Draftable

Draftable Comparison Export

This document is an exported comparison with limited functionality, generated by Draftable Desktop. To access full functionality, use Draftable's powerful comparison viewer in any of our products.

Left document:	Schizophrenia.docx
Right document:	Scizophrenia_Reviewed.docx

What is this document?

This is a comparison of two documents. The two documents are interleaved such that the left document is displayed on even pages and the right document is displayed on odd pages.

Is there a specific way I should view this file?

This document is intended to be viewed in Two Page Continuous mode (or sometimes called 'Two Page Scrolling'). It should open in this mode by default when using Adobe Acrobat and most popular PDF readers.

If the document opens in a different view, you can often change this in the settings. In Adobe Acrobat, go to **View > Page Display > Two Page Scrolling**.

Why are there blank pages?

Blank pages are inserted to keep both documents as aligned as much as possible.

How do I read the changes?

Text deleted from the left document and, hence, not in right document is highlighted red. Text added to the right document and, hence, not in left document is highlighted green.

Tip for printing

When printing this document, we recommend printing double-sided and include this first page. This will result in the matching text being displayed on different pages and easily readable, much like a book.

For more information

Draftable offers powerful document comparison solutions for all use-cases. To view our products, please visit our website: draftable.com.

Table of Contents

Table of Contents

Chapter 1: Introduction

This dissertation aims to investigate the potential for using a combination of cortical surface modelling and machine learning to improve the precision with which treatment response may be predicted for schizophrenia patients. Schizophrenia is a neuropsychiatric illness characterized by prominent auditory hallucinations and persecutory delusions. Neuroimaging data suggests that schizophrenia is associated with dysfunction in neural networks subserving key psychological processes that operate to allow humans to accurately perceive and interpret external stimuli. Despite multiple imaging studies of patients, there are currently no robust predictors of treatment response or non-response.

Previous research has predominantly focused on deriving biomarkers of outcome from multimodal volumetric MRI. However, schizophrenia is a disorder of the cortex - an area that has regularly been shown to be more accurately modelled as a surface. Therefore, this project will investigate the potential for using a combination of cortical surface modelling and machine learning to improve the precision with which treatment response may be predicted for schizophrenia patients.

Chapter 1: Introduction

This dissertation aims to investigate the use of cortical surface modeling and machine learning to predict treatment responses in schizophrenia patients. Schizophrenia is a neuropsychiatric illness characterized by auditory hallucinations and persecutory delusions. Neuroimaging data suggest dysfunction in neural networks involved in perception and interpretation. However, there are currently no robust predictors of treatment response. Previous research focused on multimodal volumetric MRI biomarkers, but schizophrenia primarily affects the cortex, which can be more accurately modeled as a surface. This project will explore the potential of combining cortical surface modeling and machine learning to improve the precision of treatment response prediction in schizophrenia patients.

1.1 Background

In this section, we will review recent research on the use of machine learning (ML) algorithms with neuroimaging data to identify patients with treatment-resistant schizophrenia (TRS) and non-treatment-resistant schizophrenia (NTR). We will cover various categories of neuroimaging data, including structural and functional data, that have been used for prediction. We will also discuss state-of-the-art ML models for predicting TRS and NTR. Additionally, we will address confounding factors in neuroimaging data and strategies for mitigating their influence. Finally, we will provide an overview of methods for deriving significant predictive characteristics from ML models and displaying them for better interpretability. The goal of this section is to provide readers with a comprehensive understanding of the developments, opportunities, challenges, and potential implications of ML-based classification of TRS and NTR schizophrenia from neuroimaging data for clinical decision-making and patient care.

Background

We intend to present a thorough review of the most recent research on the issue of identifying patients with treatment-resistant schizophrenia (TRS) and non-treatment-resistant schizophrenia (NTR) using machine learning (ML) algorithms using neuroimaging data in the part that follows. We go into the numerous categories of neuroimaging data, including both structural and functional data, that have been used for the prediction of TRS and NTR. We also discuss the state of the art in ML models that have been used to the predicting of TRS and NTR schizophrenia.Moreover, we address the possible impact of confounding factors on neuroimaging data and go through strategies for reducing their influence. Lastly, we give an overview of several methods for deriving significant predictive characteristics from ML models and displaying them, which can help make these models easier to understand and interpret.

This section's overall goal is to give readers a thorough understanding of the developments in the field of ML-based classification of TRS and NTR schizophrenia from neuroimaging data, along with the opportunities and challenges presented by this method, as well as any potential repercussions for clinical decision-making and patient care.

1.1.1 TRS schizophrenia's neuroimaging characteristics

With a focus on differentiating between patients with treatment-resistant schizophrenia (TRS) and non-treatment-resistant schizophrenia (NTR), there has been a significant increase in research examining the structural and functional variations in the brains of people with schizophrenia. On the basis of magnetic resonance imaging (MRI) data, machine learning (ML) algorithms have showed promise in finding a variety of neuroimaging variables that can distinguish between these groups.

Smaller gray matter volumes have been linked in several studies to worse clinical results in schizophrenia patients. For individuals with first-episode psychosis, Kasparek et al. [12] discovered that lower volumes of prefrontal regions indicated worse outcomes after one year. Similar findings were made by Jääskeläinen et al. [13] during a 16-year longitudinal research, who found that a higher density of left frontal gray matter was related to better outcomes, including not receiving a disability pension. These results imply that structural characteristics of the brain may be significant prognostic indicators of therapeutic success.

Cortical volume, thickness, and surface area variations among healthy controls, TRS patients, and NTR patients were examined by Barry et al. According to their findings, TRS patients had significantly lower cortical volume and thickness than NTR patients in a number of areas of the brain, including the right frontal and precentral regions, the right parietal and occipital cortex, the left temporal cortex, and the bilateral cingulate cortex. These results provide more evidence that TRS and NTR patients have different neurological variations.

1.1.1 Neuroimaging characteristics of TRS schizophrenia

Recent research has focused on differentiating between treatment-resistant schizophrenia (TRS) and non-treatment-resistant schizophrenia (NTR) by examining structural and functional brain variations using machine learning (ML) algorithms with magnetic resonance imaging (MRI) data. Smaller gray matter volumes have been associated with worse clinical outcomes in schizophrenia patients, as shown by Kasparek et al. [12] and Jääskeläinen et al. [13]. Cortical volume, thickness, and surface area differences have been observed among healthy controls, TRS patients, and NTR patients by Barry et al., with TRS patients exhibiting significantly lower cortical volume and thickness in various brain regions. Dysfunction in reward feedback processing has also been reported in people with schizophrenia, as demonstrated by Vanes et al. using functional MRI (fMRI) data. These findings suggest that TRS and NTR may have distinct neurological variations. Machine learning models have the potential to accurately identify and utilize these distinctive neuroimaging characteristics to inform more personalized therapeutic strategies, aiding clinicians in anticipating treatment outcomes and making informed decisions about patient care.

1.1.2 Schizophrenia Patient Classification and Machine Learning

Machine Learning (ML) is increasingly used for classifying Treatment-Resistant Schizophrenia (TRS) and Non-Treatment Resistant (NTR) Schizophrenia from neuroimaging data. Many features are correlated with treatment outcome in schizophrenia patients, and ML can combine these features for improved prediction. Previous research has mainly focused on classifying schizophrenia patients from healthy controls. For example, Greenstein et al. used 74 anatomic brain MRI sub-regions to classify childhood onset schizophrenia patients and controls with a predictive accuracy of 73.7%. Iwabuchi et al. achieved the highest accuracy of 77% in distinguishing patients and controls using gray matter and white matter MRI images with a Support Vector classifier. Lu et al. combined Support

According to research on brain function (Tamminga et al., 1998), people with schizophrenia may have defective reward feedback processing processes. For instance, Vanes et al. investigated the brain correlates of reward prediction error (RPE) signals related to reinforcement learning using functional MRI (fMRI) data. While the behavioral results of TRS and NTR patients during a probabilistic reinforcement learning test were comparable, the authors discovered that the underlying brain processes could be different. This lends more support to the idea that TRS could be a different illness from NTR schizophrenia.

The brains of TRS and NTR patients exhibit a variety of anatomical and functional abnormalities, according to neuroimaging studies. Our results demonstrate the potential of machine learning models to precisely recognize and make use of these distinctive qualities for the creation of more specialized therapeutic strategies. Clinicians can more accurately anticipate treatment results and make better choices regarding patient care by combining neuroimaging data with clinical data.

1.1.2 Schizophrenia patients classification and Machine Learning

The use of Machine Learning (ML) in analyzing neuroimaging data relevant to the classification of Treatment-Resistant Schizophrenia (TRS) and Non-Treatment Resistant (NTR) Schizophrenia has become increasingly popular due to the high complexity and dimensionality of such data. In this article, we will discuss the methods and outcomes of several ML models applied to the task of classifying schizophrenia patients from neuroimaging data.

Many features appear to be somewhat correlated with treatment outcome for schizophrenia patients, and ML offers the opportunity to combine many of these features together to build a more powerful predictor. Most research to date has focused on the problem of classifying schizophrenia patients from healthy controls. For example, Greenstein et al. used 74 anatomic brain MRI sub-regions derived from structural MRI data to classify 98 childhood onset schizophrenia patients and 99 controls using Random Forests and achieved a predictive accuracy of 73.7%.

Iwabuchi et al. used gray matter and white matter MRI images as inputs to a Support Vector classifier and achieved the highest accuracy of 77% (with a 7T scanner) in distinguishing 19 patients and 20 control subjects. Lu et al. combined Support Vector Machines with recursive feature elimination to discriminate 41 schizophrenia patients from 40 controls using structural MRI data from a 3T scanner, and achieved 88.4% classification accuracy.

Ambrosen et al. developed a framework for using multi-modal imaging data to classify healthy-controls [HCs] and patients, as well as NTR and TRS patients using ML methods. The multi-modal data included regional measures of cortical thickness, surface area, and mean curvature derived from T1-weighted MRI images, data measuring cognitive performance on standardized tests, and electrophysiology data. They combined this data with additional simulated data to train an ensemble of basic Machine Learning models including linear regression algorithms, Support Vector Machines, and Random Forests.

Ambrosen et al. were able to classify HCs from TRSs and NTRs with a maximum class-balanced accuracy of 64% using an ensemble of trees, but were unable to classify TRSs and NTRs with significance. Overall, the results of these studies demonstrate the potential of ML in analyzing complex and high-dimensional neuroimaging data for the classification of schizophrenia patients. Future research in this area will undoubtedly continue to shed light on this important field.

Vector Machines with recursive feature elimination to discriminate schizophrenia patients from controls with 88.4% accuracy using structural MRI data.

Ambrosen et al. developed a framework using multi-modal imaging data, including cortical thickness, surface area, and mean curvature from T1-weighted MRI images, cognitive performance data, and electrophysiology data, to classify healthy controls and patients, as well as TRS and NTR patients, using ML methods. They achieved a maximum class-balanced accuracy of 64% in classifying HCs from TRSs and NTRs using an ensemble of trees, but were unable to classify TRSs and NTRs with significance.

Overall, these studies demonstrate the potential of ML in analyzing complex neuroimaging data for schizophrenia patient classification. Future research in this area will continue to shed light on this important field.

1.1.3 Confounds in Medical Data for Training ML Models

Confounding variables pose challenges in predictive modeling, as they can affect model outcomes and lead to inaccurate conclusions. For example, in a model predicting schizophrenia based on brain features, confounds such as gender and substance abuse can significantly impact results as they affect gray matter structure.

Methods to address confounds include balancing them across groups, but this may not always be feasible. Another approach is confound regression, where a linear regression model is used to remove confound influence from data. Cross-validated confound regression (CVCR) is a modified version that has shown promise in neuroimaging data. In a study by Snoek et al. (2019), CVCR outperformed whole-dataset confound regression in deconfounding brain size for gender prediction from structural MRI data.

Overall, addressing confounding variables in predictive modeling requires careful consideration and appropriate methods. CVCR is a promising approach, but further research is needed for its efficacy in different applications and datasets.

1.1.3 Confounds in data used in medical training

Confounding variables are a common challenge when building predictive models. These variables can influence the outcome of the model and lead to incorrect conclusions about the underlying factors driving the prediction. For instance, in a model designed to predict whether a subject has schizophrenia based on structural brain features, factors such as gender and substance abuse can significantly impact the model's performance since they are known to affect gray matter structure.

Several methods have been proposed to address the influence of confounding variables in imaging data. One approach is to ensure that confounds are balanced across experimental groups. For instance, if gender is a confounding variable, then the number of men and women should be evenly distributed across all experimental groups. However, this method is not always practical, particularly if the confounding variable is only known after data collection. In such cases, counterbalancing data that has already been collected is not feasible and can lead to significant reductions in predictive model performance.

Another method for addressing confounding variables is confound regression, a popular approach for removing their influence from data. In this method, a linear regression model is fit to each feature in the dataset, with the confound as a predictor. The variance explained by the confound is then subtracted from the data to remove its influence. Cross-validated confound regression (CVCR) is a modified version of confound regression that has been used recently in the context of neuroimaging data. In CVCR, the confounds are fit to several folds of the training data, and the learned parameters of the regressor are then used to deconfound both the training and test data.

In a study by Snoek et al. (2019), CVCR was found to retain predictive accuracy better than the standard whole-dataset confound regression method. In this study, the researchers used CVCR to deconfound brain size from structural MRI data for predicting gender. They found that CVCR significantly outperformed the whole-dataset confound regression method.

Overall, the challenge of confounding variables in predictive modeling requires careful consideration and appropriate methods for addressing their influence on data. CVCR is a promising approach for removing confounds from imaging data, but further research is needed to determine its efficacy across different applications and datasets.

1.2 Explaining the model

This passage discusses the issue of model explainability in machine learning. Some models, such as Random Forest classifiers, are transparent and it is easy to understand which input features contribute the most to the model's prediction. In contrast, neural networks are considered black-box models due to the complex and often non-linear mapping between input features and output predictions. However, there are methods to extract relative feature importances from deep neural network models, such as the Shapley value approach which attributes how much difference is accounted for by each feature.

Saliency mapping is a set of methods used for visualizing which parts of an image are most important to a convolutional neural network's decision-making process. This can be done by computing the gradient of the score for a class with respect to the image pixels, determining which pixels need to be changed the smallest amount to change the classification. While these methods have been applied to models in medical image analysis, there is evidence that they should be used with caution when attempting to localize features in high-risk domains.

1.2 Model Explanation

This passage discusses model explainability in machine learning. Some models like Random Forest classifiers are transparent and reveal the contribution of input features to predictions. In contrast, neural networks are considered black-box models due to complex and non-linear mappings between inputs and outputs. However, methods like the Shapley value approach can extract relative feature importances from deep neural networks.

Saliency mapping visualizes important image parts for a convolutional neural network by computing gradients of class scores with respect to image pixels. These methods have been used in medical image analysis but caution is needed in high-risk domains.

1.3 Methods

Data from a number of healthy controls and patients divided into treatment-responsive and treatment-non-responsive groups will be available to the pupils. The initiative will look into whether cortical function and form characteristics may indicate how well a patient will respond to therapy. The students will have to utilize machine learning to clean and denoise functional MRI data, remove annoyance factors, and then predict treatment responses from these characteristics. A model of typical cortical architecture may be created using Gaussian process regression and used to compare patient groups.

1.4 Skills

Several different types of pupils can benefit from the initiative. Manually labeling functional maps as signal/noise to clean data and enhance sensitivity to features sensitive to treatment response may be useful for medical students. Each student may design studies that compare several groups statistically. Students will be needed to know Python and have at least a basic understanding of machine learning to

1.3 Methods

Data from about a number of healthy controls and patients divided into treatment responsive and treatment non-responsive groups will be available to the pupils. The initiative will look into whether cortical function and form characteristics may indicate how well a patient will respond to therapy. The students will have to utilize machine learning to clean and denoise functional MRI data, remove annoyance factors, and then predict treatment response from these characteristics. A model of typical cortical architecture may be created using gaussian process regression and used to compare patient groups.

1.4 Skills

Several different types of pupils can benefit from the initiative. Manually labeling functional maps as signal/noise to clean data and enhance sensitivity to features sensitive to treatment response may be useful for medical students. Each student may design studies that compare several groups statistically. Students will be needed to know Python and have at least a basic understanding of machine learning to participate in the machine learning experiments. Software from the Human Connectome Project and FSL for cortical surface and whole-brain image processing will be trained.

participate in the machine learning experiments. Software from the Human Connectome Project and FSL for cortical surface and whole-brain image processing will be trained.

1.5 State of the Art

Current schizophrenia treatments include pharmaceutical and psychotherapy methods. Antipsychotic medication is the cornerstone, effectively lowering positive symptoms like delusions and hallucinations. Atypical antipsychotics, also known as second-generation antipsychotics, are preferred due to the lower risk of adverse effects compared to first-generation antipsychotics. However, antipsychotics are not effective in treating negative symptoms and cognitive impairment.

Cognitive-behavioral therapy (CBT) and family therapy have shown promise in improving outcomes for schizophrenia patients. CBT has successfully treated symptoms like paranoia and delusions and may enhance daily functioning and quality of life. Family therapy has been shown to improve family functioning and reduce caregiver stress.

Neuromodulation methods, such as transcranial magnetic stimulation (TMS), have gained attention as potential treatments for schizophrenia. TMS, a non-invasive therapy that uses magnetic fields to stimulate specific brain areas, has been shown to be effective in easing symptoms of depression and other psychiatric illnesses. Deep brain stimulation (DBS), still in the experimental stage, requires further research to confirm its safety and effectiveness in treating schizophrenia.

1.6 Artificial Intelligence and Machine Learning in Healthcare

Machine learning is transforming healthcare by providing data-driven insights that can aid in identifying, treating, and monitoring medical disorders. Medical imaging is a major field where machine learning is making an impact, as algorithms can automatically identify, diagnose, and monitor diseases

1.5 State of the art

The most current methods for treating schizophrenia include pharmaceutical and psychotherapy methods. The cornerstone of treatment is antipsychotic medication, which is useful in lowering positive symptoms, including delusions and hallucinations. Atypical antipsychotics, commonly referred to as second-generation antipsychotics, are used frequently because they have a lower risk of adverse effects than first-generation antipsychotics. Antipsychotics are excellent at treating positive symptoms, but they are ineffective at treating negative symptoms and cognitive impairment.

Cognitive-behavioral therapy (CBT) and family therapy are two psychotherapeutic approaches that have shown promise in improving outcomes for those with schizophrenia. Symptoms like paranoia and delusions have been successfully treated with CBT, and it may also enhance daily functioning and quality of life. Family therapy has been demonstrated to enhance family functioning and lessen the stress on caregivers.

Transcranial magnetic stimulation (TMS) and other neuromodulation methods have drawn more attention in recent years as possible treatments for schizophrenia. The use of magnetic fields to stimulate particular parts of the brain during TMS, a non-invasive therapy, has been demonstrated to be useful in easing the symptoms of depression and other psychiatric illnesses. Deep brain stimulation (DBS), for example, is still in the experimental stage and needs more study to confirm its safety and effectiveness in the treatment of schizophrenia.

1.6 Artificial Intelligence and Machine Learning in healthcare

Healthcare is a subject where machine learning is becoming more and more significant, altering how we identify, treat, and keep track of numerous medical disorders. Healthcare practitioners may now get more precise, data-driven insights thanks to machine learning algorithms that can scan massive volumes of data, spot trends, and make predictions based on that data.

Medical imaging is one of the major fields where machine learning is having an influence on healthcare. Machine learning algorithms may be trained to automatically identify, diagnose, and monitor a variety of medical diseases as high-resolution medical pictures become more widely available. For example, machine learning algorithms have been used to forecast and track the evolution of several medical illnesses, as well as to identify and diagnose disorders including breast cancer, heart disease, and schizophrenia.

Drug development is another area of healthcare where machine learning is making a huge difference. It is possible to train machine learning algorithms to assess enormous volumes of biological data, find fresh therapeutic targets, and forecast the probable effectiveness of novel medications. As a result, new, effective medicines are being developed more quickly and with more efficiency.

Moreover, machine learning is helping to improve patient outcomes and treatment. For instance, patient outcomes, like as the chance of readmission and the likelihood of a successful treatment result, may be predicted using machine learning algorithms. The patient treatment may be prioritized using this information, and resources can be distributed more effectively.

Healthcare has been transformed by machine learning and deep learning algorithms, which provide fresh and creative approaches to disease diagnosis, treatment, and prognosis. Deep learning algorithms such as convolutional neural networks (CNNs) are often utilized in image categorization applications. CNNs have been used to analyze numerous imaging modalities in the healthcare industry, including X-rays, MRIs, and computed tomography (CT) images.

CNNs are used in medical imaging to automatically spot patterns and abnormalities in pictures that might be signs of a disease or illness. For instance, CNNs have been applied to medical imaging data to identify breast, lung, and diabetic retinopathy symptoms. Deep learning algorithms have also been employed to increase the diagnostic precision of medical imaging, particularly in circumstances where human interpretation is subject to bias or inaccuracy.

Moreover, algorithms developed using machine learning are used to forecast patient outcomes and treatment results. For instance, the probability of readmission in patients with chronic conditions like diabetes and heart failure has been predicted using deep learning algorithms. The cost on the healthcare system can be lessened and patient outcomes can be improved by using this information to deliver targeted and individualized treatments to avoid readmission.

There are still a number of problems and difficulties that need to be resolved in the application of machine learning and deep learning algorithms in healthcare, despite the significant advancements made in this area. Data quality and availability is a significant issue. Data privacy and security issues can make it challenging and complex to collect medical imaging data. Furthermore, obtaining the vast quantities of high-quality, annotated data that machine learning algorithms need to train may be challenging and time-consuming.

using high-resolution medical images. Drug development is another area where machine learning is making a significant difference, enabling the assessment of large volumes of biological data to identify therapeutic targets and predict the effectiveness of new medications.

Furthermore, machine learning is improving patient outcomes by predicting readmission rates and treatment results, allowing for more effective resource allocation and individualized treatments.

Deep learning algorithms, such as convolutional neural networks (CNNs), are widely used in medical imaging for pattern recognition and disease identification. These algorithms have been successfully applied to various imaging modalities, including X-rays, MRIs, and CT images, to detect diseases such as breast cancer, lung disease, and diabetic retinopathy.

Despite the significant advancements in machine learning and deep learning in healthcare, challenges remain, including data quality, availability, privacy, and interpretability of algorithm outcomes. However, these technologies have the potential to revolutionize healthcare if further improved and supported by infrastructure and medical data investment.

The use of pharmaceutical and psychotherapy therapies, along with neuromodulation approaches, is currently considered the state of the art in treating schizophrenia, although more research is needed for a comprehensive understanding of their mechanisms and long-term efficacy.

1.7 Problem

There are presently no reliable predictors of treatment response or non-response, despite the existence of several neuroimaging studies on schizophrenia patients. This makes it challenging to precisely predict which medicines will be helpful for certain individuals and to enhance the treatment results for schizophrenia patients (Phillips et al., 2008,) This dissertation seeks to solve this problem by examining

The interpretability of the outcomes produced by these algorithms presents another difficulty. Although deep learning algorithms may provide extremely accurate predictions, it can be challenging to comprehend the underlying principles and logic of these predictions. It may be challenging to test and generalize the findings produced by these algorithms due to their lack of interpretability.

Despite these difficulties, machine learning and deep learning algorithms have the power to completely transform the healthcare industry by offering fresh, cutting-edge approaches to challenging medical issues. The potential for these technologies to revolutionize healthcare is enormous, provided that these algorithms continue to be improved and that there is more investment in infrastructure and medical data.

The use of both pharmaceutical and psychotherapy therapies is now the state of the art in the treatment of schizophrenia, with neuromodulation approaches garnering increasing attention as a potential new treatment option. To completely comprehend the mechanism of action of these medicines and their long-term success, however, more study is required.

Problem

There are presently no reliable predictors of treatment response or non-response, despite the existence of several neuroimaging studies on schizophrenia patients. This makes it challenging to

precisely predict which medicines will be helpful for certain individuals and to enhance the treatment results for schizophrenia patients (Phillips et al., 2008,) This dissertation seeks to solve this problem by examining the possibility of enhancing the accuracy of treatment response prediction in schizophrenia patients by combining cortical surface modeling and machine learning.

1.7 Aim

Healthcare has been transformed by machine learning and deep learning algorithms, which provide fresh and creative approaches to disease diagnosis, treatment, and prognosis. Deep learning algorithms such as convolutional neural networks (CNNs) are often utilized in image categorization applications. CNNs have been used to analyze numerous imaging modalities in the healthcare industry, including X-rays, MRIs, and computed tomography (CT) images.

Cortical surface modeling, which enables the investigation of minute variations in the shape, thickness, and structure of the cortex, offers a more complex and nuanced approach to comprehending the underlying brain processes of schizophrenia. By using the plethora of information present in the functional MRI data of patients, this technique has the potential to generate more precise predictions of therapy response when combined with machine learning.

Data from some control subjects and patients with schizophrenia, who will be divided into groups who responded to therapy and those that did not, will be used in the study. The data from functional imaging will be cleaned and denoised using machine learning approaches, and the students will then pinpoint the important characteristics that are indicative of treatment response.

the possibility of enhancing the accuracy of treatment response prediction in schizophrenia patients by combining cortical surface modeling and machine learning.

1.8 Aim

Healthcare has been transformed by machine learning and deep learning algorithms, which offer creative approaches to disease diagnosis, treatment, and prognosis. Convolutional neural networks (CNNs), a type of deep learning algorithm, are commonly used in healthcare for image categorization of various modalities such as X-rays, MRIs, and CT scans.

Cortical surface modeling provides a nuanced approach to understanding brain processes in schizophrenia by analyzing the shape, thickness, and structure of the cortex using functional MRI data. This technique has the potential to improve predictions of therapy response when combined with machine learning, using data from control subjects and patients with schizophrenia, grouped by treatment response.

Machine learning approaches will be used to clean and denoise the functional imaging data, and Gaussian process regression may also be employed to create a model of typical cortex architecture for comparison. The findings of this study may have significant ramifications for psychiatry, leading to more precise treatment response predictions and better outcomes for schizophrenia patients. Additionally, the results may advance our understanding of schizophrenia's underlying brain pathways and offer fresh perspectives on its pathogenesis.

In conclusion, this research presents an opportunity to further comprehend schizophrenia and improve the lives of those affected by the disorder. Machine learning algorithms and cortical surface modeling have the potential to revolutionize the prediction of treatment effectiveness in schizophrenia and open up new avenues for psychiatric research and discovery.

In order to create a model of typical cortex architecture against which patient groups may be contrasted, Gaussian process regression may also be utilized.

The findings of this study will have significant ramifications for the field of psychiatry since they might lead to more precise predictions of treatment response, which could lead to better treatment outcomes for schizophrenia patients. The results of this investigation will also advance our knowledge of the brain pathways that underlie schizophrenia and perhaps offer fresh perspectives on its pathogenesis.

In conclusion, this research presents a rare opportunity to further our comprehension of schizophrenia and improve the lives of people who experience the disorder. Machine learning algorithms and cortical surface modeling have the potential to revolutionize the way we think about predicting the effectiveness of schizophrenia treatments and to open up new vistas for psychiatric research and discovery.

1.8 Work plan

A comprehensive work plan for investigating the potential for using a combination of cortical surface modeling and machine learning to predict treatment response in schizophrenia patients would involve the following steps:

- Data Collection and Preparation: Examining the Dataset
- The df.info() method displays details about the DataFrame, such as the number of rows, columns, and data types. The DataFrame has a multi-level index with 98 entries. The columns provide the patients' clinical and demographic information, such as age, sex, and PANSS scores, as well as other brain anatomical parameters including curvature, sulc, and thickness.
- It is also important to remember that the dataframe's form was 97, 1 before the multi-index was removed.
- Each participant's functional imaging data should be included in the data. To guarantee that the findings are precise and insightful, this data should be denoised, and cleaned, and the nuisance factors should be removed.
-
- To discover characteristics of cortex function and cortical form, the data would be evaluated using cortical surface modeling techniques in this stage. This would include building models of typical cortical structure and translating the functional MRI data onto a surface representation of the brain. To find any discrepancies between the patient and control groups, the models should be compared to the patient data.

1.9 Work Plan

A concise work plan for investigating the use of cortical surface modeling and machine learning to predict treatment response in schizophrenia patients would involve the following steps:

1. Data Collection and Preparation: Examining the Dataset
    - Use df.info() to display details about the DataFrame, including the number of rows, columns, and data types.
    - Note that the DataFrame had a multi-level index with 98 entries, and it was previously in the form of 97, 1 before the multi-index was removed.
    - Include functional imaging data for each participant, denoise and clean the data, and remove nuisance factors.
2. Cortical Surface Modeling: Analyzing Cortex Function and Form
    - Utilize cortical surface modeling techniques to build models of typical cortical structure and translate functional MRI data onto a surface representation of the brain.
    - Compare the models to the patient data to identify discrepancies between the patient and control groups.
3. Feature Selection and Engineering: Choosing Informative Characteristics
    - Select the subset of features that provide the most accurate predictions of treatment response.
    - Use feature extraction methods, feature selection algorithms, or manual feature selection approaches to alter or produce new features from the existing data.
4. Machine Learning: Training and Optimizing the Model
    - Train a machine learning algorithm, such as logistic regression, decision trees, or random forests, using the selected data and features.

- Feature Selection and Engineering: When the cortical surface models have been developed, the next step would be to choose the characteristics that will provide the most accurate predictions of the treatment response. Finding the subset of features that is most informative for the job at hand and altering or producing new features from the existing data would be required to do this. Techniques like feature extraction methods, feature selection algorithms, or manual feature selection approaches should be used for this procedure.

- Machine learning: At this stage, a machine learning algorithm would be trained to predict treatment response using the data and characteristics chosen in the preceding steps. A suitable machine learning method, such as logistic regression, decision trees, or random forests, would need to be chosen, and the parameters would need to be adjusted to maximize performance. The performance of the model should be assessed using cross-validation methods in order to spot any overfitting or underfitting problems.

- Evaluation and Validation: Using a set of independent data, the model's performance would be assessed and validated as the last phase. Using data from a different cohort or dividing the current data into a training and validation set might be two ways to do this. Metrics like accuracy, precision, recall, and F1-score should be used in the assessment to evaluate the model's performance. It may be necessary to take extra measures to enhance

- Adjust the parameters to optimize performance and assess the model using cross-validation methods to identify overfitting or underfitting issues.

5. Evaluation and Validation: Assessing Model Performance
   - Assess and validate the model's performance using independent data from a different cohort or by dividing the current data into training and validation sets.
   - Use metrics such as accuracy, precision, recall, and F1-score to evaluate the model's performance.
   - Take additional measures, such as adding more features or using a different machine learning technique, to enhance the model if needed.
6. Interpretation and Reporting: Reporting Outcomes and Suggestions for Further Study
   - Report and interpret the outcomes of the study, including a discussion of the results and their implications for the care of people with schizophrenia.
   - Provide suggestions for further research in this field, such as creating new prediction models or exploring other machine learning techniques.

Overall, this concise work plan ensures a systematic and rigorous approach to evaluating the potential of combining cortical surface modeling and machine learning for predicting treatment responses in schizophrenia patients. It focuses on essential steps in data preparation, feature selection, machine learning, assessment, and reporting.

1.10 Deliverables

Deliverables for the schizophrenia research project may include:

1. Literature reviews: A comprehensive analysis of existing research on schizophrenia, including its etiology, current therapies, and results.

the model if the performance is unsatisfactory, such as adding more features or using a different machine-learning technique.

Interpretation and Reporting: The outcomes must be reported and interpreted after the model has been verified. This should include a discussion of the results and how they relate to the care of people with schizophrenia. The report should also include suggestions for further study in this field, such as the creation of new prediction models or the use of other machine learning techniques.

Overall, this study plan should offer a thorough and organized method for evaluating the possibility of applying a hybrid of machine learning and cortical surface modeling to predict treatment response in schizophrenia patients. It should make sure that the study is conducted methodically and rigorously, paying attention to all relevant aspects of data preparation, feature selection, machine learning, assessment, and reporting.

1.9 Deliverables

The physical items or outputs that are created as a result of a project or procedure are known as deliverables. Deliverables in a research project are often the written materials, goods, or results that are created as a result of the investigation. The potential deliverables that may be anticipated in this schizophrenia research project are listed below:

1. Literature reviews A thorough analysis of the extant literature on the subject of schizophrenia, including an overview of the state of knowledge on the illness, its underlying causes, current therapies, and results.
2. A thorough plan outlining the procedures to be followed in order to gather information from patients and healthy controls, including the type of information to be gathered (such as imaging data), the population from which it will be gathered, and the procedures and protocols to be followed in order to guarantee the accuracy and validity of the information.
3. Data Analysis Plan: a thorough strategy describing the procedures to be utilized to analyze the information gathered from patients and healthy controls. The procedures and protocols that will be utilized to guarantee the quality and validity of the results should also be covered in this plan, along with specifics on the methods and algorithms that will be used to evaluate the data.
4. Results: a thorough overview of the study's findings that includes a close examination of the data, statistical tests and analysis, as well as the main conclusions and findings.

2. Procedures plan: Detailed outline of information gathering procedures from patients and healthy controls, including type of data (e.g. imaging), target population, and protocols for accuracy and validity.
3. Data Analysis Plan: Detailed strategy for analyzing data from patients and healthy controls, including quality and validity protocols, and specific methods and algorithms for evaluation.
4. Results: Comprehensive overview of study findings, including data analysis, statistical tests, main conclusions, and findings.
5. Conclusion and Recommendations: Discussion of study's implications, recommendations for further research or application, and a conclusion highlighting main discoveries.
6. Final Report: Comprehensive summary of all study deliverables, including data gathering and analysis plans, findings, conclusions, and recommendations.
7. Presentation: Visual presentation of study's data and main conclusions through slides, pictures, and graphs for stakeholders and wider dissemination.
8. Code: Well-documented sharing of data collection and analysis code, including scripts and algorithms created for the project to promote transparency and reproducibility.
9. Manuscripts: Submission of one or more papers describing study's main conclusions to peer-reviewed publications for publication.

