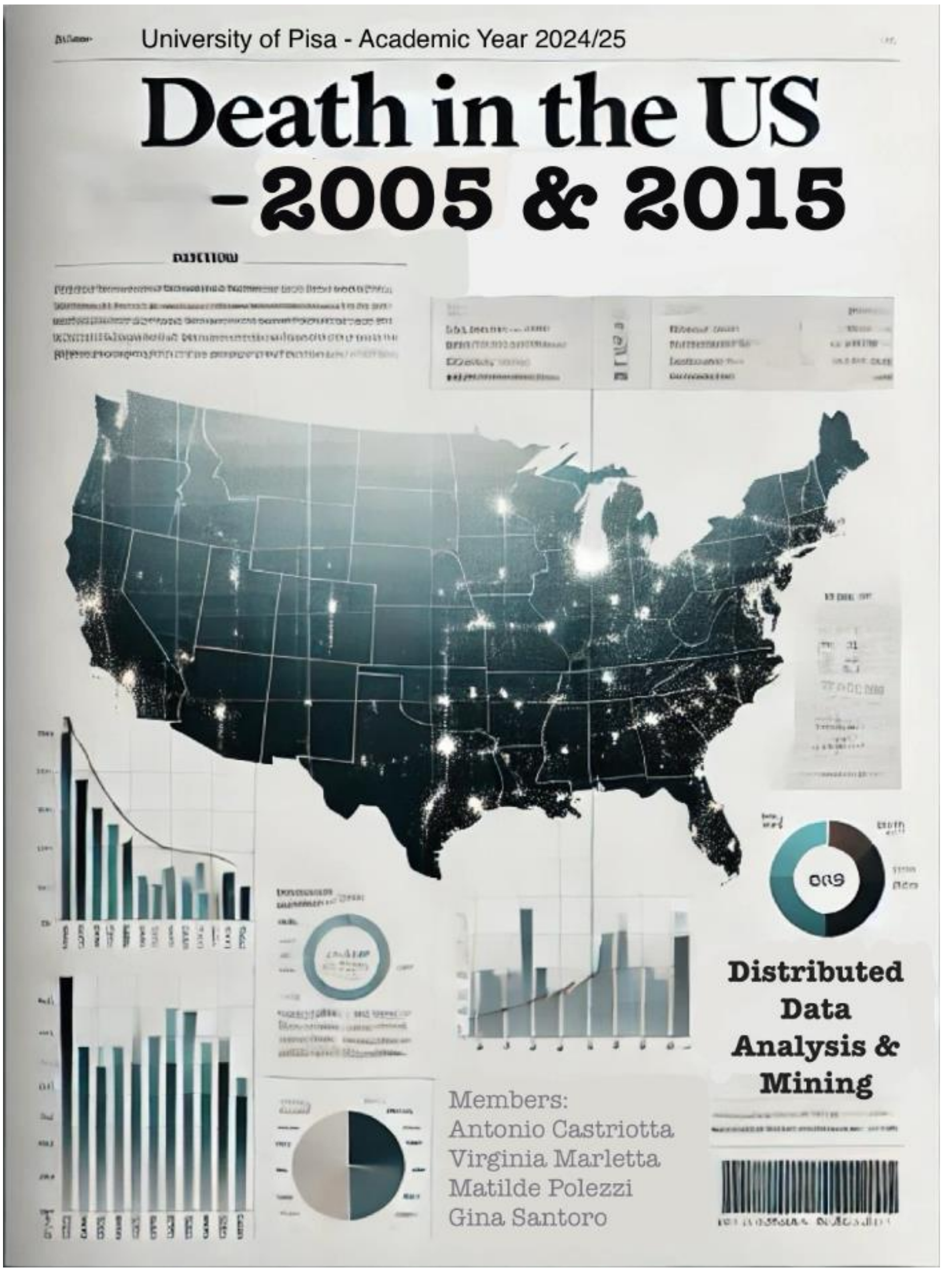


DEATH IN THE US, GROUP 7

ABSTRACT

This study examines U.S. mortality trends in 2005 and 2015, highlighting shifts in leading causes of death and demographic influences. Using classification models and SHAP analysis, it identifies key features and offers insights for public health and policymaking.

