Analyzing the Relationship Between Social Trust and the Importance of Democracy Using European Social Survey Data (2020)

Group 20: Emma ‘Ais’ Burniston, Ryan Camp, Jacqueline Sanchez, Tan Tran, ScrappyJet

Spring 2025

# Introduction (Names):

In the past decade, there has been a great deal of changes within the political realm across the world, namely pushes towards governments that are more right-wing and authoritarian in nature. The causes of this shift are of course, multi-faceted, and may express itself differently across the world. [Differences between right wing rise in turkey (religous/ethononationalist reasons), brazil (economic deprivation, conflict with indigenous amazon tribes) and/or el salvador (gang violence, disaffected voters). This sense of lacking inertia is one of the key motivators worldwide, and especially in Europe. Doing a numeric analysis on lack of interest is somewhat difficult, however report high Non-response rate. related issue of does the government do it’s job? measures of social trust]

For the purposes of our project, we will be looking at the relationship between measures of social trust and an the importance of democracy on six regions of Europe [partitions here or later to not mess up the flow?] We have used data from the European Social Survey (ESS) Round 10, conducted by the Core Scientific Team (CST) based in the University of London with several participating universities throughout the continent in 2020. Our objective is to discover if there is a meaningful/significant relationship between these predictors and importance of democracy using linear discriminant analysis and categorization decision trees (random forest?) Comparisons between our approaches and these regions may provide insight into the continental differences in political thought.

Our analysis focused on 10 predictors of social trust. [list here?] The data was cleaned and preprocessed, split it into training and testing sets, linear discriminant analysis and decision tree models used to identify relationship with reported importance of democracy We evaluated the performance of each model and discussed their relative strengths and limitations. Comparing LDA and decision tree models will hopefully allow a more complete understanding of democracy’s relative worth within Europe.

**Linear Discriminant Analysis Model (Ais, Jacqueline, ScrappyJet):**

**~~~~~~~~~~~**

Logistic regression is a simple, fast, and interpretable model that works best with linear relationships, while decision trees ofer greater fexibility, robustness to outliers, and the ability to handle complex relation- ships. However, both models have their limitations, including logistic regression’s sensitivity to outliers and assumptions, and decision trees’ tendency for overftting and instability.

Our objective is to shed light on the factors that infuence the cause of death among patients during the COVID-19 pandemic. We aim to identify factors that may increase or decrease the likelihood of death from this lethal virus. Logistic Regression models attempt to determine the probability of the response variable, date of death, in relation to the chosen predictor variables, such as pneumonia, obesity, tobacco use, pregnancy, and so on. Given that some predictors may carry more signifcance than others, employing multiple logistic regression might be necessary to develop an accurate model.

The logistic regression model formula would be:

exp(*β*0 + *β*1*X* + *...* + *βpXp*)

( ) =

*p x* 1 + exp(*β*0 + *β*1*X* + *...* + *β X* )

*p*

*p*

in which p(x) represent our response variable ‘data died’, *β*0 represents the model intercept. The predictors, *Xi* for i = 1, 2, 3, . . . , 17, will represent pneumonia, obesity, hypertension, tobacco, pregnancy, diabetes, and so on. *βi*i for i = 1, 2, 3, . . . , 17 will represent their respective coefcients. For this Logistic Regression, we are going to include every predictor that we decided to pick until we fnd any predictors that are not signifcant in fnding the relationship with our response variable, ‘data died’. Unfortunately, our response variable, ‘date died’, is considered a quantitative variable for some reason, so we had to manipulate the dataset to make ‘data died’ a qualitative variable. We added a column containing a qualitative variable called ‘SURVIVED’. If the date in the dataset shown “9999-99-99” (which means they survived), then we set the ‘SURVIVED’ factor to ‘Yes’. Otherwise, we would set the ‘SURVIVED’ factor to ‘No’, which means they did not survive the COVID-19 virus.

Model Formula:

logisticModel = glm(Covid.Data$SURVIVED ~ Covid.Data$TOBACCO+Covid.Data$OBESITY+Covid.Data$DIABETES+ Covid.Data$PNEUMONIA + Covid.Data$HIPERTENSION + Covid.Data$PREGNANT+Covid.Data$INTUBED+Covid.Data$USMER+Covid.Data$SEX+ Covid.Data$COPD+Covid.Data$PATIENT\_TYPE+Covid.Data$ASTHMA+ Covid.Data$INMSUPR+Covid.Data$OTHER\_DISEASE+ Covid.Data$CARDIOVASCULAR+Covid.Data$ICU+Covid.Data$RENAL\_CHRONIC,

family = "binomial")

We are creating the signifcance table for our multiple logistic regression model to fnd potential predictors that are not signifcant.

