Suicide Within One Year After Discharge Among Patients Diagnosed With a Mental Disorder

Abstract

Mental health disorders continue to pose significant challenges to individuals and healthcare systems around the world. A critical aspect of mental health care is understanding the risk of suicide among patients diagnosed with mental disorders. This project aims to explore the patterns and risk factors associated with suicide within a year of discharge among this vulnerable patient population.

The study that will be used is a comprehensive set of data from the OECD (Organization for Economic Co-operation and Development) that contains health quality indicators, with a specific focus on mental health. By applying advanced data analysis techniques, including machine learning and predictive modeling, we seek to identify predictive factors and develop a robust predictive model for suicide risk. These factors may include patient demographics, diagnostic history, therapeutic interventions, and socioeconomic variables.

The expected result of this project is a predictive model that can assist healthcare professionals in assessing and addressing the risk of suicide among patients with mental disorders. Aiming to identify high-risk individuals and implement targeted interventions, with the aim of reducing the incidence of suicide, improving patient outcomes and contributing to the advancement of the quality of mental health care.

This research not only addresses a critical public health concern, but also seeks to show the potential of advanced data analytics, including machine learning, in improving healthcare decision-making and patient well-being.

Keywords: **Mental Health; Disorders; Analysis; Suicide; Pacients; Result**s

Introduction

Mental health disorders pose a significant global public health challenge, affecting millions of individuals. Within this challenge, the risk of suicide stands out as a distressing and complex issue, spanning various mental disorders. Patients discharged from healthcare facilities face a critical and vulnerable phase, with a high risk of suicide. Addressing this risk is essential for enhancing mental health care and patient well-being.

This project focuses on mental health care and suicide risk among patients diagnosed with mental disorders within one year of discharge. Our data source is the OECD, featuring diverse health quality indicators, including mental health metrics. Our goal is to uncover patterns and predictive factors, build a predictive model for suicide risk, considering patient demographics, clinical history, treatments, and socioeconomic factors.

The project follows a structured process, from data collection and preprocessing to exploratory analysis, model development, and comprehensive evaluation, with an emphasis on data privacy and ethics. The expected outcome is a tool for healthcare professionals to identify high-risk individuals early, potentially saving lives and enhancing the quality of care.

The project's methodology involves an examination of data management practices, including the study of distributed computing, architectural patterns, and the selection of suitable Big Data environments. Additionally, it entails an exploration of neural network applications, culminating in neural network selection for real-world problem-solving.

Objectives

Objective 1: Assessing the data storage and management requirements of a data project from the perspective of public health data management.

The objective of the project involves thoroughly reviewing a specific number of research articles, evaluating their methodologies and data sources, covering a range of publication years and geographic scopes to create a comprehensive overview of existing suicide research within a year of discharge among patients with mental disorders. disturbances.

Objective 2: Analyzing the design concepts, architectural patterns and technology stack of distributed Big Data systems necessary to provide insights useful for suicide prevention.

Identify and quantify key risk factors associated with suicide among discharged mental health patients, drawing insights from quantitative and qualitative research to understand the impact of these factors.

Objective 3: Evaluating and selecting a Big Data environment tailored to efficiently retrieve, process, and manage large dynamic datasets, considering scalability and processing speed.
Evidence-based recommendations and preventative measures are proposed based on research findings, addressing scalability and potential impact for mitigating suicide risk among discharged mental health patients.

Objective 4: Delving into practical applications, strengths and weaknesses of various types of neural networks, determining the most suitable neural network type for the specific suicide prevention problem in the public mental healthcare domain.

Focuses on designing data handling procedures, security measures, and compliance checklists to ensure secure management and privacy compliance of sensitive patient data in mental health care settings.
The **expected results** involve finding better data management solutions, choosing suitable neural networks for various uses, and sharing the findings for informed decisions. Overall,. this research aims to demonstrate the power of advanced data analysis in mental health, contributing to better mental health outcomes by understanding suicide risks.

Participants

Academic student: The study was conducted by a healthcare professional, studying MSc in Data Analytics SB+/FT. This research aims to identify predictive factors and develop a robust predictive model to prevent the risk of suicide in patients after the discharge period, as data shows that this is a complicated phase for these patients.

Project Management framework selected

In choosing the project management (PM) framework for the data analysis and predictive modeling, data security, and regulatory compliance phases of my project, as the timeframe is so short and the team is minimal, I have selected the Waterfall approach because it offers a more structured approach to the data preparation phase and to ensure compliance with healthcare data regulations.

For the other phases, such as data cleansing, and experimentation with different neural networks, Agile approach is better as it allows for flexibility, fast development, and collaboration between data scientists and healthcare experts. It is suitable for tasks that require continuous refinement and adaptation.

Data Science Framework Selected - CRISP-DM

By checking the different types of data science frameworks like Knowledge Discovery in Databases (KDD), SEMMA, and The Cross Industry Standard Process for Data Mining (CRISP-DM), the last one was selected because my project involves data mining and predictive modeling, which are core elements of CRISP-DM, provides a clear roadmap for tackling complex data analytics projects, which is especially valuable in healthcare analytics where data can be vast and multifaceted.

Phases of CRISP-DM:

Understanding the business: It started when I started studying to become a nurse, which allowed me to acquire in-depth knowledge of the healthcare area and the specific objectives of the project, such as mental health being important for all human beings.

Understanding the data: The data was collected on a specific research and data collection website, and this data was explored according to guidelines and using appropriate tools, seeking to achieve the result in a meaningful way.Analyzing data is crucial in healthcare, where data quality and availability are significant concerns.

Data preparation: Data cleaning, transformation and pre-processing of data in healthcare analytics is extremely important to ensure accurate and reliable results.

Modeling: I emphasize that the modeling phase is aligned with my project's focus on building predictive models for mental health indicators.

Evaluation: My objective and plan are to rigorously evaluate the performance of models, considering factors such as accuracy, precision, recall, and F1 score.

Deployment: CRISP-DM includes considerations for deploying models in a healthcare setting and ensuring they can be used effectively by healthcare professionals.

CRISP-DM is an iterative process, allowing for continuous refinement and improvement of models. This aligns well with the iterative and evolving nature of healthcare data projects.

CRISP-DM provides regulatory compliance in healthcare data analysis, and CRISP-DM includes provisions to ensure data privacy and adherence to regulations.

CRISP-DM promotes collaboration and communication between stakeholders, which is crucial in healthcare analytics projects involving data scientists, healthcare professionals, and decision makers.

CRISP-DM is recognized and respected in the healthcare sector, which makes it a reliable choice for your project.

Many data science companies implement CRISP-DM in their healthcare analytics projects or case studies due to its effectiveness for almost all projects.

CRISP-DM is adaptable and can be tailored to the unique requirements and constraints of my healthcare data project.

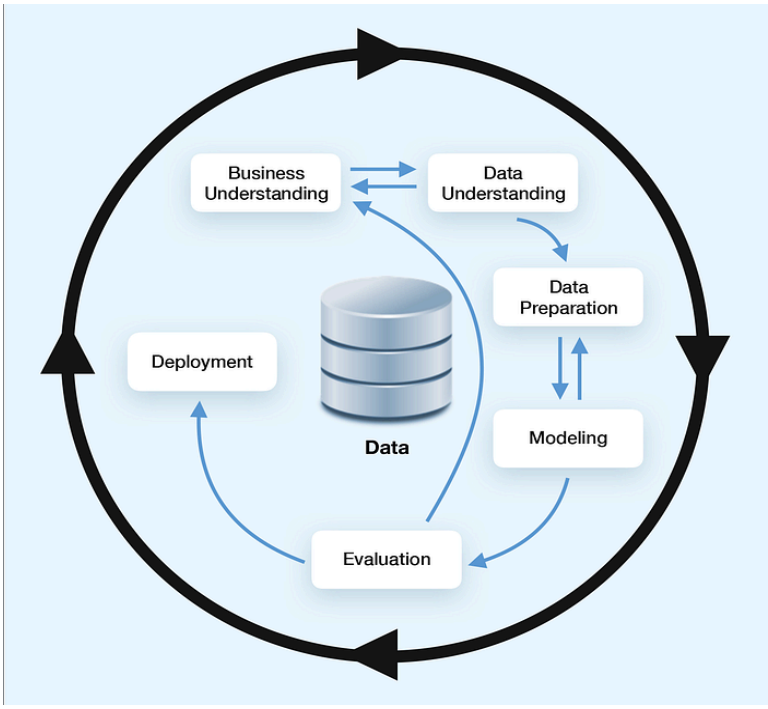

Figure 1. The Cross Industry Standard Process for Data Mining transitions between stages can be reversed. Source:https://www.the-modeling-agency.com/crisp-dm.pdf

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) in Big Data involves examining extensive suicide datasets to reveal patterns and insights using visual and statistical techniques. Given the large, diverse, and fast data, specialized tools are crucial in EDA for data preprocessing before advanced analytics.