Authors

Antonio Castriotta
Virginia Marletta
Matilde Polezzi
Gina Santoro

INTRODUCTION

Understanding mortality trends is essential for public health, policymaking, and resource allocation, hence the need for governments and organizations to identify the leading causes of death and implement strategies to mitigate them. In this study, we focus on the United States, analysing data on causes of death for the years 2005 and 2015. The dataset, sourced from Kaggle, provides comprehensive information about mortality in the U.S., enabling us to uncover trends and shifts in public health concerns over a decade. The objective of the analysis is to compare the demographic and socio-economic variables that influence the incidence of diseases and identify changes in their relevance.

DATA PREPARATION

The dataset for the year 2005 contains 2,452,506 observations, while the dataset for the year 2015 contains 2,718,198 observations, both described across 77 features, which are categorized into three main groups:

1. **Demographic Information**: including variables that describe the individual's characteristics, such as sex, age, race, marital status, resident status, and level of education. These details help to understand the background and context of each person.
2. **Details About the Death Event**: these features relate to the circumstances and specifics of the death, such as the place of death and the decedent's status, the day of the week and the month of the death, the method of disposition, the manner of death, whether an autopsy was conducted, and activity codes indicating the context in which the death happened.
3. **Causes of Death**: this group includes up to 40 columns labelled as 'entity_condition' and ‘record_condition’, which describe additional conditions or diseases that contributed to the death.

The description and types of the most relevant variables are summarized in Table 1, providing a comprehensive overview of the data structure and content for both years. Additional columns, such as ‘entity_condition’ and ‘record_condition’, have been excluded from the analysis due to the significant number of missing values.

Encoding icd_code_10th_revision

The original datasets contained more than 3700 unique diseases, based on the 10th revision of the International Classification of Diseases (ICD-10). To simplify the analysis and improve interpretability, we mapped these categories into a reduced set of 22 aggregated disease types using a lookup table.

Encoding and filling Education 2003 Revision

Upon examining the Education Level columns from 1989 and 2003, we observed that missing values in one column were present in the other. To resolve this, we aligned the evaluation metric of the 1989 data with the 2003 standard by creating a mapping dictionary; we then applied the coalesce function to combine the two columns into one, education_level, ensuring no missing values remained.

Missing values

Although we excluded several features due to the number of missing values that could not be recovered, we identified 3 that could be informative enough for the EDA and Classification task:

- Manner of Death: Assign the value 8 for "Not Specified."
- Activity Code: Assign the value 7 for "Not Applicable."
- Place of Injury for Causes W00-Y34 (except Y06.- and Y07.-): Assign the value 99.

The number of columns we retained for the EDA task was 18 out of 77.

Table 1: Selected features for further analysis.

Feature Name	Description	Type
sex	Indicates the biological sex of the deceased person (F, M).	STRING
marital_status	Specifies the marital status of the person at the time of death.	
injury_at_work	Indicates whether the death was caused by a work-related injury.	
method_of_disposition	Specifies the method by which the body was handled after death.	
autopsy	Indicates whether an autopsy was performed on the body to determine the cause of death.	
type_of_disease	Categorization of the causes of death based on aggregated data.	
resident_status	Indicates the residency status of the deceased in relation to the location where the death occurred.	INT
month_of_death	Specifies the month in which the death occurred.	
detail_age	Indicates the exact age of the deceased at the time of death, expressed in years.	
age_recode_12	Groups the age of the deceased into 12 categories.	
place_of_death_and_ decedents_status	Specifies the place where the death occurred and the status of the deceased.	
day_of_week_of_death	Specifies the day of the week when the death occurred.	
manner of_death	Indicates the manner in which the death occurred.	
activity_code	Encodes the activity the person was engaged in at the time of death (if applicable).	
place_of_injury_for_causes_w00_y34_	Specifies the place where an injury occurred.	
race_recode_5	Groups racial categories into five main groups.	
hispanic_origin race_recode	Combines information about Hispanic origin and race.	
education_level	Indicates the education level of the deceased, aggregated into categories.	