##

## Call:

## glm(formula = Covid.Data$SURVIVED ~ Covid.Data$TOBACCO + Covid.Data$OBESITY + ## Covid.Data$DIABETES + Covid.Data$PNEUMONIA + Covid.Data$HIPERTENSION +

## Covid.Data$PREGNANT + Covid.Data$INTUBED + Covid.Data$USMER + ## Covid.Data$SEX + Covid.Data$COPD + Covid.Data$PATIENT\_TYPE +

## Covid.Data$ASTHMA + Covid.Data$INMSUPR + Covid.Data$OTHER\_DISEASE +

## Covid.Data$CARDIOVASCULAR + Covid.Data$ICU + Covid.Data$RENAL\_CHRONIC, ## family = "binomial")

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ##  ## | Deviance | Residuals: |  | | | | |
| ## | Min | 1Q Median | 3Q | Max | | | |
| ## | -3.8950 | 0.1121 0.1344 | 0.1369 | 2.5021 | | | |
| ## |  |  |  |  | | | |
| ## Coefficients:  ## Estimate Std. Error z value Pr(>|z|) | | | | | | | |
| ## | (Intercept) | | 12.0933814 | 0.1531254 | 78.977 | < 2e-16 | \*\*\* |
| ## | Covid.Data$TOBACCO | | -0.0061116 | 0.0013023 | -4.693 | 2.69e-06 | \*\*\* |
| ## | Covid.Data$OBESITY | | -0.0045882 | 0.0009921 | -4.625 | 3.75e-06 | \*\*\* |
| ## | Covid.Data$DIABETES | | 0.0058633 | 0.0013770 | 4.258 | 2.06e-05 | \*\*\* |
| ## | Covid.Data$PNEUMONIA | | 0.0259081 | 0.0006265 | 41.352 | < 2e-16 | \*\*\* |
| ## | Covid.Data$HIPERTENSION | | 0.0184840 | 0.0018190 | 10.162 | < 2e-16 | \*\*\* |
| ## | Covid.Data$PREGNANT | | 0.0090592 | 0.0013303 | 6.810 | 9.77e-12 | \*\*\* |
| ## | Covid.Data$INTUBED | | 0.0103875 | 0.0021080 | 4.928 | 8.33e-07 | \*\*\* |
| ## | Covid.Data$USMER | | 0.2029168 | 0.0089968 | 22.554 | < 2e-16 | \*\*\* |
| ## | Covid.Data$SEX | | -1.2377033 | 0.1263409 | -9.797 | < 2e-16 | \*\*\* |
| ## | Covid.Data$COPD | | -0.0084990 | 0.0016178 | -5.253 | 1.49e-07 | \*\*\* |
| ## | Covid.Data$PATIENT\_TYPE | | -5.1557159 | 0.0444125 | -116.087 | < 2e-16 | \*\*\* |
| ## | Covid.Data$ASTHMA | | -0.0122138 | 0.0018043 | -6.769 | 1.29e-11 | \*\*\* |
| ## | Covid.Data$INMSUPR | | -0.0006017 | 0.0013147 | -0.458 | 0.64716 |  |

## Covid.Data$OTHER\_DISEASE 0.0011833 0.0006863 1.724 0.08469 .

## Covid.Data$CARDIOVASCULAR -0.0051003 0.0015710 -3.247 0.00117 \*\*

## Covid.Data$ICU -0.0215633 0.0020643 -10.446 < 2e-16 \*\*\*

## Covid.Data$RENAL\_CHRONIC -0.0006541 0.0017350 -0.377 0.70615 ## ---

## Signif. codes: 0 ’\*\*\*’ 0.001 ’\*\*’ 0.01 ’\*’ 0.05 ’.’ 0.1 ’ ’ 1 ##

## (Dispersion parameter for binomial family taken to be 1) ##

## Null deviance: 550061 on 1048574 degrees of freedom ## Residual deviance: 333724 on 1048557 degrees of freedom ## AIC: 333760

##

## Number of Fisher Scoring iterations: 7

The signifcance table shows that the predictors ‘INMSUPR’ (patient is immunosuppressed or not), ‘RE- NAL\_CHRONIC’ (whether the patient has chronic renal disease or not), and ‘OTHER\_DISEASE’ are not signifcant on the relationship if the patient died or not. Knowing which predictors are insignifcant, we can calculate our mean MSE with the newly found information.

