. To answer this question, three distinct methods were employed to extract feature importance scores. Through these methods, a set of 24 features was identified, which demonstrated the highest predictive value on global level based on the outcomes of all three approaches. Within these features, nine features were related to PANSS scores at baseline which is in line with other studies conducted by Li et al. (2021) and Podichetty et al. (2021). On a local level, this study employed a counterfactual explanation method and introduced a novel approach for its analysis, namely the visualization using a strip plot. With help of this strip plot, the study not only enables the extraction of feature scores, but it also provides insights into how features should be changed to achieve a desired outcome by the RF classifier. By visualizing the scale of change for each feature, the strip plot offers valuable guidance on the direction and magnitude of feature modifications necessary to potentially influence the outcome. The first

sub question in this study was: *What are the comparative performance results of Random Forest and Support Vector Machine models in predicting psychosis prognosis?*. To address this inquiry, we conducted a comparison between RF and non-linear SVM models using their default parameter settings. The dataset was stratified into training and test sets, and the models were trained using stratified k-fold cross-validation exclusively on the training set. Subsequently, the models were tested on the independent test set in 20 iterations. Our observations revealed that the RF algorithm exhibited slightly higher performance in terms of both mean AUC and mean AUPRC when compared to the non-linear SVM. The mean AUC scores differed 0.005, and the mean AUPRC results displayed a difference of 0.003. Although the average scores for mean AUC and mean AUPRC were higher for RF, the statistical analysis did not provide conclusive evidence of significant differentiation between the two models. The second

sub question in this study was: *How consistent are the results obtained from different feature importance methods in determining the most significant features for predicting psychosis prognosis?* To investigate this, three distinct feature importance methods were applied to the highest performing model identified in the first sub-question. Specifically, the RF model was evaluated using RF feature importances, permutation feature importance, and SHAP. By constructing a correlation matrix to evaluate the agreement between feature importance scores obtained from these methods, it was found that permutation feature importance and SHAP exhibited a high level of correlation. Conversely, RF feature importances displayed a weak negative correlation with both permutation feature importance and SHAP. Despite this lower negative correlation between RF and the other two methods, the Pairwise feature overlap test indicated a notable degree of similarity among the 30 most influential features across these methods.

### 6.3 *Societal and scientific contribution*

This study conducted a comparison between the performances of a RF model and a non-linear SVM model. The fact that both models performed similarly aligns with related studies that investigated the prognosis of first-phase psychosis treatment using socio-demographic and clinical features. Several studies have demonstrated that the RF algorithm yielded the best performance (Li et al., 2021; Podichetty et al., 2021), while in other studies, the SVM model outperformed other methods (Koutsouleris et al., 2016; Soldatos et al., 2022). The study aimed to enhance understanding of the factors influencing patient outcomes in psychosis prognosis prediction by investigating the features contributing to decision outcomes. This knowledge has the potential to improve the accuracy of predictions and the care provided to individuals with psychosis. To ensure robustness, the study compared different feature importance methods. Among the 30 most influential features identified for decision-making, 24 of them (0.8) were found to be common across the different methods. However, the specific feature contribution values varied between the methods. The analysis revealed a weak correlation between RF feature importance and permutation importance/SHAP, while a high correlation was observed between permutation importance and SHAP. These findings align with the research conducted by Bloch et al. (2022), where RF feature importance showed only moderate correlation with permutation importance and weak correlation with SHAP, while SHAP and permutation importance exhibited high correlation. It is noteworthy that both datasets used in the study had high cardinality, which might introduce bias in RF feature importances due to RF's tendency to favor features with high cardinality as they can create more splits in the trees (Pedregosa et al., 2011). No other studies have been conducted on this specific dataset to measure feature importance scores that encompass all static features. The analysis conducted in this study revealed a consistent observation across all three feature importance scoring methods. Among the 24 most influential features, the baseline PANSS scores (comprising a total of 9 scores) emerged as valuable predictors for psychosis prognosis. Previous related studies by Li et al. (2021) and Podichetty et al. (2021) have also highlighted the predictive significance of PANSS scores at baseline. However, it is noteworthy that alternative studies have found that psychosocial factors such as unemployment, poor education, functional deficits, and unmet psychosocial needs were more influential predictors than symptom data (Wu et al., 2020). By studying the interpretability of the algorithm used for psychosis prognosis prediction, the research aimed to address these concerns and ensure that the decision-making process is transparent and understandable. The study introduces a novel approach, the visualization using a strip plot, for analyzing counterfactual explanations and understanding feature changes. This contributes to the scientific field of XAI and interpretable ML by providing a new method for exploring and visualizing the interpretability of algorithms.