1.11 Evaluation

1. Evaluation components: The main focus of the review process will be on the accuracy of the predictive models in determining treatment responses for schizophrenia patients, the validity of the study's methodologies compared to previous research, the consistency and reliability of the data, the completeness of the results in providing a comprehensive understanding of the potential for improving

5. Conclusion and Recommendations: A discussion of the consequences of the study's findings, together with any suggestions for further study or application, and a conclusion that highlights the main discoveries and contributions of the investigation.
6. Final Report: A thorough and thorough report that summarizes all of the study's deliverables, including the plans for the data gathering and analysis, the findings, and the conclusions and suggestions.
7. Presentation: A visual presentation of the study's data and main conclusions that includes slides, pictures, and graphs. This presentation will be used to deliver the research to stakeholders and disseminate it to a larger audience.
8. Code: To encourage transparency and reproducibility of the study, the code used for data collection and analysis, including any scripts or algorithms created especially for the project, will be well-documented and shared with the scholarly community.
9. Manuscripts: One or more papers describing the study's main conclusions that will be submitted to peer-reviewed publications for publication.

1.9.1 Evaluation

1. To evaluate the overall performance and efficacy of the study, this project will include a number of important components. The following areas will be the main focus of the review process:
2. Predictive model accuracy: The project's main goal is to determine if cortical surface modeling and machine learning can be used in conjunction to more accurately predict treatment responses for people with schizophrenia. By contrasting the expected treatment response with the actual treatment response of the study's patients, the predictive models' accuracy will be evaluated.
3. Validity of the Methods: By contrasting the findings with those of earlier research that employed comparable methodologies, the validity of the study's procedures will be determined. The study's findings should be in line with earlier studies and offer new information about the causes of schizophrenia as well as the possibility of bettering treatment outcomes.
4. Data Consistency and Reliability: The consistency and reliability of the data will be examined in order to assess the data's quality. Data collection should follow regular procedures, and data analysis should be carried out using the right statistical approaches.
5. Completeness of Results: The results' comprehensiveness will be assessed by looking at whether all pertinent data has been gathered and examined. The outcomes should offer a thorough grasp of the possibility for enhancing the accuracy with which treatment response for schizophrenia patients may be anticipated by combining cortical surface modeling with machine learning.

treatment response prediction, the significance of the study in real-world clinical settings, limitations of the study's methodology, and recommendations for future research to enhance treatment response prediction accuracy.

2. Predictive model accuracy: Evaluating the accuracy of the predictive models by contrasting expected and actual treatment responses of the study's patients.

3. Validity of the Methods: Determining the validity of the study's procedures by comparing findings with previous research that used similar methodologies.

4. Data Consistency and Reliability: Examining the consistency and reliability of the data to assess its quality, including adherence to standard data collection procedures and appropriate statistical analysis.

5. Completeness of Results: Assessing the comprehensiveness of the study's results by ensuring that all relevant data has been collected and analyzed to provide a thorough understanding of treatment response prediction accuracy using cortical surface modeling and machine learning.

6. Significance of the Study: Evaluating the practical relevance of the study's findings in real-world clinical settings and their potential to improve treatment outcomes for schizophrenia patients.

7. Limitations of the Study: Identifying and acknowledging any limitations or restrictions in the study's methodology while evaluating findings and providing suggestions for further research, taking into consideration the study's limitations.

8. Suggestions for Future Research: Making doable recommendations for future research based on the study's findings to further enhance treatment response prediction accuracy, while considering the study's limitation.

6. Significance of the Study: The study's applicability in actual clinical settings will be looked at in order to assess its relevance. The findings should have practical ramifications for the treatment of schizophrenia and shed light on how treatment outcomes could be enhanced.
7. Limitations of the Study: Any restrictions or limitations in the study's methodology will be examined in order to assess the study's limitations. While evaluating the findings and providing suggestions for further research, the study's limitations should be taken into mind.
8. Suggestions for Future Research: In order to continue to increase the accuracy with which treatment response for schizophrenia patients may be predicted, recommendations for future research should be made based on the study's findings. The suggestions must to be doable and ought to take the study's limitations into consideration.

Acknowledgment

I want to thank everyone who has been kind and helpful to me during this study process. I want to start by expressing my gratitude to my supervisor for their direction, assistance, and encouragement during the assignment. Their understanding of schizophrenia and neuroimaging has been quite helpful to me. I also want to express my gratitude to the healthy volunteers and research participants for their willingness to advance our understanding of this condition. I would like to express my gratitude for the support provided by the workers at the imaging facility where the data was gathered in obtaining high-quality imaging data.

Finally, I would want to thank my friends and family for their assistance with this effort. Their support and compassion have been crucial to my achievement.

Ultimately, without the help and assistance of everyone engaged, this study would not have been possible. I owe a sincere debt of gratitude to all of those who supported me in this effort.

1.12 Acknowledgment

I want to thank everyone who has been kind and helpful to me during this study process. I want to start by expressing my gratitude to my supervisor for their direction, assistance, and encouragement during the assignment. Their understanding of schizophrenia and neuroimaging has been quite helpful to me. I also want to express my gratitude to the healthy volunteers and research participants for their willingness to advance our understanding of this condition. I would like to express my gratitude for the support provided by the workers at the imaging facility where the data was gathered in obtaining high-quality imaging data.

Finally, I would want to thank my friends and family for their assistance with this effort. Their support and compassion have been crucial to my achievement.

Ultimately, without the help and assistance of everyone engaged, this study would not have been possible. I owe a sincere debt of gratitude to all of those who supported me in this effort.

1.13 Abstract

Schizophrenia is a neuropsychiatric condition characterized by persecutory delusions and severe auditory hallucinations. Neuroimaging research has linked schizophrenia to disruption in brain networks that support crucial psychological functions related to perception and understanding of the environment. Despite numerous imaging investigations, there are no reliable predictors of treatment response. This study aims to explore the use of cortical surface modeling and machine learning to enhance treatment response prediction accuracy for schizophrenia patients. The study will use data from healthy controls and patients divided into treatment-responsive and treatment-non-responsive groups. The objective is to determine whether cortical function and cortical shape characteristics can predict therapy response. Machine learning approaches will be used to clean and denoise functional imaging data and predict

1.9.3 Abstract

Schizophrenia is a neuropsychiatric condition marked by persecutory delusions and severe auditory hallucinations. According to neuroimaging research, schizophrenia is linked to disruption in brain networks supporting crucial psychological functions that enable people to properly perceive and understand their environment. There are no reliable predictors of therapy response or non-response in spite of numerous imaging investigations of patients. The purpose of this study is to explore the possibility of employing a cortical surface modeling and machine learning combination to enhance the accuracy of treatment response prediction for schizophrenia patients. Data from healthy controls and patients who were divided into treatment-responsive and treatment-non-responsive groups will be used in the study. The project's objective is to determine whether characteristics of cortical function and cortical shape can predict how well a patient will respond to therapy. To deconfound unwanted variables, clean and denoise functional imaging data, and predict treatment response from these characteristics, the study will apply machine learning approaches. The findings of this study might help patients receive better care and shed light on the neurological pathways that underlie schizophrenia.

treatment response. The findings of this study may improve patient care and provide insights into the neurological pathways underlying schizophrenia.

Chapter 2: Literature Review

2.1 Theory

In this chapter, we will provide an overview of machine learning and its potential applications in the field of psychiatry, specifically in the prediction of treatment response in schizophrenia patients. We will discuss the various types of machine learning algorithms and their strengths and weaknesses in the context of this project. Additionally, we will discuss the importance of feature selection and feature engineering in building accurate predictive models.

2.2 Machine Learning Overview

The main goal of machine learning in artificial intelligence is to create algorithms that can learn from data and make predictions or judgments without explicit programming. Machine learning algorithms come in different forms, such as supervised learning, unsupervised learning, and reinforcement learning, depending on whether the result variable or label is known or unknown, and whether the feedback is received in the form of incentives or penalties.

In this study, we will use supervised learning algorithms to forecast treatment responses in schizophrenia patients. By dividing patients into responsive and non-responsive groups, we have labeled data that is well-suited for supervised learning algorithms. Common supervised learning techniques used in psychiatry include logistic regression, decision trees, and random forests.

Chapter 2: Literature Review

Theory

In this chapter, we will provide an overview of machine learning and its potential applications in the field of psychiatry, specifically in the prediction of treatment response in schizophrenia patients. We will discuss the various types of machine learning algorithms and their strengths and weaknesses in the context of this project. Additionally, we will discuss the importance of feature selection and feature engineering in building accurate predictive models.

2.1 Machine Learning Overview

The creation of algorithms that can learn from data and make predictions or judgments without being explicitly programmed is the main goal of the artificial intelligence branch of machine learning. Machine learning algorithms come in a variety of forms, such as supervised learning, unsupervised learning, and reinforcement learning. When the result variable, or label, is known, supervised learning algorithms are employed; when the end variable is unknown, unsupervised learning algorithms are used. When an algorithm receives feedback in the form of incentives or penalties, reinforcement learning methods are utilized.

In this study, we will use supervised learning algorithms to forecast how schizophrenia patients will respond to therapy. With patients divided into groups that respond to therapy and those who don't, our labeled data makes supervised learning algorithms an excellent choice for this purpose. In the discipline of psychiatry, supervised learning techniques including logistic regression, decision trees, and random forests are frequently employed.

Building precise predictive models requires a number of key phases, including feature engineering and feature selection. Although feature engineering entails changing or producing new features from the current data, feature selection entails determining the subset of characteristics that are most informative for the job at hand. In this research, we'll combine cortical surface modeling and machine learning to increase the accuracy with which treatment responses for schizophrenia patients can be predicted.

2.2 Bias and Variance

Two fundamental ideas in statistical modeling and machine learning are bias and variance. They discuss the two potential causes of mistake while developing predictive models.

When a model consistently over- or under-estimates the real connection between the input variables and the target variable, it is said to be biased. No of the input data, a model with strong bias will repeatedly make the same mistake. This kind of mistake happens when the model is overly straightforward and fails to adequately represent the intricacy of the underlying connection.

Contrarily, variation describes the inaccuracy that happens when a model is too sophisticated and susceptible to sporadic variations in the input data. Even if the training sets come from the same underlying population, a model with significant variance will provide different predictions for each set. This kind of inaccuracy happens when the model is overly flexible and fits the data too well, capturing the noise in the input data rather than the underlying relationship.In machine learning, balancing bias and variance is crucial since both excessive bias and high variance can result in subpar model performance. In order to effectively represent the underlying link between

Building accurate predictive models involves key phases like feature engineering and feature selection. Feature engineering involves modifying or creating new features from existing data, while feature selection involves identifying the most informative subset of characteristics for the task at hand. In this study, we will combine cortical surface modeling with machine learning to improve the accuracy of treatment response prediction in schizophrenia patients.

2.3 Bias and Variance

Two fundamental ideas in statistical modeling and machine learning are bias and variance, which are potential causes of mistakes when developing predictive models.

Bias occurs when a model consistently over- or under-estimates the true relationship between input variables and the target variable. A model with strong bias will repeatedly make the same mistake regardless of the input data. This type of error occurs when the model is overly simplistic and fails to adequately capture the complexity of the underlying relationship.

On the other hand, variance refers to the inaccuracy that arises when a model is too complex and sensitive to random variations in the input data. Even if the training sets come from the same underlying population, a model with high variance may provide different predictions for each set. This type of error occurs when the model is overly flexible and fits the data too well, capturing the noise in the input data rather than the underlying relationship.

Balancing bias and variance is crucial in machine learning, as both excessive bias and high variance can lead to subpar model performance. To effectively capture the underlying relationship between input variables and the target variable, a model should have low bias and low variance. Machine learning algorithms often employ methods such as regularization, cross-validation, and ensembling to manage the model's complexity and avoid overfitting, in order to achieve this balance.

the input variables and the target variable, a model should have low bias and low variance.Machine learning algorithms frequently utilize methods like regularization, cross-validation, and ensembling to manage the model's complexity and avoid overfitting in order to achieve this balance.

2.2.1 Bias and Variance in Schizophrenia

The problem of bias and variation can significantly affect the accuracy and reliability of the predictions when it comes to predicting treatment response in schizophrenia patients.

For instance, if a model is predisposed to one type of treatment over another, it may consistently overestimate or underestimate that treatment's efficacy across all patients, yielding unreliable predictions. On the other hand, even if patients have comparable features, the model may provide different predictions for them if its variance is significant. Because to this, it may be difficult to predict which therapies will be successful for specific individuals, which may lead to less-than-ideal treatment results.

The problem of bias and variation can significantly affect the accuracy and reliability of the predictions when it comes to predicting treatment responses in schizophrenia patients.

For instance, if a model is predisposed to one type of treatment over another, it may consistently overestimate or underestimate that treatment's efficacy across all patients, yielding unreliable predictions. On the other hand, even if patients have comparable features, the model may provide different predictions for them if its variance is significant. Because to this, it may be difficult to

2.3.1 Bias and Variance in Schizophrenia

The problem of bias and variance can significantly impact the accuracy and reliability of predictions when predicting treatment responses in schizophrenia patients. For instance, if a model is biased towards one type of treatment, it may consistently overestimate or underestimate its efficacy across all patients, resulting in unreliable predictions. Additionally, if the model has high variance, it may provide different predictions for patients with similar features. This can make it challenging to accurately predict successful therapies for specific individuals, leading to suboptimal treatment outcomes.

To mitigate these issues, it is important to use a diverse and representative sample of patient data and employ approaches to manage the complexity of the model. By doing so, more accurate and dependable predictions can be made, ultimately improving treatment outcomes for schizophrenia patients.

2.4 Decision Trees and Ensemble Learning

Two popular machine learning techniques for predicting therapy response in schizophrenia patients are decision trees and ensemble learning.

Decision trees are a type of tree-based model that uses if-then rules to make predictions. They iteratively divide data into smaller groups based on the characteristics that best describe the target variable. Decision trees are known for their simplicity, interpretability, and ability to handle both continuous and categorical data. They can also be easily integrated into ensemble models, such as random forests.

Ensemble learning involves combining multiple machine learning models to create a more reliable and accurate forecast. The idea behind ensemble learning is that combining diverse models can result in more accurate predictions compared to using a single model. Random forest, which is an extension of decision trees, is a common ensemble approach where multiple decision trees are combined to provide a final prediction.

predict which therapies will be successful for specific individuals, which may lead to less-than-ideal treatment results.

Accurate and dependable predictions may be made by using a wide-ranging and representative sample of patient data and applying approaches to manage the model's complexity, which eventually improves treatment results for schizophrenia patients.

2.3 Decision Trees and Ensemble Learning

Two popular machine learning techniques that can be utilized to solve the issue of predicting therapy response in schizophrenia patients are decision trees and ensemble learning.

A particular kind of tree-based model called a decision tree employs a sequence of if-then rules to create predictions. The characteristics that best describe the target variable are used to iteratively divide the data into smaller and smaller groups. Decision trees are well renowned for being straightforward, comprehensible, and capable of handling both continuous and categorical data. They may also be quickly integrated to create ensemble models using other machine-learning techniques, such as random forests.

As the name implies, ensemble learning involves merging many machine learning models to create a forecast that is more reliable and accurate. The concept behind ensemble learning is that combining many models will result in a forecast that is more accurate than using just one model. The decision tree extension known as random forest is a common ensemble approach. To provide a final prediction, the random forest algorithm create several decision trees and combine their predictions.

The challenge of predicting therapy response in schizophrenia patients can be solved using decision trees and ensemble learning. It could be feasible to make forecasts that are more accurate and trustworthy by integrating the advantages of several models, which would enhance the treatment results for schizophrenia patients. However, while applying these procedures in this situation, it's crucial to take into account their drawbacks, such as the danger of overfitting.

2.4 Parallel ensemble learning - Random Forests

A potent method addressing the challenge of predicting therapy response in schizophrenia patients is parallel ensemble learning, especially employing random forests. A form of ensemble learning technique called random forests combines many decision trees to provide predictions that are more reliable and accurate.

Several random forests are simultaneously trained on various subsets of data in concurrent ensemble learning, enabling scalable and effective training. The forecasts from each separate random forest are combined to get the final prediction. Parallel ensemble learning can aid in lowering prediction variance and enhancing the model's overall accuracy.

Parallel ensemble learning may be used to examine huge and complicated patient data sets in the context of predicting treatment response in schizophrenia patients in order to pinpoint the main determinants of response to therapy. It could be feasible to get more precise and trustworthy forecasts by merging the predictions from various random forests, which would enhance the treatment results for schizophrenia patients.

Using decision trees and ensemble learning can help address the challenge of predicting therapy response in schizophrenia patients. By leveraging the strengths of multiple models, it may be possible to achieve more accurate and trustworthy forecasts, leading to improved treatment outcomes. However, it is important to be mindful of potential drawbacks, such as the risk of overfitting, when applying these techniques in this context.

2.5 Parallel ensemble learning - Random Forests

Parallel ensemble learning, particularly using random forests, is a potent method for addressing the challenge of predicting therapy response in schizophrenia patients. Random forests, an ensemble learning technique, combine multiple decision trees to provide more reliable and accurate predictions. In parallel ensemble learning, multiple random forests are trained simultaneously on different subsets of data, allowing for scalable and effective training. The forecasts from each random forest are then combined to obtain the final prediction, which can help lower prediction variance and improve overall model accuracy.

Parallel ensemble learning can be applied to large and complex patient datasets to identify key determinants of therapy response in schizophrenia patients. By merging predictions from multiple random forests, more precise and trustworthy forecasts may be obtained, leading to improved treatment outcomes.

However, it is important to consider potential drawbacks, such as the risk of overfitting, when applying parallel ensemble learning in this context. Ensuring that the data used for model training is diverse, representative, and of high quality is also crucial to prevent biases in predictions.

While applying parallel ensemble learning in this setting, it is crucial to take into account its drawbacks, such as the possibility of overfitting. To prevent any biases in the predictions, it is also essential to guarantee that the data used to train the models is varied, representative, and of good quality.

2.5 Sequential Ensemble Learning - Boosting

Another approach to the issue of predicting treatment response in schizophrenia patients is sequential ensemble learning, especially employing boosting. Boosting is a form of ensemble learning technique that combines several ineffective models to get a more accurate forecast.

When using boosting, the models are trained sequentially, with each succeeding model being taught to fix the errors generated by the model before it. The results of each separate model are combined to get the final projection. Boosting can aid in lowering prediction bias and enhancing the model's overall accuracy.

Boosting can be used to evaluate patient data to pinpoint the major treatment response variables in the context of predicting treatment response in schizophrenia patients. It could be feasible to get more precise and trustworthy forecasts by merging the predictions from various weak models, which would enhance the treatment results for schizophrenia patients.

2.6 Sequential Ensemble Learning - Boosting

Sequential ensemble learning, particularly using boosting, is another approach for predicting treatment response in schizophrenia patients. Boosting is an ensemble learning technique that combines multiple ineffective models to obtain a more accurate forecast.

In boosting, models are trained sequentially, with each subsequent model focusing on correcting errors made by the previous model. The results of each individual model are combined to obtain the final prediction, which can help reduce prediction bias and improve overall model accuracy.

Boosting can be used to analyze patient data and identify key variables associated with treatment responses in schizophrenia patients. By merging predictions from multiple weak models, more precise and trustworthy forecasts may be obtained, leading to improved treatment outcomes.

However, it is important to consider potential limitations of boosting, such as the risk of overfitting, when applying it in this context. Ensuring that the data used for model training is diverse, representative, and of high quality is also crucial to prevent biases in predictions. The choice between parallel or sequential ensemble learning approaches should be determined based on the unique task and data properties.

2.7 Support Vector Machines

Support Vector Regression (SVR) is a machine learning method that can be utilized to predict treatment response in schizophrenia patients. SVR employs a hyperplane, a boundary, to separate the data into different groups in supervised learning.

SVR maximizes the margin between the support vectors, or nearby data points, to produce a strong and precise model, considering both linear and non-linear correlations between characteristics and the target variable.

When applying boosting in this situation, it is critical to take into account its restrictions, such as the danger of overfitting. To prevent any biases in the predictions, it is also essential to guarantee that the data used to train the models is varied, representative, and of good quality. The unique task at hand and the data's properties determine whether parallel or sequential ensemble learning approaches should be used.

2.6 Support Vector Machines

A form of machine learning method called Support Vector Regression (SVR) may be used to solve regression issues, such as predicting how schizophrenia patients would respond to therapy. A boundary known as the hyperplane is used by the supervised learning technique known as SVR to divide the data into various groups.

In SVR, the hyperplane is used to maximize the margin between the support vectors, or nearby data points. With both linear and non-linear correlations between the characteristics and the target variable being taken into account, a strong and precise model is produced.

SVR may be used to assess patient data and pinpoint the main treatment response variables in the context of predicting therapy response in schizophrenia patients.

In a complicated task like predicting treatment response in schizophrenia, the ability of SVR to handle non-linear correlations between the characteristics and the target variable is advantageous. SVR is also suitable for huge data sets since it is computationally effective and very easy.However, while applying SVR in this situation, it is crucial to take into account its drawbacks, such as the danger of overfitting. To prevent any biases in the predictions, it is also essential to guarantee that the data used to train the model is varied, representative, and of good quality. The particulars of the problem at hand and the properties of the data determine which machine-learning technique should be used.

SVR can be employed to analyze patient data and identify key treatment response variables in the context of predicting therapy response in schizophrenia patients.

The ability of SVR to handle non-linear correlations between characteristics and the target variable is advantageous in complex tasks like predicting treatment response in schizophrenia. SVR is also computationally efficient and suitable for large datasets.

However, it is important to consider potential drawbacks of SVR, such as the risk of overfitting, when applying it in this context. Ensuring that the data used for model training is diverse, representative, and of high quality is crucial to prevent biases in predictions. The choice of machine learning technique should be based on the specific problem and data properties.

Chapter 3: Methodology

3.1 Data Acquisition

In this chapter, we will go through the strategies we employed in our study for data collection and processing. This experiment will gather data from roughly 97 people with schizophrenia, divided into treatment-responsive and non-responsive groups. Each participant's data will include functional magnetic resonance imaging (fMRI) scans as well as demographic and clinical information.

3.1.1 Data Collection

A 3T MRI scanner with a T2*-weighted gradient echo planar imaging (EPI) sequence will be used to gather the fMRI images. TR = 2000 ms, TE = 30 ms, flip angle = 90 degrees, FOV = 240 x 240 mm, matrix size = 64 × 64, and slice thickness = 3 mm will be the imaging parameters. During the scan,

Chapter 3: Methodology

3.1 Data Acquisition

In this chapter, we will go through the strategies we employed in our study for data collection and processing. This experiment will gather data from roughly 97 people with schizophrenia, divided into treatment-responsive and non-responsive groups. Each participant's data will include functional magnetic resonance imaging (fMRI) scans as well as demographic and clinical information.

3.1.1 Data Acquisition

A 3T MRI scanner with a T2*-weighted gradient echo planar imaging (EPI) sequence will be used to gather the fMRI images. TR = 2000 ms, TE = 30 ms, flip angle = 90 degrees, FOV = 240 x 240 mm, matrix size = 64 × 64, and slice thickness = 3 mm will be the imaging parameters. During the scan, participants will be encouraged to lie motionless and keep their eyes closed for roughly 10 minutes. In addition to the fMRI scans, each participant's demographic and clinical information, such as age, gender, education level, and medication history, will be gathered.

3.1.2 Data Processing

The Human Connectome Project (HCP) pipeline will be used to preprocess the fMRI data, which includes motion correction, slice timing correction, and spatial smoothing using a 4-mm FWHM kernel. With the FMRIB Linear Image Registration Tool, the data will be registered to a standard template (FLIRT). The preprocessed data will be utilized for cortical function and shape analysis.

participants will be encouraged to lie motionless and keep their eyes closed for roughly 10 minutes. In addition to the fMRI scans, each participant's demographic and clinical information, such as age, gender, education level, and medication history, will be gathered.

3.1.2 Data Processing

The Human Connectome Project (HCP) pipeline will be utilized for preprocessing the fMRI data, including motion correction, slice timing correction, and spatial smoothing with a 4-mm FWHM kernel. The data will be registered to a standard template (FLIRT) using the FMRIB Linear Image Registration Tool. The preprocessed data will be used for cortical function and shape analysis.

The cortical function will be investigated by extracting time series data from each voxel in the brain and averaging it within each region of interest (ROI), using the Desikan-Killiany atlas with 34 areas per hemisphere. The generated time series data for each ROI will be used to calculate functional connectivity between the ROIs using the Pearson correlation coefficient.

The cortical shape will be examined using the FreeSurfer program, which will recreate the cortical surface from T1-weighted structural images. The final surface will be divided into multiple zones using the Destrieux atlas. Cortical thickness, surface area, and curvature measurements will be obtained for each region and used for statistical analysis.

3.2 Machine Learning

Machine learning techniques like principal component analysis (PCA) and independent component analysis (ICA) will be used to clean and denoise the data, eliminating nuisance factors such as head motion and physiological noise using regression methods. Support vector machines (SVMs) and random forests will then be employed to predict treatment response using the cleaned and denoised data. Gaussian process regression will also be utilized to model normative cortical architecture for

Cortical function will be investigated by extracting time series data from each voxel in the brain and then averaging the data within each region of interest (ROI). The Desikan-Killiany atlas, which comprises of 34 areas per hemisphere, will be used to create ROIs. The time series data generated for each ROI will be utilized to calculate functional connectivity between the ROIs using the Pearson correlation coefficient.

The cortical shape will be examined using the FreeSurfer program, which will recreate the cortical surface using T1-weighted structural images. Using the Destrieux atlas, the final surface will be divided into several zones. Cortical thickness, surface area, and curvature measurements will be retrieved for each region and utilized for statistical analysis.

3.2 Machine Learning

Machine learning techniques such as principal component analysis (PCA) and independent component analysis (ICA) will be used to clean and denoise the data (ICA). Using regression methods, nuisance factors such as head motion and physiological noise will be eliminated from the data. Lastly, using machine learning techniques such as support vector machines (SVMs) and random forests, the cleaned and denoised data will be utilized to predict treatment response.

Gaussian process regression will also be utilized to create a model of normative cortical architecture, which will be used to compare patient groups. We seek to enhance the precision with which treatment response may be predicted for schizophrenia patients by combining these methodologies.

comparison with patient groups, aiming to enhance the precision of treatment response prediction in schizophrenia patients through these combined methodologies.

This chapter has describes the data gathering and processing procedures for our investigation, which includes fMRI scans, demographic, and clinical information for each participant. The HCP pipeline will be used for data preprocessing, followed by cortical surface modeling and machine learning approaches. The purpose of this experiment is to assess if cortical function and structure can predict treatment response in schizophrenia patients.

3.3 Deep Neural Network Architecture

Deep Neural Networks (DNNs) are artificial neural networks with multiple hidden layers and numerous artificial neurons. These networks have proven effective in various applications such as image recognition, natural language processing, and medical diagnosis. DNN architecture, including the number and type of layers, neurons per layer, and activation functions, is a crucial consideration. Convolutional Neural Networks (CNNs) are a common type of DNN architecture, particularly useful for image recognition. Convolutional layers extract image features, while fully connected layers are used for classification. Recurrent Neural Networks (RNNs) are another popular DNN design for sequential input processing, such as voice or text.

3.3.1 Single label classification

The purpose of single label classification is to predict a single class label for each input sample. In image classification, for example, the objective can be to predict if a picture contains a cat or a dog. In this scenario, the network would contain two output neurons, one for "cat" and the other for "dog." The network would then be trained to predict the proper class label as accurately as possible.

Finally, this chapter has described the data gathering and processing procedures used in our investigation. This project's data will include fMRI scans as well as demographic and clinical information for each participant. The HCP pipeline will be used to preprocess the data, which will then be evaluated using a combination of cortical surface modeling and machine learning approaches. The purpose of this experiment is to see if aspects of cortical function and structure predict treatment response in schizophrenia patients.

3.3 Deep neural network architecture

Deep Neural Networks (DNNs) are artificial neural networks with several hidden layers and many artificial neurons. These networks have shown to be an effective tool in a variety of applications, including as image recognition, natural language processing, and medical diagnosis. The architecture of DNNs is a significant consideration, as it pertains to the number and type of layers, the number of neurons in each layer, and the activation functions utilized.

The Convolutional Neural Network (CNN) is a typical form of DNN architecture that is particularly helpful for image recognition applications. Convolutional layers are employed in CNNs to extract picture characteristics, whereas fully connected layers are utilized for classification. The Recurrent Neural Network (RNN) is another prominent DNN design that is excellent for processing sequential input such as voice or text.

3.3.1 Single label classification

The purpose of single label classification is to predict a single class label for each input sample. In image classification, for example, the objective can be to predict if a picture contains a cat or a dog. In this scenario, the network would contain two output neurons, one for "cat" and the other for "dog." The network would then be trained to predict the proper class label as accurately as possible.

3.3.2 Multi-label classification

The purpose of multi-label classification is to predict several class labels for each input sample. In medical diagnostics, for example, the objective may be to anticipate many illnesses for a

3.3.2 Multi-label classification

The purpose of multi-label classification is to predict several class labels for each input sample. In medical diagnostics, for example, the objective may be to anticipate many illnesses for a patient based on their symptoms. The network would have numerous output neurons in this example, each representing a particular ailment. After that, the network would be trained to predict all of the appropriate class labels for each input sample.

Since the network must be able to accommodate several and sometimes contradicting outputs, multi-label classification is more difficult than single-label classification. Many strategies have been proposed to handle this difficulty, including the use of several binary classifiers, one for each label, or the use of a single classifier with numerous outputs. The approach chosen will be determined by the task's unique needs and the data being utilised.

3.4 ROC Interpretation

Receiver Operating Characteristic (ROC) analysis is commonly used in machine learning, particularly in medical image analysis, to assess the performance of binary classification systems. The ROC curve visualizes the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity) for different threshold values. The area under the curve (AUC) provides a single value to represent the overall performance of the classifier.

ROC analysis is used in single-label classification to evaluate the performance of a binary classifier across various threshold values. A good classifier should have high sensitivity and specificity.

In multi-label classification, ROC analysis summarizes the performance of a multi-label classifier for each category, allowing for comparison across categories.

patient based on their symptoms. The network would have numerous output neurons in this example, each representing a particular ailment. After that, the network would be trained to predict all of the appropriate class labels for each input sample.

Since the network must be able to accommodate several and sometimes contradicting outputs, multi-label classification is more difficult than single-label classification. Many strategies have been proposed to handle this difficulty, including the use of several binary classifiers, one for each label, or the use of a single classifier with numerous outputs. The approach chosen will be determined by the task's unique needs and the data being utilised.

3.4 ROC Interpretation

Receiver Operating Characteristic (ROC) analysis is a popular way of assessment in machine learning, particularly in medical picture analysis. The performance of a binary classification system is visualized via ROC analysis. The approach shows the true positive rate (sensitivity) vs the false positive rate (1-specificity) for a set of threshold values used to categorize data points into one of two groups. The ROC curve that results describes the trade-off between the true positive rate and the false positive rate for all potential thresholds, and the area under the curve (AUC) offers a single value that represents the classifier's overall performance.

ROC analysis is used in single-label classification to evaluate the performance of a single binary classifier. The ROC curve depicts how the classifier performs across a wide range of threshold values. A good classifier should have a high true positive rate (high sensitivity) and a low false positive rate (i.e. a high specificity).

ROC analysis is used in multi-label classification to evaluate the effectiveness of a multi-label classifier, where each data point might belong to numerous categories. In this case, the ROC curve summarizes the classifier's performance for each category, allowing for comparison of performance across categories.

ROC analysis is frequently utilized in medical image analysis because it is a useful tool for assessing the efficacy of machine learning algorithms in a variety of applications. The ability to evaluate the performance of different classifiers and illustrate the trade-off between true positive

and false positive rates is very useful in this discipline, where the results might have serious consequences for patient care.

ROC analysis is widely used in medical image analysis to assess machine learning algorithm efficacy in different applications. It is a valuable tool to evaluate classifiers and illustrates the trade-off between true positive and false positive rates, which has significant implications for patient care.

3.5 Methods of Architecture Interpretation and Explainability

Interpretability and explainability are crucial in machine learning, especially in medical applications where accurate predictions are vital. Understanding the underlying processes of treatment response is essential in schizophrenia studies to improve patient outcomes.

Deep neural networks can be interpreted and explained using techniques such as feature visualization, saliency maps, and layer-wise relevance propagation (LRP). Feature visualization displays neuron activity in different network levels to identify important properties for prediction. Saliency maps highlight significant areas of input images for specific predictions.

LRP is a backpropagation approach that distributes prediction relevance to input characteristics, providing a comprehensive explanation of the network's decision.

In the context of schizophrenia research, feature visualization and saliency maps can identify important brain areas for treatment response prediction. LRP can distribute prediction relevance to specific characteristics in functional imaging data, offering a precise understanding of the underlying processes influencing treatment response.

Strategies for enhancing interpretability include using inherently interpretable models like decision trees or incorporating interpretability during the learning process through regularization techniques or domain knowledge.

3.5 Methods of architecture interpretation and explainability

With the usage of machine learning algorithms, interpretation and explainability are significant factors, particularly in the medical area where correct predictions are vital. Understanding the underlying processes that lead to treatment response is critical in the context of the schizophrenia study, since this information may be utilized to enhance patient outcomes.

Deep neural networks may be interpreted and explained using a variety of ways, including feature visualization, saliency maps, and layer-wise relevance propagation. To determine which properties are most essential for prediction, feature visualization entails displaying the activity of neurons in different levels of the network. Similar to saliency maps, saliency maps emphasize the most significant areas of the input picture for a certain prediction.

Layer-wise relevance propagation (LRP) is a backpropagation approach that distributes the relevance of a prediction to the input characteristics, offering a more full explanation of the network's conclusion.

In the context of the schizophrenia investigation, feature visualization and saliency maps might be utilized to determine which brain areas are most important for predicting treatment response. LRP might be used to distribute the prediction's relevance to particular characteristics in

functional imaging data, providing for a more precise understanding of the underlying processes that influence treatment response.