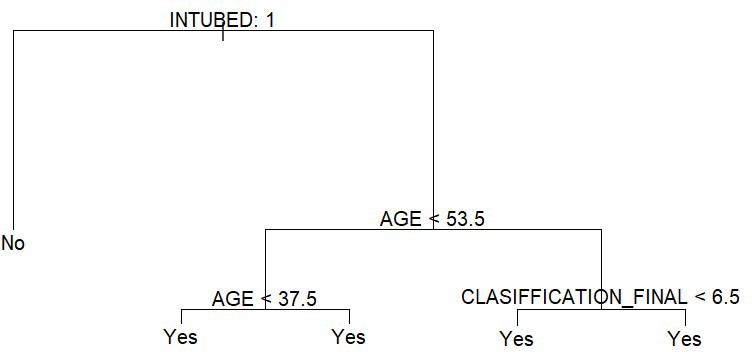
The mean squared error (MSE) was computed with 10 iterations. The average MSE across the 10 iterations came out to be 0.0728 with values ranging from 0.0720 to 0.0737.

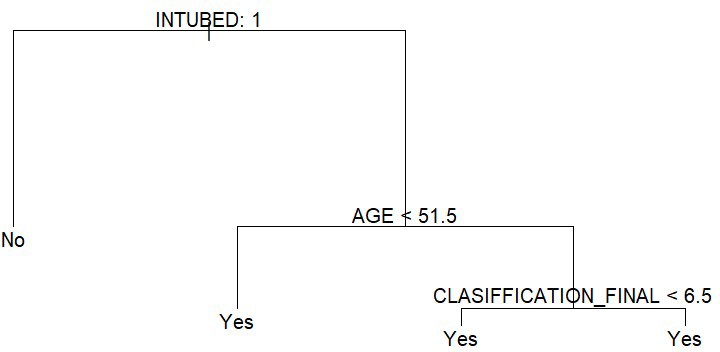
The variables we found that are most signifcant in the Logistic Regression model came out to be:

* PNEUMONIA (whether you had pneumonia or not), PATIENT\_TYPE (what type of care the patient received), and USMER (whether the patient treated medical units of the frst, second or third level)
* For example, although the “PNEUMONIA” predictor is very signifcant, a very small percentage of people had pneumonia and died with it. This is because the model is susceptible to outliers, and as a result we have many predictors that were identifed as signifcant. Consequently, we decided to implement a Decision Tree to narrow it down to specifc predictors.

**Decision Tree Model (Names):**

Using a decision tree to analyze COVID-19 deaths ofers several advantages, including interpretability, the ability to capture non-linear relationships and multi-factor interactions, robustness to outliers, and efective handling of categorical data.



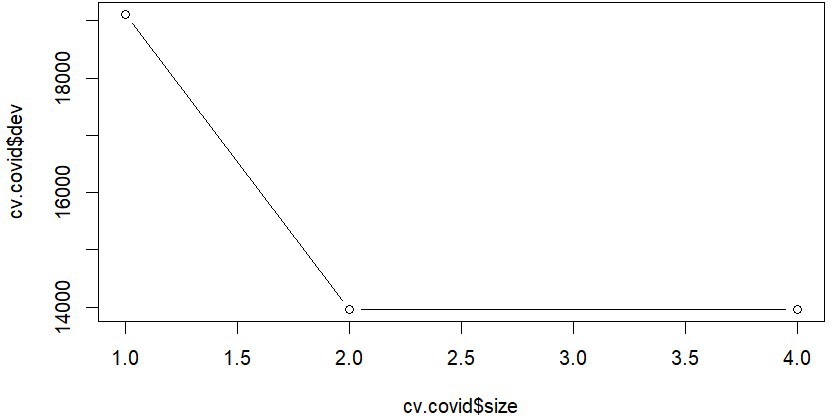


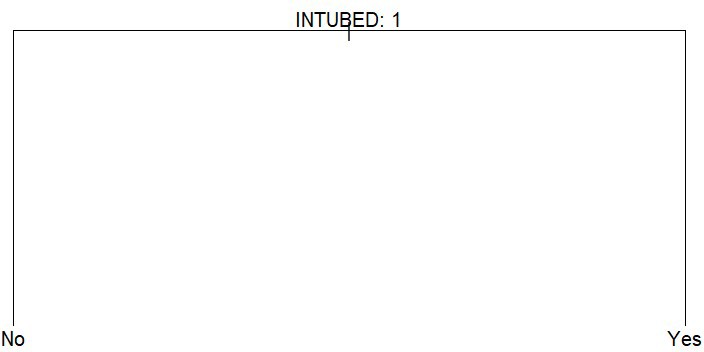
The present study employed a decision tree algorithm in R to predict patient survival based on a set of 17 predictors including “USMER”, “SEX”, “PATIENT\_TYPE”, “INTUBED”, “PNEUMONIA”, “PREG-

NANT”, “DIABETES”, “COPD”, “ASTHMA”, “INMSUPR”, “HIPERTENSION”, “OTHER\_DISEASE”, “CARDIOVASCULAR”, “OBESITY”, “RENAL\_CHRONIC”, “TOBACCO”, and “ICU”.

