Introduction

Hello everyone my name is Vithun and i am joined by ZI jun and today we will be discussing our project which proposes a suicide risk assessment.

Slide 2 -

Before we dive into the specifics of our project, let me give you a quick overview of what we'll be covering. First, we'll talk about the aim of our project. Then, we'll discuss the data wrangling process, including data cleaning, preprocessing, and exploratory data analysis. Next, we'll go over the methodology we used, which includes several machine learning models. After that, we'll discuss our experiments and conclusions, including an analysis and evaluation of our models.

Slide 3 - Introduction

Now let's get started by discussing the objectives of our project.Suicide is a serious public health concern, and it is important to identify individuals who are at higher risk of suicide so that they can receive the help they need. Our project aims to do just that, by exploring key indicators of suicide and developing models that can help identify those at risk. By doing so, we hope to better allocate resources to those at risk, influence government policies on suicide prevention, and raise awareness of those at risk.

Slide 4 - Datasets

To accomplish our objectives, we used two datasets: the Suicide Rates Overview dataset from 1985 to 2021, and the Country Mapping dataset which includes ISO, continent, and region information. These datasets provided us with valuable information on suicide rates and demographic factors that can contribute to suicide risk.

Slide 5 -

Regarding Data cleaning, we dropped unnecessary Columns in the regional Data dataframe, and renamed some column names in our dataframes for merging and easier accessing. This allowed us to work with a more streamlined dataset.

Slide 6 -

NExt we dropped Countries with undefined value, the Health Development Index as its largely undefined with its associated country-year column. We also removed years after 2016 as the data was inaccurate, with population figures being erroneous.

Slide 7 -

One of the tools we learned and utilized was Geopandas, which allowed us to access shapefiles and initialize a world map. We then check if the names of the countries exist and match in both datasets, discovering countries which have different names on the Geopandas shapefile which we then renamed to match for eventual merging of datasets.

Slide 8 -

Following this, we Created a new binary categorical variable called suicide_risk, which allowed us to categorize individuals who need help and those who do not. High risk was defined as having a suicide rate higher than the mean, with high being labeled as 1 and low being labeled as 0.

Slide 9:

The next step in data preprocessing is removing unnecessary columns to reduce noise. We'll drop total GDP and population as GDP per capita is derived from both. We also drop suicides/100kpop, suicide numbers and population as suicide risk is dereived from these. numbers. We'll also remove the year variable as we are predicting future values. This helps reduce noise in our dataset eventually used for the models.

Slide 10:

In order to input our categorical variables into all the models, we used one-hot encoding which will transform our categories into numerical form and improve our prediction accuracy, by using MinMaxScaler. Once we've done this, we'll create two dataframes - one with our target variable, suicide risk, and another with our predictors.

Slide 11:

We split our data into training and testing datasets. We'll use a ratio of 8:2 for the train-test split. This allows us to validate our model accurately by seeing how it will perform with new data, and prevents overfitting.

Slide 12:

Now that we've cleaned our data, preprocessed it, and created our required datasets, lets move on to exploratory data analysis and take a closer look at the correlation values between our variables and suicide/100K by plotting a heatmap. There does not seem to be a clear correlation. We've found that population are not good indicators once normalized, and GDP per capita has the lowest correlation.

Slide 13:

While we haven't found a clear correlation between our variables regarding suicide risk, we have discovered that as age increases, suicide risk per 100k population increases as well. We see a positive correlation between suicide rates and age, which tells us that age may be an important predictor in our models.

Slide 14:
We've also discovered that males are about four times more likely to commit suicide than females. This trend is consistent over the years and may be due to social constructs.
Slide 15:

Now, let's take a look at the regional distribution of suicide risk. Using a barplot, we've found that Europe, Asia, and the Americas have similar numbers of countries in our dataset, but Europe seems to have the highest suicide rates.

Slide 16:

To better visualize this data, we used Geopandas to create a choropleth map. This tool allows us to see the distribution of suicide risk across different regions of the world. However, we should note that there is a lack of data in Africa and Asia, which may affect the representativeness of our models in these regions.