DATA UNDERSTANDING

In this section the aim is to analyse the data in a general way, considering both variables individually and in relation to others.

Life Expectancy

The line chart shows the distribution of age at death for 2005 and 2015. The 2015 curve (orange) shifts slightly to the right, indicating an increase in the average age at death, reflecting improvements in longevity. Both curves peak between ages 70 and 85, with the 2015 curve showing a more prominent tail beyond 90 years, suggesting more people reached advanced ages in 2015 compared to 2005. The 2005 curve (blue) represents slightly lower average ages at death.

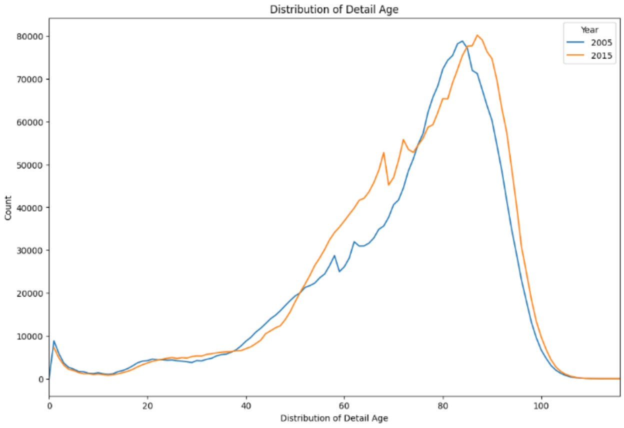


Figure 1: Comparison of Detail Age over 2005 and 2015.

Changes in Leading Causes of Death

Over the decade from 2005 to 2015, notable shifts were observed in the leading causes of death, with 2015 also gaining a new class of diseases being "Special Purposes"[1].

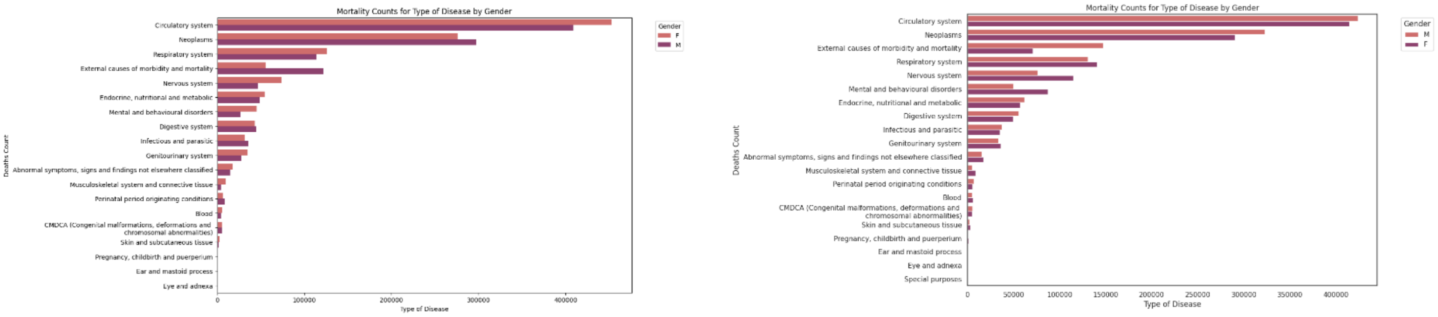


Figure 2: Distribution of diseases, 2005 (left) and 2015 (right)

- ***Cardiovascular Diseases*** remained the leading cause of death in both years; however, their relative impact showed a slight decrease in 2015, suggesting some progress in prevention or treatment strategies.
- ***Neoplasms*** exhibited a significant surge in mortality rates by 2015, becoming one of the most rapidly growing causes of death, especially among men.
- ***External Causes of Morbidity and Mortality*** surpassed Respiratory system diseases as the third leading cause of the death over a decade; this category includes transportation accidents, falls, accidental poisoning, self-harm, assault, providing US officials with key areas of intervention.