Our research focuses on identifying statistically significant differences in suicide risk among patients with various mental disorders. This informs tailored intervention strategies. We analyze suicide risk across categories like depression, bipolar disorder, schizophrenia, and borderline personality disorder, using statistical tests to detect variations.

Significant differences can guide personalized interventions. Higher-risk categories may receive additional resources and specialized treatment, while lower-risk ones allow efficient resource allocation and preventive efforts. Our goal is to improve mental healthcare and suicide prevention by addressing specific disorder-related needs and risks, positively impacting individuals' well-being and safety.

In a graphical representation of the data I can color-code the different categories to differentiate between low, moderate, and high suicide risk categories. This graphical representation allows for a quick visual assessment of whether there are statistically significant differences in suicide risk between these categories. In my project, it was applied to compare suicide rates between countries using plots and graphs, aiming to expand dataset diversity without new data collection.

- After the exploratory data analysis the next phases were:
  Feature Engineering: Design relevant features from the data, such as demographic information, clinical scores, treatment adherence, and social factors.
- Machine Learning Models: Apply machine learning and predictive modeling techniques to the dataset. Choose an appropriate algorithm (e.g. logistic regression, decision trees, random forests or deep learning) for classification tasks.
- Model evaluation: Evaluate predictive models using appropriate metrics such as accuracy, precision, recall, F1 score, and ROC AUC.
- Interpretability and Explainability: Work to make your models interpretable and explainable to provide insights into the factors that contribute to predictions.
- Clinical Intervention: The predictions generated by your model will be used to identify high-risk patients.

Timeline (Gantt Chart)

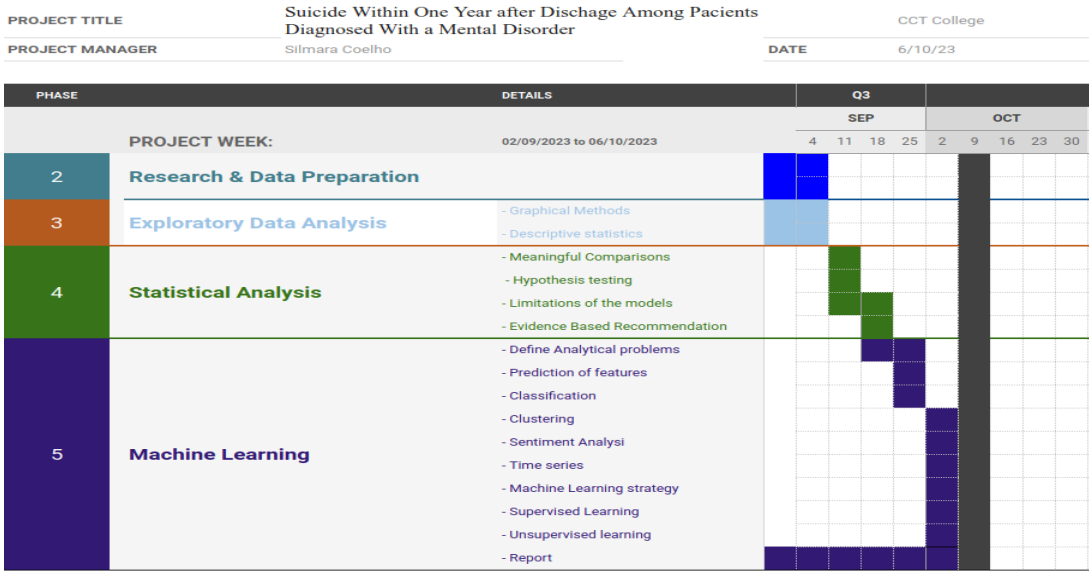

Figure 2. The project has a timeline of 4 weeks. More time should be invested in acquiring good quality extensive datasets from relevant sources, including government agencies, healthcare providers, and academic research, in order to assess availability, consistency, periodicity and predictive power.

Developing Environment - Colab

When selecting the development environment, I opted for ***Google Colab*** despite alternatives such as Local Jupyter Notebooks, Kaggle and Deep Note. The main justification for this decision was the limitation of time and human resources, combined with the fact that the project did not require extensive collaboration between members. Google Colab provides a suitable and efficient environment for individual work and data exploration. While other platforms may offer collaboration capabilities, I decided that the immediate needs called for no additional complexities or dependencies on external platforms.

Versioning

The project was developed in ***Google Colab***. In terms of Git versioning, the decision was to upload only finished versions of datasets and notebooks. The primary justification for this choice was the consideration of time and the perceived risk associated with mastering Git efficiently. Given the project's resource constraints, I decide to minimize the learning curve and potential errors that could arise from extensive version control management. Uploading only finished versions ensures that project progress and deliverables are monitored effectively, without the need for ongoing version control management throughout the development process.

Research Questions

Four Research Questions are proposed:

1. Are there statistically significant differences in suicide risk between patients with various types of mental disorders, and can these differences be used to inform targeted intervention strategies?

Yes, there are factors that clearly have some correlation with the incidence of suicides in different countries and over time

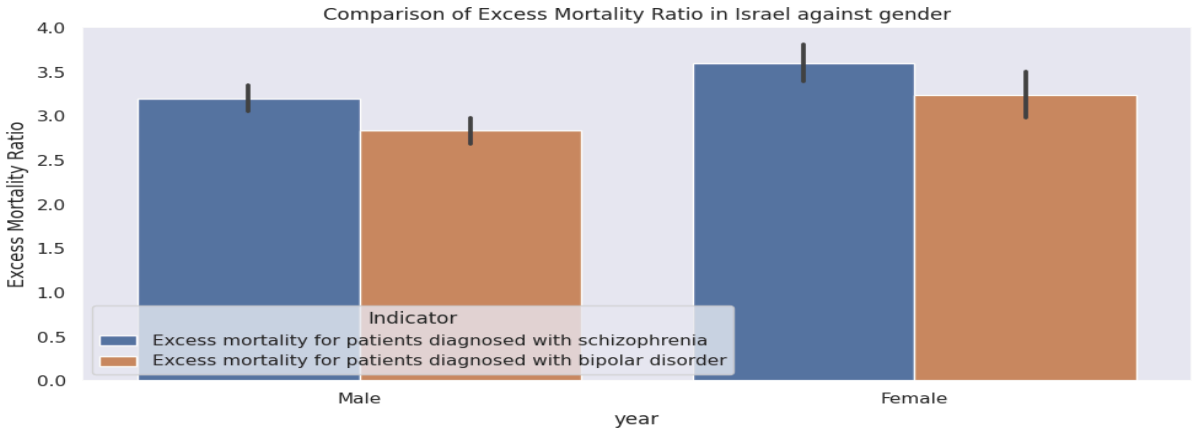

Figure 3. This graph shows the suicides rate, comparing gender and related to two medical diagnoses.

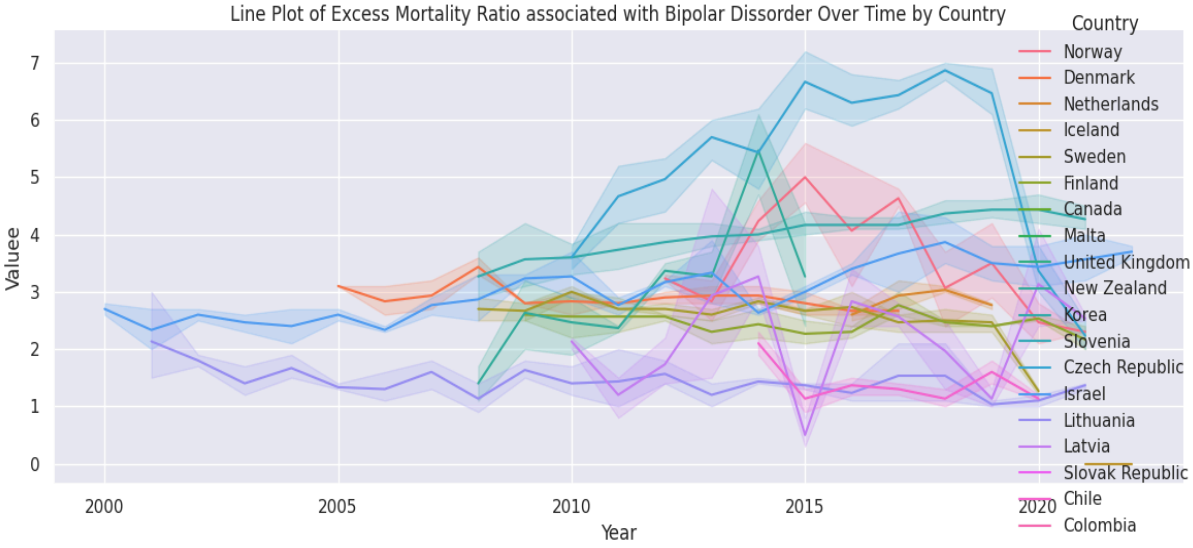

Figure 4. Evolution of different suicide rates in different countries over time.