Our model formula was: decision\_tree <- tree(SURVIVED ~ ., data = train\_data).

The algorithm was run 10 times, resulting in only two distinct decision trees. The frst tree had “intubed” as the main node, which then split into “no” and “Age < 53.5”. The latter split further into “yes” and “yes”. The second tree also had “intubed” as the main node, which then split into “no” and “Age < 53.5”. The latter split further into “yes” and “Age < 37.5”, which then split into “yes” and “yes”. The mean squared error (MSE) was computed for 10 iterations, and the average MSE across all models was found to be 0.2281382 with a range from 0.2243118 to 0.2340730.



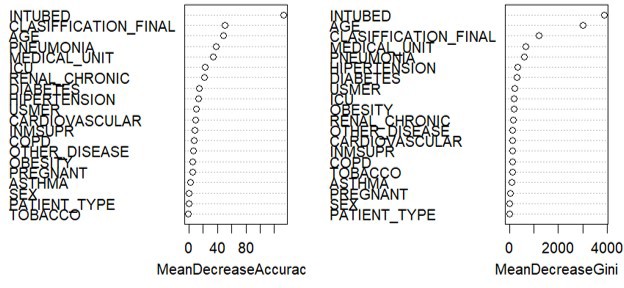


The unpruned tree had four terminal nodes while the pruned only had two. The unpruned tree was split by more variables because of node purity so the pruning process eliminated the other two. However, it did not reduce the classifcation error.

Overall, the present fndings suggest that “intubed” may be a crucial predictor in determining patient survival, as it was the primary node in both trees generated by the algorithm. Additionally, the limited variability in the resulting decision trees across multiple iterations may indicate a potential lack of robustness in the model, which warrants further investigation. Finally, the average MSE provides an estimate of the error in the model’s predictions and may serve as a useful metric for evaluating the algorithm’s performance.

**Random Forest Model (Names):**

Decision trees struggle with prediction accuracy and sufer from high variance because subsets of the same data could yield diferent results. Based on the results of the single decision trees, we decided to also implement the random forest algorithm since it will not only reduces the variance but could possibly give us an opportunity to analyze the predictors in depth as to why “Intubed” was the only predictor appearing in the pruned tree. We used the same model formula as the decision tree which set the predictors against our created Survived response variable.



Since the model performed well and the mtry value was small enough, we did not have to tamper with any parameters and used the default 500 trees and mtry of sqrt(predictors) to confgure the function. We obtained an average test MSE of 0.189998 across the 10 iterations with range from 0.1857877 to approximately 0.2090844.

The variables we found to be most efective in determining survival were the following.

* Intubation: The results indicate that intubation plays a signifcant role in the likelihood of survival. Patients who were intubated may have had more severe cases of COVID-19, requiring mechanical ventilation to support their breathing. This suggests that early identifcation and management of respiratory distress could be crucial for improving the chances of survival.
* Age: Age was found to be another important factor in determining survival. As age increases, the likelihood of survival from COVID-19 decreases. This is consistent with previous studies showing that older individuals are at a higher risk of complications and mortality due to the virus. Public health measures should continue to prioritize the protection of older populations, including vaccination and targeted interventions.
* Classifcation Final: The classifcation of the fnal variable, which may represent the severity or stage of the disease, also plays a signifcant role in the likelihood of survival. Patients with a more severe classifcation of COVID-19 have lower chances of survival. This fnding underscores the importance of early detection and intervention to prevent the progression of the disease to more severe stages.

|  |  |  |
| --- | --- | --- |
| **Conclusion (Names):** |  | |
|  | Model | TestError |
|  | Logistic Regression | 0.0728 |
|  | Decision Tree | 0.2281 |
|  | Random Forest | 0.1899 |

In conclusion, our comprehensive analysis of the COVID-19 dataset has provided valuable insights into the factors that signifcantly impact the severity of the disease. Our fndings revealed that intubation emerged as the most infuential variable, highlighting its critical role in determining patient outcomes during the pandemic.