### 6.4 *Limitations and future work*

Despite diligent efforts to ensure the reliability and validity of the results, this study is subject to certain limitations that should be taken into account when interpreting the findings. First of all, two ML algorithms are compared before extracting feature importance scores. Based on mean scores the RF is chosen to further analyze in this study. Also, the dataset used in this study is small with 376 clients, this may impact the stability and reliability of the results. A larger dataset could provide more robust conclusions and better generalize findings (Hendrycks et al., 2021). To mitigate this challenge, stratified hold-out cross-validation is employed. Each feature importance method may have inherent limitations and assumptions. It is essential to acknowledge the specific limitations associated with each method.

The DICE method is relatively new and, like any method, has its limitations, which have been acknowledged by the authors (Mothilal et al., 2020). Firstly, it should be noted that previous benchmarking studies of the DICE method have primarily utilized datasets with a small number of features. However, when applied to datasets with a large number of features, the method becomes computationally expensive and time-consuming for generating counterfactual explanations. Hence, in this study, a limitation was imposed on the number of features that could be modified in order to accommodate the computational constraints of the DICE method. The DICE method incorporates a local feature importance attribute, which is intended to provide valuable insights into the contribution of individual features. However, due to the large number of features involved, an unintended issue arises where a feature attribution value of 0.01 is assigned to all features. This discrepancy has been recognized as a bug by the authors of the method. Due to the inherent class imbalance in the dataset, there is a potential risk of ML algorithms overfitting the majority class, leading to minor changes, made within the DICE algorithm, having a significant impact on the outcome. Therefore, further investigations are recommended to evaluate the efficacy of this visualization method in the context of counterfactual explanations. However, the author maintains an optimistic perspective regarding the potential of strip plots, as initial indications suggest promising results.

It is recommended for future research to employ alternative ML algorithms on the dataset and extract feature importance scores. Comparing these scores with the feature importance scores obtained in this study can provide insights into the degree to which different algorithms with comparable performance rely on the same set of features for classification. This comparative analysis can shed light on the underlying factors contributing to the classification process across various ML approaches. Additionally, it is recommended to conduct further investigations into alternative counterfactual explanations for the same dataset, utilizing the RF classifier as employed in this study. By comparing the outcomes of these alternative explanations with the results of the present study, valuable insights can be gained regarding the consistency and direction of variation for the same set of features. This comparative analysis will contribute to a more comprehensive understanding of how different counterfactual explanations elucidate the behavior of the RF classifier on the given dataset. Furthermore, it is of particular interest to explore the impact of changing the desired outcome from 'in remission' to 'not in remission.' By doing so, it becomes possible to investigate whether the features should be altered in the opposite direction to achieve the desired outcome.

# 7 CONCLUSION

In this study, feature importance scores were assessed for a RF classifier with the aim of determining the predictive value of different features in psychosis prognosis prediction using data from first-episode psychosis. The study employed RF feature importances, permutation feature importance, and SHAP methods. These approaches were utilized to evaluate and compare the importance scores attributed to each feature, shedding light on the features that exhibited the highest predictive value in the RF classifier for psychosis prognosis prediction. Among the total of 208 features examined, all three different methods consistently identified 9 items related to the PANSS scores as being among the top 30 most influential features, with PANSS-N4 (indicating passive/apathetic social withdrawal) as most influential factor for psychosis prognosis prediction. Also, it is evident that, at present, the DICE method is not entirely viable for generating meaningful counterfactual explanations that can be thoroughly analyzed. This limitation stems from several minor bugs present in the method. As a result, the reliability and accuracy of the generated counterfactual explanations may be compromised, hindering their usefulness in practical applications. Further improvements and bug fixes are necessary to enhance the feasibility and effectiveness of the DICE method for generating interpretable counterfactual explanations. Taking a forward-looking perspective, there is a noticeable shift towards explaining ML algorithms (XAI) which might lead to improvements in predictions and overall acceptance. As a result, we might anticipate a future where healthcare decisions are increasingly informed by data-driven insights, leading to improved patient outcomes and personalized treatment approaches.