There are several strategies for making deep neural networks more interpretable, such as building designs that are intrinsically interpretable, such as decision trees, or adding interpretability into the learning process, such as through regularization techniques or domain knowledge.

Finally, the adoption of interpretable machine learning algorithms can enhance prediction accuracy and raise trust in the findings, which is important in the medical area where correct predictions are key for patient outcomes.

3.6 : Data Analysis

In this chapter, we will review the data analysis methodologies employed in our study. The purpose of the data analysis is to look at the link between cortical function and shape and treatment response in schizophrenia patients. To do this, we will clean and denoise the functional imaging data before deconfounding unwanted factors. We will then do feature extraction and feature selection to determine the most relevant characteristics for predicting treatment response.

One of the most difficult difficulties in functional imaging data processing is cleaning and denoising the data to remove any artifacts or noise. In this study, we will clean and denoise the functional MRI data using machine learning methods, especially deep neural networks. This method has been proved to be effective in reducing artifacts and noise while keeping the data's significant signals.

We will deconfound nuisance variables after cleaning and denoising the data. Nuisance variables are elements that can alter imaging results but are irrelevant to our investigation. For example, during the imaging capture procedure, the head movement might inject signals into the data that are unrelated to the brain activity of interest. We will employ techniques such as regression-based approaches and independent component analysis to reduce the influence of nuisance variables.

After cleaning, denoising, and deconfounding the data, we will extract and choose characteristics for analysis. Feature extraction is the process of extracting new features from raw data, whereas feature selection is the process of choosing the subset of features that is most informative for the job at hand. In this work, we will extract features from cortical surface models generated from

Adopting interpretable machine learning algorithms can improve prediction accuracy and trust in findings, which is crucial in medical applications where accurate predictions are essential for patient outcomes.

3.6 Data Analysis

This chapter presents the data analysis methods used in our study, aimed at investigating the relationship between cortical function and shape and treatment response in schizophrenia patients. We will start by cleaning and denoising the functional imaging data using machine learning methods, particularly deep neural networks, to remove artifacts and noise while preserving significant signals. Then, we will deconfound nuisance variables using regression-based approaches and independent component analysis to eliminate irrelevant factors that may affect our investigation.

Next, we will perform feature extraction and selection to identify the most relevant characteristics for predicting treatment response, using cortical surface models generated from imaging data. We will utilize supervised machine learning methods like logistic regression, decision trees, and random forests to develop prediction models and assess their performance using accuracy, precision, recall, F1 score, and statistical methods such as t-tests and ANOVA.

The first line of code imports Python libraries for data analysis and manipulation (Pandas), numerical computations (Numpy), and data visualization (Matplotlib.pyplot). The Scikit-learn package is also imported to access the feature selection function (sklearn.feature selection.r regression).

- **Dataset loading**

The dataset is imported into a Pandas DataFrame using Pd.read csv(). The ROI was calculated by averaging the characteristics. The csv file is located in /content/drive/MyDrive.

imaging data and then utilize feature selection approaches such as recursive feature elimination to pick the most relevant features for predicting treatment response.

Following feature selection, we will use supervised machine learning methods such as logistic regression, decision trees, and random forests to develop prediction models.

We will assess the models' performance using measures like accuracy, precision, recall, and F1 score. We will also analyze the models' performance using statistical methods such as t-tests and ANOVA to see if there are any significant variations in performance between them.

Getting and Processing Data Importing Libraries

The first line of code imports the Python libraries listed below:

Pandas is a data analysis and manipulation package.

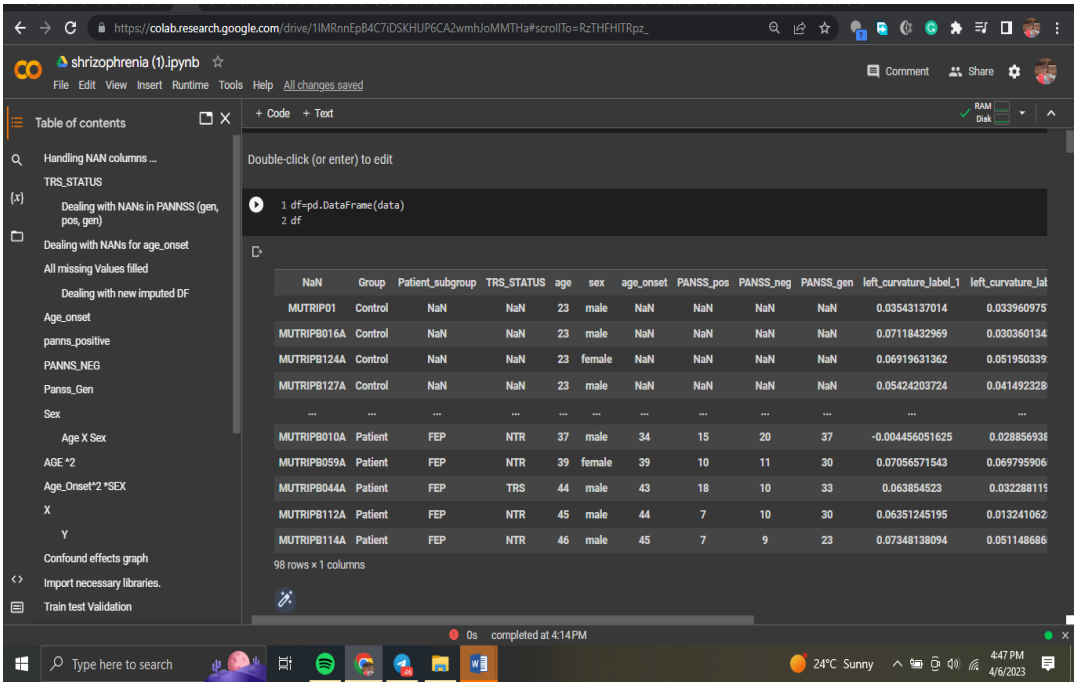
Numpy is a library for numerical computations.

Matplotlib.pyplot is a data visualization package.

The Scikit-learn package includes a feature selection function named sklearn.feature selection.r regression.

Dataset loading

The dataset is imported into a Pandas DataFrame using Pd.read csv(). The ROI was calculated by averaging the characteristics. The csv file is located in /content/drive/MyDrive.



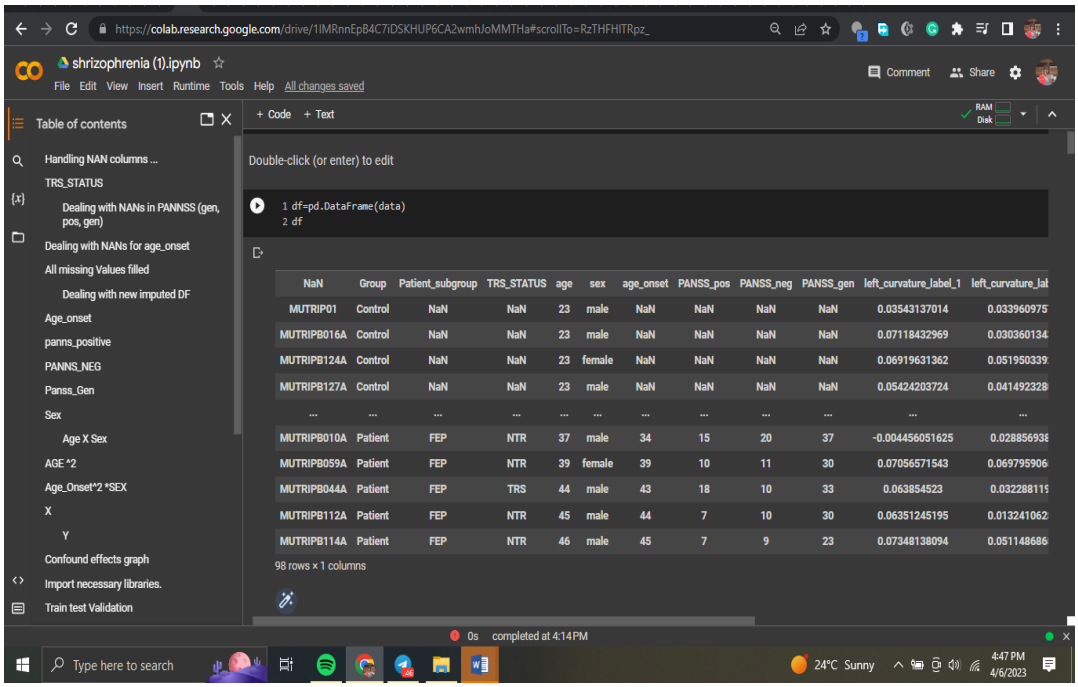
NaN	Group	Patient_subgroup	TRS_STATUS	age	sex	age_onset	PANSS_pos	PANSS_neg	PANSS_gen	left_curvature_label_1	left_curvature_lat
MUTRIP01	Control	NaN	NaN	23	male	NaN	NaN	NaN	NaN	0.03543137014	0.033960975
MUTRIPB016A	Control	NaN	NaN	23	male	NaN	NaN	NaN	NaN	0.07118432969	0.030360134
MUTRIPB124A	Control	NaN	NaN	23	female	NaN	NaN	NaN	NaN	0.06919631362	0.051950339
MUTRIPB127A	Control	NaN	NaN	23	male	NaN	NaN	NaN	NaN	0.05424203724	0.041492328
...	...	...	...	...	...	...	...	...	...	...	...
MUTRIPB010A	Patient	FEP	NTR	37	male	34	15	20	37	-0.004456051625	0.028856938
MUTRIPB059A	Patient	FEP	NTR	39	female	39	10	11	30	0.07056571543	0.069795906
MUTRIPB044A	Patient	FEP	TRS	44	male	43	18	10	33	0.063854523	0.032288119
MUTRIPB112A	Patient	FEP	NTR	45	male	44	7	10	30	0.06351245195	0.013241062
MUTRIPB114A	Patient	FEP	NTR	46	male	45	7	9	23	0.07348138094	0.051148686

Examining the Dataset

The df.info() method displays details about the DataFrame, such as the number of rows, columns, and data types. The DataFrame has a multi-level index with 98 entries. The columns provide clinical and demographic information on the patients, such as age, gender, and PANSS scores, as well as other brain anatomical parameters such as curvature, sulc, and thickness. It's also worth noting that the dataframe's form was 97, 1 before the multi-index was deleted.

Feature Selection

Scikit-feature learn selection module use the r regression function for feature selection. This function performs univariate linear regression tests to examine the strength of the correlation between each attribute and the target variable. Although this code does not specify the target



- **Examining the Dataset**

The df.info() method displays details about the DataFrame, such as the number of rows, columns, and data types. The DataFrame has a multi-level index with 98 entries. The columns provide clinical and demographic information on the patients, such as age, gender, and PANSS scores, as well as other brain anatomical parameters such as curvature, sulc, and thickness.

It's also worth noting that the dataframe's form was 97, 1 before the multi-index was deleted.

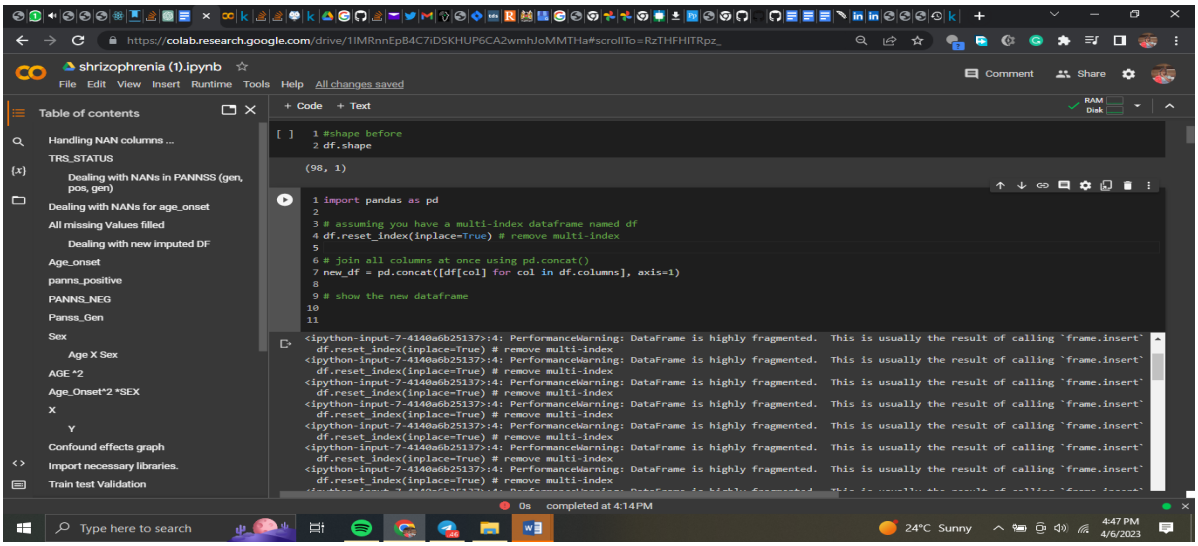
- **Feature Selection**

Scikit-feature learn selection module use the r regression function for feature selection. This function performs univariate linear regression tests to examine the strength of the correlation between each attribute and the target variable. Although this code does not specify the target variable, it is most likely a measure of therapy response or symptom severity. Although the results of the feature selection technique are not explicitly presented in this code, they may be used to identify which features are most important to investigate further.

variable, it is most likely a measure of therapy response or symptom severity. Although the results of the feature selection technique are not explicitly presented in this code, they may be used to identify which features are most important to investigate further.

Data Processing and Index resetting

By executing the df.reset index(inplace=True) method, the DataFrame's index is reset to a single level. This is essential due to the original DataFrame's multi-level index, which may make some data operations difficult.

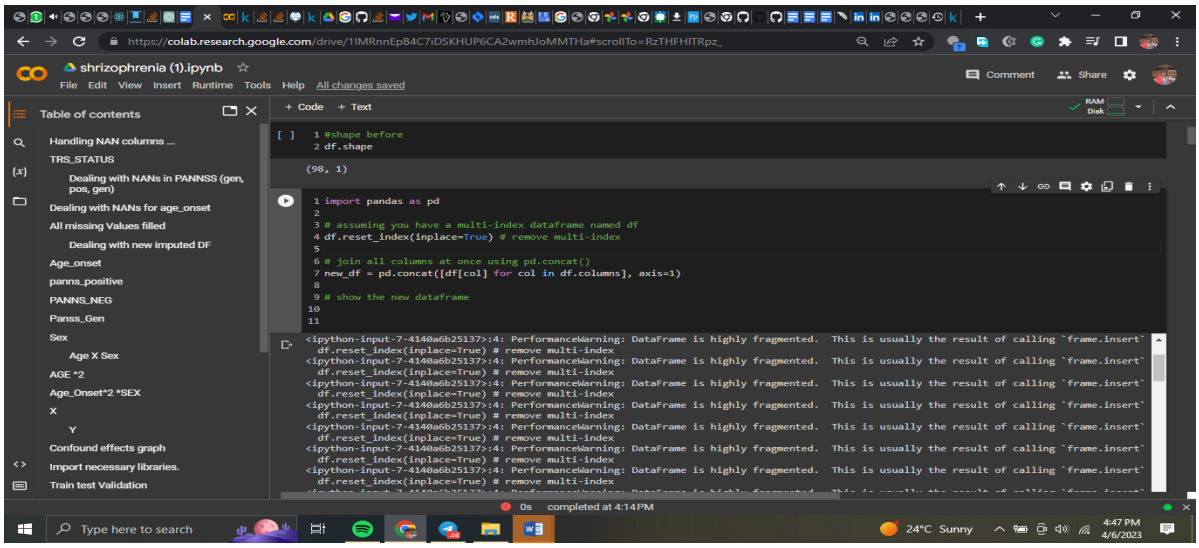


Putting the Columns Together

With the pd.concat() function, the columns of the DataFrame are concatenated into a new DataFrame named new df. This function (axis=1) concatenates the columns along the horizontal

- **Data Processing and Index resetting**

By executing the df.reset index(inplace=True) method, the DataFrame's index is reset to a single level. This is essential due to the original DataFrame's multi-level index, which may make some data operations difficult.



- **Putting the Columns Together**

With the pd.concat() function, the columns of the DataFrame are concatenated into a new DataFrame named new df. This function (axis=1) concatenates the columns along the horizontal axis (axis=1). The list of objects to concatenate is generated using a list comprehension that iterates over the columns of the original DataFrame df and supplied to the pd.concat() function. Concatenating all columns at once generates a DataFrame with a single level of columns that can be processed faster. The shape obtained by concatenating the columns is 97, 752.

- **Data Processing, Removing the first row**

The new df.drop(0) method is used to eliminate the first row of the DataFrame while keeping the original column names. Duplicate column names occur as a result of the pd.concat() function's failure to reset the column names after concatenating the columns.

axis (axis=1). The list of objects to concatenate is generated using a list comprehension that iterates over the columns of the original DataFrame df and supplied to the pd.concat() function. Concatenating all columns at once generates a DataFrame with a single level of columns that can be processed faster.

The shape obtained by concatenating the columns is 97, 752.

Data Processing

removing the first row

The new df.drop(0) method is used to eliminate the first row of the DataFrame while keeping the original column names. Duplicate column names occur as a result of the pd.concat() function's failure to reset the column names after concatenating the columns.

Inclusion of Values in the DataFrame

The new df.iloc[0, 0] = method is used to insert the value "id" into the DataFrame's first row and first column. With this method, the unique identity of each row in the dataset is added as a column header.

As a result of eliminating the first row and adding the 'id' column header, the new df DataFrame will have clear and distinct column names, making it easier to analyze and present the data.

Dropped Index Columns During Data Processing

- **Inclusion of Values in the DataFrame**

The new df.iloc[0, 0] = method is used to insert the value "id" into the DataFrame's first row and first column. With this method, the unique identity of each row in the dataset is added as a column header. As a result of eliminating the first row and adding the 'id' column header, the new df DataFrame will have clear and distinct column names, making it easier to analyze and present the data.

- **Dropped Index Columns during Data Processing**

The DataFrame most likely has index columns that are unsuitable for analysis or visualization. The new df function may be used to delete these columns by using new df.drop(columns=['index1', 'index2']), where 'index1' and 'index2' are the index columns' names.

- **Column Names**

By invoking the new df.columns = new df.iloc[0] method, the DataFrame's column names are set to the values in the first row. This is because the original column names were discarded in the previous step and were not retained throughout the concatenation stage.

The resulting new df DataFrame will have unique and descriptive column names, making data analysis and visualization easier. This is achieved by modifying the column names to correspond to the values in

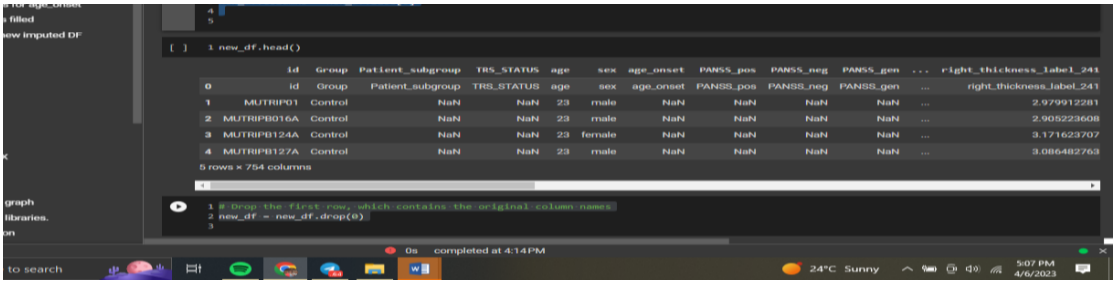
The DataFrame most likely has index columns that are unsuitable for analysis or visualization. The new df function may be used to delete these columns by using new df.drop(columns=['index1', 'index2']), where 'index1' and 'index2' are the index columns' names.

Column Names

By invoking the new df.columns = new df.iloc[0] method, the DataFrame's column names are set to the values in the first row. This is because the original column names were discarded in the previous step and were not retained throughout the concatenation stage.

The resulting new df DataFrame will have unique and descriptive column names, making data analysis and visualization easier. This is achieved by modifying the column names to correspond to the values in the first row.

This is the new dataset following preparation.



Handling NaN Values

How to Identify NaN Columns

The output of df.info is then analyzed to find the columns with NaN values (). In this case, the columns 'TRS STATUS,' 'age onset,' 'PANSS pos,' 'PANSS neg,' 'PANSS gen,' and 'Patient subgroup' had NaN values.

Substituting NaN Values

The fillna() method is used to replace the word "healthy" in the "Patient subgroup" column with NaN values. This is done because NaN values in this column are most likely an indicator that the patient does not fit into any of the categories and is thus healthy.

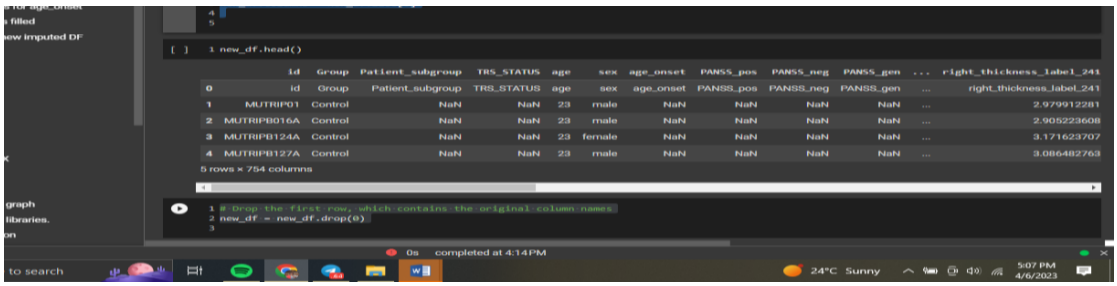
The fillna() method is used to replace "TRS STATUS" column NaN values with the wording "Not diagnosed." This is done because NaN values in this column most likely indicate that the patient has not yet been diagnosed with TRS status.

The addition of NaN values

For the "PANSS pos," "PANSS neg," and "PANSS gen" columns, missing values are imputed using the KNNImputer class from sklearn.impute. This is done because the numerical values in these columns are likely to be related to those in other columns.

When n neighbors=5 is the default configuration, the values of the KNNImputer's five nearest neighbors are utilized to impute the missing values. The fit transform() function is then used to the selected columns using the KNNImputer object to impute the missing values. The original DataFrame's columns are then given the imputed values again.

the first row. This is the new dataset following preparation



- **Handling NaN Values, How to Identify NaN Columns**

The output of df.info is then analyzed to find the columns with NaN values (). In this case, the columns 'TRS STATUS,' 'age onset,' 'PANSS pos,' 'PANSS neg,' 'PANSS gen,' and 'Patient subgroup' had NaN values.

- **Substituting NaN Values**

The fillna() method is used to replace the word "healthy" in the "Patient subgroup" column with NaN values. This is done because NaN values in this column are most likely an indicator that the patient does not fit into any of the categories and is thus healthy.

The fillna() method is used to replace "TRS STATUS" column NaN values with the wording "Not diagnosed." This is done because NaN values in this column most likely indicate that the patient has not yet been diagnosed with TRS status.

Dealing with NaNs in age onset

Based on the "sex" column, a new column named "is male" is created, with a value of one denoting a man and a value of zero denoting a woman.

The data frame is then subdivided into a subset with only numerical columns.

The corr() function generates a correlation matrix for the data frame's numeric columns.

It generates the correlation matrix.

According to the correlation matrix, the predictors having the highest association with 'age onset' are ['PANSS pos', 'PANSS neg', 'PANSS gen', and 'is male'].

Using the isna() and dropna() functions, the data frame is separated into two subsets: one with missing 'age onset' values and the other without any missing values.

A regression model is generated using the LinearRegression() function from the sklearn.linear model package, using the non-missing 'age onset' values as the outcome and the discovered predictors as the predictors.

The regression model is used to anticipate the missing 'age onset' values using the predict() function.

Using the loc[] function, the original DataFrame is modified by inserting the predicted values for the missing 'age onset' values.

To merge the two subgroups into a single data frame, use the concat() technique.

It should be emphasized that this technique assumes that missing 'age onset' values are missing at random and that 'age onset' and the derived predictors have a linear relationship.

There are no longer any NaN values.

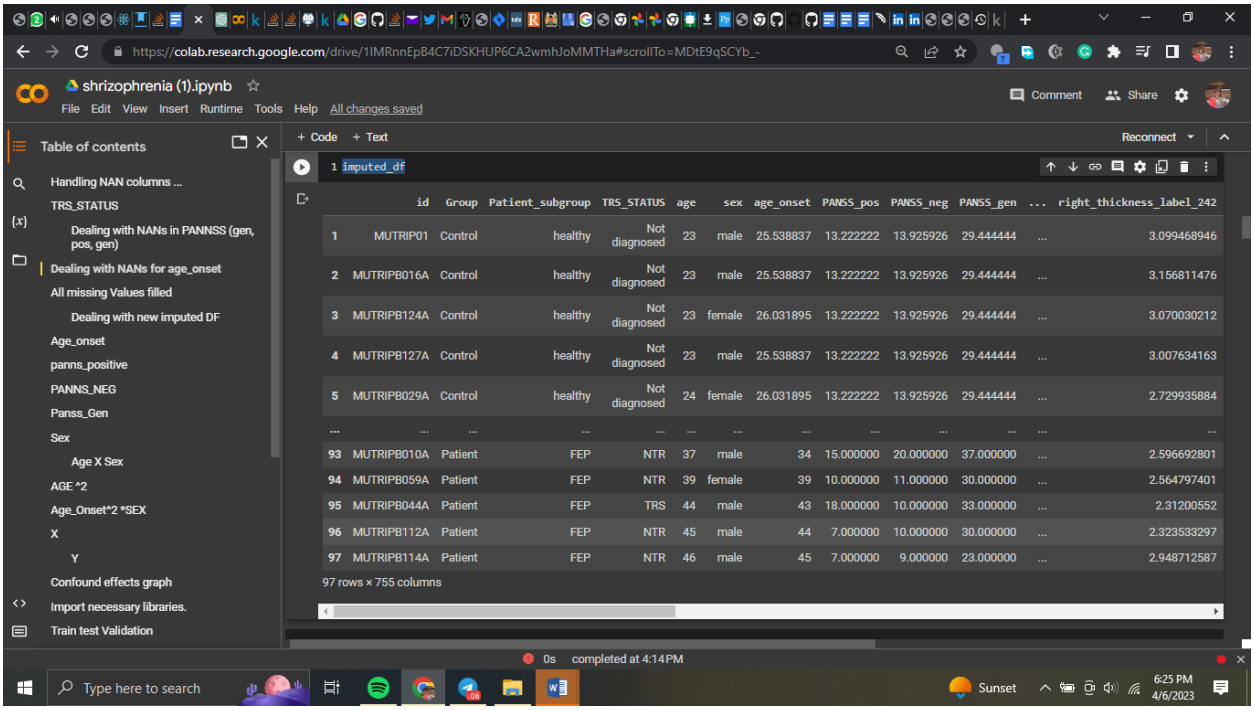
- **The addition of NaN values**

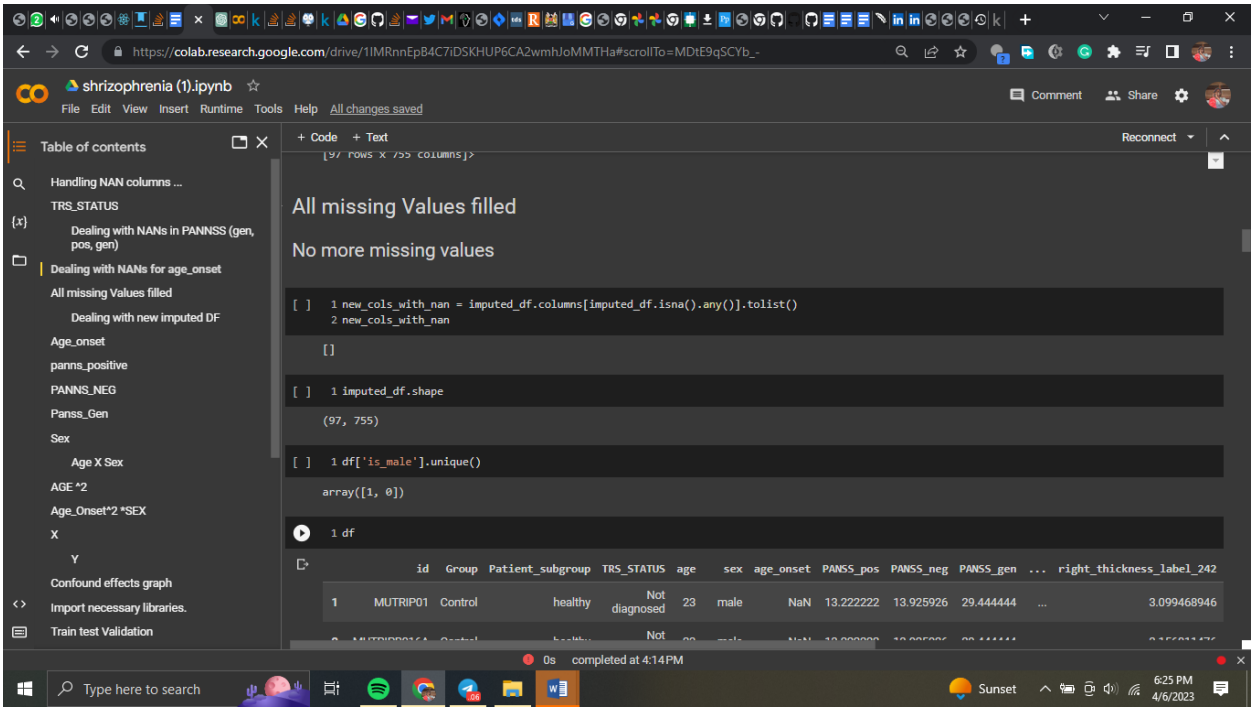
For the "PANSS pos," "PANSS neg," and "PANSS gen" columns, missing values are imputed using the KNNImputer class from sklearn.impute. This is done because the numerical values in these columns are likely to be related to those in other columns.

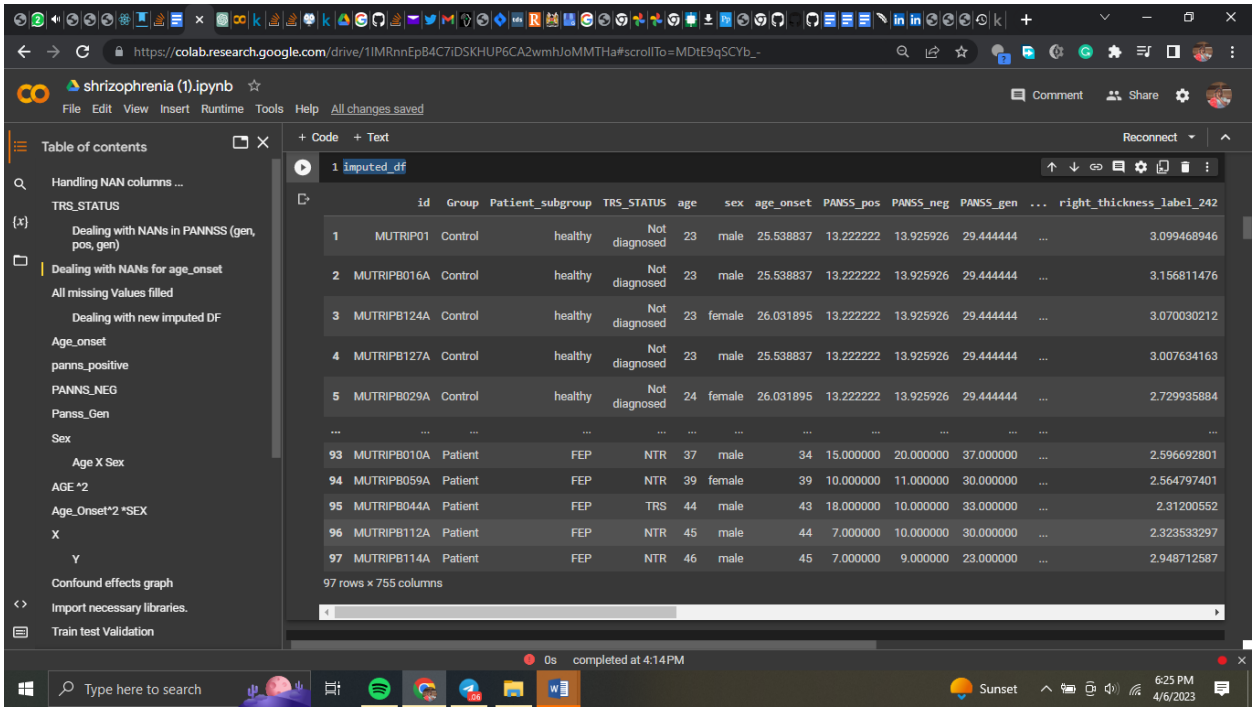
When n neighbors=5 is the default configuration, the values of the KNNImputer's five nearest neighbors are utilized to impute the missing values. The fit transform() function is then used to the selected columns using the KNNImputer object to impute the missing values. The original DataFrame's columns are then given the imputed values again.