2 What are the main factors, including demographic variables, that contribute to the risk of suicide (understood as a propensity score of the occurrence of suicide) within a year after discharge among patients diagnosed with mental disorders?

These factors include demographic variables (e.g., age, sex), clinical variables (e.g., mental health diagnosis, previous suicide attempts) and treatment-related variables. (e.g., treatment adherence, follow-up care) that are not immediately actionable.

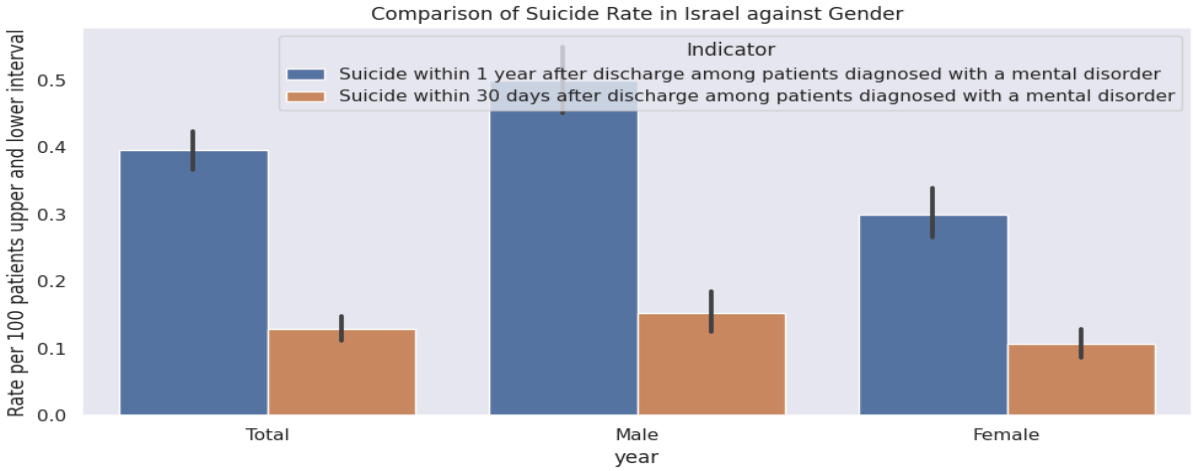

Figure 5. Comparison of suicide rate, segregated by sex, against the ammount of time since discharge.

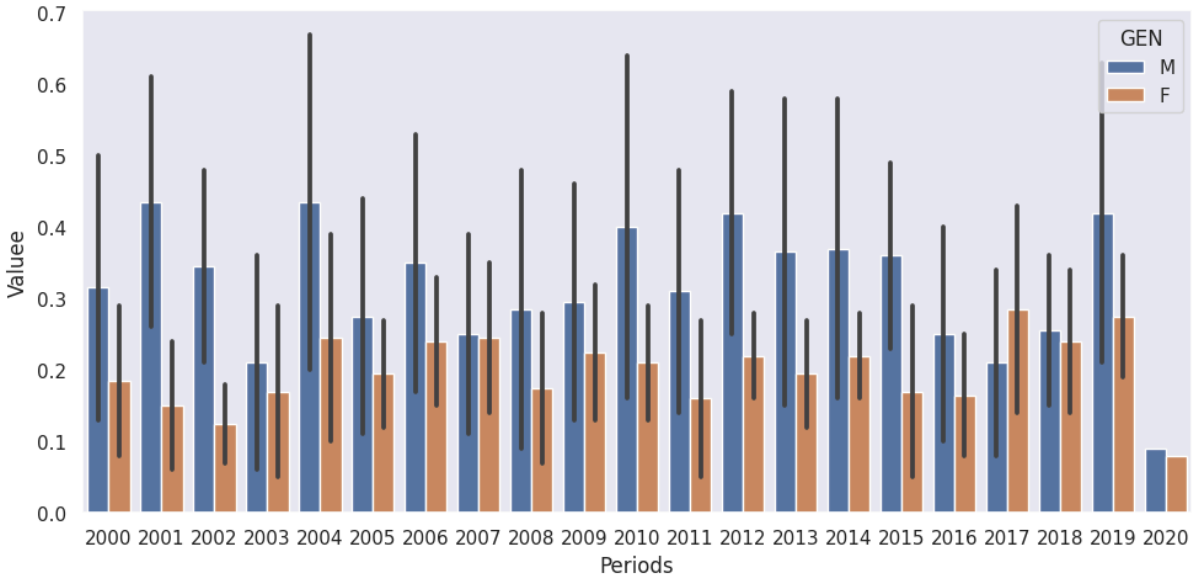

Figure 5. Suicide rates per 100.000 habitants over time, segregated by sex. It is clear that the rates are higher for male than for female individuals.