[1] U is the code, as per the ICD 10th code revision, reserved for both provisional assignment of diseases of uncertain origin and for tracking specific experimental categories.

Focusing more on gender trends, the differences in mortality rates between genders remained consistent across both years. While men continued to exhibit higher rates of death from external and cardiovascular causes, women experienced a noticeable increase in cancer-related mortality, emphasizing the need for gender-specific health interventions.

Increase in Mortality Among Older Age Groups

A significant increase in overall mortality was observed in the age groups above 65 years in 2015 compared to 2005, reflecting the general aging of the population over the decade. While the distribution of deaths across age groups remains similar, the mortality peak for individuals aged over 85 appears slightly more pronounced in 2015, which is led by the increase in average mortality for men.

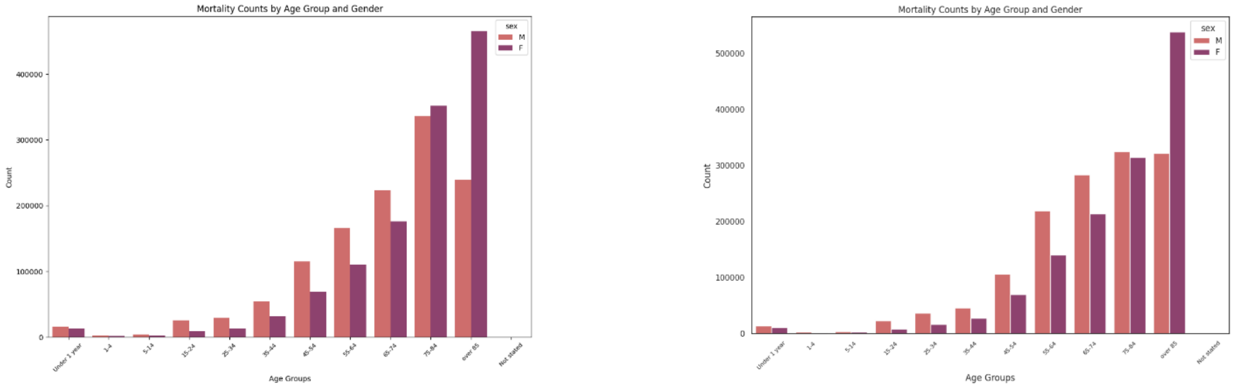


Figure 3: Mortality over Age Groups, 2005 (left) and 2015 (right)

Distribution of Mortality Among Race and Education Level

Lower levels of education are associated with higher mortality rates, regardless of race, in both years. White individuals appear to dominate numerically across all education categories, indicating a significant disparity in mortality.