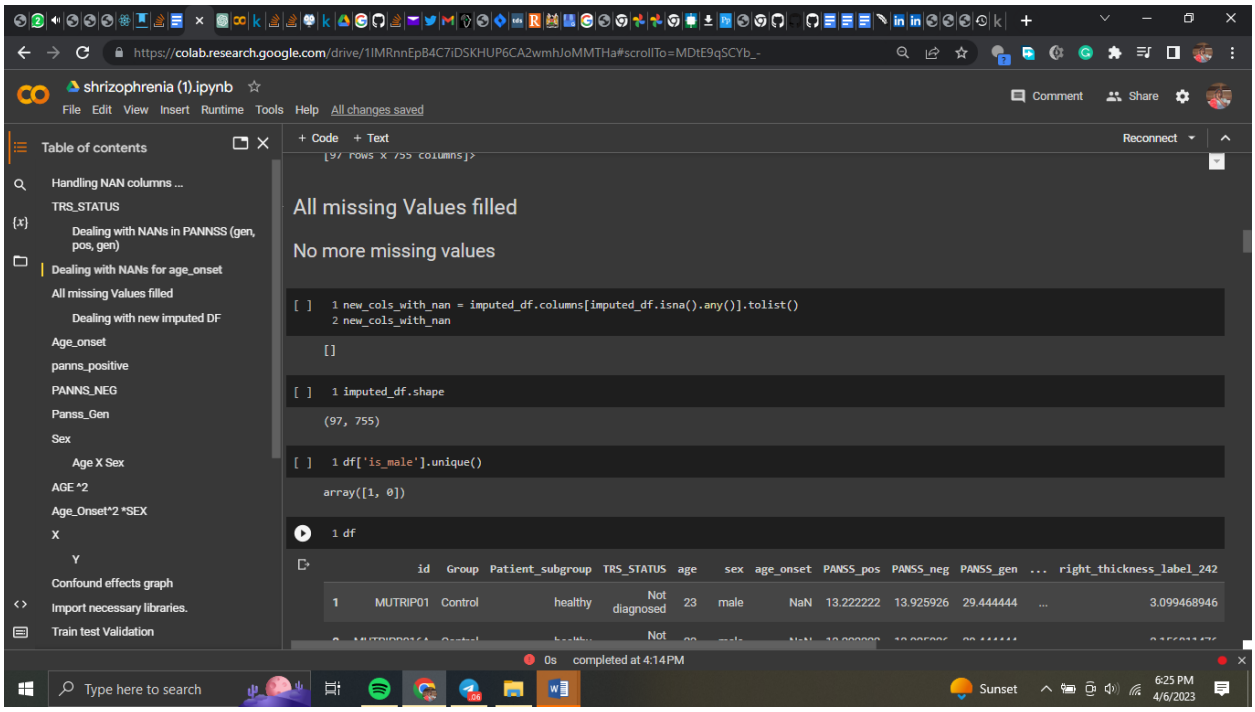
- **Dealing with NaNs in age onset**

To handle NaNs in 'age onset', a new column named "is male" is created based on the 'sex' column. The data frame is then divided into a subset with only numerical columns, and a correlation matrix is generated using the corr() function. The highest associated predictors with 'age onset' are identified as ['PANSS pos', 'PANSS neg', 'PANSS gen', and 'is male']. The data frame is separated into two subsets using isna() and dropna() functions - one with missing 'age onset' values and the other without any missing values. A regression model is created with the LinearRegression() function, using non-missing 'age onset' values as the outcome and the discovered predictors as the predictors. The missing 'age onset' values are anticipated using the predict() function, and the original DataFrame is updated with the predicted values using the loc[] function. The two subgroups are merged into a single data frame using the concat() technique. This technique assumes that missing 'age onset' values are missing at random and that 'age onset' and derived predictors have a linear relationship.


shrizophrenia (1).ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment Share
Table of contents
+ Code + Text
Reconnect
1 imputed_df
id Group Patient_subgroup TRS_STATUS age sex age_onset PANSS_pos PANSS_neg PANSS_gen ... right_thickness_label_242
1 MUTRIP01 Control healthy Not diagnosed 23 male 25.538837 13.222222 13.925926 29.444444 ... 3.099468946
2 MUTRIPB016A Control healthy Not diagnosed 23 male 25.538837 13.222222 13.925926 29.444444 ... 3.156811476
3 MUTRIPB124A Control healthy Not diagnosed 23 female 26.031895 13.222222 13.925926 29.444444 ... 3.070030212
4 MUTRIPB127A Control healthy Not diagnosed 23 male 25.538837 13.222222 13.925926 29.444444 ... 3.007634163
5 MUTRIPB029A Control healthy Not diagnosed 24 female 26.031895 13.222222 13.925926 29.444444 ... 2.729935884
93 MUTRIPB010A Patient FEP NTR 37 male 34 15.000000 20.000000 37.000000 ... 2.596692801
94 MUTRIPB059A Patient FEP NTR 39 female 39 10.000000 11.000000 30.000000 ... 2.564797401
95 MUTRIPB044A Patient FEP TRS 44 male 43 18.000000 10.000000 33.000000 ... 2.31200552
96 MUTRIPB112A Patient FEP NTR 45 male 44 7.000000 10.000000 30.000000 ... 2.323533297
97 MUTRIPB114A Patient FEP NTR 46 male 45 7.000000 9.000000 23.000000 ... 2.948712587
97 rows × 755 columns
Handling NAN columns ...
TRS_STATUS
Dealing with NANs in PANNSS (gen, pos, gen)
Dealing with NANs for age_onset
All missing Values filled
Dealing with new imputed DF
Age_onset
panns_positive
PANNS_NEG
Panss_Gen
Sex
Age X Sex
AGE ^2
Age_Onset^2 *SEX
X
Y
Confound effects graph
Import necessary libraries.
Train test Validation
0s completed at 4:14PM
Type here to search
Sunset
6:25 PM
4/6/2023


shrizophrenia (1).ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment Share
Table of contents
+ Code + Text
Reconnect
All missing Values filled
No more missing values
1 new_cols_with_nan = imputed_df.columns[imputed_df.isna().any()].tolist()
2 new_cols_with_nan
[]
1 imputed_df.shape
(97, 755)
1 df['is_male'].unique()
array([1, 0])
1 df
id Group Patient_subgroup TRS_STATUS age sex age_onset PANSS_pos PANSS_neg PANSS_gen ... right_thickness_label_242
1 MUTRIP01 Control healthy Not diagnosed 23 male NaN 13.222222 13.925926 29.444444 ... 3.099468946
Handling NAN columns ...
TRS_STATUS
Dealing with NANs in PANNSS (gen, pos, gen)
Dealing with NANs for age_onset
All missing Values filled
Dealing with new imputed DF
Age_onset
panns_positive
PANNS_NEG
Panss_Gen
Sex
Age X Sex
AGE ^2
Age_Onset^2 *SEX
X
Y
Confound effects graph
Import necessary libraries.
Train test Validation
0s completed at 4:14PM
Type here to search
Sunset
6:25 PM
4/6/2023





- **How to Handle NANs Due to Age**

The purpose of the code is to apply a regression model to fill in missing values for the variable 'age onset' in a pandas DataFrame. First, the strongest predictors for

How to Handle NANs Due to Age

The goal of this code is to apply a regression model to fill in missing values for the variable 'age onset' in a pandas DataFrame. The first stage is to identify the predictors with the strongest association to "age onset." This is performed by creating a correlation matrix for the numerical columns of the DataFrame, which is then saved in the corr matrix variable.

Each participant's gender is then recorded as a binary variable in a new column named "is male," with 1 indicating male and 0 indicating female. This is accomplished using the following line of code:

```
df['is_male'] = (df['sex'] == 'male').astype(int)
```

The select dtypes() method is then used to generate a subset of the DataFrame with just numeric columns, which is then saved in the variable numeric df.

To print the correlation matrix, use the following code:

```
print(corr matrix)
```

As a consequence, a square DataFrame is produced, with the correlation coefficient between each pair of columns acting as the values and the numeric columns serving as both the row and column indices.

The system then identifies the factors that are highly related to "age onset." In this case, the predictors are "PANSS pos," "PANSS neg," "PANSS gen," and "is male." These are retained in the variable predictors.

"age onset" are identified by creating a correlation matrix for the numerical columns of the DataFrame. Each participant's gender is then recorded as a binary variable in a new column named "is male." This is accomplished using the following line of code:

```
df['is_male'] = (df['sex'] == 'male').astype(int)
```

The select dtypes() method is then used to generate a subset of the DataFrame with just numeric columns, which is then saved in the variable numeric df. To print the correlation matrix, use the following code:

```
print(corr matrix)
```

As a consequence, a square DataFrame is produced, with the correlation coefficient between each pair of columns acting as the values and the numeric columns serving as both the row and column indices.

The system then identifies the factors that are highly related to "age onset." In this case, the predictors are "PANSS pos," "PANSS neg," "PANSS gen," and "is male." These are retained in the variable predictors. The DataFrame is then separated into two subsets: one with missing values for "age onset" and one with present data for "age onset." To do this, use the following code:

```
missing age onset = df
[df['age onset'].isna()]
non missing age onset = df.dropna(subset=['age onset'])
```

To avoid a SettingWithCopyWarning, the copy() method is used to create a duplicate of the chosen rows if the column "age onset" is missing (NaN). The subset option specifies that only rows with "age onset" should be chosen, and the dropna() method selects rows with "age onset" that are not missing.

A regression model is then built utilizing the non-missing 'age onset' values as the outcome and the identified factors as predictors. Scikit-LinearRegression() Learn's class is used for this:

The DataFrame is then separated into two subsets: one with missing values for "age onset" and one with present data for "age onset." To do this, use the following code:

missing age onset = df

[df['age onset'].isna()]

non missing age onset = df.dropna(subset=['age onset'])

To avoid a SettingWithCopyWarning, the copy() method is used to create a duplicate of the chosen rows if the column "age onset" is missing (NaN). The subset option specifies that only rows with "age onset" should be chosen, and the dropna() method selects rows with "age onset" that are not missing.

A regression model is then built utilizing the non-missing 'age onset' values as the outcome and the identified factors as predictors. Scikit-LinearRegression() Learn's class is used for this:

import sklearn.linear model Reg model LinearRegression = LinearRegression (). non missing age onset[predictors], non missing age onset['age onset'] fit(non missing age onset['age onset'])

The final model is held by the variable reg model.

The regression model is then used to fill in the blanks for the missing 'age onset' values. To do this, the following code is used:

reg model.predict(missing age onset[predictors]) = missing age onset.loc[:, 'age onset']

The predict() function of the LinearRegression object is used to anticipate the missing "age onset" values for the subset of the DataFrame when "age onset" is absent.

The.loc accessor is used to assign the produced predictions to the missing age onset DataFrame slice's 'age onset' column. The existing DataFrame is adjusted as a result of this.

Finally, concat() is used to unite the two DataFrame portions into a single DataFrame:

df imputed = pd.concat ([missing age onset, non missing age onset])

import sklearn.linear model Reg model LinearRegression = LinearRegression (). non missing age onset[predictors], non missing age onset['age onset'] fit(non missing age onset['age onset'])

The final model is held by the variable reg model. The regression model is then used to fill in the blanks for the missing 'age onset' values. To do this, the following code is used:

reg model.predict(missing age onset[predictors]) = missing age onset.loc[:, 'age onset']

The predict() function of the LinearRegression object is used to anticipate the missing "age onset" values for the subset of the DataFrame when "age onset" is absent.

The.loc accessor is used to assign the produced predictions to the missing age onset DataFrame slice's 'age onset' column. The existing DataFrame is adjusted as a result of this.

Finally, concat() is used to unite the two DataFrame portions into a single DataFrame:

df imputed = pd.concat ([missing age onset, non missing age onset])

The generated DataFrame, imputed df, contains the original DataFrame as well as values for the missing field "age onset" that were substituted using a linear regression model.

- **Preparation of data**

The code first constructs a numpy array called age onset array, which is then filled with data from the 'age onset' column of a pandas DataFrame called df. This is accomplished using the DataFrame.values attribute and the np.array() function because the values in the age onset array are imported as strings, the next line utilizes the pd.to numeric() function to convert them to numbers. The errors='coerce' argument converts all non-numeric values to NaN. (Not a Number). The age onset array is then divided into five similarly sized sub-arrays using the np.array split() technique. These sub-arrays, which each include a sample of the data, are assigned to a variable (split1, split2, etc.). Lastly, using the np.save() function and a different file name, each sub-array is stored as a distinct file. This is done to make it simple to import the data and utilize it for future research. To prepare the data for

The generated DataFrame, imputed df, contains the original DataFrame as well as values for the missing field "age onset" that were substituted using a linear regression model.

Preparation of data

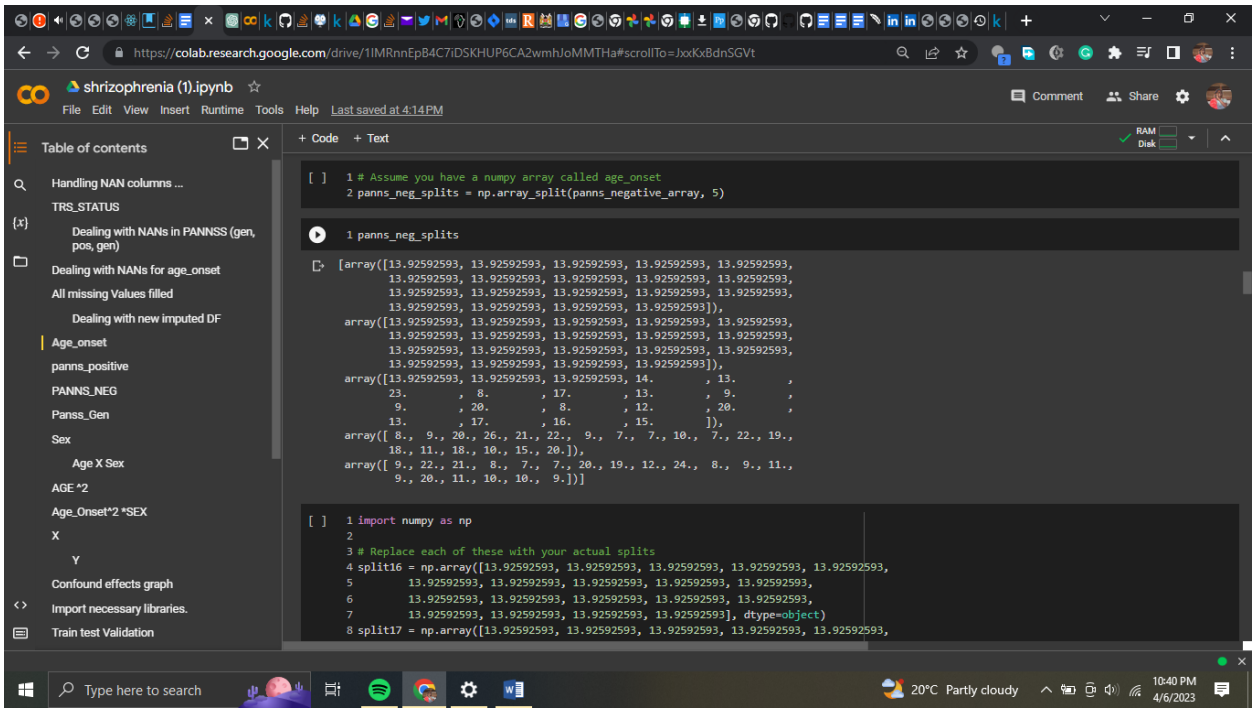
The code first constructs a numpy array called age onset array, which is then filled with data from the 'age onset' column of a pandas DataFrame called df. This is accomplished using the DataFrame.values attribute and the np.array() function.

Because the values in the age onset array are imported as strings, the next line utilizes the pd.to numeric() function to convert them to numbers. The errors='coerce' argument converts all non-numeric values to NaN. (Not a Number).

The age onset array is then divided into five similarly sized sub-arrays using the np.array split() technique. These sub-arrays, which each include a sample of the data, are assigned to a variable (split1, split2, etc.).

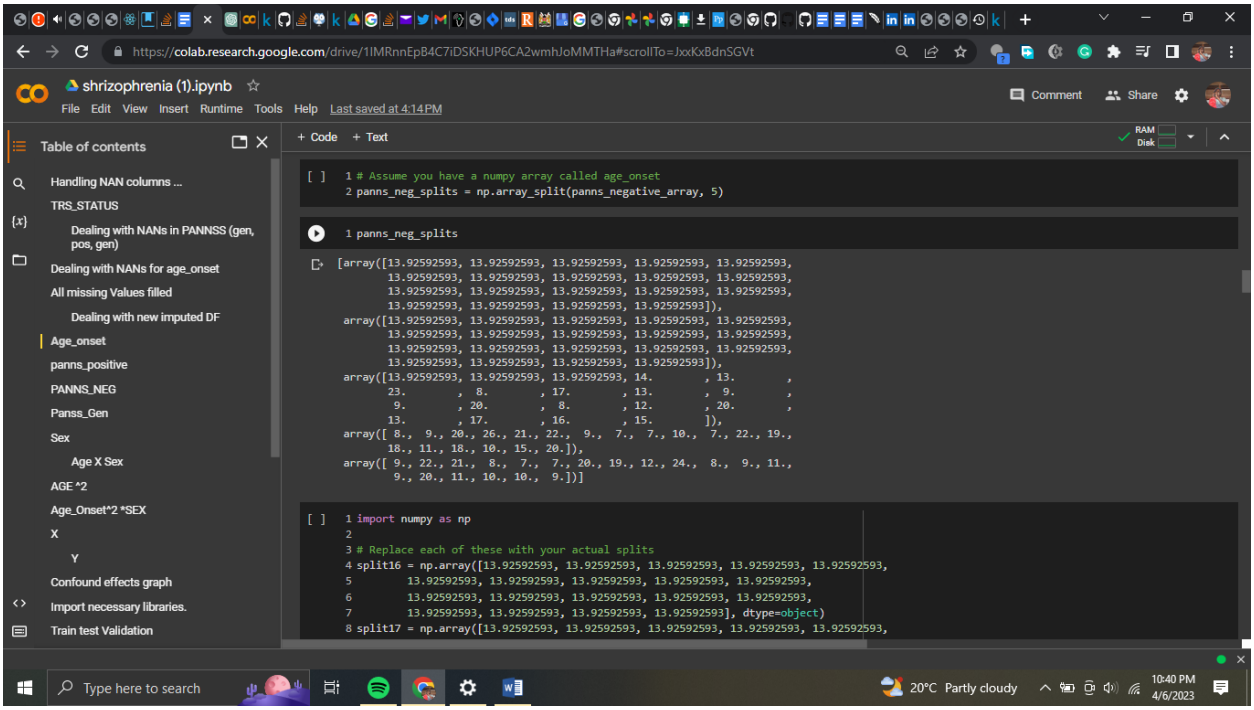
Lastly, using the np.save() function and a different file name, each sub-array is stored as a distinct file. This is done to make it simple to import the data and utilize it for future research. To prepare the data for future analysis, this procedure is done for the variables "panss neg," "panss gen," "panss pos," "age," "sex," "X," "y," "age squared," "age squared sex," and "age sex." By giving readily available subsets of the data, the data are divided into sub-arrays and saved as separate files to simplify future research.

future analysis, this procedure is done for the variables "panss neg," "panss gen," "panss pos," "age," "sex," "X," "y," "age squared," "age squared sex," and "age sex." By giving readily available subsets of the data, the data are divided into sub-arrays and saved as separate files to simplify future research.



- **Visualizations**

The heatmap displays the correlation coefficients between each pair of features (variables) in the dataset. The correlation coefficient measures the strength of the linear association between two variables, with values ranging from -1 (completely negative correlation) to 1 (perfectly positive correlation) and 0 (no correlation). The correlation coefficient values are presented on a scale of -1 to 1, with negative values representing negative correlations, positive values representing positive correlations, and 0 representing no link. The colours of the heatmap, which vary from blue (negative correlation) to red (positive correlation) and white (zero correlation), show the values of the correlation coefficients (no correlation). The numbers in each heatmap box reflect the correlation coefficient between the two variables corresponding to that row and column. A value of 0.256 in the box at the



Visualizations

The heatmap displays the correlation coefficients between each pair of features (variables) in the dataset. The correlation coefficient measures the strength of the linear association between two variables, with values ranging from -1 (completely negative correlation) to 1 (perfectly positive correlation) and 0 (no correlation).

The correlation coefficient values are presented on a scale of -1 to 1, with negative values representing negative correlations, positive values representing positive correlations, and 0 representing no link.

The colours of the heatmap, which vary from blue (negative correlation) to red (positive correlation) and white (zero correlation), show the values of the correlation coefficients (no correlation).

The numbers in each heatmap box reflect the correlation coefficient between the two variables corresponding to that row and column. A value of 0.256 in the box at the intersection of the variables "age onset" and "PANSS pos," for example, indicates a weakly positive correlation between the age at which schizophrenia first emerges and the positive symptoms of schizophrenia as measured by the PANSS scale.

Age onset scatter plot PANNS gen

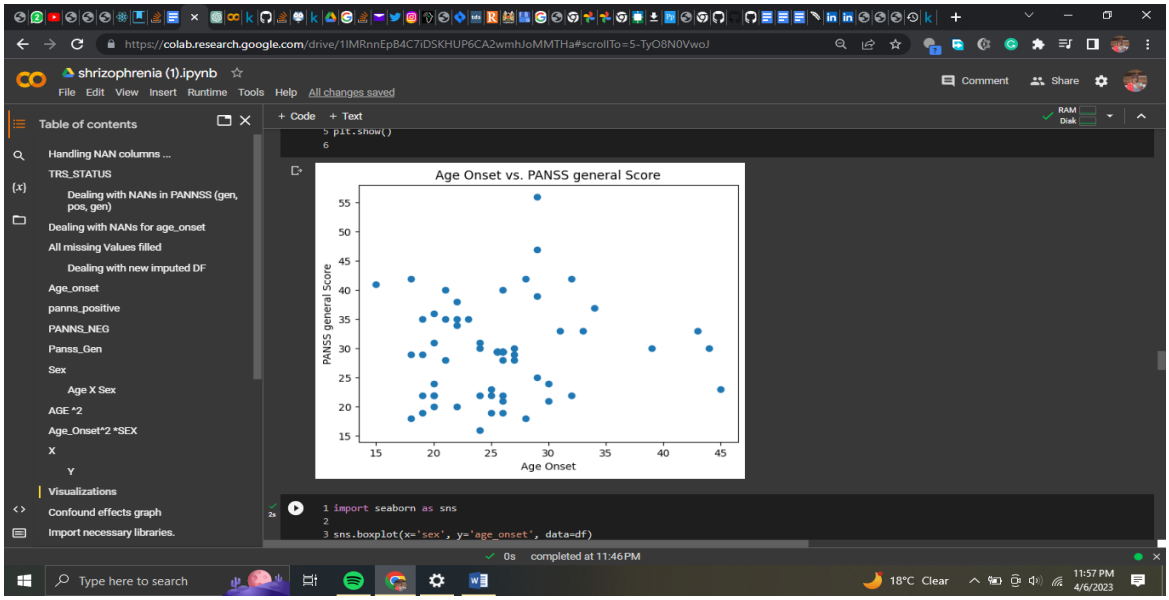
The scatter plot between age of onset and PANSS general score may show any potential relationship between these two parameters. If there is a positive connection, the data points should be distributed in a diagonal line extending from the bottom left to the top right of the plot. If there is a negative connection, the data points would be distributed in a diagonal line from top left to bottom right.

As we can see, there does not appear to be a substantial association in this case between age of onset and PANSS general score. Instead, the data points are spread and fail to create a discernible diagonal line in any direction. Nevertheless, when analyzing the scatter plot, we must also consider other factors such as outliers, data point distribution, and sample size.

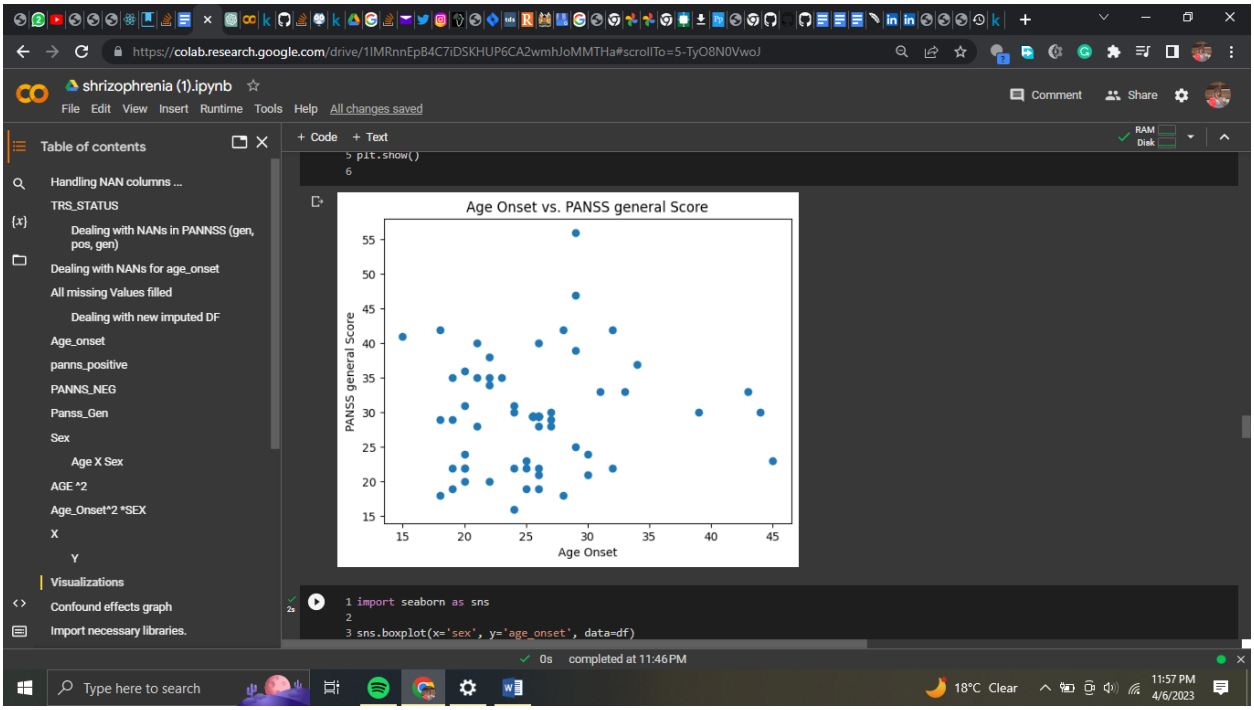
intersection of the variables "age onset" and "PANSS pos," for example, indicates a weakly positive correlation between the age at which schizophrenia first emerges and the positive symptoms of schizophrenia as measured by the PANSS scale.

- **Age onset scatter plot PANNS gen**

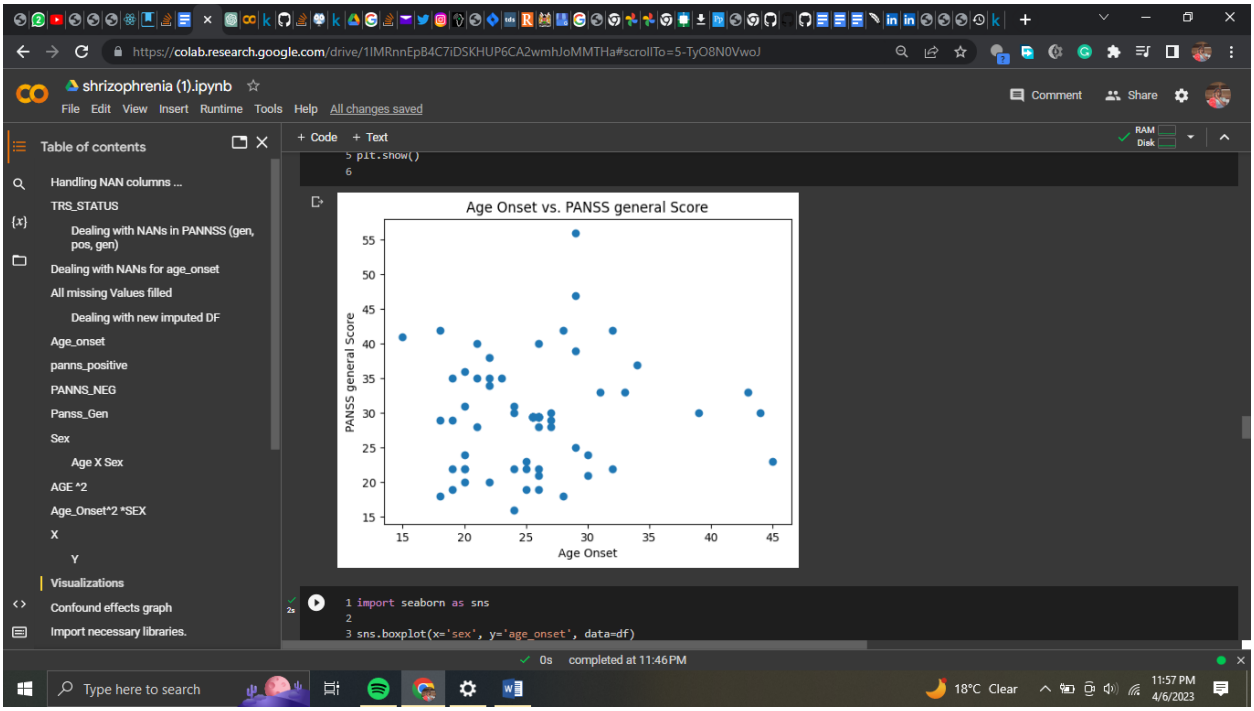
The scatter plot between age of onset and PANSS general score may show any potential relationship between these two parameters. If there is a positive connection, the data points should be distributed in a diagonal line extending from the bottom left to the top right of the plot. If there is a negative connection, the data points would be distributed in a diagonal line from top left to bottom right. As we can see, there does not appear to be a substantial association in this case between age of onset and PANSS general score. Instead, the data points are spread and fail to create a discernible diagonal line in any direction. Nevertheless, when analyzing the scatter plot, we must also consider other factors such as outliers, data point distribution, and sample size.



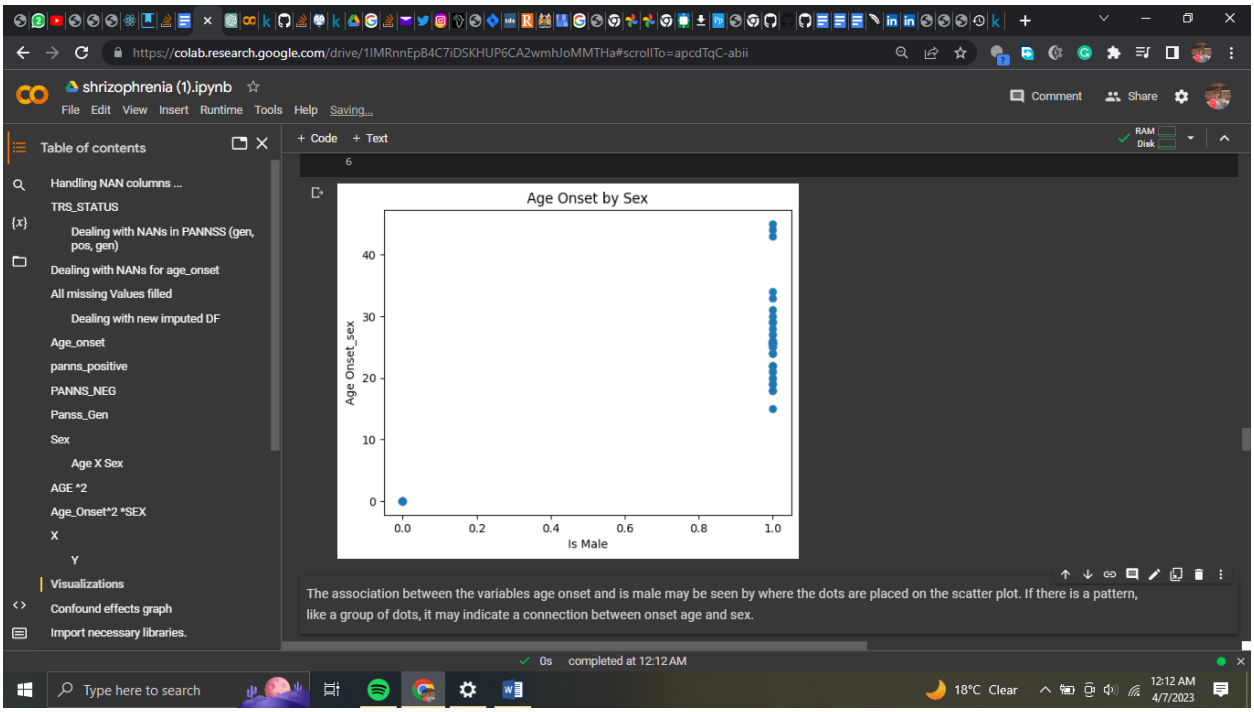
Age_onset PANNS positive was the same



Age_onset PANNS positive was the same

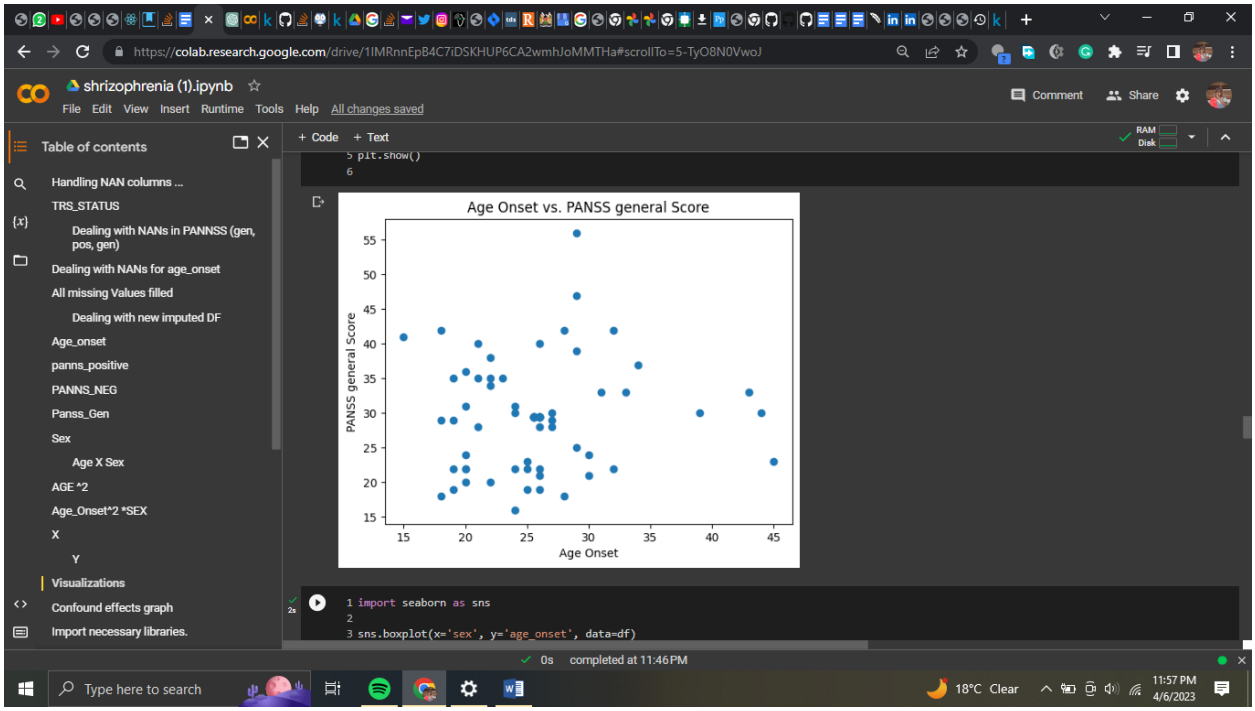


Age-onset_sex and gender

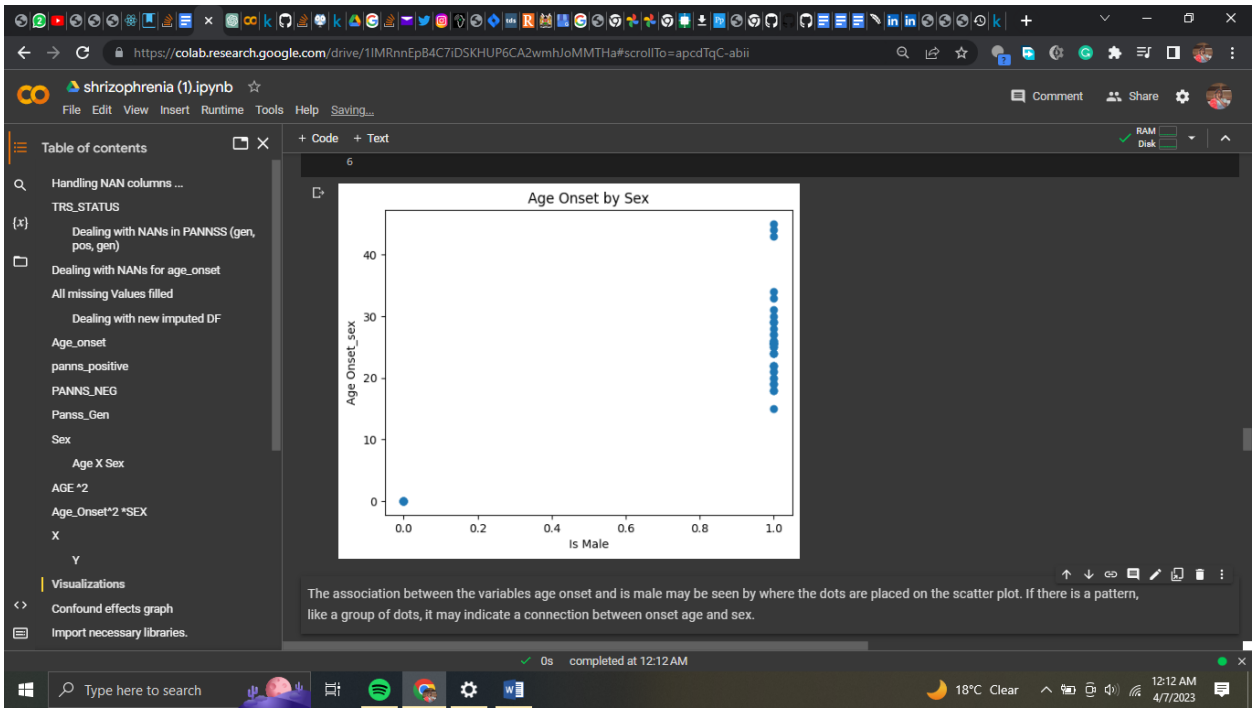


The location of the dots on the scatter plot indicates the relationship between the variables age onset and male. If there is a pattern, such as a cluster of dots, it may imply a link between onset age and gender.