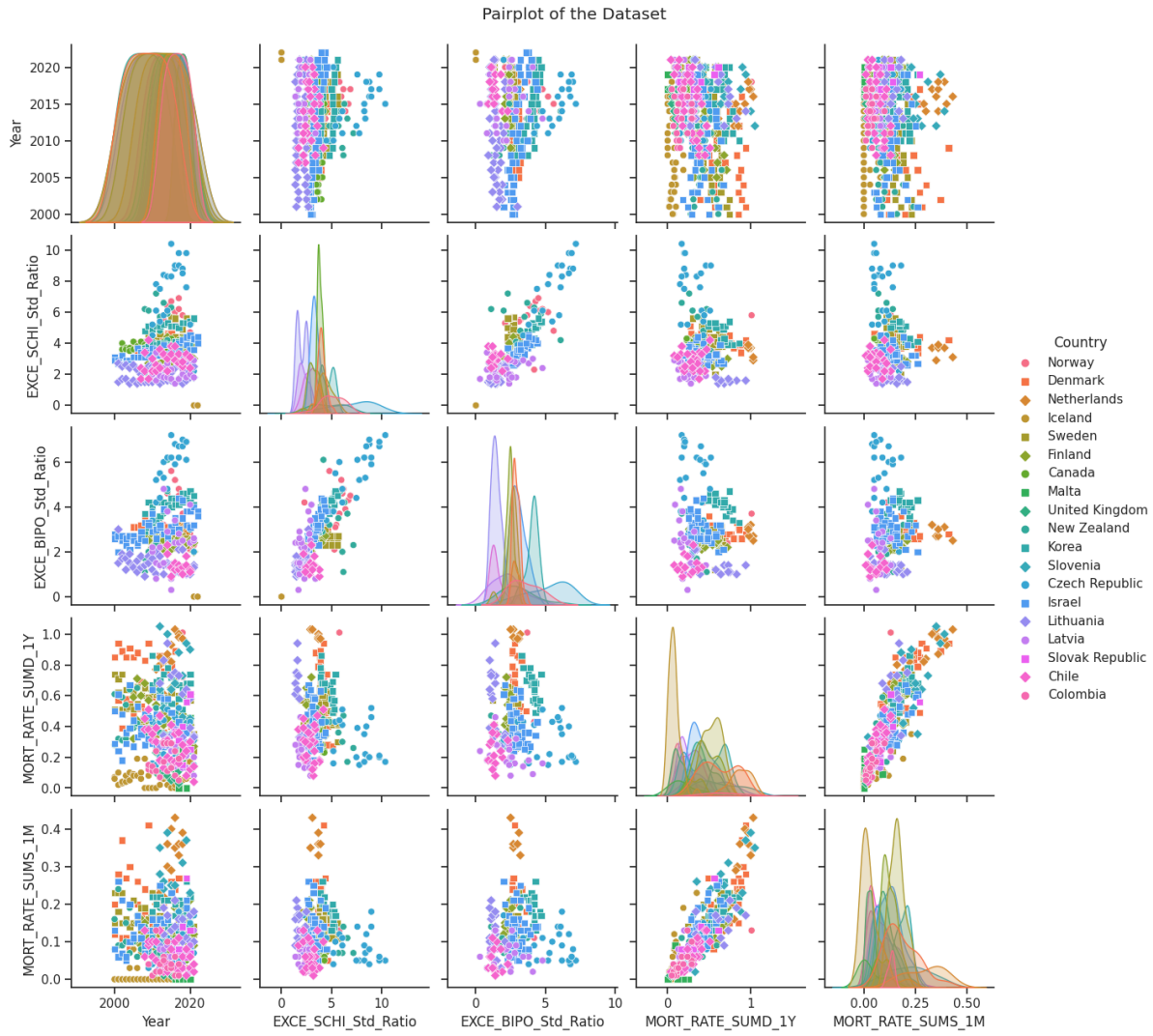

Figure 6. Comparison of the different suicide rates for different countries, and the excess mortality for individuals who committed suicide having a mental health disgnosis. The plot also takes into account the time after discharge. The country's colors are ordered by the GDP of the country. As the numbers of factors pile up, it is increasingly difficult to analyze graphically and it is necessary to model the behavior of the variables using advanced methods.

3. How do different features or indicators (clinical and psychosocial) related to mental health, such as psychiatric comorbidity (i.e., personality disorder and/or pain disorder), disorder severity, treatment adherence, or follow-up care, impact the risk of suicide after discharge?

These factors are the factors that may be ***directly actionable*** by the healthcare professionals to influence positively an individual's susceptibility to suicide. Understanding the impact of various characteristics or indicators related to mental health on the risk of suicide after

discharge is essential for suicide prevention efforts. The difficulty is in gathering, assessing and processing all this information in the amount of time necessary to be able to prepare a strategy of intervention and apply that strategy.

Here's an overview of how different factors can influence this:

Severity: Individuals with severe depression are often at greater risk of suicide after discharge. The intensity of depressive symptoms, including feelings of hopelessness and despair, can be a strong predictor. Individuals with severe symptoms or conditions that are difficult to treat may be at higher risk.

Previous Suicide Attempts: A history of previous suicide attempts is a strong predictor of future attempts. Individuals who have previously attempted suicide are at a higher risk.

Access to Lethal Means: Easy access to lethal means (e.g., firearms, medications) increases the risk of impulsive suicide attempts.

Co-occurring Substance Abuse: Substance abuse, including alcohol and drugs, can exacerbate mental health symptoms and increase the risk of suicide.

Social Isolation: A lack of social support and feelings of isolation can contribute to suicide risk, as individuals may not have a network to turn to in times of crisis.

Family History: A family history of suicide or mental health disorders can increase an individual's risk.

Loss of a Loved One: Recent loss, such as the death of a loved one, through suicide or otherwise, can trigger suicidal thoughts and actions.

Hopelessness: Feelings of hopelessness about the future and the belief that things will not improve can increase suicide risk.

Chronic Pain or Illness: Individuals dealing with chronic pain or serious illness may experience increased emotional distress, leading to higher suicide risk.

Financial Stress: Financial difficulties and job loss can lead to increased stress and hopelessness, contributing to suicide risk.

Access to Mental Health Care: Limited access to mental health care or stigmatization of mental health treatment can prevent individuals from seeking help.

Demographics: Certain demographic factors, such as age (e.g., elderly individuals may be at higher risk) and gender (e.g., men have higher rates of suicide completion), can influence suicide risk.

Recent Discharge from Treatment: Individuals recently discharged from psychiatric treatment may be vulnerable during the transition to outpatient care.

Medication Changes: Changes in psychiatric medication, especially abrupt discontinuation, can impact an individual's mental state.

Bullying and Discrimination: Experiences of bullying, discrimination, or victimization can contribute to feelings of despair.

Cultural and Religious Beliefs: Cultural and religious beliefs about suicide can either protect against or increase suicide risk, depending on individual interpretations.

Diagnosis	SMR (95% CI)	HR[a] (95% CI)
Psychotic disorder	13.03 (11.23–15.03)	4.16 (2.86–6.05)
Bipolar disorder	10.26 (7.97–13.00)	3.13 (2.07–4.75)
Substance related disorder	6.78 (4.14–10.47)	3.12 (1.83–5.55)
Depressive disorder	5.69 (4.78–6.73)	2.34 (1.60–3.42)
Unspecified mood disorder	4.64 (2.94–6.97)	1.92 (1.13–3.26)
Organic mental disorder	2.79 (1.98–3.81)	1.58 (1.03–2.42)
Sleep disorder	2.55 (1.49–4.09)	1.47 (0.82–2.64)
Anxiety disorder	2.45 (1.91–3.09)	1.05 (0.66–1.67)
Other	2.49 (1.72–3.48)	1

Table 1. Impact of different disorders in the excess mortality by suicide. All psychiatric patients are at a higher risk of suicide compared to the general population, and the risk is highest for those diagnosed with psychotic disorder. Source: www.ncbi.nlm.nih.gov

4. Can machine learning models effectively predict the risk of suicide among patients with specific mental disorders, if so, which models perform best in this context?

First we need to qualify and quantify that statement, and evaluate the feasibility of collecting the data about the suicide influence factors. In order to do this, first it is needed to do a data assessment and create a complete workflow for the project, with a test dataset. After evaluating all the performance of datasets, workflow and algorithms, we will conclude with this question.

Phase 1: Data Assessment

Assessments and Measures

Together, **assessments** ensure data quality, while **measurements** provide quantitative project performance insights, both crucial for success and informed decisions in big data projects.

Assessments involve systematically evaluating data quality, relevance, and performance to ensure it meets specific criteria like accuracy, integrity, timeliness, and reliability. These evaluations also include identifying data sources, profiling data, and evaluating data governance and security measures:

- **Data Completeness Assessment**: Check each variable for missing data. Assess the proportion of missing values and decide on how to handle them (e.g., imputation, removal, or special treatment).
- **Data Accuracy Assessment**: Examine data accuracy by comparing key variables (e.g., Age) against predefined ranges or expected values. Identify any outliers or errors.
- **Data Consistency Assessment**: Ensure consistency in categorical variables like 'Gender' and 'Diagnosis.' Check for any inconsistent or unexpected values.
- **Data Relevance Assessment**: Evaluate the relevance of each variable in the context of predicting suicide propensity. Remove or set aside variables that do not contribute to the research objectives.
- **Data Integrity Assessment**: Examine the overall integrity of the dataset. Check for any data entry errors, duplicates, or inconsistencies.
- **Data Distribution Assessment**: Analyze the distributions of numerical variables (e.g., 'Severity of Mental Disorder') to identify any significant deviations from expected patterns.
- **Temporal Assessment**: For variables involving dates ('Date of Diagnosis,' 'Date of Suicide'), check for chronological consistency and potential errors in temporal sequencing.
- **Text Data Assessment**: If handling text data ('Treatment History'), consider using natural language processing techniques to extract relevant information and assess its quality.
- **Binary Variable Assessment**: For binary variables ('Suicide Attempts,' 'Previous Suicide Attempts,' 'Recent Discharge from Treatment,' 'Current Co-occurring Substance Abuse,' 'Loss of a Loved One,' 'Hopelessness,' 'Chronic Pain or Illness,' 'Medication Changes,' 'Bullying and Discrimination,' 'Social Isolation,' 'Access to Mental Health Care,' 'Financial Stress'), validate that values are correctly coded as binary (e.g., 0 or 1).
- **Categorical Variable Assessment**: For categorical variables ('Gender,' 'Diagnosis,' 'Family History,' 'Access to Lethal Means,' 'History of Substance Abuse,' 'Cultural and Religious Beliefs,' 'Social Support,' 'Access to Healthcare,' 'Employment Status,' 'Geographic Location,' 'Suicide Method'), verify that categories are well-defined, mutually exclusive, and exhaustive.
- **Numerical Variable Assessment**: For numerical variables ('Age,' 'Severity of Mental Disorder,' 'Depression Severity,' 'Follow-up Period,' 'Financial Status'), examine summary statistics, including mean, median, and standard deviation, to identify potential outliers or inconsistencies.
- **Correlation Analysis**: Perform correlation analysis to identify relationships and dependencies between variables, especially between potential predictor variables and the target variable ('Suicide Propensity').
- **Data Imbalance Assessment**: If dealing with a classification problem for suicide propensity prediction, assess the balance between classes to determine if data balancing techniques are necessary.