We employed three diferent machine learning techniques—logistic regression, decision trees, and random forests—to model the relationship between the predictor variables and the target variable. These techniques allowed us to capture and interpret the complex interactions between the variables efectively. Among the models we explored, logistic regression demonstrated superior performance, yielding the lowest mean squared error (MSE) compared to the other two methods. Random forests came in second, providing a robust and relatively accurate model. In contrast, decision trees, while ofering an easily interpretable model, had the highest MSE, indicating a lower predictive accuracy compared to the other techniques.

Our study underscores the importance of intubation in the context of COVID-19 and provides a solid foundation for further research in this area. The application of diferent modeling techniques has enhanced our understanding of the data and allowed for a comprehensive evaluation of the factors impacting COVID- 19 severity. These fndings may help inform clinical decision-making and public health policies aimed at improving patient care and mitigating the impact of future outbreaks.

# Bibliography

* <https://www.kaggle.com/datasets/meirnizri/covid19-dataset>
* Examples from Canavs

# Source Code

**Logistic Regression**

Covid.Data = read.csv("Covid Data.csv")

Covid.Data$SURVIVED <- factor(ifelse(Covid.Data$DATE\_DIED == "9999-99-99", "Yes", "No"))

Covid.Data$DATE\_DIED <- NULL

*#Converting predictors into factors*

cols\_to\_factor <- c("USMER", "SEX", "PATIENT\_TYPE", "INTUBED", "PNEUMONIA", "PREGNANT", "DIABETES", "COPD", "ASTHMA", "INMSUPR", "HIPERTENSION", "OTHER\_DISEASE", "CARDIOVASCULAR", "OBESITY", "RENAL\_CHRONIC", "TOBACCO", "ICU")

*#Splitting training and test data running for 10 iterations*

mse = rep(0,10) **for** (i **in** 1:10){ set.seed(iˆ2)

train\_index = sample(1:nrow(Covid.Data), size = 0.8 \* nrow(Covid.Data)) train\_data = Covid.Data[train\_index,]

test\_data = Covid.Data[-train\_index,]

model = glm(SURVIVED ~ TOBACCO+PREGNANT+OBESITY+DIABETES+PNEUMONIA+HIPERTENSION+SEX+ COPD+PATIENT\_TYPE+ASTHMA+INTUBED+USMER+CARDIOVASCULAR+ICU,

data=train\_data, family="binomial")

predictions = predict(model, newdata=test\_data, type="response") mse[i] = mean((test\_data$SURVIVED == "Yes") != (predictions > 0.5))

}

mean\_mse = mean(mse) summary(logisticModel)

**Decision Tree**

*#Splitting training and test data running for 10 iterations*

library(tree) MSE = rep(0,10)

**for** (i **in** 1:10) { set.seed(iˆ2)

train\_index <- sample(1:nrow(covid\_data), size = 0.8 \* nrow(covid\_data)) train\_data <- covid\_data[train\_index, ]

test\_data <- covid\_data[-train\_index, ]

decision\_tree <- tree(SURVIVED ~ ., data = train\_data) plot(decision\_tree)

text(decision\_tree, pretty = 0)

yhat = predict(decision\_tree, newdata=test\_data, type = "class") MSE[i] = mean((as.numeric(test\_data$SURVIVED) - as.numeric(yhat))ˆ2)

} MSE

mean(MSE)

*#Pruning the tree using cross validation*

**Random Forest**

library(randomForest)

*#Splitting into training and test data running for 10 iterations*

MSE = rep(0,10)

p = ncol(covid\_data)-1 B = 500

**for** (i **in** 1:10) cv.covid = cv.tree(decision\_tree, FUN = prune.misclass) plot(cv.covid$size,cv.covid$dev,type='b')

prune.covid = prune.misclass(decision\_tree, best = 2) plot(prune.covid)

text(prune.covid, pretty = 0){ set.seed(iˆ2)

train\_index <- sample(1:nrow(covid\_data), size = 0.8 \* nrow(covid\_data)) Survived.test = covid\_data[-train\_index,"SURVIVED"]

X.test = covid\_data[-train\_index,-21]

rf.model = randomForest(SURVIVED ~., data = covid\_data,

subset = train\_index,

xtest = X.test, ytest = Survived.test, ntree = B,

mtry = sqrt(p), importance = TRUE, keep.forest = TRUE)

yhat = predict(rf.model, newdata=test\_data, type = "class")

MSE[i] = mean((as.numeric(test\_data$SURVIVED) - as.numeric(yhat))ˆ2)

} MSE

mean(MSE) varImpPlot(rf.model)