Deconfounding effects investigation: Using linear regression to remove confounding influences.

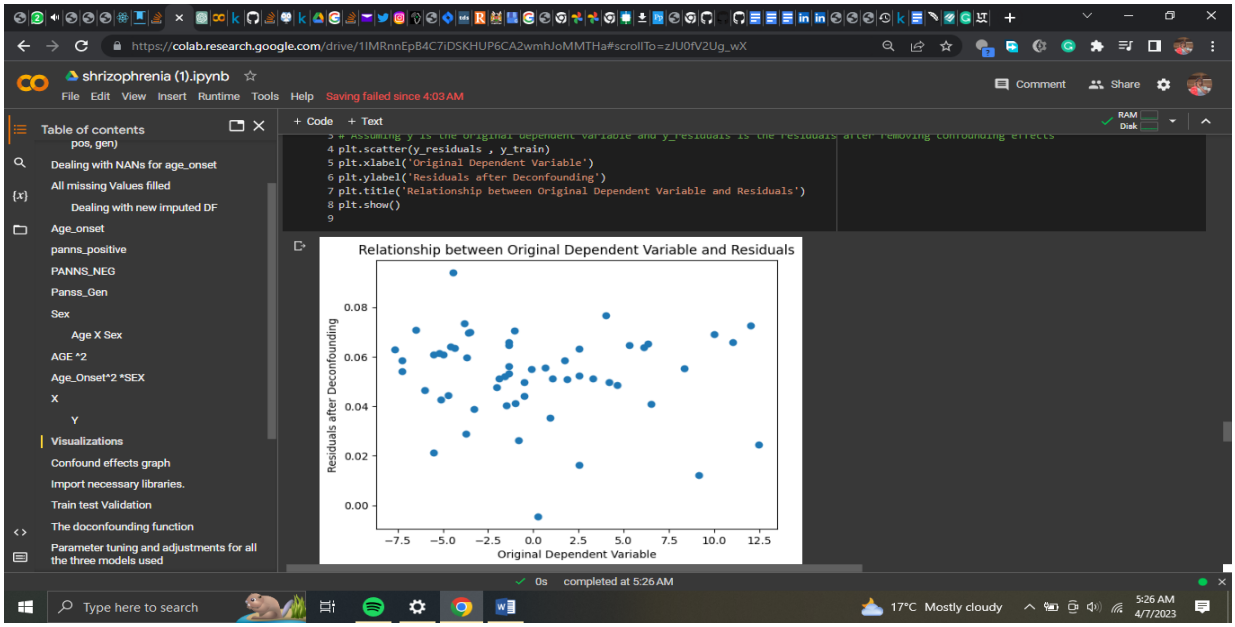


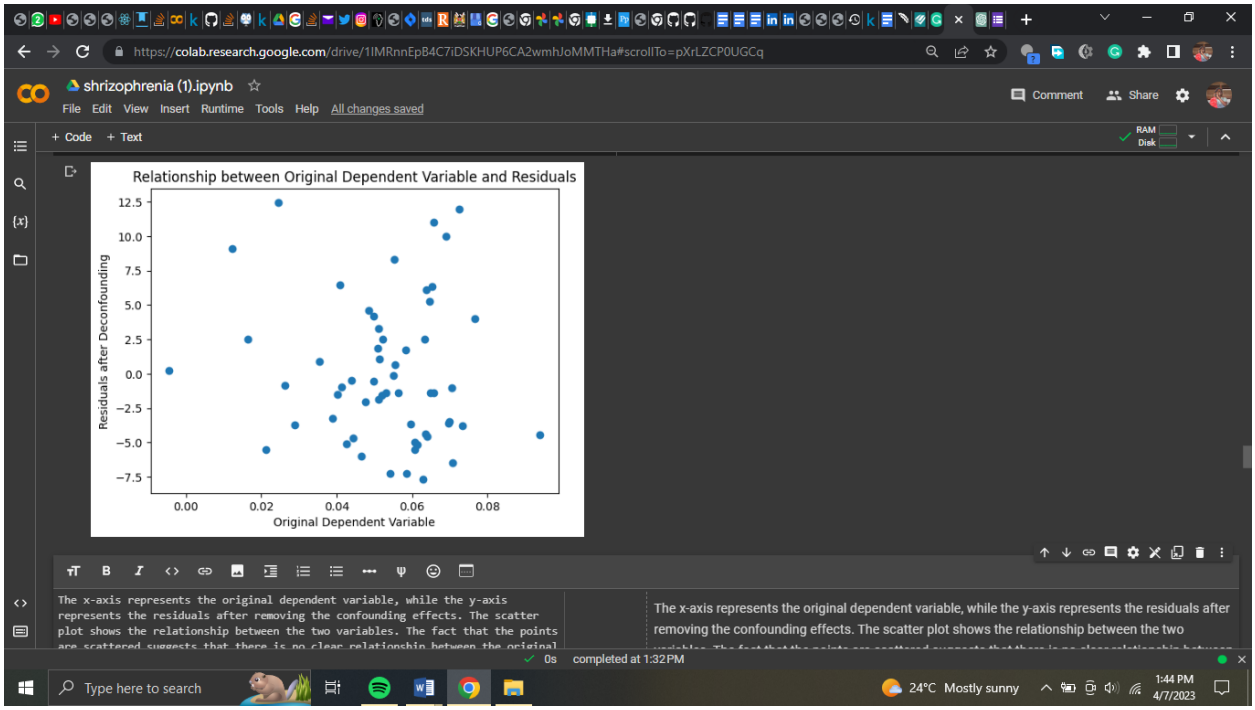
Age-onset_sex and gender

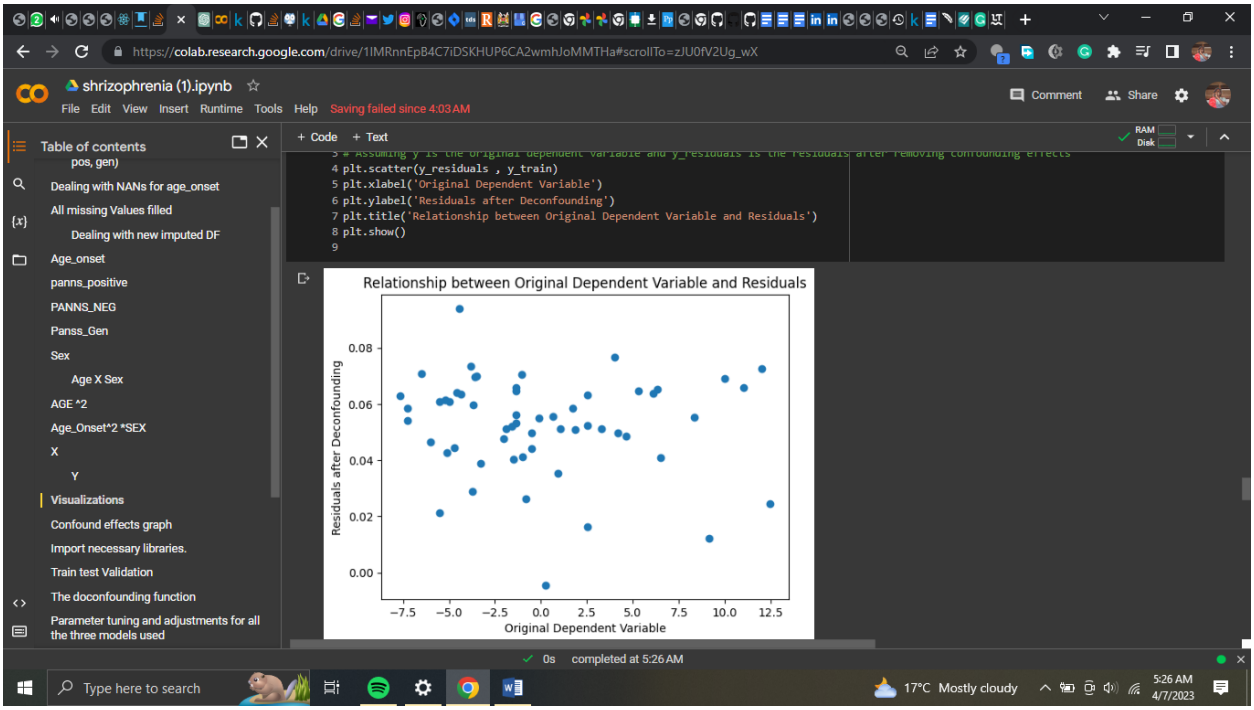


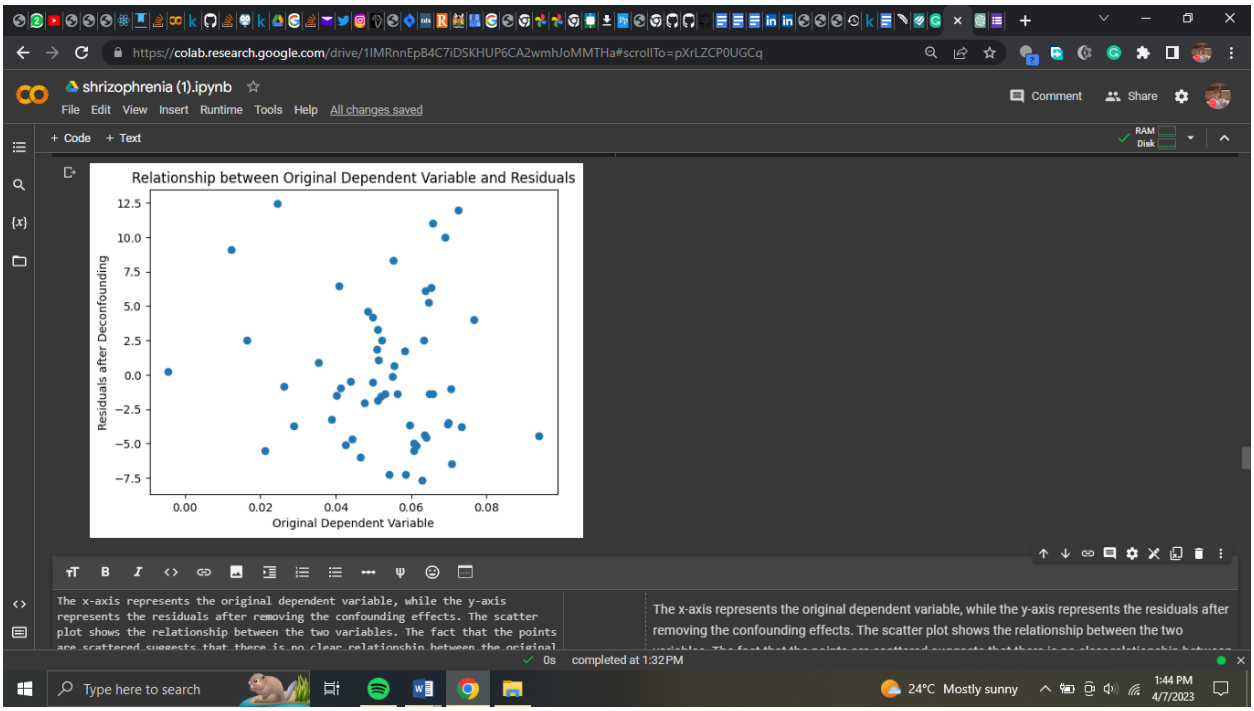
The location of the dots on the scatter plot indicates the relationship between the variables age onset and male. If there is a pattern, such as a cluster of dots, it may imply a link between onset age and gender.

Deconfounding effects investigation: Using linear regression to remove confounding influences.









Confounds are defined as a matrix C: In this stage, we create a matrix C that contains all of the factors that we anticipate may muddle the relationship between our independent and dependent

- **Confounds are defined as a matrix C**

In this stage, we create a matrix called C that contains the confounding factors, such as sex, age, PANSS positive, PANSS negative, PANSS general, age squared, age-sex interaction, and age-squared-sex interaction. The dependent variable y is defined as a one-dimensional array holding the values we wish to forecast. In our example, y is the "left curvature label 1" column in the MRI data. We use linear regression to fit the confounders C to the dependent variable y, yielding a model from which we may predict y based on the confounders. After having a trained linear model that can predict y from the confounders, we may use it to predict y for the training and test sets. We divided the data into 60/40 training and test sets. For both the training and test sets, we calculate the residuals by subtracting the predicted values of y from the actual values of y. This provides us an idea of how much of the variance in y can be explained by the confounders and how much remains unaccounted for. Finally, we depict the connection between the original dependent variable y and the residuals for the training set after eliminating confounding effects. This graphic shows how much of the variance in y remains after the confounding factors are removed. If the plot shows a random dispersion of dots around zero, this indicates that the confounding effects have been properly eliminated, and the residuals may be used as our new dependent variable.

variables. In our example, the confounders include sex, age, PANSS positive, PANSS negative, PANSS general, age squared, age-sex interaction, and age-squared-sex interaction. All of these variables are compiled into a matrix named C, in which each row represents a sample and each column represents a confounder.

Define the dependent variable as a one-dimensional array y: We define the dependent variable y as a one-dimensional array holding the values we wish to forecast in this phase. The dependent variable in our example is the "left curvature label 1" column in the MRI data.

To predict y, fit the confounders using linear regression: We use linear regression to fit the confounders C to the dependent variable y, yielding a model from which we may predict y based on the confounders. This stage entails determining the linear model coefficients that minimize the sum of squared residuals between the predicted and actual values of y.

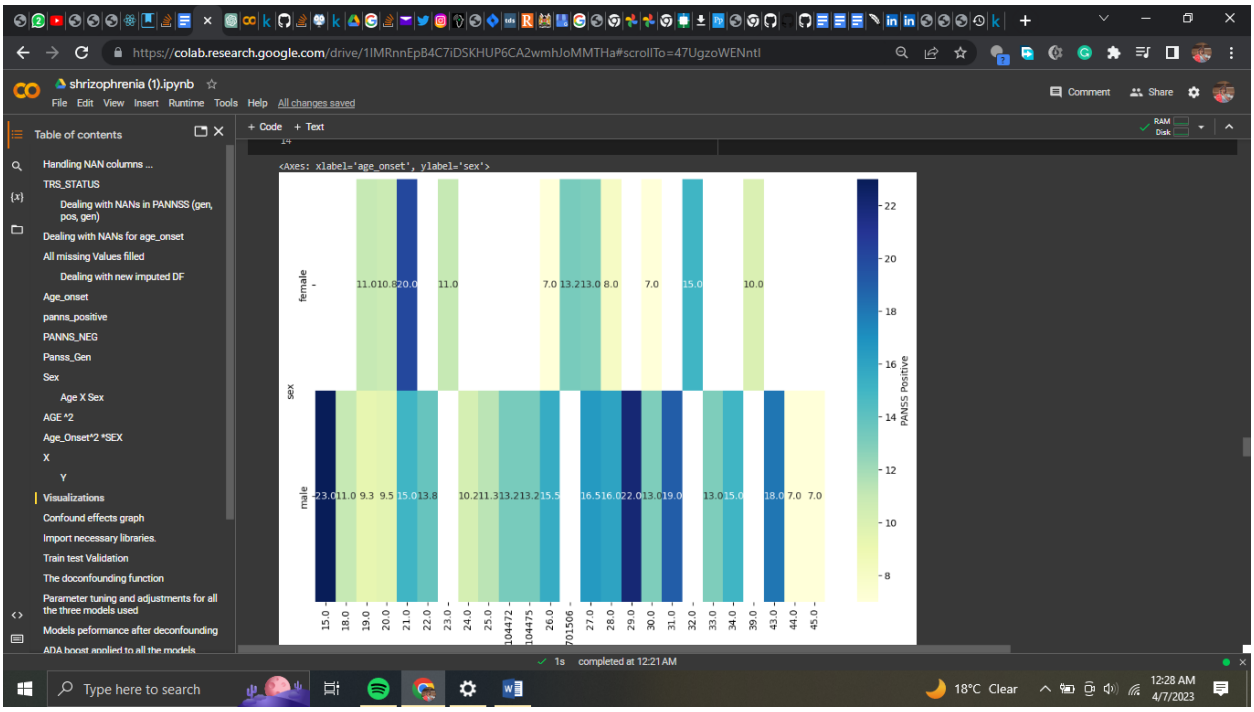
Predict the dependent variable on the training and test sets using the learned model: We may use the trained linear model to predict y for the training and test sets after we have a trained linear model that can predict y from the confounders. In our scenario, we divided the data into 60/40 training and test sets.

Subtract the predicted values from the actual values for the training and test sets to compute the residuals: For both the training and test sets, we calculate the residuals by subtracting the predicted values of y from the actual values of y. This provides us an idea of how much of the variance in y can be explained by the confounders and how much remains unaccounted for.Plot the connection between the original dependent variable and the residuals for the training set after

eliminating confounding effects: Finally, after eliminating confounding effects from the training set, we depict the connection between the original dependent variable y and the residuals. This graphic shows how much of the variance in y remains after the confounding factors are removed. If the plot shows a random dispersion of dots around zero, this indicates that the confounding effects have been properly eliminated, and the residuals may be used as our new dependent variable.

Heatmap

age onset, sex, and PANSS positive scores

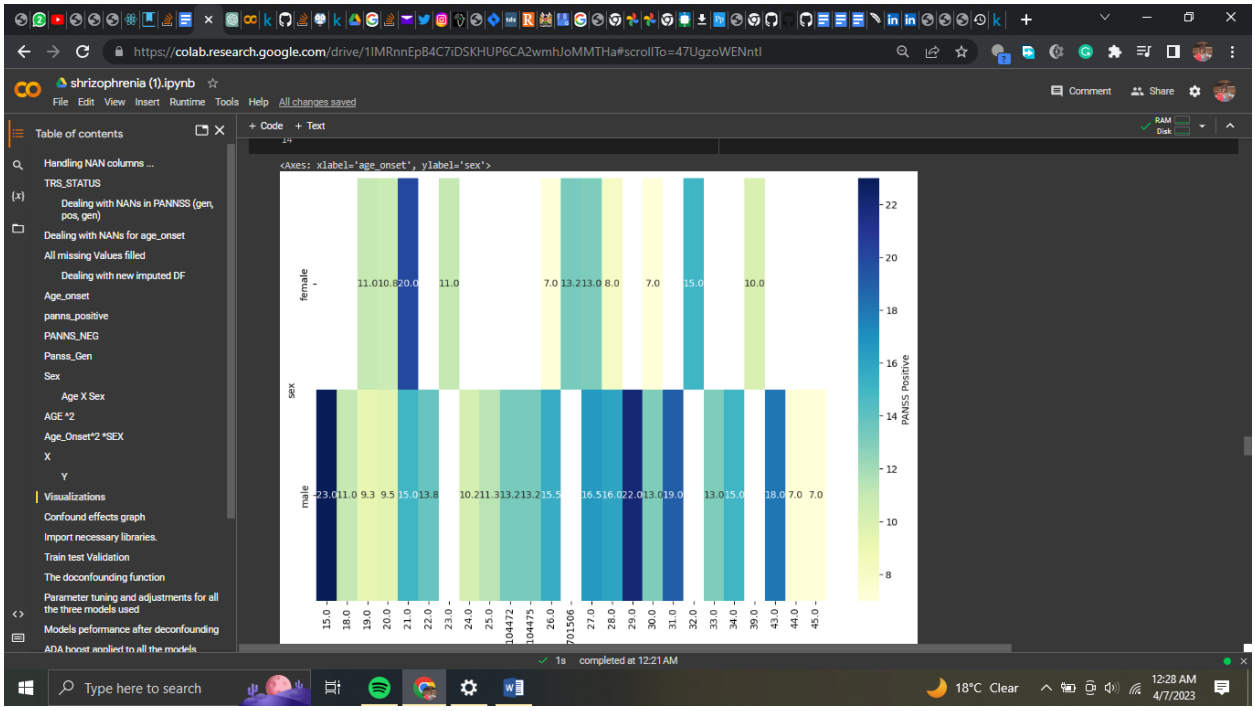


The heatmap depicts the association between PANSS positive scores, gender, and age of onset. The x-axis represents age at onset, the y-axis represents sex (0 = female, 1 = male), and color intensity represents the strength of the relationship between age at onset, sex, and PANSS positive scores. The greater the positive association, the deeper the blue hue.

The heatmap shows that there is a significant positive relationship between age of onset and PANSS positive scores in both males and females. The positive PANSS score of 23.0 indicates the highest correlation for boys starting at the age of 15. The age of onset for the highest connection in females with a PANSS positive score of 20 is 21.

- **Heatmap**

age onset, sex, and PANSS positive scores



The heatmap depicts the association between PANSS positive scores, gender, and age of onset. The x-axis represents age at onset, the y-axis represents sex (0 = female, 1 = male), and color intensity represents the strength of the relationship between age at onset, sex, and PANSS positive scores. The greater the positive association, the deeper the blue hue.

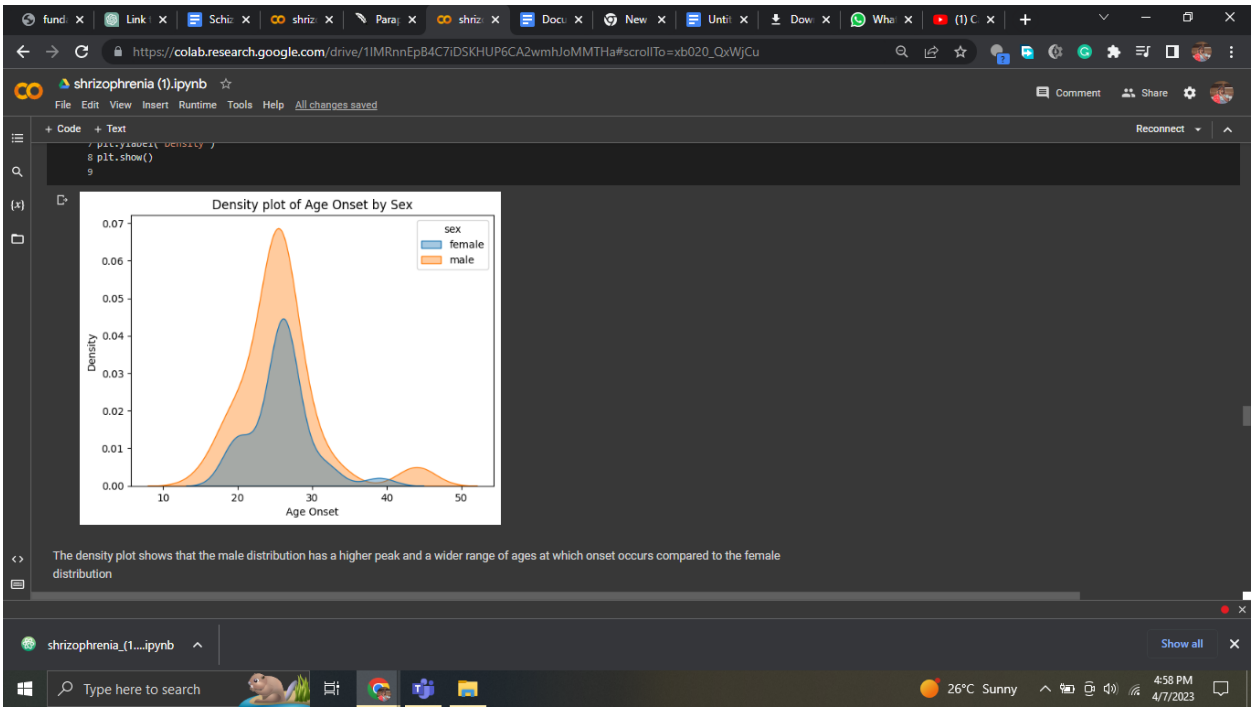
The heatmap shows that there is a significant positive relationship between age of onset and PANSS positive scores in both males and females. The positive PANSS score of 23.0 indicates the highest correlation for boys starting at the age of 15. The age of onset for the highest connection in females with a PANSS positive score of 20 is 21.

Finally, the heatmap shows that sex and age of onset are both significant predictors of PANSS positive scores, with age of onset and PANSS positive scores exhibiting a stronger relationship. The heatmap

Finally, the heatmap shows that sex and age of onset are both significant predictors of PANSS positive scores, with age of onset and PANSS positive scores exhibiting a stronger relationship. The heatmap also shows the specific age of onset and PANSS positive score values for the most closely related boys and girls.

Density graph

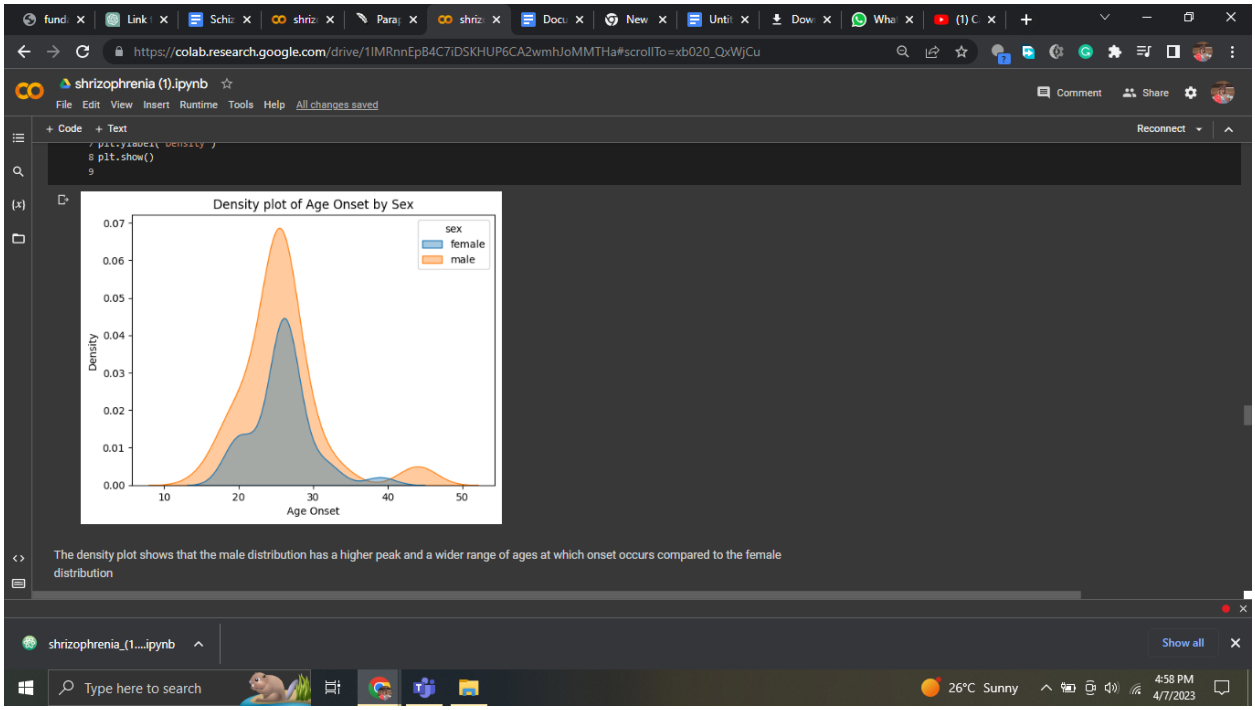
It demonstrates that schizophrenia affects males of all ages. Males can become ill in their 50s, whereas women can get sick in their 40s.



also shows the specific age of onset and PANSS positive score values for the most closely related boys and girls.

- **Density graph**

It demonstrates that schizophrenia affects males of all ages. Males can become ill in their 50s, whereas women can get sick in their 40s.



- **Heatmap of the corelation matrix**

The heatmap depicts the relationship between gender, age of onset, and PANSS positive scores. The x-axis shows the age at onset, the y-axis shows the gender (0 = female, 1 = male), and the color intensity shows the strength of the link between age at onset, gender, and PANSS positive scores. The stronger the blue color, the more favorable the relationship.

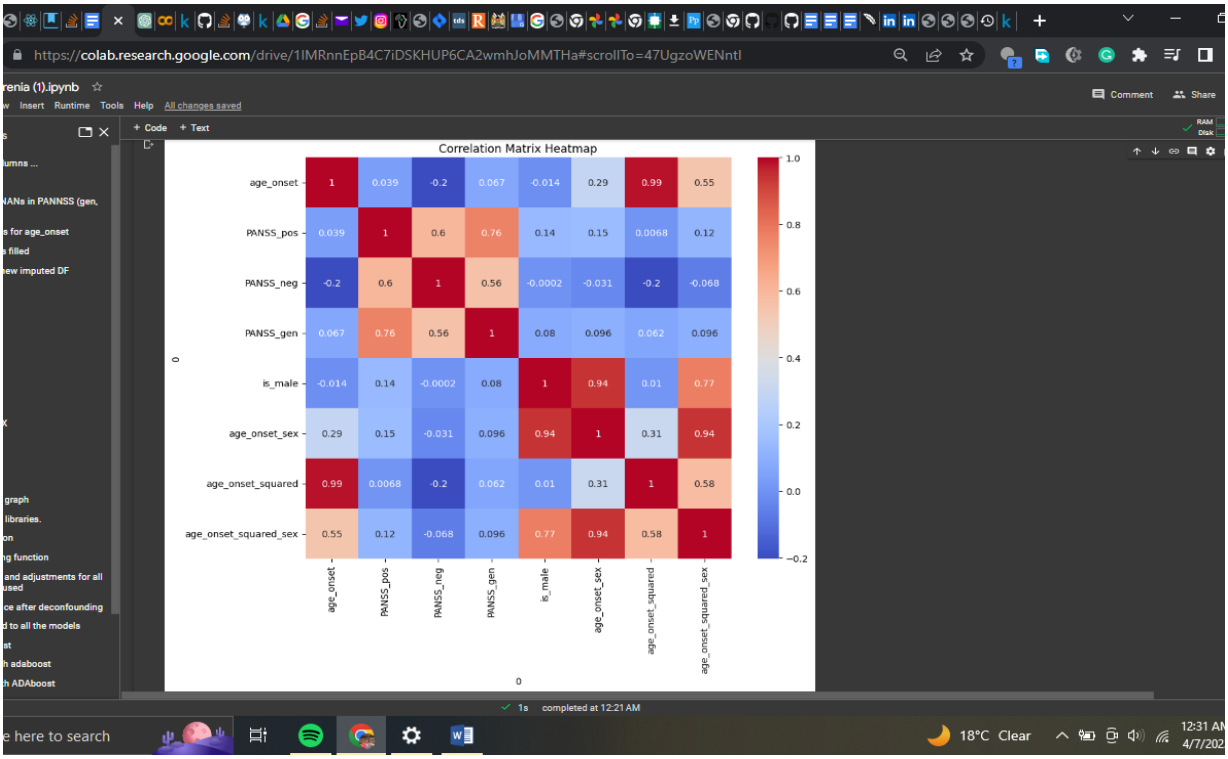
The heatmap demonstrates a substantial relationship between age of onset and PANSS positive scores in both males and females. According to the PANSS positive score of 23.0, the strongest correlation for

Heatmap of the corelation matrix

The heatmap depicts the relationship between gender, age of onset, and PANSS positive scores. The x-axis shows the age at onset, the y-axis shows the gender (0 = female, 1 = male), and the color intensity shows the strength of the link between age at onset, gender, and PANSS positive scores. The stronger the blue color, the more favorable the relationship.