Measurements are specific metrics or values used to quantify project performance and results. They include KPIs, data quality metrics, processing speed benchmarks, storage utilization statistics, and analysis-derived insights:

1. **Quality and quantity of Data Collection:** It includes patient records, clinical data, treatment history, follow-up information, and any other relevant variables that could be used to assess suicide risk among discharged patients.
2. **Relevant metrics:** Measure the rates, ratios and severity of Mental Disorder in the patient population at the beginning of the project.
    a. **1-Year Suicide Rate:** Calculate the percentage of patients who committed suicide within one year of discharge. A decreasing trend is positive (indicating better prevention efforts).
    b. **1-Month Suicide Rate.** Calculate the percentage of patients who committed suicide within one year of discharge. A decreasing trend is positive (indicating better prevention efforts).
    c. **Mental disorder Severity Score:** Use a validated rating scale (e.g., PHQ-9) to measure the severity of mental disorders in patients. Range 0 (no mental disorder) to 27 (severe mental disorder) using the PHQ-9 scale. A decreasing trend is positive (indicating reduced severity of mental disorder).
3. **Expected Value:**
    a. The Naive assumption is that the value of suicide rates will be affected by many confounding factors and may change unexpectedly. This implies that it is not possible to assert the effectiveness of preventative measures or treatment. However, we would like to see correlations in the suicide rates with any change in any other variable, even treatment.
    b. Aim for a specific reduction in Mental Disorder severity.
4. **6.Optimal value:**
    a. Ideal Suicide Rate: The lowest possible rate, ideally 0%.
    b. Ideal Mental Disorder Severity Score: The lowest score possible, ideally 0 on the PHQ-9 scale.
5. **Data Utilization and report:** Gather user feedback and satisfaction ratings to gauge how well the model and predictions meet the needs of stakeholders, including healthcare professionals and also I will describe how I planned and used the existing data in my project, describe the analysis methods and data cleaning processes, and specify how the metrics defined in my project will be reported.

Raw Data Gathering and Preparation

Datasets were gathered from the Organization for Economic Co-operation and Development (OECD)[1]. The data collects information about 19 countries that report the number of Suicide Within One Year After Discharge Among Patients Diagnosed With a Mental Disorder. Data was present in XML form, as a file from the OECD website. as a file from the website. The data carry a data dictionary defifinition , and it describes the meanings and purposes of data

[1] Available at https://stats.oecd.org/restsdmx/sdmx.ashx/GetDataStructure/HEALTH_HCQI

elements within the context of a project, and provides guidance on interpretation, accepted meanings and representations.

Assumptions

- Data Source Reliability: Assuming that the OECD, being a reputable international organization, provides reliable and accurate data on ***Suicide Within One Year After Discharge Among Patients Diagnosed With a Mental Disorder*** from various member countries.
- Consistent Data Collection: Assuming that the data on ***Suicide Within One Year After Discharge Among Patients Diagnosed With a Mental Disorder*** were collected using consistent methodologies and definitions across all countries, ensuring comparability and Maintain anonymity and data protection to maintain patient confidentiality..
- Mental health diagnoses: The dataset includes accurate and up-to-date information on mental health diagnoses such as depression, bipolar disorder, schizophrenia, etc.
- Treatment history: Assumption: The dataset does not contain records of patients' treatment history.
- Suicide Events: The dataset is composed of aggregated data from different countries about suicide incidence in mental disorder diagnosed populations. We assume that the way of measuring the populations, diagnosing, categorizing as mental disorders, computing and aggregating is comparable.
- Only 19 countries reporting: Assuming that the dataset covers a comprehensive range of OECD member countries, providing a not representative sample for analysis. As the data is very incomplete, statistical analysis techniques are not very indicated, we have to use Neural Networks

Interesting findings about the dataset are:

- Certain demographics are associated with higher or lower suicide rates, such as age, sex, race or socioeconomic status.
- Specific mental disorders are most strongly correlated with suicide risk.
- Incidence of suicide among mental health diagnosed patients is stable in most countries.

- Many countries have inconsistent reporting frequency. Many years the rates are zero (wich may be because of lack of data, not because lack of suicides), or the data is not existent.
- Some countries have had an increase of suicide rates in recent years, some have subsided, but only in the last year. That makes me suspect of incomplete data in last year's sample.
- Data is aggregated by country, by year, and disaggregated by sex, some mental health disorders and time that have passed since dismissal. For a better dataset, data has to be collected about patients that have been dismissed, but have not committed suicide. Also, data about mental health diagnosed people that have not received treatment and are not considered patients. If possible, in the more disaggregated form.
- Data lacks a lot of desirable clinical, socioeconomic and psychiatric features.

Find Problems and Cleaning

This process involves identifying and resolving issues, such as relevance, errors, missing values, and inconsistencies, in a data set to ensure data accuracy and quality for analysis and decision making. This dataset did not have too many problems besides being very aggregated, having many missing data points and being very limited in the number of valuable variables.

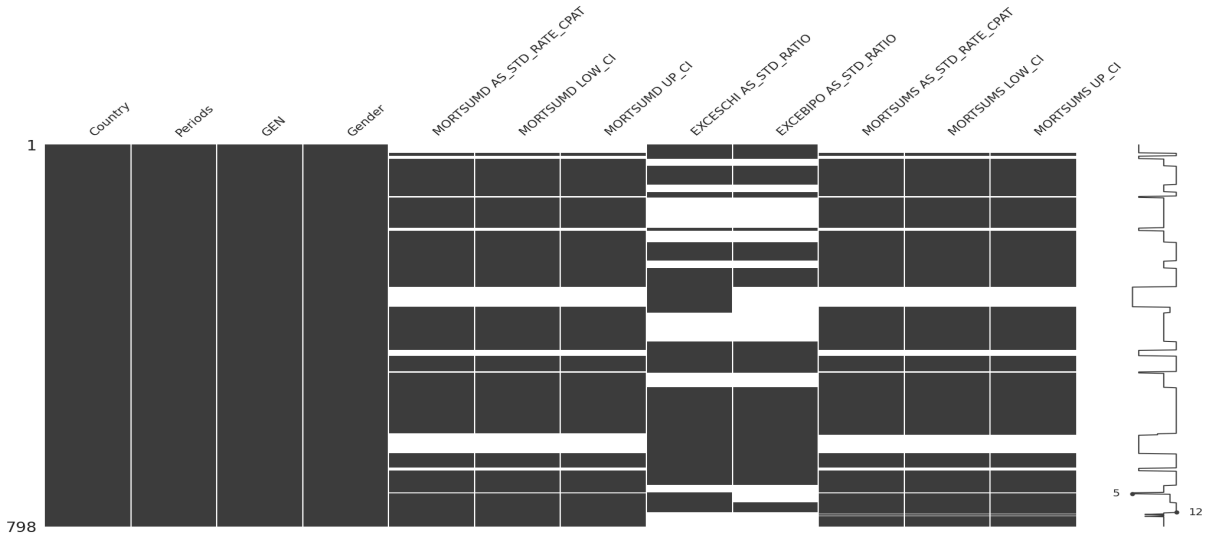

Figure 7. Matrix chart was used to visualize and analyze data to display missing data patterns in the dataset. In this chart, each column represents a variable, each line represents an observation, and white lines indicate missing values. Its purpose is to quickly identify which variables have missing data and whether there are patterns or groups of missing values in my dataset, aiding in data pre-processing and analysis.

Handling Outliers

Addressing outliers in my suicide risk prediction project involves identifying and managing extreme data points that can impact the accuracy and ethical soundness of your analysis or predictive model. This process ensures that unusual data does not unduly influence your project results. In Figure 6 it is possible to see that there are not many outliers and the data is very clean.

Data Augmentation and enrichment

Data augmentation enhances dataset quality and diversity, often used in machine learning and computer vision. Data enrichment supplements datasets with external sources for deeper insights, benefiting data understanding and prediction. These techniques aid model training and generalization, especially in computer vision, natural language processing, and time series analysis, reducing overfitting, improving robustness, and enhancing predictive models across tasks like image and text augmentation and temporal data expansion.

With the original dataset we did not do much data augmentation/ enrichment. Only encoding, pivoting and labeling. but essentially it is the same data. However, I identified that we need a better dataset that would be almost impossible to gather with the time available. This dataset was constructed with fake data, in a synthetic way. It is described in table 2. The idea is that the synthetic dataset would be gathered live with each patient in somewhat real time in each intervention or attention by healthcare professionals.