## REFERENCES

Adadi, A., & Berrada, M. (2020). Explainable ai for healthcare: From black box to interpretable models. *Embedded Systems and Artificial Intelligence: Proceedings of ESAI 2019, Fez, Morocco*, 327–337.

Adlung, L., Cohen, Y., Mor, U., & Elinav, E. (2021). Machine learning in clinical decision making. *Med*, *2*(6), 642–665.

Andreasen, N. C., Carpenter Jr, W. T., Kane, J. M., Lasser, R. A., Marder, S. R., & Weinberger, D. R. (2005). Remission in schizophrenia: Proposed criteria and rationale for consensus. *American Journal of Psychiatry*, *162*(3), 441–449.

Bloch, L., Friedrich, C. M., & Initiative, A. D. N. (2022). Machine learning workflow to explain black-box models for early alzheimer's disease classification evaluated for multiple datasets. *SN Computer Science*, *3*(6), 509.

Bracher-Smith, M., Rees, E., Menzies, G., Walters, J. T., O'Donovan, M. C., Owen, M. J., Kirov, G., & Escott-Price, V. (2022). Machine learning for prediction of schizophrenia using genetic and demographic factors in the uk biobank. *Schizophrenia Research*, *246*, 156–164.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*, 273–297.

D MCGORRY, P. (2002). The recognition and optimal management of early psychosis: An evidence-based reform. *World Psychiatry*, *1*(2), 76.

Del Fabro, L., Bondi, E., Serio, F., Maggioni, E., D'Agostino, A., & Brambilla, P. (2023). Machine learning methods to predict outcomes of pharmacological treatment in psychosis. *Translational Psychiatry*, *13*(1), 75.

Demircioğlu, A. (2022). Benchmarking feature selection methods in radiomics. *Investigative Radiology*, *57*(7), 433–443.

de Oliveira, R. M. B., & Martens, D. (2021). A framework and benchmarking study for counterfactual generating methods on tabular data. *Applied Sciences*, *11*(16), 7274.

Dey, S., Chakraborty, P., Kwon, B. C., Dhurandhar, A., Ghalwash, M., Saiz, F. J. S., Ng, K., Sow, D., Varshney, K. R., & Meyer, P. (2022). Human-centered explainability for life sciences, healthcare, and medical informatics. *Patterns*, *3*(5), 100493.

European Commission. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive

95/46/EC (General Data Protection Regulation) (Text with EEA relevance). https://eur-lex.europa.eu/eli/reg/2016/679/oj

Fusar-Poli, P., McGorry, P. D., & Kane, J. M. (2017). Improving outcomes of first-episode psychosis: An overview. *World psychiatry*, *16*(3), 251–265.

Graham, J. W., Cumsille, P. E., & Shevock, A. E. (2013). Methods for handling missing data.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). Xai—explainable artificial intelligence. *Science robotics*, *4*(37), eaay7120.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. (2021). The many faces of robustness: A critical analysis of out-of-distribution generalization. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8340–8349.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Kahn, R. S., van Rossum, I. W., Leucht, S., McGuire, P., Lewis, S. W., Leboyer, M., Arango, C., Dazzan, P., Drake, R., Heres, S., et al. (2018). Amisulpride and olanzapine followed by open-label treatment with clozapine in first-episode schizophrenia and schizophreniform disorder (optimise): A three-phase switching study. *The Lancet Psychiatry*, *5*(10), 797–807.

Kay, S. R., Fiszbein, A., & Opler, L. A. (1987). The positive and negative syndrome scale (panss) for schizophrenia. *Schizophrenia bulletin*, *13*(2), 261–276.

Koutsouleris, N., Kahn, R. S., Chekroud, A. M., Leucht, S., Falkai, P., Wobrock, T., Derks, E. M., Fleischhacker, W. W., & Hasan, A. (2016). Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: A machine learning approach. *The Lancet Psychiatry*, *3*(10), 935–946.

Kusiak, A. (2001). Feature transformation methods in data mining. *IEEE Transactions on Electronics packaging manufacturing*, *24*(3), 214–221.

Li, Y., Zhang, L., Zhang, Y., Wen, H., Huang, J., Shen, Y., & Li, H. (2021). A random forest model for predicting social functional improvement in chinese patients with schizophrenia after 3 months of atypical antipsychotic monopharmacy: A cohort study. *Neuropsychiatric Disease and Treatment*, 847–857.