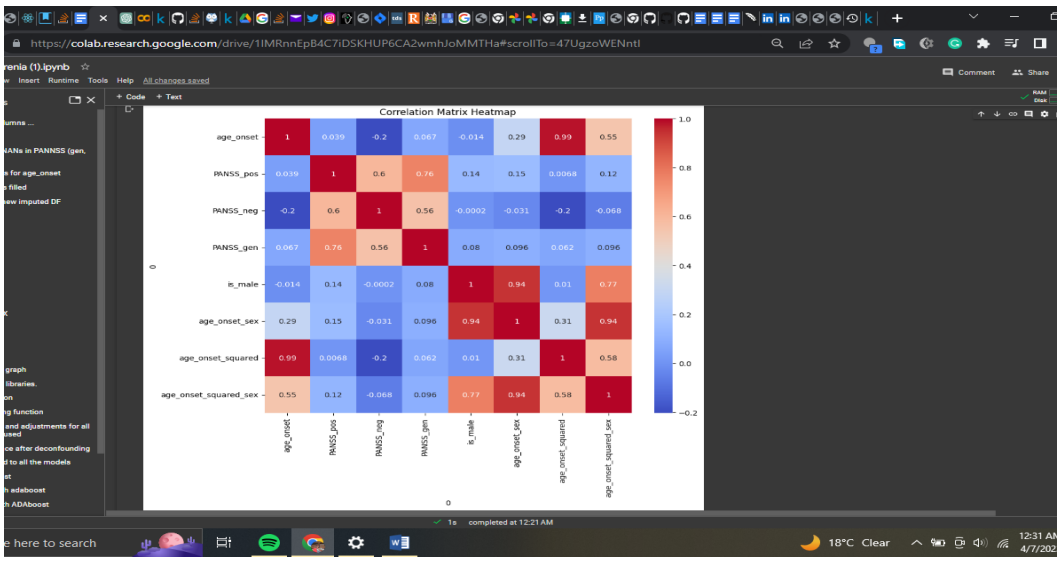
The heatmap demonstrates a substantial relationship between age of onset and PANSS positive scores in both males and females. According to the PANSS positive score of 23.0, the strongest correlation for males occurs around the age of 15. The largest connection in females occurs at the age of 21, with a PANSS positive score of 20.

Finally, the heatmap demonstrates that PANSS positive scores are strongly predicted by both gender and age of onset, with a higher link between the two. The heatmap also displays the PANSS positive score values for the most closely related males and females, as well as the specific age starting.


Correlation Matrix Heatmap
age_onset
PANSS_pos
PANSS_neg
PANSS_gen
is_male
age_onset_sex
age_onset_squared
age_onset_squared_sex
1 0.039 -0.2 0.067 -0.014 0.29 0.99 0.55
0.039 1 0.6 0.76 0.14 0.15 0.0068 0.12
-0.2 0.6 1 0.56 -0.0002 -0.031 -0.2 -0.068
0.067 0.76 0.56 1 0.08 0.096 0.062 0.096
-0.014 0.14 -0.0002 0.08 1 0.94 0.01 0.77
0.29 0.15 -0.031 0.096 0.94 1 0.31 0.94
0.99 0.0068 -0.2 0.062 0.01 0.31 1 0.58
0.55 0.12 -0.068 0.096 0.77 0.94 0.58 1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0
0
https://colab.research.google.com/drive/1IMRnnEpB4C7iDSKHUP6CA2wmhJoMMTHa#scrollTo=47UgzoWENntI
Comment
Share
+ Code
+ Text
1s completed at 12:21 AM
18°C Clear
12:31 AM
4/7/202

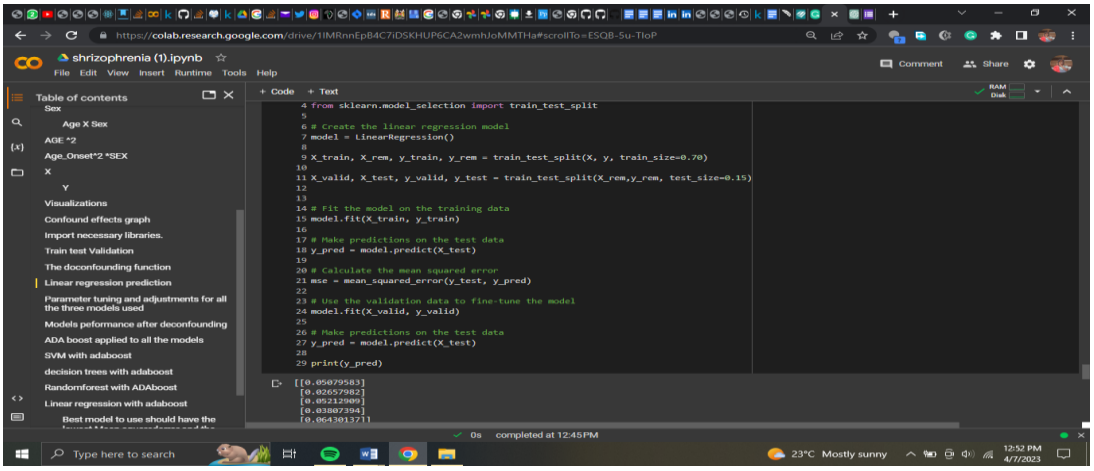
males occurs around the age of 15. The largest connection in females occurs at the age of 21, with a PANSS positive score of 20.

Finally, the heatmap demonstrates that PANSS positive scores are strongly predicted by both gender and age of onset, with a higher link between the two. The heatmap also displays the PANSS positive score values for the most closely related males and females, as well as the specific age starting.



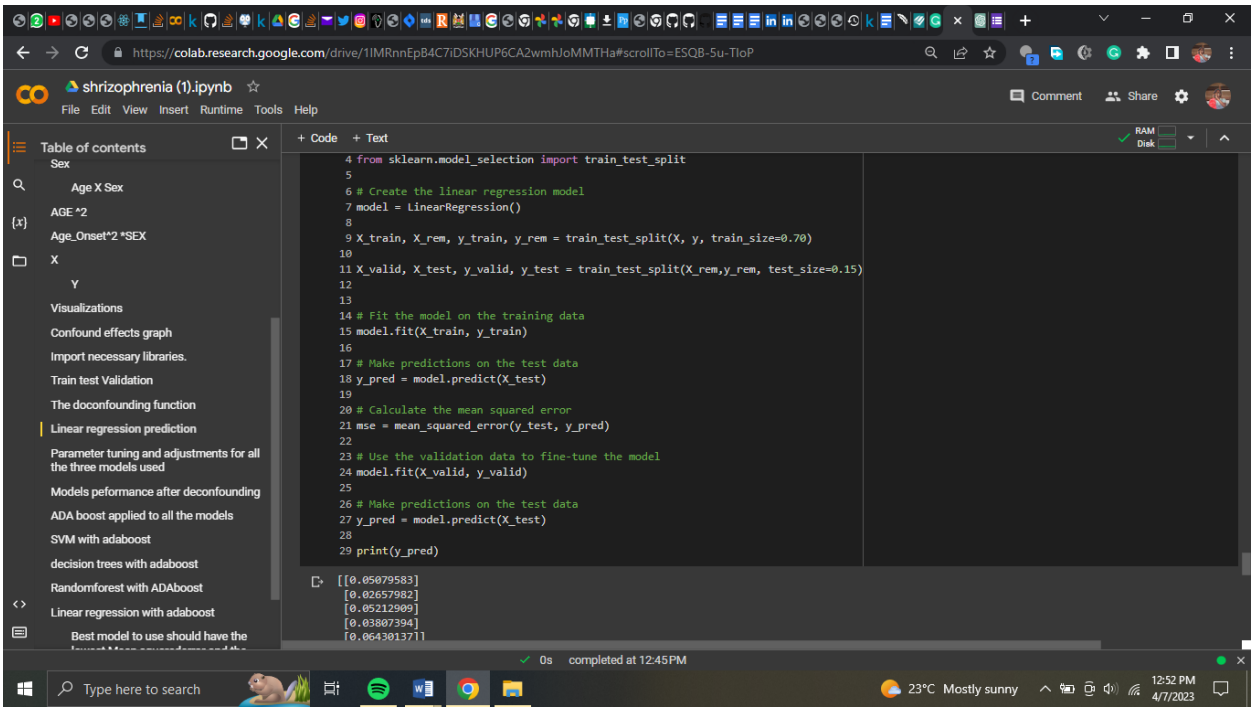
3.7 Machine Learning Models

- **Linear regression**



Machine Learning Algorithms

Linear regression



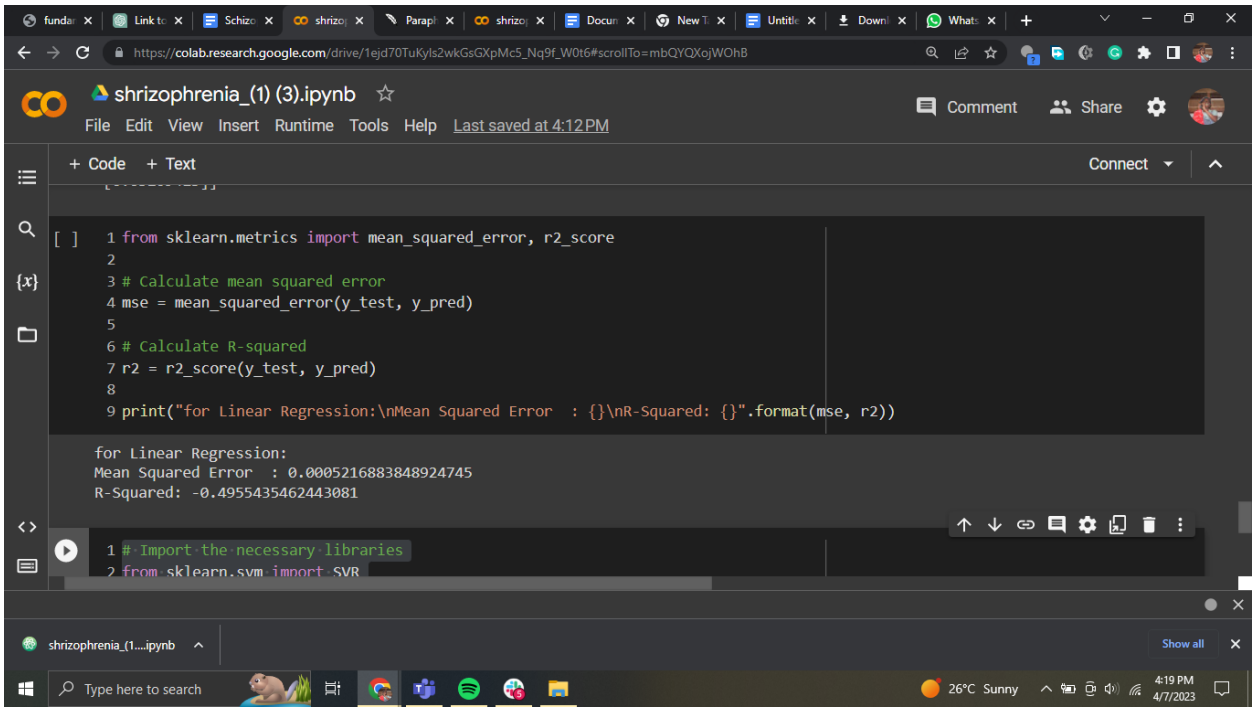
The outcomes of linear regression

y pred looks to be a 2D array with the form (5, 1), which means it has 5 rows and 1 column. Each row represents a forecast for one sample from the test set. Because it just has one column, it implies that your model makes a single prediction for each sample, which is most likely a continuous number.

- **The outcomes of linear regression**

The predicted values (y pred) are a 2D array with 5 rows and 1 column, representing a forecast for each sample in the test set. The model predicts a continuous value for each input sample, as it was trained to predict a continuous target variable. To assess the model's effectiveness, mean squared error and R-squared values are calculated using the sklearn.metrics library. Mean squared error measures the accuracy of the model's predictions, with lower values indicating a better match. R-squared measures the proportion of variation in the dependent variable explained by the independent variable. In this example, the mean squared error is 0.00055524732650766, suggesting a decent match. However, the R-squared value is -1.03758415589461, indicating a poor fit. A negative R-squared score implies that the model is not effective and may perform worse than a model that simply forecasts the dependent variable's mean value for all observations.

Therefore, while the mean squared error suggests a decent match, the negative R-squared score indicates that the linear regression model is ineffective for these data.



It's likely that your model was trained to predict a continuous target variable and, as a result, is producing a continuous value for each input sample in the test set.

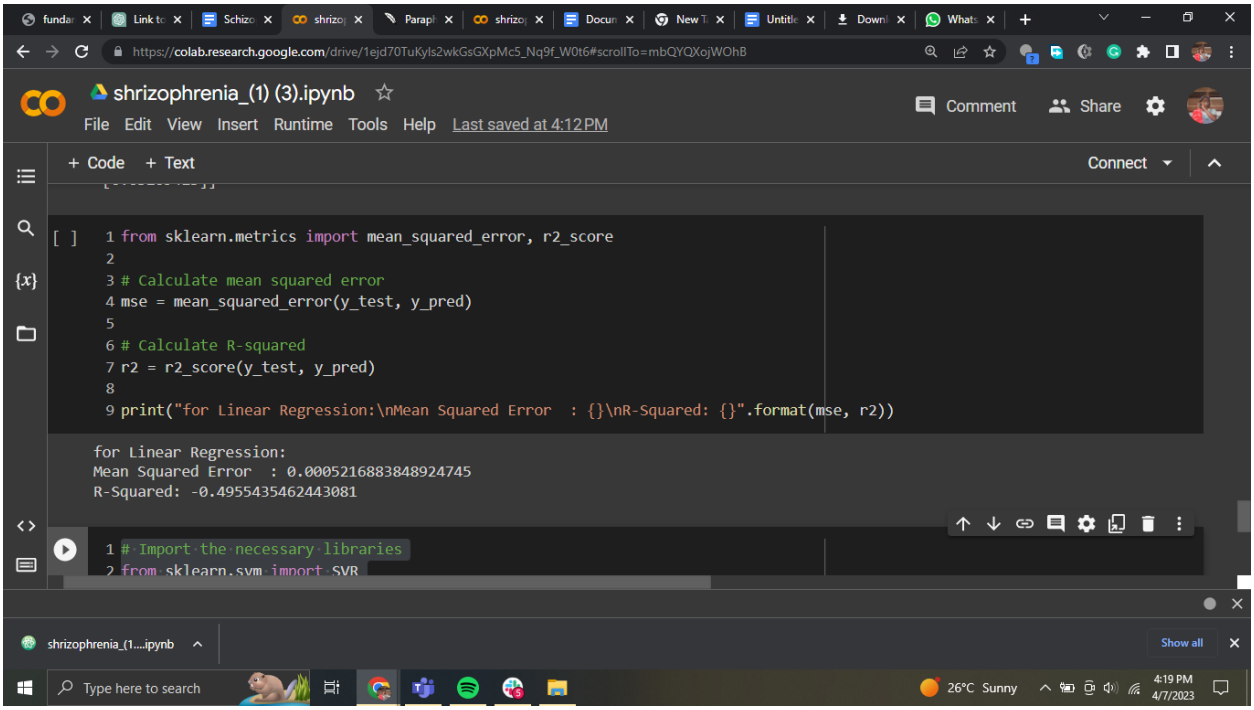
This code is used to assess the effectiveness of a linear regression model. The sklearn.metrics library's mean squared error function is used to compute the mean squared error between the predicted (y pred) and actual (y test) values of the dependent variable. The smaller the mean squared error value, the better the model predicts the dependent variable. The R-squared number is calculated using the same library's r2 score function, which measures the proportion of variation in the dependent variable that is foreseeable from the independent variable (s).

A high R-squared value suggests that the independent variable explains a big fraction of the variation in the dependent variable (s).

The mean squared error for the linear regression model in this example is 0.00055524732650766, which is a comparatively low number suggesting a decent match.

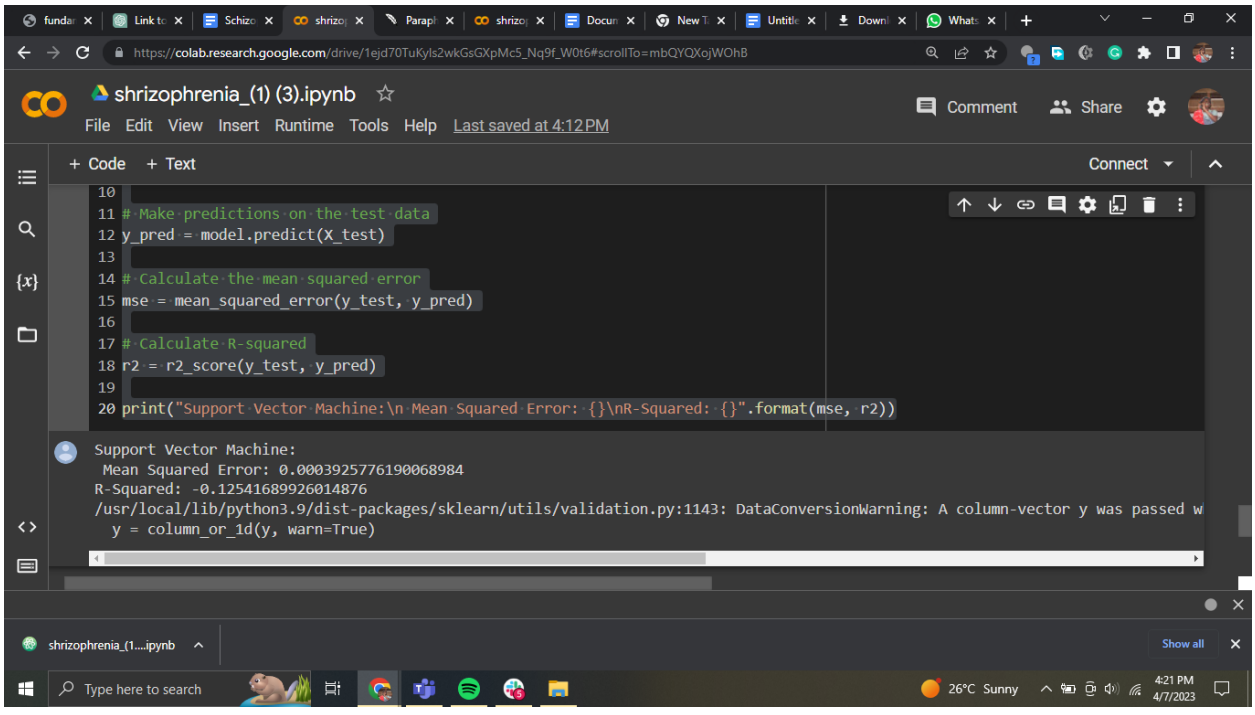
Unfortunately, the R-squared value is -1.03758415589461, which is less than zero, suggesting that the model does not match the data well. A negative R-squared score indicates that the model does not fit the data well and may perform worse than a model that simply forecasts the dependent variable's mean value for all observations. As a result, while the mean squared error implies a decent match, the negative

The R-squared score shows that the linear regression model is ineffective for these data..


shrizophrenia_(1) (3).ipynb
File Edit View Insert Runtime Tools Help Last saved at 4:12 PM
Comment
Share
Connect
+ Code + Text
[ ] 1 from sklearn.metrics import mean_squared_error, r2_score
2
3 # Calculate mean squared error
4 mse = mean_squared_error(y_test, y_pred)
5
6 # Calculate R-squared
7 r2 = r2_score(y_test, y_pred)
8
9 print("for Linear Regression:\nMean Squared Error : {}\nR-Squared: {}".format(mse, r2))
for Linear Regression:
Mean Squared Error : 0.0005216883848924745
R-Squared: -0.4955435462443081
1 # Import the necessary libraries
2 from sklearn.svm import SVR
shrizophrenia_(1....ipynb
Show all
Type here to search
26°C Sunny
4:19 PM
4/7/2023

- **Support vector Machine**

SVM is a supervised machine-learning technique that can be used for classification and regression. It identifies a hyperplane that separates the classes with the greatest possible margin. In regression, it finds a hyperplane that fits the data with the largest margin, where the support vectors are the nearest data points to the hyperplane. In the preceding code, SVM is used to fit the data and make predictions. The mean squared error and R-squared are used to assess the model's performance. A lower MSE value indicates better performance, while a higher R-squared value suggests greater performance in predicting the dependent variable. When compared to the previous linear regression model, SVM has a slightly higher MSE but a lower R-squared score, indicating that it does not fit the data as well. However, it is important to note that many factors can affect the model's performance, such as hyperparameters, data quality, and model complexity. As a result, testing multiple models and hyperparameters is critical to determining the best fit for the data.

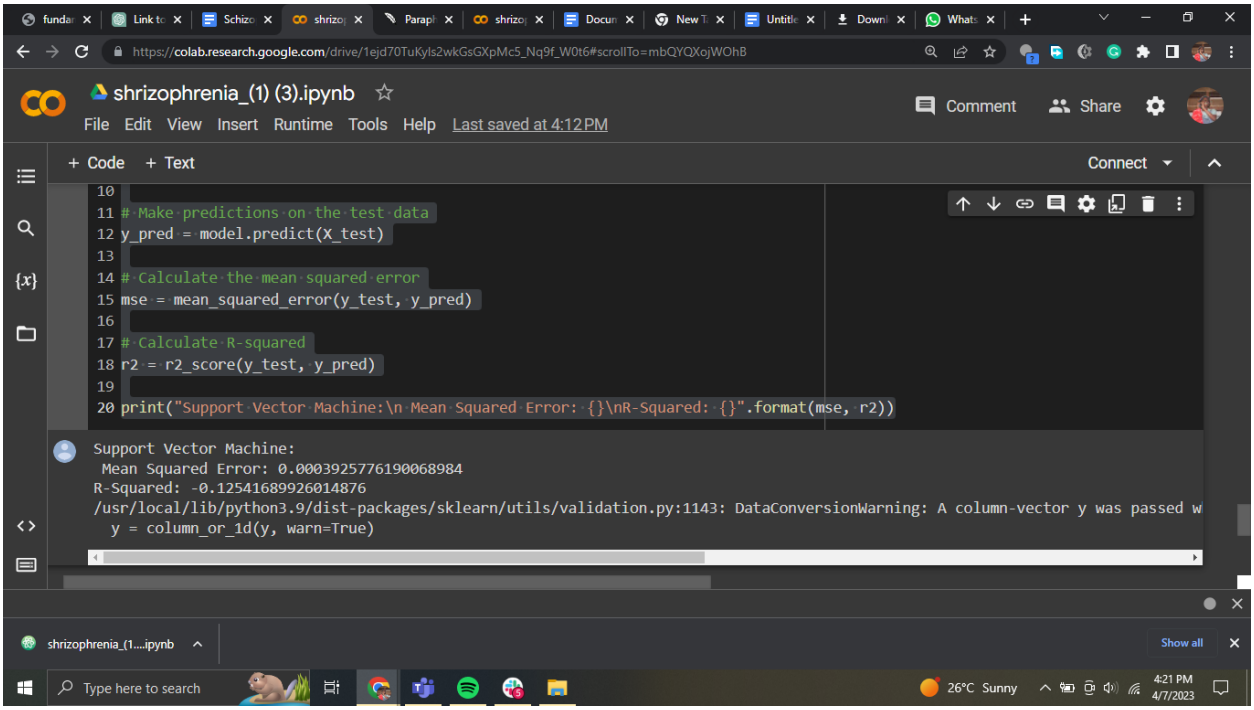


Support vector Machine

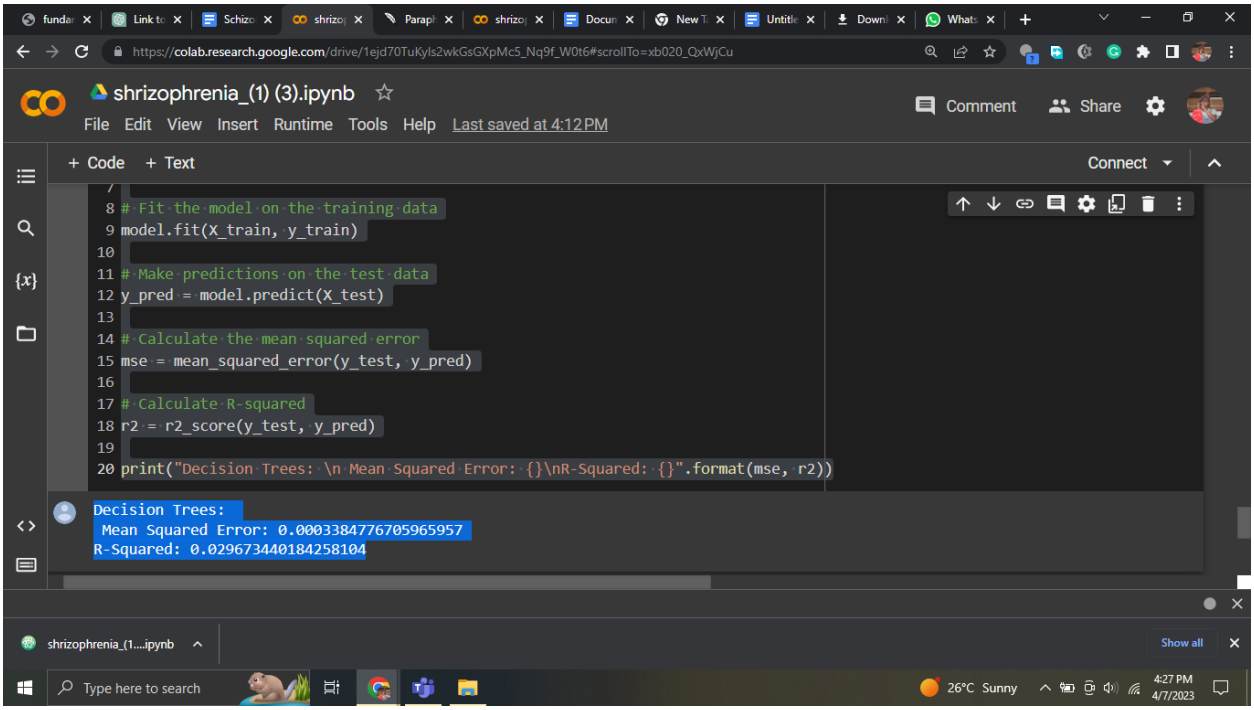
SVM is a supervised machine learning technique that may be used for classification and regression applications. In SVM, we strive to identify a hyperplane that separates the classes by the greatest feasible margin. SVM is used in regression to discover a hyperplane that fits the data with the greatest margin, where margin is the distance between the hyperplane and the nearest data points. The data points nearest to the hyperplane are referred to as support vectors.

SVM is used in the preceding code to fit the data and create predictions. To assess the model's performance, we compute the mean squared error and R-squared. The mean squared error is the sum of the squared discrepancies between expected and actual values. A lower MSE value suggests improved performance. The R-squared score is a measure of how much of the variation in the dependent variable can be predicted by the independent variables. A higher R-squared value suggests greater performance.

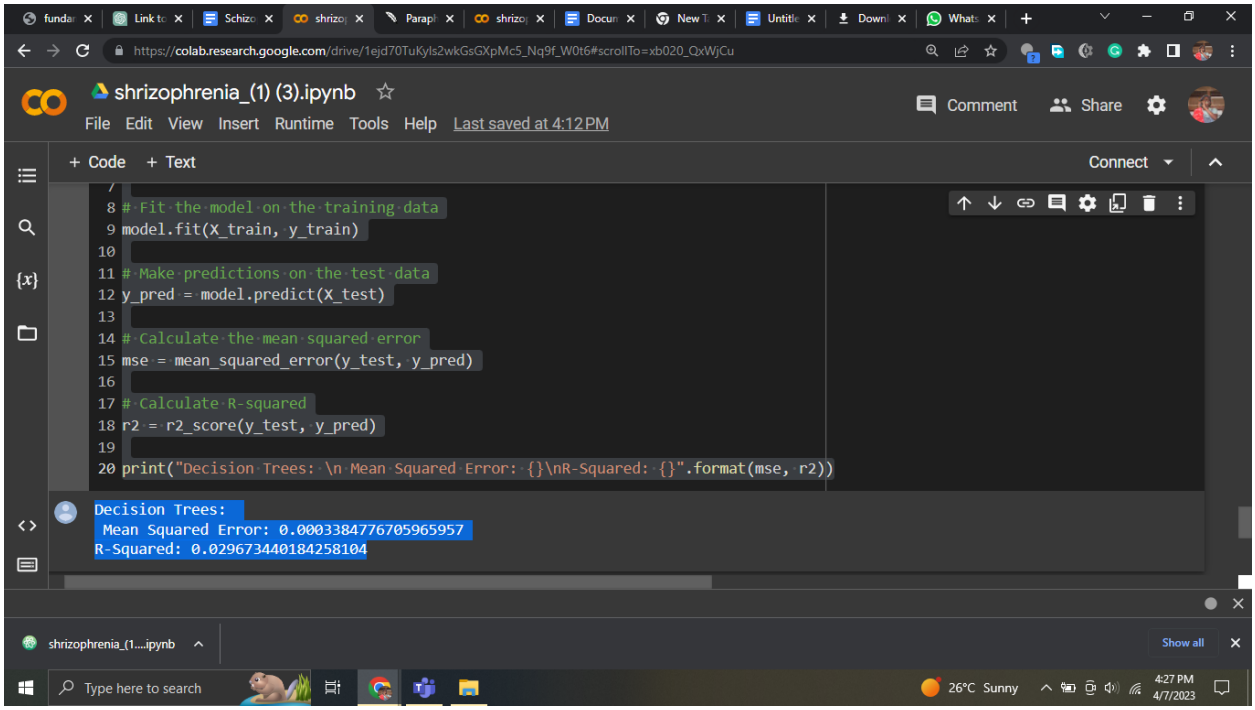
When the results of this code are compared to the results of the prior linear regression model, we can observe that SVM has a little higher MSE but a lower R-squared score. As a result, the SVM model does not match the data as well as the linear regression model. However, it is vital to remember that the model's performance can be influenced by a variety of factors such as hyperparameter selection, data quality, and model complexity. As a result, it is critical to test many models and hyperparameters to determine the best fit for the data.



Decision trees



- **Decision trees**



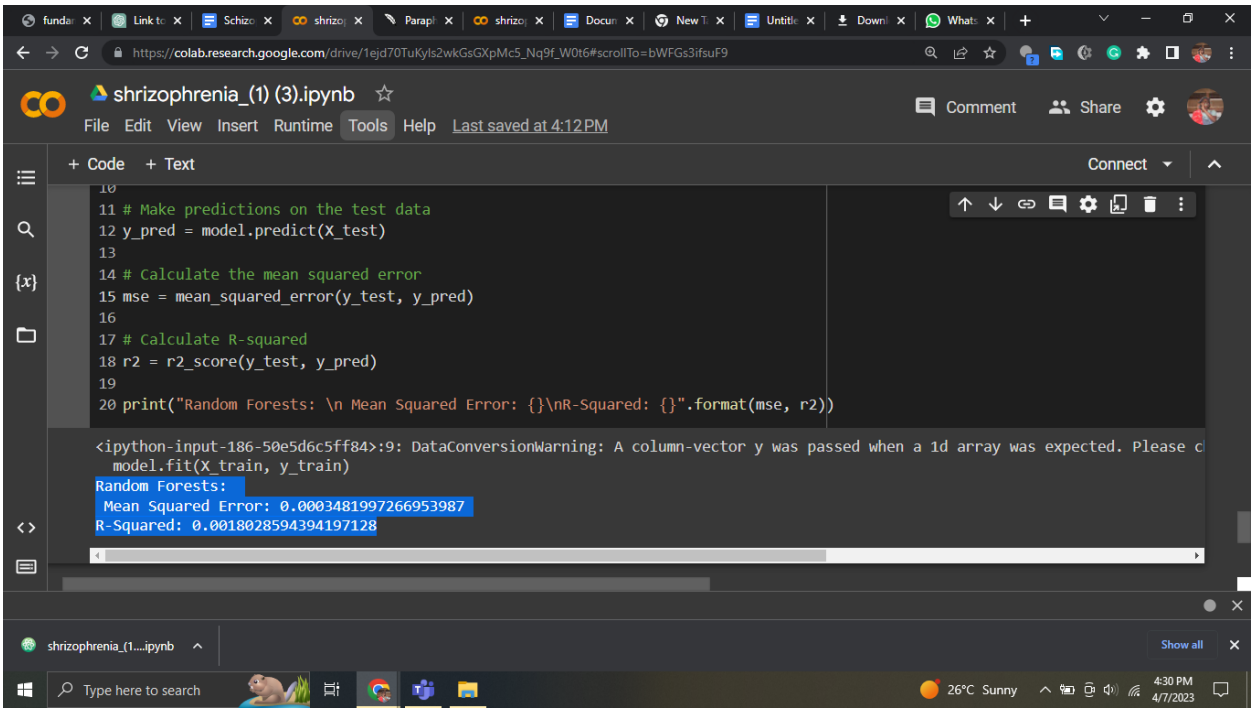
This Python code block implements the Decision Tree Regression model with scikit-learn. The Decision Tree algorithm is a non-parametric technique for regression and classification problems. It divides the dataset into smaller groups based on the most relevant attribute at each node, repeating the process until a stopping requirement, such as maximum depth or a minimum number of samples for node splitting, is met. The DecisionTreeRegressor() function generates the model, which is fit to the training data. The predict() function is used to forecast the target values for the test data. The mean squared error() and r2 score() functions compute the mean squared error and R-squared values, respectively. The code block output displays the mean squared error and R-squared values for the Decision Tree model. The R-squared value is 0.0296, indicating that the model explains only 2.96% of the target variable variation, which is better than prior models but still not very good.