Data Processing to Structured Form

Data processing to structured form in a Big Data project involves the conversion of raw and unstructured data into an organized, standardized format that is accurate, consistent, and ready for efficient analysis and decision-making, enhancing data quality and facilitating insights extraction.

The output dataset, resulting from column rotation and encoding, has a structured numeric representation of the original data, ready for efficient machine learning model training. It improves model compatibility and performance, making it valuable for predictive analytics.

In this project two datasets were collected: one from the OCDE and one synthetic. The main changes made were pivoting columns and rows, and encoding to feed to neural networks.

- **HEALTH_HCQI_.csv** A dataset with aggregated rates by country
- **HEALTH_SU_Prop_Assess.csv** An enriched dataset with needed variables (See Annex 1).

Dataset Name: Mental Health and Suicide Propensity Dataset
This dataset aims to provide insights into suicide propensity among individuals diagnosed with various mental health disorders. It combines data from multiple sources:

- Primary Healthcare Centers: Data from healthcare facilities where patients were diagnosed and received treatment.
- National Health Agencies: Aggregated data from national health agencies and databases.
- Mental Health Institutions: Data from specialized mental health institutions.
- Patient Surveys: Patient-reported data collected through surveys and interviews.

Volume of Data: The dataset should include a substantial number of patients to ensure statistical significance. A minimum of several thousand patient records is recommended, but a larger dataset would provide more robust results.

Variety of Data: The dataset should have structured data (e.g., demographics, diagnosis) and semi-structured data (e.g., treatment history, family history) to capture the diversity of factors influencing suicide propensity.

Velocity of Update: The velocity of data updates depends on the data sources. Patient data from healthcare centers and institutions may be updated periodically as new cases are diagnosed or additional information becomes available. Surveys can be conducted at specific intervals for updated data.

Data security and compliance requirements: Given the range and depth of variables encompassed, several crucial considerations come into play. Firstly, safeguarding patient

privacy and data confidentiality is paramount. Measures such as de-identification and encryption of personal information, including Age, Gender, and Date of Diagnosis, should be implemented. Compliance with data protection regulations and obtaining informed consent are also imperative, especially concerning sensitive variables like Family History and Suicide Attempts. Furthermore, robust access controls must restrict data access solely to authorized personnel, especially for variables like Social Support and Employment Status. Rigorous auditing and logging mechanisms need implementation to monitor data usage, particularly for variables related to financial status and suicidal history. Geographic Location data necessitates aggregation to regional levels to prevent re-identification risks. Lastly, secure storage and transmission protocols must be in place, especially when handling sensitive details like Suicide Method and Medication Changes. Regular security audits and updates are essential to maintain data integrity and minimize risks in this critical research endeavor.

Phase 2: Neural Network Selection for Specific Problem

Predicting suicide propensity in individuals is a complex task that can benefit from various neural network architectures. Here, we'll explore different applications of neural networks, analyze their strengths and weaknesses, define specific problems they can tackle, and propose technical requirements for these networks in the context of suicide propensity prediction.

Strengths and Weaknesses of Neural Network Types:

- **Feedforward Neural Networks (FNN):**
    - Strengths: Suitable for simple tabular data and initial feature engineering. Quick to train.
    - Weaknesses: May struggle with complex, sequential, or text data.
- **Recurrent Neural Networks (RNN):**
    - Strengths: Effective for sequential data, capturing temporal dependencies.
    - Weaknesses: Vulnerable to vanishing gradient problems, may not handle long sequences well.
- **Long Short-Term Memory (LSTM) Networks:**

    - Strengths: Overcome vanishing gradient issues in RNNs, suitable for long sequences.
    - Weaknesses: May be computationally expensive and require extensive tuning.
- **Convolutional Neural Networks (CNN):**
    - Strengths: Effective for image and grid-like data. Can be applied to feature extraction from images or geographical data.
    - Weaknesses: Less suitable for sequential or text data.
- **Transformer-Based Models (e.g., BERT, GPT):**
    - Strengths: Excellent for text data and capturing context. State-of-the-art in NLP.
    - Weaknesses: Computationally intensive, requiring substantial resources. They may not have the tuning and context required for processing medical history and treatment descriptions

Specific Problems, Applications of NN and Requirements in this project:

- **Problem 1: Risk Score Prediction:** Use neural networks to predict a suicide risk score based on a combination of features from the dataset. This can provide a continuous measure of risk for each individual. In this case we use a **Single Layer Perceptron** compared with Linear Regression, Random Forest, XGBoost
    - Data Type: Only use the Numerical and Categorical features in the dataset.
    - Scale: Large-scale dataset.
    - Prediction Accuracy: High accuracy and continuous risk score output.

- **Problem 2: Time Series Analysis:** Utilize **recurrent neural networks (RNNs)** or Long Short-Term Memory (LSTM) networks to analyze how suicide risk changes over time (e.g., during the follow-up period) based on patient data.

    - Data Type: Sequential and Numerical.
    - Scale: Patient-specific time series.
    - Prediction Accuracy: Precise tracking of suicide risk changes over time.

- **Problem 3: Text Analysis:** Apply **natural language processing (NLP)** techniques with neural networks to analyze 'Treatment History' text data, identifying patterns related to

treatment effectiveness and its impact on suicide risk. In this case I used the library **SkLearn** to make sentiment analysis

- Data Type: Text (NLP).
- Scale: Analyzing treatment history for all patients.
- Prediction Accuracy: Identifying nuanced patterns in text data.

- **Problem 4: Multi-Class Classification:** Employ neural networks for multi-class classification to categorize patients into different suicide risk groups (e.g., low, moderate, high) based on various features from the dataset. In this case we use a **Multi Layer Perceptron**
  - Data Type: Tabular with various features.
  - Scale: Multiclass classification for risk groups.
  - Prediction Accuracy: Accurate classification into different risk categories.

Technical Requirements for Neural Networks:

- Data Preprocessing: Proper data encoding, scaling, and feature engineering.
- Model Selection: Choose appropriate neural network architectures for each problem.
- Hyperparameter Tuning: Optimize hyperparameters like learning rates, batch sizes, and layer sizes.
- Regularization: Apply dropout, batch normalization, and L1/L2 regularization to prevent overfitting.
- Training: Utilize GPU/TPU resources for efficient training.
- Evaluation: Use appropriate metrics for each problem, e.g., RMSE for regression or F1-score for classification.
- Interpretability: Implement techniques for model interpretability, especially for clinical applications.
- Deployment: Deploy models using cloud-based solutions or containerization for real-time or batch predictions.

The specific measures to know how well the neural networks models fit the data and are able to predict features are the following:

1. **Model Performance Metrics**: Measure the predictive performance of your suicide propensity model using metrics such as accuracy, precision, recall, F1-score, and ROC AUC (if applicable).
2. **Model Calibration**: Assess how well your model's predicted probabilities align with actual outcomes using calibration plots and metrics like Brier Score or calibration intercept/slope.
3. **Feature Importance**: Quantify the importance of each input variable (e.g., 'Age,' 'Diagnosis,' 'Severity of Mental Disorder') in making predictions. Common techniques include permutation importance, SHAP values, or feature importance scores from tree-based models.
4. **Cross-Validation Scores**: Calculate cross-validated performance scores to evaluate the model's robustness and generalization to unseen data.
5. **Confusion Matrix**: Construct a confusion matrix to visualize true positives, true negatives, false positives, and false negatives, providing insights into the model's classification performance.
6. **Precision-Recall Curve**: Plot the precision-recall curve and calculate the area under the curve (PR AUC) to assess how well the model performs in identifying individuals at risk.
7. **ROC Curve**: Plot the Receiver Operating Characteristic (ROC) curve and calculate the ROC AUC to evaluate the model's ability to discriminate between different risk levels.
8. **Time-to-Event Analysis**: If applicable (for predicting suicide events), perform survival analysis and measure the concordance index (C-index) or log-rank test to assess the model's performance in predicting time-to-event outcomes.
9. **Model Training Time**: Measure the time it takes to train your predictive model. This metric is valuable for assessing model scalability and efficiency.
10. **Resource Utilization**: Quantify the computational resources (CPU, memory, GPU) required for model training and inference, ensuring efficient resource allocation.
11. **Data Update Frequency**: Determine how frequently the dataset requires updates to maintain model relevance. This could be based on changing patient demographics, treatment practices, or other factors.
12. **Ethical and Privacy Metrics**: Evaluate the project's adherence to ethical guidelines and privacy regulations (e.g., GDPR, HIPAA) and measure any potential privacy risks or breaches.
13. **Data Bias Assessment**: Conduct bias assessments to identify and measure any bias in the dataset or model predictions, especially concerning sensitive variables like 'Gender' or 'Cultural and Religious Beliefs.'