Lundberg, S. M., & Lee, S.-I. (2017a). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances* *in neural information processing systems 30* (pp. 4765–4774). Curran Associates, Inc. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

Lundberg, S. M., & Lee, S.-I. (2017b). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 607–617.

Oliphant, T. E., et al. (2006). *A guide to numpy* (Vol. 1). Trelgol Publishing USA.

pandas development team, T. (2020). *Pandas-dev/pandas: Pandas* (Version latest). Zenodo. https://doi.org/10.5281/zenodo.3509134

Patel, K. R., Cherian, J., Gohil, K., & Atkinson, D. (2014). Schizophrenia: Overview and treatment options. *Pharmacy and Therapeutics*, *39*(9), 638.

Pawelczyk, M., Bielawski, S., Heuvel, J. v. d., Richter, T., & Kasneci, G. (2021). Carla: A python library to benchmark algorithmic recourse and counterfactual explanation algorithms. *arXiv preprint arXiv:2108.00783*.

Pawelczyk, M., Broelemann, K., & Kasneci, G. (2020). Learning model-agnostic counterfactual explanations for tabular data. *Proceedings of The Web Conference 2020*, 3126–3132.

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

Podichetty, J. T., Silvola, R. M., Rodriguez-Romero, V., Bergstrom, R. F., Vakilynejad, M., Bies, R. R., & Stratford Jr, R. E. (2021). Application of machine learning to predict reduction in total panss score and enrich enrollment in schizophrenia clinical trials. *Clinical and Translational Science*, *14*(5), 1864–1874.

Pólya, G., Szegö, G., & Szegő, G. (1951). *Isoperimetric inequalities in mathematical physics*. Princeton University Press.

Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., & Flach, P. (2020). Face: Feasible and actionable counterfactual explanations. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 344–350.

Quillbot [v14.135.1]. (2023). https://www.quillbot.com

Ricciardi, A., McAllister, V., & Dazzan, P. (2008). Is early intervention in psychosis effective? *Epidemiology and Psychiatric Sciences*, *17*(3), 227–235. https://doi.org/10.1017/S1121189X00001329

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal*, *20*(1), 3–29.

Soldatos, R. F., Cearns, M., Nielsen, M. Ø., Kollias, C., Xenaki, L.-A., Stefanatou, P., Ralli, I., Dimitrakopoulos, S., Hatzimanolis, A., Kosteletos, I., et al. (2022). Prediction of early symptom remission in two independent samples of first-episode psychosis patients using machine learning. *Schizophrenia Bulletin*, *48*(1), 122–133.

Stepin, I., Alonso, J. M., Catala, A., & Pereira-Fariña, M. (2021). A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, *9*, 11974–12001. https://doi.org/10.1109/ACCESS.2021.3051315

Van Rossum, G., & Drake Jr, F. L. (1995). *Python tutorial* (Vol. 620). Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands.

Verma, S., Dickerson, J., & Hines, K. (2020). Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*.

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Wu, C.-S., Luedtke, A. R., Sadikova, E., Tsai, H.-J., Liao, S.-C., Liu, C.-C., Gau, S. S.-F., VanderWeele, T. J., & Kessler, R. C. (2020). Development and validation of a machine learning individualized treatment rule in first-episode schizophrenia. *JAMA network open*, *3*(2), e1921660–e1921660.

# APPENDIX A

| **Features** |
|---|
| V1_PROGNOSIS |
| V2_PANNS_N2 |
| V2_PANNS_P3 |
| V2_PANNS_N5 |
| V2_PANNS_P2 |
| avg_dosage |
| V1_PE_WAIST_CM |
| V2_PANNS_N4 |
| CCL2 |
| V2_PSP_SCORE |
| V2_CGI_SEVERITY |
| V2_PANNS_G10 |
| V1_DOV_AGE |
| V2_SWN_20 |
| V1_INCOME_SRC |
| IL-15 |
| V2_PANNS_N1 |
| V2_PANNS_N6 |
| CRP |
| V1_OCC_TYPE |
| V1_MINI_M_SCHIZOPH_CURR |
| V2_PANNS_N7 |
| V2_SWN_5 |
| TNF-a |

Table 4: List of Features with highest predictive value