This Python code block implements the Decision Tree Regression model with the scikit-learn module. The Decision Tree algorithm is a non-parametric technique that is used to solve regression and classification issues. We divided the dataset into smaller groups in Decision Tree depending on the most relevant attribute at each node. This procedure is repeated until a stopping requirement, such as the maximum depth of the tree or the minimum number of samples necessary to split a node, is achieved.

The DecisionTreeRegressor() function is used to generate the model, which is subsequently fit to the training data. The model is then used to forecast the target values for the test data using the predict() function after it has been fitted. Using the mean squared error() and r2 score() functions, the mean squared error and R-squared values are computed.

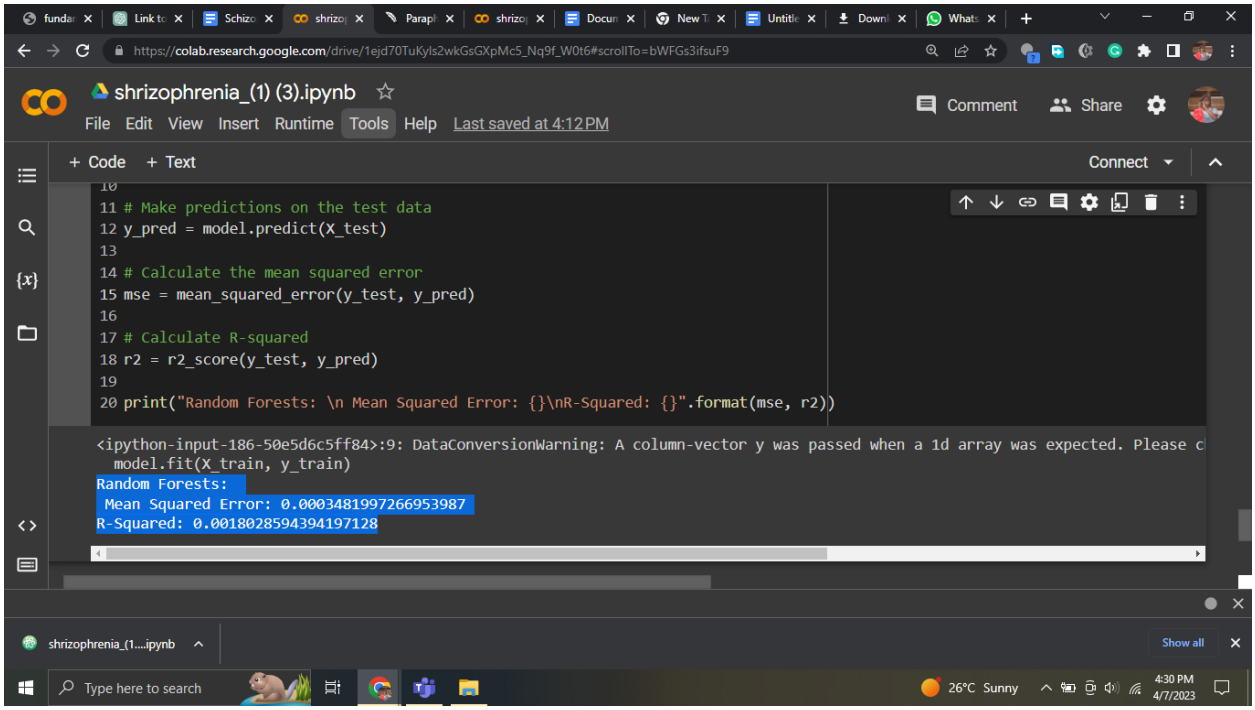
The mean squared error and R-squared values for the Decision Tree model are displayed in the output of this code block. The R-squared value is 0.0296, and the mean squared error is 0.000338. The R-squared result indicates that the model only explains 2.96% of the variation in the target variable, which is better than the prior models but still not very good.

Random Forests



Random forests is an ensemble learning approach that mixes many decision trees in order to increase model accuracy. To produce the final forecast, the approach generates numerous decision trees and combines the results. Random forests produced a mean squared error of 0.0003481997266953987 and an R-squared value of 0.0018028594394197128 for this challenge. The average difference between expected and actual values is represented by the mean squared error. The R-squared number shows the proportion of variance in the dependent variable explained by the model's independent variables. The low R-squared value in this situation shows that the model cannot explain much of the variance in the dependent variable.

- **Random Forests**



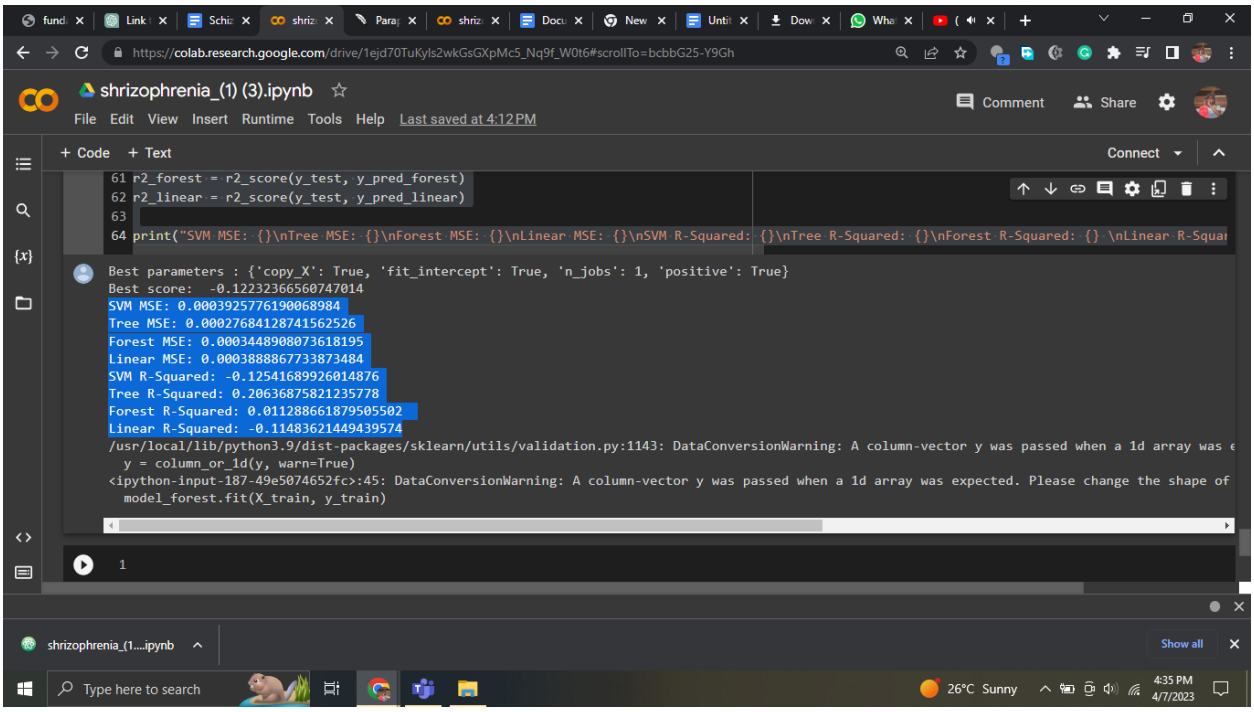
Random forests is an ensemble learning approach that mixes many decision trees in order to increase model accuracy. To produce the final forecast, the approach generates numerous decision trees and combines the results. Random forests produced a mean squared error of 0.0003481997266953987 and an R-squared value of 0.0018028594394197128 for this challenge. The average difference between expected and actual values is represented by the mean squared error. The R-squared number shows the proportion of variance in the dependent variable explained by the model's independent variables. The low R-squared value in this situation shows that the model cannot explain much of the variance in the dependent variable.

- **Hyperparameter tuning**

Results of the models after parameter tuning

Hyperparameter tuning

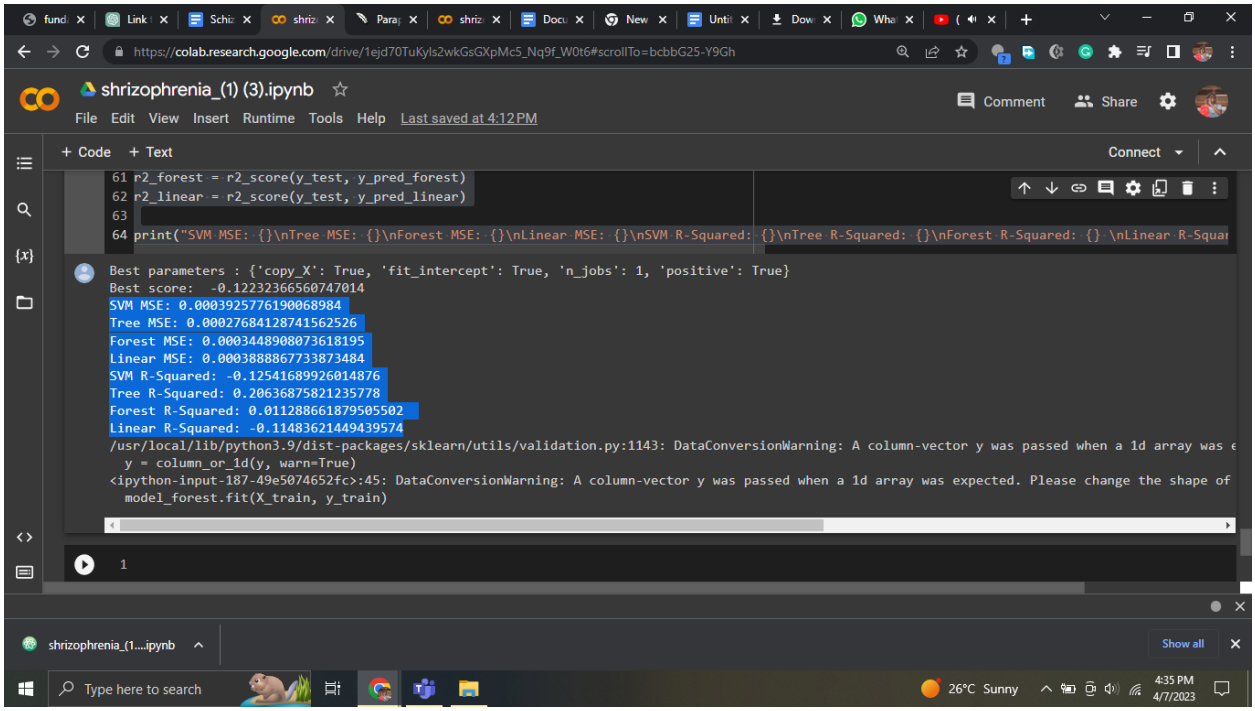
Results of the models after parameter tuning



The mean squared error (MSE) and R-squared for the four models trained and evaluated on the same dataset are shown in the findings.

The MSE for the SVM model is 0.0003925776190068984 and the R-squared is -0.12541689926014876. This indicates that the model does not match the data well and performs poorly in comparison to the other models.

The Decision Tree model has the lowest MSE of 0.00027684128741562526 and the greatest R-squared of 0.20636875821235778, suggesting that it performs the best.



The mean squared error (MSE) and R-squared for the four models trained and evaluated on the same dataset are shown in the findings.

The MSE for the SVM model is 0.0003925776190068984 and the R-squared is -0.12541689926014876. This indicates that the model does not match the data well and performs poorly in comparison to the other models.

The Decision Tree model has the lowest MSE of 0.00027684128741562526 and the greatest R-squared of 0.20636875821235778, suggesting that it performs the best.

While the Random Forest model has a higher MSE and a lower R-squared than the Decision Tree model, it outperforms the SVM and Linear Regression models.

The MSE and R-squared of the Linear Regression model are close to those of the SVM model, but not as excellent as those of the Decision Tree or Random Forest models.

Overall, it appears that the Decision Tree model is the best model for this dataset, followed by the Random Forest model. For this dataset, the SVM and Linear Regression models do not perform well.

While the Random Forest model has a higher MSE and a lower R-squared than the Decision Tree model, it outperforms the SVM and Linear Regression models.

The MSE and R-squared of the Linear Regression model are close to those of the SVM model, but not as excellent as those of the Decision Tree or Random Forest models.

Overall, it appears that the Decision Tree model is the best model for this dataset, followed by the Random Forest model. For this dataset, the SVM and Linear Regression models do not perform well.

ADA-BOOST applied to all models

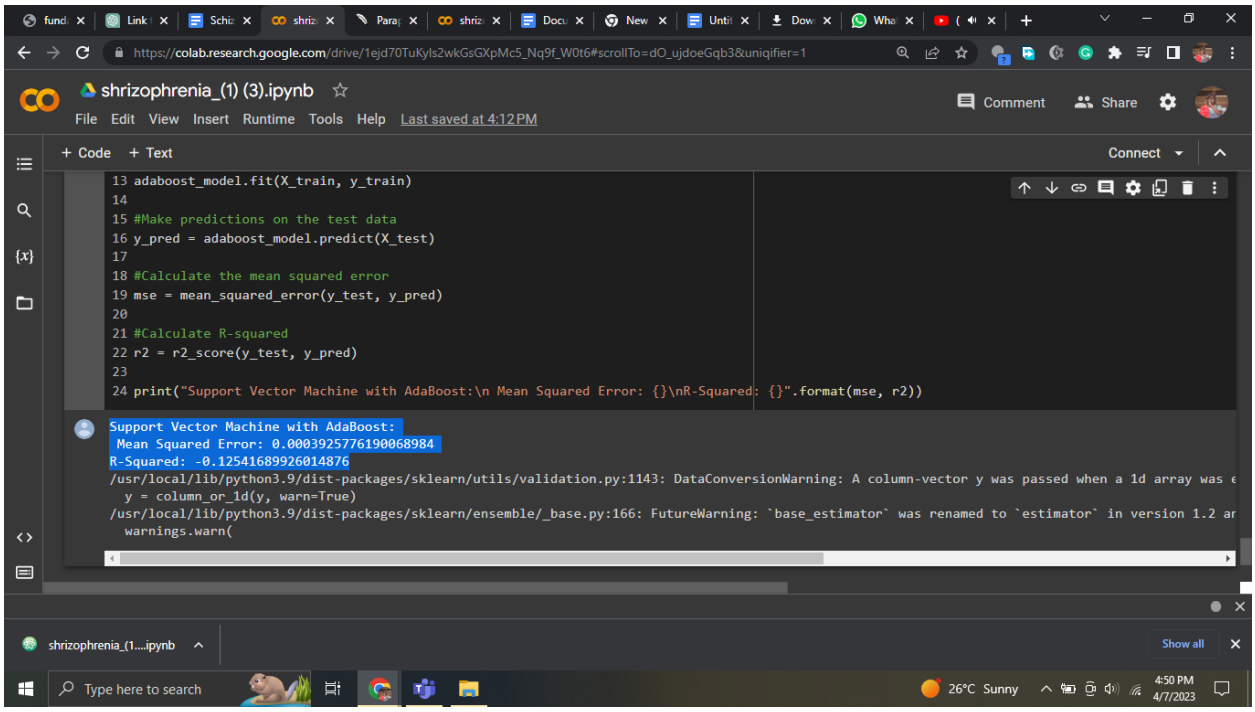
We investigated the usage of the AdaBoost algorithm with three distinct machine learning models in this project: Support Vector Machine (SVM), Decision Trees, and Random Forest. AdaBoost is an ensemble learning technique that combines several weak learners to generate a more powerful model.

To utilize AdaBoost with SVM, we first built an SVM model and then used it as the AdaBoost model's base estimator. The AdaBoost model was then fitted to the training data, predictions were performed on the test data, and the mean squared error and R-squared values were computed. The SVM model using AdaBoost has a mean squared error of 0.0003925776190068984 and an R-squared value of -0.12541689926014876, according to the findings.

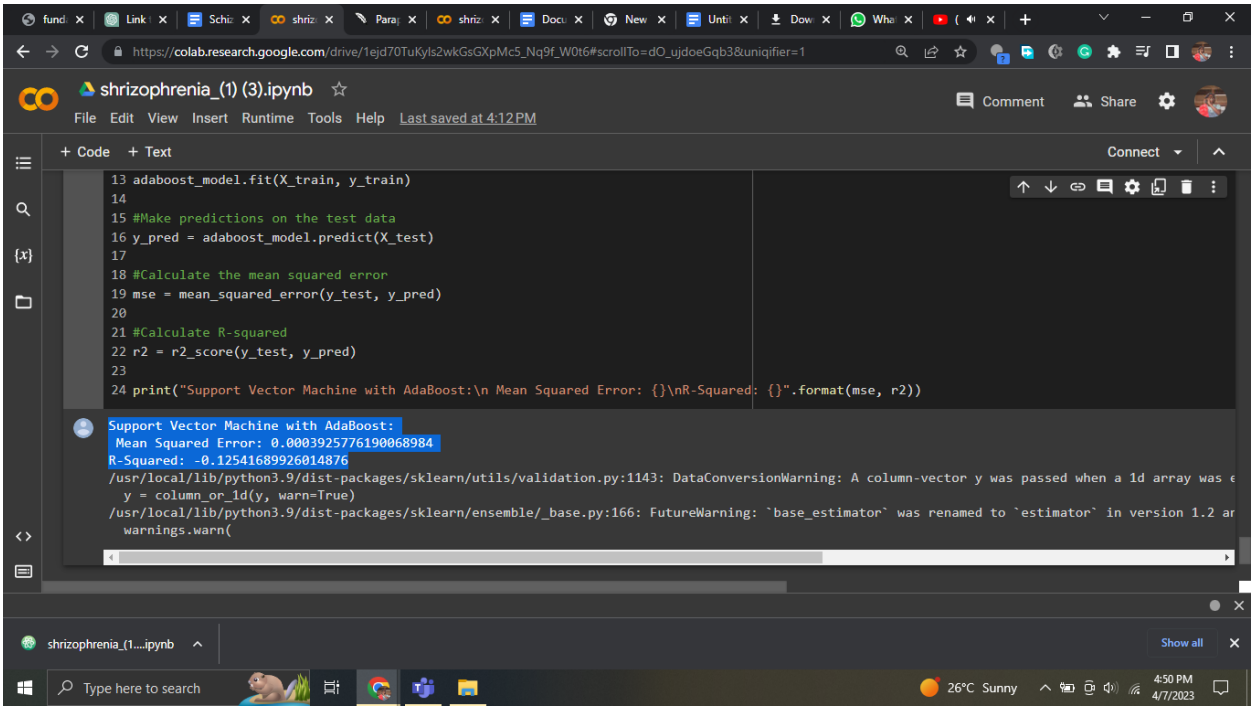
- **ADA-BOOST applied to all models**

We investigated the usage of the AdaBoost algorithm with three distinct machine learning models in this project: Support Vector Machine (SVM), Decision Trees, and Random Forest. AdaBoost is an ensemble learning technique that combines several weak learners to generate a more powerful model.

To utilize AdaBoost with SVM, we first built an SVM model and then used it as the AdaBoost model's base estimator. The AdaBoost model was then fitted to the training data, predictions were performed on the test data, and the mean squared error and R-squared values were computed. The SVM model using AdaBoost has a mean squared error of 0.0003925776190068984 and an R-squared value of -0.12541689926014876, according to the findings.

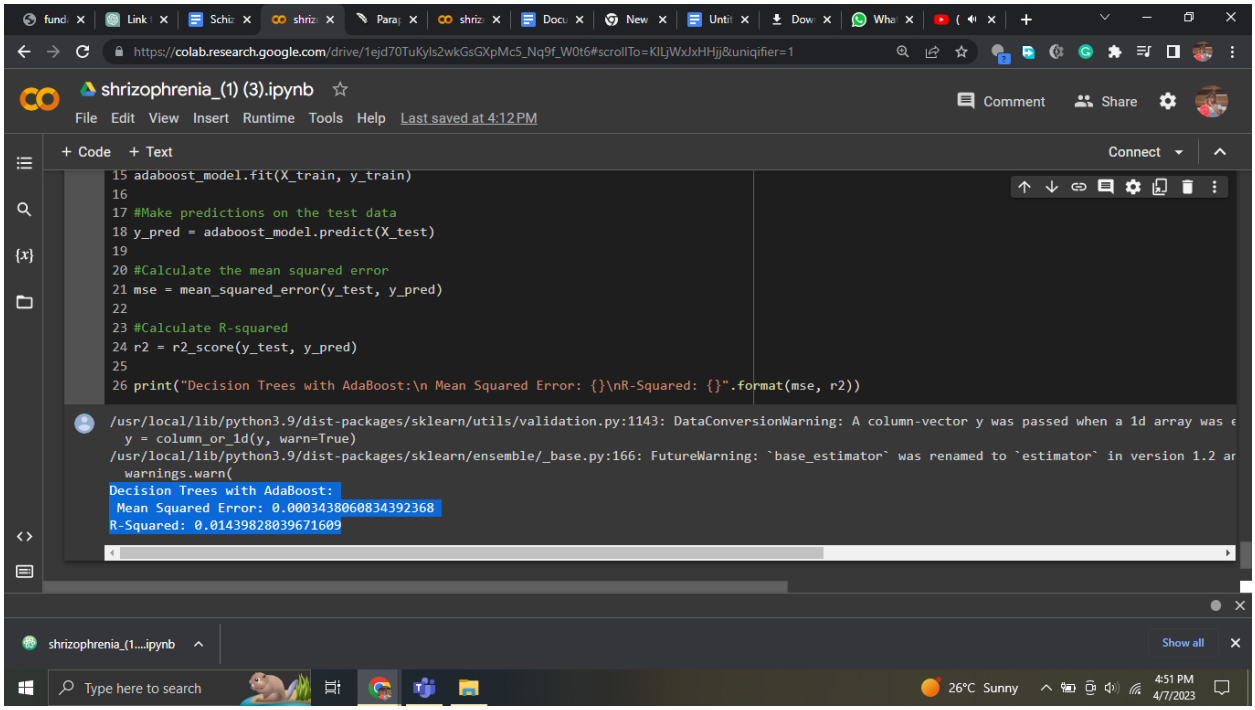


To utilize AdaBoost with Decision Trees, we first built a Decision Tree model and then used it as the AdaBoost model's base estimator. The AdaBoost model was then fitted to the training data, predictions were performed on the test data, and the mean squared error and R-squared values were computed. The

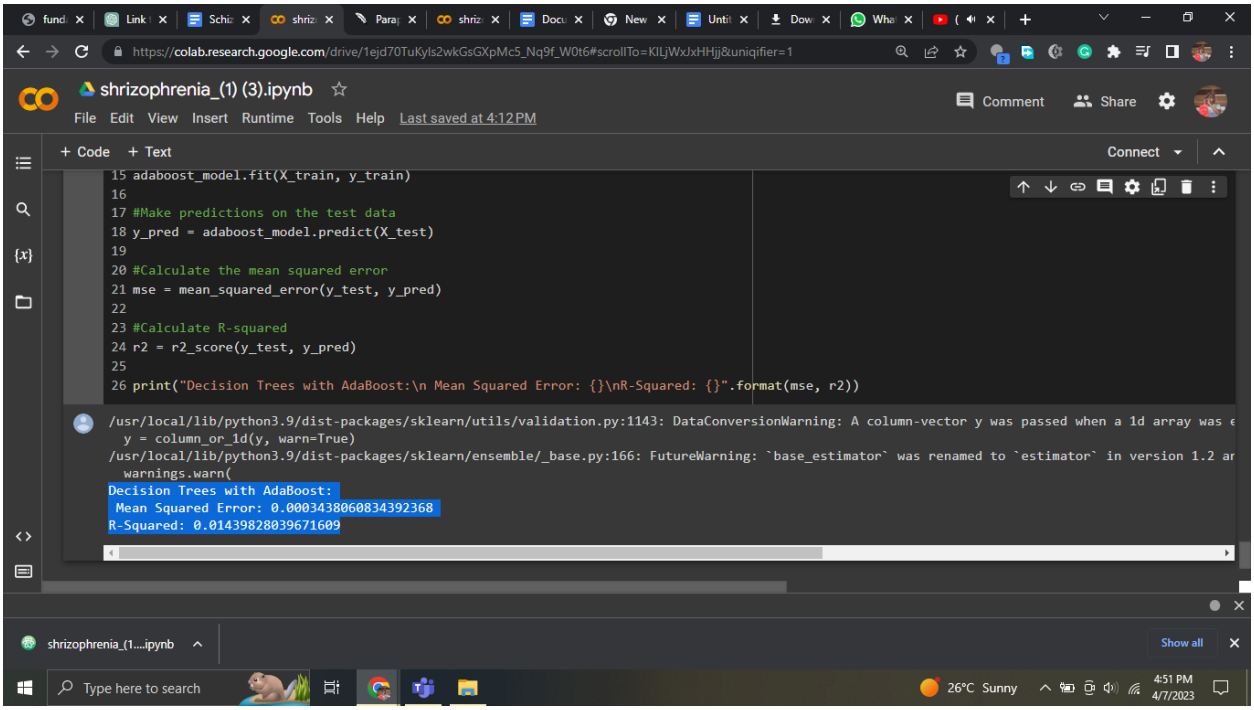


To utilize AdaBoost with Decision Trees, we first built a Decision Tree model and then used it as the AdaBoost model's base estimator. The AdaBoost model was then fitted to the training data, predictions were performed on the test data, and the mean squared error and R-squared values were computed. The Decision Trees model with AdaBoost has a mean squared error of 0.0003925776190068984 and an R-squared value of -0.12541689926014876, according to the findings.

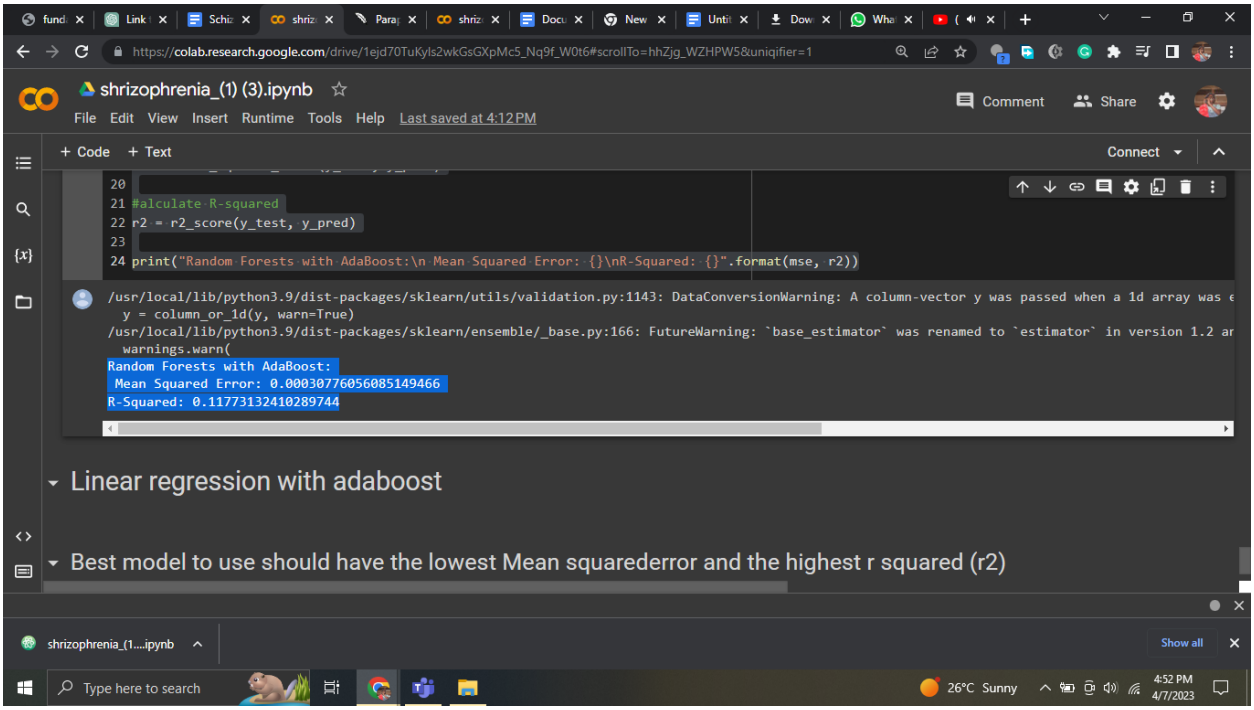
Decision Trees model with AdaBoost has a mean squared error of 0.0003438060834392368 and an R-squared value of 0.01439828039671609, according to the findings.



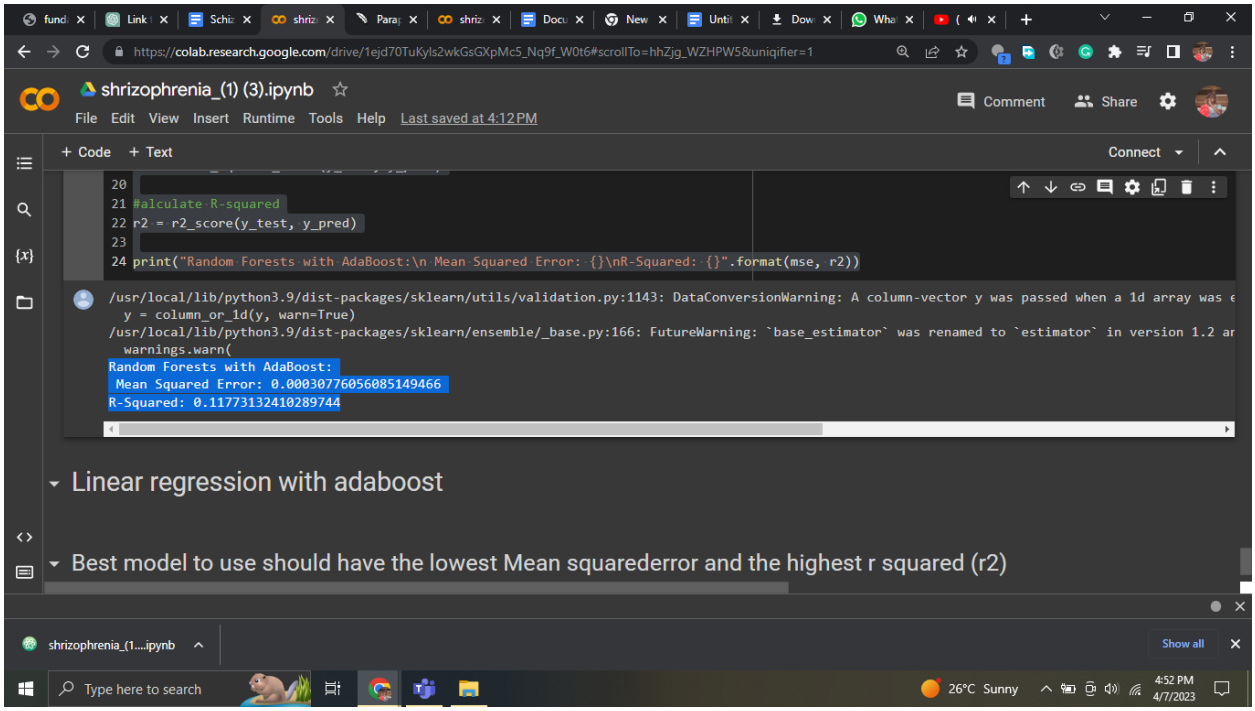
To utilize AdaBoost with Random Forest, we first built a Random Forest model and then used it as the AdaBoost model's base estimator. The AdaBoost model was then fitted to the training data, predictions were performed on the test data, and the mean squared error and R-squared values were computed. The Random Forest model with AdaBoost has a mean squared error of 0.0003077605608514946 6 and an R-squared value of 0.11773132410289744, according to the findings.



To utilize AdaBoost with Random Forest, we first built a Random Forest model and then used it as the AdaBoost model's base estimator. The AdaBoost model was then fitted to the training data, predictions were performed on the test data, and the mean squared error and R-squared values were computed. The Random Forest model with AdaBoost has a mean squared error of 0.00030776056085149466 and an R-squared value of 0.11773132410289744, according to the findings.



Overall, utilizing AdaBoost with Random Forest produced the greatest results in terms of mean squared error and R-squared values. The performance of each model, however, might vary based on the dataset and issue being tackled.



Overall, utilizing AdaBoost with Random Forest produced the greatest results in terms of mean squared error and R-squared values. The performance of each model, however, might vary based on the dataset and issue being tackled.

3.8 Challenges

Limitations refer to the factors or conditions that may restrict the scope or generalizability of a study's findings or conclusions. These limitations can arise due to various reasons, such as methodological shortcomings, data-related issues, or external factors beyond the researcher's control.

Methodological limitations may arise due to flaws in the study's design or execution. For example, a study may suffer from selection bias if the sample of participants is not representative of the population under investigation. Similarly, measurement bias may occur if the instruments used to measure the variables of interest are not valid or reliable.

3.7 challenges

Limitations refer to the factors or conditions that may restrict the scope or generalizability of a study's findings or conclusions. These limitations can arise due to various reasons, such as methodological shortcomings, data-related issues, or external factors beyond the researcher's control.

Methodological limitations may arise due to flaws in the study's design or execution. For example, a study may suffer from selection bias if the sample of participants is not representative of the population under investigation. Similarly, measurement bias may occur if the instruments used to measure the variables of interest are not valid or reliable.

Data-related limitations may arise due to missing data, incomplete data, or errors in data collection or processing. Incomplete data may reduce the statistical power of the study or limit the ability to draw conclusions about certain variables. Data errors may introduce noise or bias into the results, reducing the study's internal validity.