Next ill pass it over to zi jun who will carry on with the rest of the presentation.

Slide 17:

For utilization of machine learning to solve our problem, we chose 3 models – Decision Tree, Random Forest and XGBoost.

Decision Tree is used as a baseline comparison, while Random Forest and XGBoost are used as they are advanced models with greater accuracy.

Our input for training all models come from the same set of training data obtained from the 8:2 split previously performed, and will be backtested against the test data.

As this is a binary categorical classification problem, we will be measuring the performance of the model through classification accuracy, true positive rate and false positive rate. The implications behind these values will be shared later in our evaluation.

Slide 18:

Here we have a decision tree - a supervised learning algorithm. Important hyperparameter of note is depth, where it is set to 4 to avoid overfitting in deep trees, but still be accurate in classifying.

Slide 19:

Next we have random forest – an ensemble learning algorithm. It combines the output of multiple decision trees to reach a single result. Important hyperparameter of note is trees and bootstrap. Bootstrap performs statistical resampling on the individual trees, thus even with high number of trees overfitting does not occur. Trees are set to 100 so that it achieves good accuracy, while not too computationally intensive.

Slide 20:

Next, we have XGBoost – a very popular machine learning algorithm right now. It is a scalable and highly accurate form of a gradient boosting algorithm. Each subsequent tree learns from the previous trees and gives more weight to features that perform better, unlike how random forest works in assigning equal weight for all trees. Hyper parameters of note

are: Objective, eta and max_depth. Max_depth of 3 to prevent overfitting, and eta of 0.3 to fit model slower to prevent overfitting.

Slide 21:

Now we evaluate our models. We use .score() function to calculate classification accuracy, and true and false positive rates using a custom function. The exact calculations can be found in the notebook. From our results, we notice that our baseline, Decision Tree, is overall the worst performing model in all aspects. Random forest and XGBoost both performed similarly, with Random Forest taking the edge in classification accuracy and true positive rate, and XGBoost in false positive rate.

Slide 22:

So which is the best model for our problem given these metrics? I will explain these metrics in relevance to our problem: Classification Accuracy represents the overall accuracy in predicting suicide risk. True positive rate represents the percentage of those of high suicide risk being classified correctly. False positive rate represents those of low suicide risk classified wrongly as of high suicide risk. We would want to focus most importantly on true positive rate – as lives matter most, we do not want anyone miscategorized and not receive the aid they need, resulting in their suicide. The difference in false positive rates between XGBoost and random forest are both considerably small, thus the idea of save marginal resources at the expense of greater miscategorization and thus suicides does not seem worth it. And thus - our model of choice is Random Forest, as it provides the lowest probability of miscategorization.

Slide 23:

Now for data-driven insights. We measure the importance of features using feature_importances_ function. GDP per capita, sex, age, and Europe as a region are the most common factors with high importance. We can use this insight. For gdp, we can focus on improving the economy of countries to reduce suicide rates. For age, we can focus on providing suicide aid or information to age groups that are more susceptible to suicide. For europe, the world could come together to provide more global suicide aid to them as they are of higher risk. And for gender, we should get rid of toxic social constructs based on gender for males, to reduce suicide rates.

Slide 24:

To conclude, we reflect on our project. Limitations of Random Forest is that large number of trees can make the algorithm too slow and ineffective for real-time predictions. Our problem does not require this, but if we do, XGBoost would a possible alternative as they do not face this constrain.

Also, there are areas we could have improved on. Firstly, use hyperparameter tuning algorithms to determine optimal hyperparameters. With optimal hyperparameters, overall accuracy of models will likely increase. Secondly, as dataset is decent but not very large, we could repeat on multiples different train-test split data and take the average to reduce variance. Thirdly, there is imbalanced data. There are approximately 2 times more Low than

High in suicide_risk. We could perform oversampling or undersampling to balance data, so that accuracy score is more representative. Lastly, instead of using mean, we could use 90th percentile of suicide rates as a benchmark - as a better representation of those who are more likely of suicide, as suicides are rare.

Slide 25:

Thank you