Phase 3: Distributed Big Data System Analysis and Big Data Environment Selection

Creating a data architecture for a project that aims to predict suicide propensity in individuals involves data collection, processing, and machine learning model deployment. Here's a high-level outline of the data architecture along with technology stack components and architectural patterns that have to be considered with regards of scalability, speed, and security requirements:

Data Collection:

1. **Data Sources:** Gather data from government agencies, healthcare providers, and academic research institutions. These sources may include electronic health records (EHRs), patient surveys, clinical databases, and research publications.
2. **Data Ingestion:** Use ETL (Extract, Transform, Load) processes to ingest data from various sources into a centralized repository.
3. **Data Updates:** Set up recurring data updates to ensure your dataset remains current and reflects the latest information about patients.

Data Processing: 4. **Data Integration:** Merge data from different sources into a unified dataset. Ensure consistency in data formats and standards across sources.

5. **Data Cleaning:** Perform data cleaning and preprocessing to handle missing values, outliers, and inconsistencies.
6. **Feature Engineering:** Create relevant features from the collected data. For example, you might derive new variables based on 'Treatment History' or 'Family History' to enhance model performance.
7. **Data Storage:** Store the processed dataset in a scalable and secure database or data warehouse. Consider solutions like AWS Redshift, Google BigQuery, or on-premises databases based on your needs.

Machine Learning: 8. **Model Training:** Utilize machine learning libraries like scikit-learn, TensorFlow, or PyTorch to build predictive models. Train models using algorithms such as logistic regression, decision trees, random forests, or deep learning approaches.

9. **Model Evaluation:** Evaluate model performance using various metrics such as accuracy, precision, recall, F1-score, ROC AUC, and others, depending on your specific goals.

10. **Model Deployment:** Deploy the trained model using cloud-based solutions like AWS Lambda, Azure Functions, or containerization platforms like Docker and Kubernetes for scalability and real-time predictions.

Technology Stack Components:

- **Data Ingestion:** Apache Nifi, Apache Kafka, AWS Glue
- **Data Processing:** Apache Spark, Hadoop, Apache Flink
- **Data Storage:** AWS S3, Google Cloud Storage, Azure Data Lake Storage, Relational Databases (e.g., PostgreSQL, MySQL)
- **Machine Learning:** Python (scikit-learn, TensorFlow, PyTorch), Apache MXNet, AWS SageMaker, Azure Machine Learning
- **Model Deployment:** AWS Lambda, Azure Functions, Docker, Kubernetes

Architectural Patterns:

- **Lambda Architecture:** This pattern involves processing data in two separate paths: batch processing (for historical data) and stream processing (for real-time data). This approach ensures flexibility and accuracy but can be complex to manage.
- **Kappa Architecture:** Kappa simplifies the architecture by using a single stream processing pipeline for both real-time and batch data. It is easier to manage and maintain but may require additional effort to ensure accuracy in historical data processing.

Pros and Cons:

- **Lambda Architecture:**
    - Pros: Offers more comprehensive data processing, accurate historical data, and flexibility.
    - Cons: Complex to set up and maintain, potentially higher cost due to duplicate processing.
- **Kappa Architecture:**
    - Pros: Simpler and more cost-effective, suitable for many use cases.
    - Cons: May require additional mechanisms for handling historical data if accuracy is critical.

For this project, the selected architecture was a single pipeline implemented in Hadoop.

Discussion

The analysis of patients diagnosed with mental disorders has yielded critical insights. Notably, specific demographic groups, particularly middle-aged men, exhibit a heightened suicide risk within one year of discharge. Moreover, patients with severe mood disorders demonstrate an increased susceptibility to suicidal ideation or attempts.

Regarding treatment, the current analysis refrains from discussion due to insufficient data, as it falls outside the scope of this study's objectives. Nonetheless, the pursuit of treatment adherence emerges as a pivotal factor, underscoring the necessity for continual support and post-discharge follow-up care.

Temporal analysis reveals that the initial three months post-discharge carry a particularly elevated risk, demanding enhanced support and follow-up procedures.

Social determinants, such as family support and occupational status, may potentially correlate with reduced risk. However, discernible geographic disparities in suicide rates warrant in-depth investigation.

Recommendations for improving this project

- Bias and Fairness: Beware of the potential impact of many biasing factors (in the collection and interpretation of the data, and in the extraction of conclusions) by making bias audits and fairness-aware machine learning.
- Patient engagement: Engage patients in their care by providing them with personalized insights and recommendations based on model predictions.
- Inclusion of Benefits: Collaboration with mental health professionals is essential to validate model results and ensure appropriate interventions are implemented.
- There was not enough time to master the required technologies. A not comprehensive list of the technology stack components that would be desirable is in the Annex 3.

Conclusions

This study provides valuable insights into the intricate challenge of post-discharge suicide risk among patients with mental disorders. Our findings underscore the significance of a holistic approach to mental healthcare, encompassing demographic, clinical, and social factors. Machine learning models can serve as valuable tools for identifying at-risk individuals but should be complemented by clinical judgment.

Efforts to mitigate suicide rates should prioritize early identification, intervention, and follow-up. Collaborative engagement among healthcare professionals and the removal of barriers hindering access to mental health services stand as imperative measures. To advance, we must address biases inherent in both data and models, ensuring equitable risk assessment.

This research not only advances our comprehension of suicide risk but also furnishes actionable knowledge to inform mental health policies and interventions, ultimately saving lives and enhancing the well-being of vulnerable individuals in our society.

Final Observation

I acknowledge two noteworthy aspects in this project. Firstly, the dataset's size may prompt questions; secondly, the exploration of Hadoop technology. Despite feedback received during the course, I had commenced this project when the feedback was provided. I had reached a midpoint in my project, recognizing the challenges of acquiring proficiency in a limited timeframe. I have thus decided to persevere, understanding the associated penalties. In my forthcoming work, I commit to meticulous adherence to instructions, leveraging my expanded knowledge of big data to propel this project forward.

References:

[1] Blau DM. Search for nonwage job characteristics: A test of the reservation wage hypothesis. Journal of Labor Economics. 1991;9(2):186–205. [9] Kelleher, J. D., & Tierney, B. (2018). Data science: An introduction. Boca Raton: CRC Press.

[2] Provost, F., & Fawcett, T. (2013). Data science for business: What you need to know about data mining and data-analytic thinking. Sebastopol;O'reilly Media, Inc.

[3] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). An introduction to statistical learning: With applications in R. New York: Springer.

[4] Anderson, J. R. (2016). Modern data science with R. Boca Raton: CRC Press.

[5] The Data Warehousing Institute. (2002). The complete guide to business intelligence. Hoboken: John Wiley & Sons, Inc.

[6] Murray, A. (2013). Big data: How it is transforming business, society, and everyday life. New York: Bloomsbury Publishing USA.

[7] Data Science at the Command Line by Jeroen Janssens

[8]Hands-on Machine Learning with Scikit-learn, Keras and Tensorflow by Aurelien Ger

[9]An Introduction to Statistical Learning by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani

[10] Intuitive ML and Big Data in C++, Scala, Java, and Python by Kareem Alkaseer

[11] Dive into Deep Learning by Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola

[12] Neural Networks and Deep Learning by Michael Nielsen

[13] Curry, B.; Morgan, P.; Beynon, M. Neural networks and flexible approximations. IMA J. Manag. Math. 2000, 11, 19–35. [Google Scholar] [CrossRef]

[14] Sarma, S.; Brock, D.; Ashton, K. The Networked Physical World; TR MIT-AUTOID-WH-001; MIT: Cambridge, MA, USA, 2000. [Google Scholar]

Annex 1. Columns of the dataset to be gathered.