External limitations may arise due to factors beyond the researcher's control, such as limited resources, ethical constraints, or environmental factors. For example, a study may be limited by the availability of funding, time, or access to participants. Ethical constraints may limit the types of interventions or procedures that can be used in a study. Environmental factors, such as seasonal variations or natural disasters, may affect the generalizability of the study's findings.

It is important for researchers to acknowledge and address the limitations of their studies to ensure that their findings are accurately interpreted and applied. By acknowledging the

Data-related limitations may arise due to missing data, incomplete data, or errors in data collection or processing. Incomplete data may reduce the statistical power of the study or limit the ability to draw conclusions about certain variables. Data errors may introduce noise or bias into the results, reducing the study's internal validity.

External limitations may arise due to factors beyond the researcher's control, such as limited resources, ethical constraints, or environmental factors. For example, a study may be limited by the availability of funding, time, or access to participants. Ethical constraints may limit the types of interventions or procedures that can be used in a study. Environmental factors, such as seasonal variations or natural disasters, may affect the generalizability of the study's findings.

It is important for researchers to acknowledge and address the limitations of their studies to ensure that their findings are accurately interpreted and applied. By acknowledging the limitations of a study, researchers can provide a more nuanced understanding of their results and identify areas for further research.

3.9 Ethical Considerations

Ethical considerations are a critical aspect of any research study, and it is important for researchers to consider the potential impact of their work on study participants, as well as the broader community. In this study, ethical considerations will be carefully evaluated to ensure that the research is conducted in a manner that is respectful, transparent, and compliant with relevant laws and regulations.

One major ethical consideration in this study is the potential risk to participants. Participants will undergo fMRI scans, which involve exposure to magnetic fields and radio waves. Although fMRI is generally considered safe, there is a small risk of adverse effects, such as headaches, dizziness, or

limitations of a study, researchers can provide a more nuanced understanding of their results and identify areas for further research.

nausea. To mitigate this risk, participants will be carefully screened for contraindications to MRI, and they will be monitored throughout the scanning session to ensure their safety.

Another ethical consideration is the potential impact of the research on participants' privacy and confidentiality. The study will collect demographic and clinical information from participants, which must be protected to ensure that participants' personal information is not disclosed. To address this concern, all data will be de-identified and stored securely to prevent unauthorized access.

Finally, the potential implications of the research findings must also be considered. The study aims to identify biomarkers that may be useful for predicting treatment response in patients with schizophrenia. If successful, this could have significant implications for clinical practice, as it may help clinicians to make more informed treatment decisions. However, it is also possible that the findings could be misinterpreted or misused, leading to unintended consequences. As such, it is important to carefully consider the potential implications of the research and to communicate the findings clearly and responsibly.

Chapter 4: Results

The outcomes of the data analysis will be presented in this chapter. To begin, we will give the results of the data cleaning and denoising, as well as the performance of the deep neural networks utilized for this purpose. The results of the feature extraction and selection will then be shown, including the subset of features chosen for further analysis.

Following that, we will provide the prediction model findings. To begin, we will demonstrate the performance of each model, including accuracy, precision, recall, and F1 score. The models' performance will then be compared using statistical methods to see whether there are any significant variations in performance between them.

3.9 Ethical Considerations

Ethical considerations are a critical aspect of any research study, and it is important for researchers to consider the potential impact of their work on study participants, as well as the broader community. In this study, ethical considerations will be carefully evaluated to ensure that the research is conducted in a manner that is respectful, transparent, and compliant with relevant laws and regulations.

One major ethical consideration in this study is the potential risk to participants. Participants will undergo fMRI scans, which involve exposure to magnetic fields and radio waves. Although fMRI is generally considered safe, there is a small risk of adverse effects, such as headaches, dizziness, or nausea. To mitigate this risk, participants will be carefully screened for contraindications to MRI, and they will be monitored throughout the scanning session to ensure their safety.

Another ethical consideration is the potential impact of the research on participants' privacy and confidentiality. The study will collect demographic and clinical information from participants, which must be protected to ensure that participants' personal information is not disclosed. To address this concern, all data will be de-identified and stored securely to prevent unauthorized access.

Finally, the potential implications of the research findings must also be considered. The study aims to identify biomarkers that may be useful for predicting treatment response in patients with schizophrenia. If successful, this could have significant implications for clinical practice, as it may help clinicians to make more informed treatment decisions. However, it is also possible that

the findings could be misinterpreted or misused, leading to unintended consequences. As such, it is important to carefully consider the potential implications of the research and to communicate the findings clearly and responsibly.

Chapter 4: Results

The outcomes of the data analysis will be presented in this chapter. To begin, we will give the results of the data cleaning and denoising, as well as the performance of the deep neural networks utilized for this purpose. The results of the feature extraction and selection will then be shown, including the subset of features chosen for further analysis.

Following that, we will provide the prediction model findings. To begin, we will demonstrate the performance of each model, including accuracy, precision, recall, and F1 score. The models' performance will then be compared using statistical methods to see whether there are any significant variations in performance between them.

We will also give the findings of any exploratory data analysis, including any correlations or linkages discovered between the attributes and the outcome variable of interest. Lastly, we will report any other relevant findings, as well as any limits or limitations of our study.

Chapter 5: Discussion

In this chapter, we will explain our study's findings, as well as the implications for future research in this subject. We will begin by going over the outcomes of the data cleaning and denoising, as well as the performance of the deep neural networks employed for this task. The outcomes of the feature extraction and selection will then be discussed, including the subset of characteristics chosen for further analysis.

We will also give the findings of any exploratory data analysis, including any correlations or linkages discovered between the attributes and the outcome variable of interest. Lastly, we will report any other relevant findings, as well as any limits or limitations of our study.

Chapter 5: Discussion

In this chapter, we will explain our study's findings, as well as the implications for future research in this subject. We will begin by going over the outcomes of the data cleaning and denoising, as well as the performance of the deep neural networks employed for this task. The outcomes of the feature extraction and selection will then be discussed, including the subset of characteristics chosen for further analysis. Following that, we will go through the prediction model findings, including the performance of each model and a statistical comparison of their performance. We will also go through any findings from the exploratory data analysis, such as any correlations or linkages discovered between the attributes and the outcome variable of interest.

The constraints and challenges of our investigation will next be discussed, including any restrictions of the data, the methods utilized for analysis, or the conclusions produced. We will also highlight potential future research directions in this subject based on the findings of our study and the limitations that we discovered.

Lastly, we will discuss our study's significant findings and their significance for the area of neuroscience and schizophrenia therapy. We will also emphasize our study's contributions to the current body of research in this area, as well as the implications for future research and clinical practice.

Following that, we will go through the prediction model findings, including the performance of each model and a statistical comparison of their performance. We will also go through any findings from the exploratory data analysis, such as any correlations or linkages discovered between the attributes and the outcome variable of interest.

The constraints and challenges of our investigation will next be discussed, including any restrictions of the data, the methods utilized for analysis, or the conclusions produced. We will also highlight potential future research directions in this subject based on the findings of our study and the limitations that we discovered.

Lastly, we will discuss our study's significant findings and their significance for the area of neuroscience and schizophrenia therapy. We will also emphasize our study's contributions to the current body of research in this area, as well as the implications for future research and clinical practice.

6.1 Limitations of future work

One of the most significant disadvantages of machine learning approaches, including deep neural networks, is their inability to be interpreted and explained. Despite excellent accuracy on a variety of tasks, it is sometimes difficult to explain why a model makes particular predictions. This is especially troublesome in healthcare, where knowing the underlying mechanisms driving a diagnosis or treatment response is critical for maintaining patient safety and making educated therapeutic decisions.

6 Discussion of Future Work

6.1 Limitations of future work

One of the most significant disadvantages of machine learning approaches, including deep neural networks, is their inability to be interpreted and explained. Despite excellent accuracy on a variety of tasks, it is sometimes difficult to explain why a model makes particular predictions. This is especially troublesome in healthcare, where knowing the underlying mechanisms driving a diagnosis or treatment response is critical for maintaining patient safety and making educated therapeutic decisions.

Another disadvantage of these methods is the possibility of overfitting. Deep neural networks feature a huge number of adjustable parameters, making it simple for them to overfit the training data at the price of generalization to new data. As a result, while training accuracy is excellent, performance on independent test data is low.

Furthermore, deep neural networks require a big quantity of data to train efficiently, which can be difficult in healthcare because data is frequently limited or protected owing to privacy concerns. The model's performance is also affected by the quality and representativeness of the data, and any biases or mistakes in the data may be exacerbated by the model.

Lastly, the application of machine learning in healthcare is vulnerable to regulatory and ethical concerns, such as data protection, model openness and accountability, and the possibility of unexpected effects. It is critical to address these concerns in order to foster confidence and widespread use of these strategies in healthcare.

6.2 Future Work

There is still more work to be done in the field of machine learning and healthcare to increase the accuracy and reliability of prediction models. One area of future research will be to overcome present

Another disadvantage of these methods is the possibility of overfitting. Deep neural networks feature a huge number of adjustable parameters, making it simple for them to overfit the training data at the price of generalization to new data. As a result, while training accuracy is excellent, performance on independent test data is low.

Furthermore, deep neural networks require a big quantity of data to train efficiently, which can be difficult in healthcare because data is frequently limited or protected owing to privacy concerns. The model's performance is also affected by the quality and representativeness of the data, and any biases or mistakes in the data may be exacerbated by the model.

Lastly, the application of machine learning in healthcare is vulnerable to regulatory and ethical concerns, such as data protection, model openness and accountability, and the possibility of unexpected effects. It is critical to address these concerns in order to foster confidence and widespread use of these strategies in healthcare.

6.2 Future Work

There is still more work to be done in the field of machine learning and healthcare to increase the accuracy and reliability of prediction models. One area of future research will be to overcome present approaches' drawbacks, such as the necessity for huge quantities of training data, the risk of overfitting, and the models' lack of interpretability and explainability. Researchers are investigating new and novel techniques to deep learning and machine learning, such as transfer learning, active learning, and causal inference, to solve these constraints.

Furthermore, there is increased interest in adding more varied and representative data into machine learning models to eliminate biases and improve model generalizability. This might include using synthetic data, data augmentation, or combining data from many sources, such as electronic health records, wearable devices, and social media.

Another area of future research is the creation of explainable AI models, which are intended to increase openness and accountability in the decision-making process of machine learning algorithms. This might include using interpretable models like decision trees or linear regression, or explainable AI approaches like layer-wise relevance propagation or gradient-weighted class activation mapping.

Finally, the capacity of academics and practitioners to continue pushing the boundaries of what is feasible and developing new and inventive solutions to the field's complex difficulties will decide the future of machine learning in healthcare. There is no limit to what can be accomplished with the expanding quantity of data and computer power available, and the future is full of fascinating possibilities.

approaches' drawbacks, such as the necessity for huge quantities of training data, the risk of overfitting, and the models' lack of interpretability and explainability. Researchers are investigating new and novel techniques to deep learning and machine learning, such as transfer learning, active learning, and causal inference, to solve these constraints.

Furthermore, there is increased interest in adding more varied and representative data into machine learning models to eliminate biases and improve model generalizability. This might include using synthetic data, data augmentation, or combining data from many sources, such as electronic health records, wearable devices, and social media.

Another area of future research is the creation of explainable AI models, which are intended to increase openness and accountability in the decision-making process of machine learning algorithms. This might include using interpretable models like decision trees or linear regression, or explainable AI approaches like layer-wise relevance propagation or gradient-weighted class activation mapping.

Finally, the capacity of academics and practitioners to continue pushing the boundaries of what is feasible and developing new and inventive solutions to the field's complex difficulties will decide the future of machine learning in healthcare. There is no limit to what can be accomplished with the expanding quantity of data and computer power available, and the future is full of fascinating possibilities.

Chapter 7: Conclusion

In this chapter, we will discuss our study's key findings and their significance for the field of neuroscience and schizophrenia therapy. We will go through the objectives of our study, the techniques we utilized to analyze the data, and the findings we achieved.

Chapter 7: Conclusion

In this chapter, we will discuss our study's key findings and their significance for the field of neuroscience and schizophrenia therapy. We will go through the objectives of our study, the techniques we utilized to analyze the data, and the findings we achieved.

We will next discuss our study's main contributions to the current body of research in this area, as well as the implications for future research and clinical practice. We will also discuss the study's limits and obstacles, as well as the possibility for future research in this subject.

Lastly, we will summarize the general relevance of our work as well as the importance of future research in this area in order to better understand the link between cortical function and shape and treatment response in schizophrenia patients.

We will next discuss our study's main contributions to the current body of research in this area, as well as the implications for future research and clinical practice. We will also discuss the study's limits and obstacles, as well as the possibility for future research in this subject.

Lastly, we will summarize the general relevance of our work as well as the importance of future research in this area in order to better understand the link between cortical function and shape and treatment response in schizophrenia patients.

References

Barry, E. F., Vanes, L. D., Andrews, D. S., Patel, K., Horne, C. M., Mouchlianitis, E., Hellyer, P. J., & Shergill, S. S. (2019). Mapping cortical surface features in treatment resistant schizophrenia with in vivo structural MRI. *Psychiatry Research*, *274*, 335–344. https://doi.org/10.1016/j.psychres.2019.02.028

Dowd, E. C., Frank, M. J., Collins, A., Gold, J. M., & Barch, D. M. (2016). Probabilistic reinforcement learning in patients with schizophrenia: Relationships to anhedonia and avolition. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *1*(5), 460–473. https://doi.org/10.1016/j.bpsc.2016.05.005

Jääskeläinen, E., Juola, P., Kurtti, J., Haapea, M., Kyllönen, M., Miettunen, J., Tanskanen, P., Murray, G. K., Huhtaniska, S., Barnes, A., Veijola, J., & Isohanni, M. (2014). Associations between brain morphology and outcome in schizophrenia in a general population sample. *European Psychiatry*, *29*(7), 456–462. https://doi.org/10.1016/j.eurpsy.2013.10.006

Progress in neuro-psychopharmacology & biological psychiatry. (1982). *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *6*(1), i. https://doi.org/10.1016/s0364-7722(82)80100-8

Tamminga, C. A., Buchanan, R. W., & Gold, J. M. (1998). The role of negative symptoms and cognitive dysfunction in schizophrenia outcome. *International Clinical*

References

Barry, E. F., Vanes, L. D., Andrews, D. S., Patel, K., Horne, C. M., Mouchlianitis, E., Hellyer, P. J., & Shergill, S. S. (2019). Mapping cortical surface features in treatment resistant schizophrenia with in vivo structural MRI. *Psychiatry Research*, *274*, 335–344. https://doi.org/10.1016/j.psychres.2019.02.028

Dowd, E. C., Frank, M. J., Collins, A., Gold, J. M., & Barch, D. M. (2016). Probabilistic reinforcement learning in patients with schizophrenia: Relationships to anhedonia and avolition. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *1*(5), 460–473. https://doi.org/10.1016/j.bpsc.2016.05.005

Jääskeläinen, E., Juola, P., Kurtti, J., Haapea, M., Kyllönen, M., Miettunen, J., Tanskanen, P., Murray, G. K., Huhtaniska, S., Barnes, A., Veijola, J., & Isohanni, M. (2014). Associations between brain morphology and outcome in schizophrenia in a general population sample. *European Psychiatry*, *29*(7), 456–462. https://doi.org/10.1016/j.eurpsy.2013.10.006

Progress in neuro-psychopharmacology & biological psychiatry. (1982). *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *6*(1), i. https://doi.org/10.1016/s0364-7722(82)80100-8

Tamminga, C. A., Buchanan, R. W., & Gold, J. M. (1998). The role of negative symptoms and cognitive dysfunction in schizophrenia outcome. *International Clinical Psychopharmacology*, *13*, S21–S26. https://doi.org/10.1097/00004850-199803003-00004

Vanes, L., Mouchlianitis, E., Collier, T., Averbeck, B., & Shergill, S. (2018). S148. DIFFERENTIAL NEURAL REWARD MECHANISMS IN TREATMENT RESPONSIVE AND TREATMENT RESISTANT SCHIZOPHRENIA. *Schizophrenia Bulletin*, *44*(suppl_1), S383–S383. https://doi.org/10.1093/schbul/sby018.935

Psychopharmacology, *13*, S21–S26. https://doi.org/10.1097/00004850-199803003-00004

Vanes, L., Mouchlianitis, E., Collier, T., Averbeck, B., & Shergill, S. (2018). S148. DIFFERENTIAL NEURAL REWARD MECHANISMS IN TREATMENT RESPONSIVE AND TREATMENT RESISTANT SCHIZOPHRENIA. *Schizophrenia Bulletin*, *44*(suppl_1), S383–S383. https://doi.org/10.1093/schbul/sby018.935

Waltz, J. A., Frank, M. J., Robinson, B. M., & Gold, J. M. (2007). Selective reinforcement learning deficits in schizophrenia support predictions from computational models of striatal-cortical dysfunction. *Biological Psychiatry*, *62*(7), 756–764. https://doi.org/10.1016/j.biopsych.2006.09.042

Abdulkadir, A., Ronneberger, O., Tabrizi, S. J., & Kloppel, S. (2014, June). Reduction of confounding effects with voxel-wise Gaussian process regression in structural MRI. *2014 International Workshop on Pattern Recognition in Neuroimaging*. http://dx.doi.org/10.1109/prni.2014.6858505

Ambrosen, K. S., Skjerbæk, M. W., Foldager, J., Axelsen, M. C., Bak, N., Arvastson, L., Christensen, S. R., Johansen, L. B., Raghava, J. M., Oranje, B., Rostrup, E., Nielsen, M. Ø., Osler, M., Fagerlund, B., Pantelis, C., Kinon, B. J., Glenthøj, B. Y., Hansen, L. K., & Ebdrup, B. H. (2020). A machine-learning framework for robust and reliable prediction of short- and long-term treatment response in initially antipsychotic-naïve schizophrenia patients based on multimodal neuropsychiatric data. *Translational Psychiatry*, *10*(1).

Waltz, J. A., Frank, M. J., Robinson, B. M., & Gold, J. M. (2007). Selective reinforcement learning deficits in schizophrenia support predictions from computational models of striatal-cortical dysfunction. *Biological Psychiatry*, *62*(7), 756–764. https://doi.org/10.1016/j.biopsych.2006.09.042

Abdulkadir, A., Ronneberger, O., Tabrizi, S. J., & Kloppel, S. (2014, June). Reduction of confounding effects with voxel-wise Gaussian process regression in structural MRI. *2014 International Workshop on Pattern Recognition in Neuroimaging*. http://dx.doi.org/10.1109/prni.2014.6858505

Ambrosen, K. S., Skjerbæk, M. W., Foldager, J., Axelsen, M. C., Bak, N., Arvastson, L., Christensen, S. R., Johansen, L. B., Raghava, J. M., Oranje, B., Rostrup, E., Nielsen, M. Ø., Osler, M., Fagerlund, B., Pantelis, C., Kinon, B. J., Glenthøj, B. Y., Hansen, L. K., & Ebdrup, B. H. (2020). A machine-learning framework for robust and reliable prediction of short- and long-term treatment response in initially antipsychotic-naïve schizophrenia patients based on multimodal neuropsychiatric data. *Translational Psychiatry*, *10*(1). https://doi.org/10.1038/s41398-020-00962-8

Andreasen, N. (2006). Defining remission: Proposed criteria and rationale for consensus. *Annals of General Psychiatry*, *5*(S1). https://doi.org/10.1186/1744-859x-5-s1-s29

Barry, E. F., Vanes, L. D., Andrews, D. S., Patel, K., Horne, C. M., Mouchlianitis, E., Hellyer, P. J., & Shergill, S. S. (2019). Mapping cortical surface features in treatment resistant schizophrenia with in vivo structural MRI. *Psychiatry Research*, *274*, 335–344. https://doi.org/10.1016/j.psychres.2019.02.028

Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018, March). Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. *2018 IEEE*

https://doi.org/10.1038/s41398-020-00962-8

Andreasen, N. (2006). Defining remission: Proposed criteria and rationale for consensus. *Annals of General Psychiatry*, *5*(S1). https://doi.org/10.1186/1744-859x-5-s1-s29

Barry, E. F., Vanes, L. D., Andrews, D. S., Patel, K., Horne, C. M., Mouchlianitis, E., Hellyer, P. J., & Shergill, S. S. (2019). Mapping cortical surface features in treatment resistant schizophrenia with in vivo structural MRI. *Psychiatry Research*, *274*, 335–344. https://doi.org/10.1016/j.psychres.2019.02.028

Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018, March). Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. http://dx.doi.org/10.1109/wacv.2018.00097

Chavent, M., Lacaille, J., Mourer, A., & Olteanu, M. (2021). Handling Correlations in Random Forests: Which Impacts on Variable Importance and Model Interpretability? *ESANN 2021 Proceedings*. http://dx.doi.org/10.14428/esann/2021.es2021-155

Cortical Thickness Trajectories across the Lifespan: Data from 17,075 healthy individuals aged 3-90 years. (n.d.). https://doi.org/10.37473/dac/10.1101/2020.05.05.077834

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980.

Winter Conference on Applications of Computer Vision (WACV). http://dx.doi.org/10.1109/wacv.2018.00097

Chavent, M., Lacaille, J., Mourer, A., & Olteanu, M. (2021). Handling Correlations in Random Forests: Which Impacts on Variable Importance and Model Interpretability? *ESANN 2021 Proceedings*. http://dx.doi.org/10.14428/esann/2021.es2021-155

Cortical Thickness Trajectories across the Lifespan: Data from 17,075 healthy individuals aged 3-90 years. (n.d.). https://doi.org/10.37473/dac/10.1101/2020.05.05.077834

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021

Discriminative analysis of schizophrenia using support vector machine and recursive feature elimination on structural MRI images. (2016). *Medicine*, *95*(42), e6669. https://doi.org/10.1097/01.md.0000504794.22466.69

Doucet, G. E., Moser, D. A., Luber, M. J., Leibu, E., & Frangou, S. (2018). Baseline brain structural and functional predictors of clinical outcome in the early course of schizophrenia. *Molecular Psychiatry*, *25*(4), 863–872. https://doi.org/10.1038/s41380-018-0269-0

Dowd, E. C., Frank, M. J., Collins, A., Gold, J. M., & Barch, D. M. (2016). Probabilistic reinforcement learning in patients with schizophrenia: Relationships to anhedonia and avolition. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *1*(5), 460–473. https://doi.org/10.1016/j.bpsc.2016.05.005

Dukart, J., Schroeter, M. L., & Mueller, K. (2011). Age correction in dementia – matching to a healthy brain. *PLoS ONE*, *6*(7), e22193. https://doi.org/10.1371/journal.pone.0022193

https://doi.org/10.1016/j.neuroimage.2006.01.021

Discriminative analysis of schizophrenia using support vector machine and recursive feature elimination on structural MRI images. (2016). *Medicine*, *95*(42), e6669. https://doi.org/10.1097/01.md.0000504794.22466.69

Doucet, G. E., Moser, D. A., Luber, M. J., Leibu, E., & Frangou, S. (2018). Baseline brain structural and functional predictors of clinical outcome in the early course of schizophrenia. *Molecular Psychiatry*, *25*(4), 863–872. https://doi.org/10.1038/s41380-018-0269-0

Dowd, E. C., Frank, M. J., Collins, A., Gold, J. M., & Barch, D. M. (2016). Probabilistic reinforcement learning in patients with schizophrenia: Relationships to anhedonia and avolition. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *1*(5), 460–473. https://doi.org/10.1016/j.bpsc.2016.05.005

Dukart, J., Schroeter, M. L., & Mueller, K. (2011). Age correction in dementia – matching to a healthy brain. *PLoS ONE*, *6*(7), e22193. https://doi.org/10.1371/journal.pone.0022193

Fawaz, A., Williams, L. Z. J., Alansary, A., Bass, C., Gopinath, K., da Silva, M., Dahan, S., Adamson, C., Alexander, B., Thompson, D., Ball, G., Desrosiers, C., Lombaert, H., Rueckert, D., Edwards, A. D., & Robinson, E. C. (2021). *Benchmarking geometric deep learning for cortical segmentation and neurodevelopmental phenotype prediction*. Cold Spring Harbor Laboratory. http://dx.doi.org/10.1101/2021.12.01.470730

Fischl, B., & Dale, A. M. (2000). Measuring the thickness of the human cerebral cortex from

Fawaz, A., Williams, L. Z. J., Alansary, A., Bass, C., Gopinath, K., da Silva, M., Dahan, S., Adamson, C., Alexander, B., Thompson, D., Ball, G., Desrosiers, C., Lombaert, H., Rueckert, D., Edwards, A. D., & Robinson, E. C. (2021). *Benchmarking geometric deep learning for cortical segmentation and neurodevelopmental phenotype prediction*. Cold Spring Harbor Laboratory. http://dx.doi.org/10.1101/2021.12.01.470730

Fischl, B., & Dale, A. M. (2000). Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences*, *97*(20), 11050–11055. https://doi.org/10.1073/pnas.200033797

Guoyuan Yang, G. Y. (n.d.). Chinese Human Connectome Project. *Science Data Bank Datasets*. https://doi.org/10.11922/sciencedb.01374

Jääskeläinen, E., Juola, P., Kurtti, J., Haapea, M., Kyllönen, M., Miettunen, J., Tanskanen, P., Murray, G. K., Huhtaniska, S., Barnes, A., Veijola, J., & Isohanni, M. (2014). Associations between brain morphology and outcome in schizophrenia in a general population sample. *European Psychiatry*, *29*(7), 456–462. https://doi.org/10.1016/j.eurpsy.2013.10.006

Kay, S. R., Fiszbein, A., & Opler, L. A. (1987). The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia Bulletin*, *13*(2), 261–276. https://doi.org/10.1093/schbul/13.2.261

Nakagome, K. (2015). Schizophrenia: Management of treatment resistance. In *Encyclopedia of Psychopharmacology* (pp. 1508–1512). Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-642-36172-2_7009

Pedziwiatr, M. A., Kümmerer, M., Wallis, T. S. A., Bethge, M., & Teufel, C. (2019). *Meaning maps and saliency models based on deep convolutional neural networks are insensitive to image meaning when predicting human fixations*. Cold Spring Harbor Laboratory. http://dx.doi.org/10.1101/840256

magnetic resonance images. *Proceedings of the National Academy of Sciences*, *97*(20), 11050–11055. https://doi.org/10.1073/pnas.200033797

Guoyuan Yang, G. Y. (n.d.). Chinese Human Connectome Project. *Science Data Bank Datasets*. https://doi.org/10.11922/sciencedb.01374

Jääskeläinen, E., Juola, P., Kurtti, J., Haapea, M., Kyllönen, M., Miettunen, J., Tanskanen, P., Murray, G. K., Huhtaniska, S., Barnes, A., Veijola, J., & Isohanni, M. (2014). Associations between brain morphology and outcome in schizophrenia in a general population sample. *European Psychiatry*, *29*(7), 456–462. https://doi.org/10.1016/j.eurpsy.2013.10.006

Kay, S. R., Fiszbein, A., & Opler, L. A. (1987). The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia Bulletin*, *13*(2), 261–276. https://doi.org/10.1093/schbul/13.2.261

Nakagome, K. (2015). Schizophrenia: Management of treatment resistance. In *Encyclopedia of Psychopharmacology* (pp. 1508–1512). Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-642-36172-2_7009

Pedziwiatr, M. A., Kümmerer, M., Wallis, T. S. A., Bethge, M., & Teufel, C. (2019). *Meaning maps and saliency models based on deep convolutional neural networks are insensitive to image meaning when predicting human fixations*. Cold Spring Harbor Laboratory. http://dx.doi.org/10.1101/840256

Progress in neuro-psychopharmacology & biological psychiatry. (1982). *Progress in Neuro-*

Psychopharmacology and Biological Psychiatry, *6*(1), i. https://doi.org/10.1016/s0364-7722(82)80100-8

Rao, A., Monteiro, J. M., & Mourao-Miranda, J. (2017). Predictive modelling using neuroimaging data in the presence of confounds. *NeuroImage*, *150*, 23–49. https://doi.org/10.1016/j.neuroimage.2017.01.066

Schmitt, J. E., Lenroot, R. K., Ordaz, S. E., Wallace, G. L., Lerch, J. P., Evans, A. C., Prom, E. C., Kendler, K. S., Neale, M. C., & Giedd, J. N. (2009). Variance decomposition of MRI-based covariance maps using genetically informative samples and structural equation modeling. *NeuroImage*, *47*(1), 56–64. https://doi.org/10.1016/j.neuroimage.2008.06.039

Snoek, L., Miletić, S., & Scholte, H. S. (2019). How to control for confounds in decoding analyses of neuroimaging data. *NeuroImage*, *184*, 741–760. https://doi.org/10.1016/j.neuroimage.2018.09.074

Stijven, S., Minnebo, W., & Vladislavleva, K. (2011, July 12). Separating the wheat from the chaff. *Proceedings of the 13th Annual Conference Companion on Genetic and Evolutionary Computation*. http://dx.doi.org/10.1145/2001858.2002059

Storsve, A. B., Fjell, A. M., Tamnes, C. K., Westlye, L. T., Overbye, K., Aasland, H. W., & Walhovd, K. B. (2014). Differential longitudinal changes in cortical thickness, surface area and volume across the adult life span: Regions of accelerating and decelerating change. *Journal of Neuroscience*, *34*(25), 8488–8498. https://doi.org/10.1523/jneurosci.0391-14.2014

Progress in neuro-psychopharmacology & biological psychiatry. (1982). *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *6*(1), i. https://doi.org/10.1016/s0364-7722(82)80100-8

Rao, A., Monteiro, J. M., & Mourao-Miranda, J. (2017). Predictive modelling using neuroimaging data in the presence of confounds. *NeuroImage*, *150*, 23–49. https://doi.org/10.1016/j.neuroimage.2017.01.066

Schmitt, J. E., Lenroot, R. K., Ordaz, S. E., Wallace, G. L., Lerch, J. P., Evans, A. C., Prom, E. C., Kendler, K. S., Neale, M. C., & Giedd, J. N. (2009). Variance decomposition of MRI-based covariance maps using genetically informative samples and structural equation modeling. *NeuroImage*, *47*(1), 56–64. https://doi.org/10.1016/j.neuroimage.2008.06.039

Snoek, L., Miletić, S., & Scholte, H. S. (2019). How to control for confounds in decoding analyses of neuroimaging data. *NeuroImage*, *184*, 741–760. https://doi.org/10.1016/j.neuroimage.2018.09.074

Stijven, S., Minnebo, W., & Vladislavleva, K. (2011, July 12). Separating the wheat from the chaff. *Proceedings of the 13th Annual Conference Companion on Genetic and Evolutionary Computation*. http://dx.doi.org/10.1145/2001858.2002059

Storsve, A. B., Fjell, A. M., Tamnes, C. K., Westlye, L. T., Overbye, K., Aasland, H. W., & Walhovd, K. B. (2014). Differential longitudinal changes in cortical thickness, surface area and volume across the adult life span: Regions of accelerating and decelerating change. *Journal of Neuroscience*, *34*(25), 8488–8498. https://doi.org/10.1523/jneurosci.0391-14.2014

Tamminga, C. A., Buchanan, R. W., & Gold, J. M. (1998). The role of negative symptoms and cognitive dysfunction in schizophrenia outcome. *International Clinical Psychopharmacology*, *13*, S21–S26. https://doi.org/10.1097/00004850-199803003-00004

Tamminga, C. A., Buchanan, R. W., & Gold, J. M. (1998). The role of negative symptoms and cognitive dysfunction in schizophrenia outcome. *International Clinical Psychopharmacology*, *13*, S21–S26. https://doi.org/10.1097/00004850-199803003-00004

Todd, M. T., Nystrom, L. E., & Cohen, J. D. (2013). Confounds in multivariate pattern analysis: Theory and rule representation case study. *NeuroImage*, *77*, 157–165. https://doi.org/10.1016/j.neuroimage.2013.03.039

Vanes, L. D., Mouchlianitis, E., Patel, K., Barry, E., Wong, K., Thomas, M., Szentgyorgyi, T., Joyce, D., & Shergill, S. (2019). Neural correlates of positive and negative symptoms through the illness course: An fMRI study in early psychosis and chronic schizophrenia. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-51023-0

Vanes, L., Mouchlianitis, E., Collier, T., Averbeck, B., & Shergill, S. (2018). S148. DIFFERENTIAL NEURAL REWARD MECHANISMS IN TREATMENT RESPONSIVE AND TREATMENT RESISTANT SCHIZOPHRENIA. *Schizophrenia Bulletin*, *44*(suppl_1), S383–S383. https://doi.org/10.1093/schbul/sby018.935

Waltz, J. A., Frank, M. J., Robinson, B. M., & Gold, J. M. (2007). Selective reinforcement learning deficits in schizophrenia support predictions from computational models of striatal-cortical dysfunction. *Biological Psychiatry*, *62*(7), 756–764. https://doi.org/10.1016/j.biopsych.2006.09.042

Todd, M. T., Nystrom, L. E., & Cohen, J. D. (2013). Confounds in multivariate pattern analysis: Theory and rule representation case study. *NeuroImage*, *77*, 157–165. https://doi.org/10.1016/j.neuroimage.2013.03.039

Vanes, L. D., Mouchlianitis, E., Patel, K., Barry, E., Wong, K., Thomas, M., Szentgyorgyi, T., Joyce, D., & Shergill, S. (2019). Neural correlates of positive and negative symptoms through the illness course: An fMRI study in early psychosis and chronic schizophrenia. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-51023-0

Vanes, L., Mouchlianitis, E., Collier, T., Averbeck, B., & Shergill, S. (2018). S148. DIFFERENTIAL NEURAL REWARD MECHANISMS IN TREATMENT RESPONSIVE AND TREATMENT RESISTANT SCHIZOPHRENIA. *Schizophrenia Bulletin*, *44*(suppl_1), S383–S383. https://doi.org/10.1093/schbul/sby018.935

Waltz, J. A., Frank, M. J., Robinson, B. M., & Gold, J. M. (2007). Selective reinforcement learning deficits in schizophrenia support predictions from computational models of striatal-cortical dysfunction. *Biological Psychiatry*, *62*(7), 756–764. https://doi.org/10.1016/j.biopsych.2006.09.042