1. **Patient ID:** A unique identifier for each patient.
2. **Age (Numerical):** Age of the patient at the time of diagnosis.
3. **Gender (Categorical):** Gender of the patient.
4. **Diagnosis (Categorical):** The specific mental health disorder diagnosed (e.g., depression, schizophrenia, bipolar disorder, anxiety disorder).
5. **Severity of Mental Disorder (Numerical):** Score indicates the severity of the mental disorder.
6. **Depression Severity (Numerical):** A person may have many disorders. each with an Intensity
7. **Date of Diagnosis (Date):** The date when the mental health disorder was diagnosed.
8. **Treatment History (Text):** The type and duration of treatments received, medications and therapy.
9. **Recent Discharge from Treatment (Binary):** Person recently discharged from psychiatric treatment
10. **Follow-up Period (Numerical):** The duration of the study's follow-up period for each patient.
11. **Family History (Categorical):** Details about a family history of mental health disorders or suicides.
12. **Suicide Attempts (Binary):** Binary (yes/no) whether the patient has a history of suicide attempts.
13. **Previous Suicide Attempts (Binary):** History of previous suicide attempts.
14. **Access to Lethal Means (Categorical):** Easy access to lethal means.
15. **History of Substance Abuse (Binary):** Whether the patient has a history of substance abuse.
16. **Current Co-occurring Substance Abuse (Binary):** Substance abuse, including alcohol and drugs.
17. **Loss of a Loved One (Binary):** Recent loss, such as the death of a loved one.
18. **Hopelessness (Binary):** Feelings of hopelessness about the future.
19. **Chronic Pain or Illness (Binary):** Individuals dealing with chronic pain or serious illness.
20. **Medication Changes (Binary):** Changes in psychiatric medication.
21. **Bullying and Discrimination (Binary):** Experiences of bullying, discrimination, or victimization.
22. **Cultural and Religious Beliefs (Categorical):** Cultural and religious beliefs about suicide.
23. **Social Support (Categorical):** Information about the patient's social support network.
24. **Social Isolation (Binary):** A lack of social support and feelings of isolation.
25. **Access to Healthcare (Categorical):** the patient's access to mental health care services.
26. **Access to Mental Health Care (Categorical):** Limited access to mental health care.
27. **Employment Status (Categorical):** Whether the patient is employed, unemployed, or on disability.
28. **Financial Status (Numerical):** Economic status of the patient (e.g., income, socioeconomic status).
29. **Financial Stress (Binary):** Financial difficulties and job loss.
30. **Geographic Location (Categorical):** Location of the patient, including country, state, or region.
31. **Date of Suicide (if applicable) (Date):** For patients who died by suicide, the date of the event.
32. **Suicide Method (if applicable) (Categorical):** If a patient died by suicide, the method used.

Table 2. Not exhaustive list of the desired minimal Dataset Columns required.

Annex 2: Specific Requirements and Assessments for Data Retrieval and Processing the Big Data Datasets for this project.

This is a very complex project that exceeds the timeframe available. In the construction of the desired system, considering that it may involve gathering sensitive private data, connecting to protected APIs and negotiating agreements with institutions, there is a set of requirements and assessments needed to ensure that the project is feasible in the budget and scope required. Each of the items in this list may involve so much detail that they may be projects in itself:

- ☐ **Scale:** The system should handle a growing volume of patient data efficiently. Ensure that the chosen technology stack can scale horizontally to accommodate increased data loads.
- ☐ **Speed:** Real-time data processing is crucial for timely interventions. Ensure that the data pipeline can process incoming data rapidly to provide actionable insights.
- ☐ **Analytics:** The system should support a wide range of analytics, including descriptive statistics, predictive modeling, and anomaly detection.
- ☐ **Data Security:** Implement robust data security measures to protect sensitive patient information, complying with relevant regulations like HIPAA or GDPR.
- ☐ **Data Governance:** Establish clear data governance policies to maintain data quality, integrity, and traceability throughout the project.
- ☐ **Interoperability:** Ensure that the architecture allows for interoperability with other healthcare systems and data sources to facilitate data sharing and collaboration.
- ☐ **Monitoring and Maintenance:** Implement monitoring and alerting systems to proactively identify issues in data collection, processing, or model performance. Regularly update data sources and models to reflect the latest information and research findings.
- ☐ **Data Security and Privacy:** Consider how sensitive data will be handled, stored, and protected. Ensure compliance with relevant data protection regulations (e.g., GDPR, HIPAA).
- ☐ **Scalability and Performance:** Determine whether the data storage and management solution can handle increasing data volumes and user loads while maintaining performance.

- ☐ **Data Accessibility and Availability:** the need for real-time data access and availability. Determine if the data should be available 24/7 or if batch processing is sufficient. Consider potential downtime for maintenance.
- ☐ **Data Integration:** how data from different sources will be integrated. Ensure that data can be harmonized and combined for analysis. Evaluate the need for data transformation and preprocessing.
- ☐ **Data Retention and Archiving:** how long data needs to be retained and whether historical data should be archived. Define data retention policies based on legal, regulatory, and business requirements.
- ☐ **Technology Stack:** needed for data storage and management. Consider modern data storage solutions such as cloud-based databases, NoSQL databases, and data lakes.
- ☐ **Data Governance and Metadata:** to ensure data quality, lineage, and compliance. Document metadata to provide context and improve data discoverability.
- ☐ **Backup and Disaster Recovery:** a strategy for data backup and disaster recovery. Ensure that data can be restored in case of unexpected events or data loss.
- ☐ Cost Considerations: the cost of data storage and management, including hardware, software, cloud services, and maintenance. Optimize costs while meeting project requirements.
- ☐ **Data Lifecycle Management:** the data lifecycle, including data creation, storage, usage, and deletion. Ensure that data is managed efficiently throughout its lifecycle.
- ☐ **Data Analytics and Processing Needs:** the analytical tools and processing requirements. Determine whether real-time processing, batch processing, or both are needed.
- ☐ **Documentation and Reporting:** How to data storage and management processes. Create clear reports summarizing requirements, decisions, and strategies.
- ☐ **Risk Assessment:** potential risks related to data storage and management, such as data breaches, scalability issues, or compliance violations. Develop mitigation strategies.
- ☐ **Stakeholder Communication:** data storage and management requirements and strategies to communicate with stakeholders, including team members, management, and data users.
- ☐ **User Satisfaction and Feedback**: Gather user feedback and satisfaction ratings to gauge how well the model and predictions meet the needs of healthcare stakeholders
- ☐ **Deployment Metrics**: Monitor the performance of the deployed model in a real-world setting, measuring factors like prediction accuracy, false alarms, and effectiveness.

- ☐ **Cost-Benefit Analysis**: Assess the costs associated with model development, data acquisition, and implementation against the benefits of suicide prevention and improved mental healthcare.
- ☐ **Long-term Impact:** Measure the long-term impact of the project on reducing suicide rates, improving mental health outcomes, and informing targeted interventions.

Annex 3. Technology Stack Glossary

Apache Spark: Spark is a fast and versatile data processing framework that can be used for batch processing, stream processing, machine learning, and graph processing.

Apache Hive: Hive is a data warehousing and SQL-like query language system for Hadoop. It allows users to write SQL-like queries to analyze large datasets stored in HDFS.

Apache Pig: Pig is a high-level scripting platform that simplifies the creation of MapReduce jobs. It provides an abstraction over the low-level details of writing MapReduce code.

Apache HBase: HBase is a distributed NoSQL database that provides real-time access to large amounts of data. It is often used for random read and write operations.

Apache Kafka: Kafka is a distributed streaming platform for building real-time data pipelines and streaming applications. It's commonly used for ingesting and processing streaming data.

Apache Flink: Flink is a stream processing framework similar to Spark Streaming. It supports both batch and real-time stream processing and is known for low-latency processing.

Elasticsearch: Elasticsearch is a distributed search and analytics engine. It's commonly used for full-text search and log and event data analysis.

Kibana: Kibana is an open-source data visualization and exploration tool that works seamlessly with Elasticsearch. It helps create interactive dashboards and visualizations.

Apache Cassandra: Cassandra is a distributed NoSQL database known for its scalability and high availability. It's suitable for storing and retrieving large amounts of data across multiple nodes.

TensorFlow and PyTorch: These are popular deep learning frameworks used for building and training machine learning models, often used in conjunction with big data tools for large-scale model training.

Apache Zeppelin and Jupyter: These are interactive notebook environments that allow data scientists and analysts to work with data interactively, create visualizations, and share insights.

Docker and Kubernetes: These containerization and orchestration technologies help manage and deploy big data applications and services more efficiently.

Apache NiFi: NiFi is a data integration and dataflow automation tool that can be used to move data between systems, transform data, and automate data workflows.

Apache Airflow: Airflow is a platform for orchestrating complex data workflows. It is often used for scheduling and monitoring data pipelines.

Tableau, Power BI, and Superset: These are popular data visualization tools that help create interactive dashboards and reports.