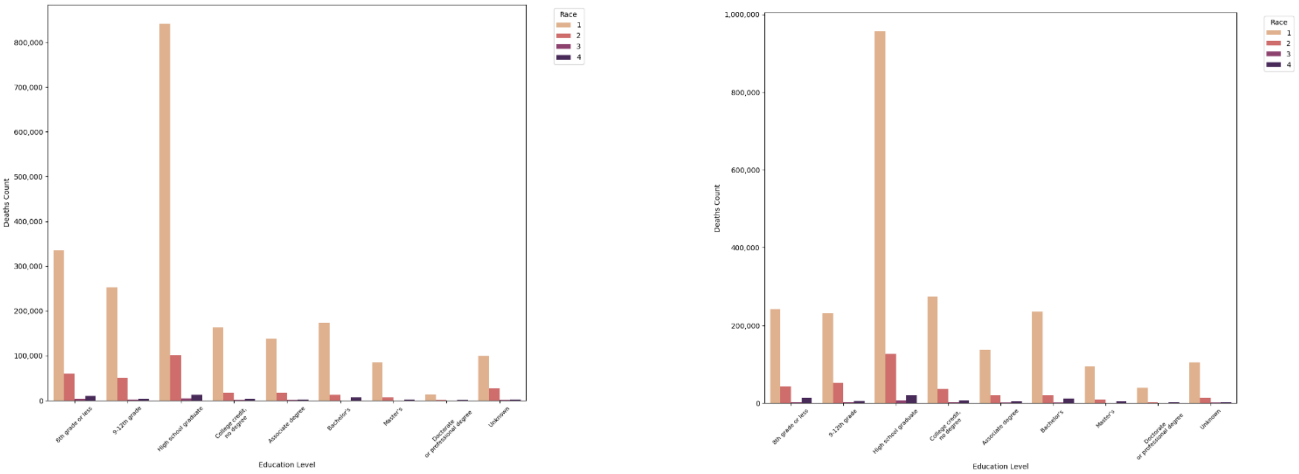


Figure 4: Mortality by education level over races 2005 (left) and 2015 (right)

CLASSIFICATION

Extended Version

In this section we applied several classification models to our dataset to assess how variables such as gender and education influence the incidence of disease, hence we wanted to identify any patterns particularly in disadvantaged social groups (e.g., low education and income levels) within the context of the U.S healthcare system. Since the datasets were severely unbalanced, we also tested the models using weights inversely proportional to a class frequency, to ensure rarer classes were represented fairly during prediction time; nonetheless, the goal of the analysis was not to build the best predictive model overall, but to examine the trends concurring across years 2005 and 2015 with respect to the features involved.

Before delving into the analysis, we decided to drop the columns manner_of_death, age_recode_12, detail_age, day_of_week_of_death and month_of_death, thus reducing the number of predictive features to 12.

The initial classification task focused on predicting type_of_disease as a label, thus a pipeline was implemented to address the multiclassification task ahead following these steps:

- ***StringIndexer*** was used to convert categorical features and the label into numerical values.
- ***VectorAssembler*** was used to combine numerical and indexed features into a single vector, used as input for classification models (the label was not included).

Note that, due to technical issues in training the model and extracting the optimal parameters for the models, only half of the original datasets [2] were used in the following analysis and a randomized search was implemented instead of the standard Grid Search, however stratified sampling was performed to further ensure consistency over the label type of disease's severe imbalance. This choice reflects our intention to preserve the natural distribution of deaths, where more common diseases appear more frequently than rarer ones.

One Vs Rest Model

We decided to adopt the One Vs Rest Model to effectively analyse how the classes relate to each other with respect to being easily distinguishable by the classifier; the classifier we used as a base was the Logistic Regression, whose parameters were tuned using 2-Fold Cross Validation and applied first to the data without weights and then by considering class weights for imbalance treating.

[2] respectively 1,226, 476 and 1,359,912 records for 2005 and 2015.

Random Forest Model

The Random Forest model was chosen as the primary tool for our analysis, while also integrating the insights we got during OVR testing, to specifically delve into the importance of the features during prediction time, as it coincidentally produces the feature importance plot.

As already mentioned with the OVR model, the Random Forest Classifier, whose optimal parameter configuration was also ensured by 2-Fold CV, was applied first to the raw data, then was fitted using the weightCol argument provided by the model itself while passing the weight column already defined.

While the results we obtained by the OVR and Random Forest Classifier, when including weights, were better in the scope of the models being able to identify rarer classes, the overall evaluation metric favoured the unweighted versions of the models, as shown by the evaluation metrics computed, that were selected as the best alternative.

	2005			*2015*		
	Precision	*Recall*	*F1*	*Precision*	*Recall*	*F1*
OVR no weights	0.29	0.42	0.28	0.26	0.39	0.25
OVR with weights	0.38	0.26	0.23	0.41	0.16	0.16
RF no weights	0.21	0.30	0.23	0.17	0.26	0.19
RF with weights	0.23	0.10	0.11	0.17	0.09	0.09

Table 2: Classification results for 22-label task.

By inspecting further, the Feature Importance plots for the RF of both years showcase the effect that applying weights reflects onto the features themselves.

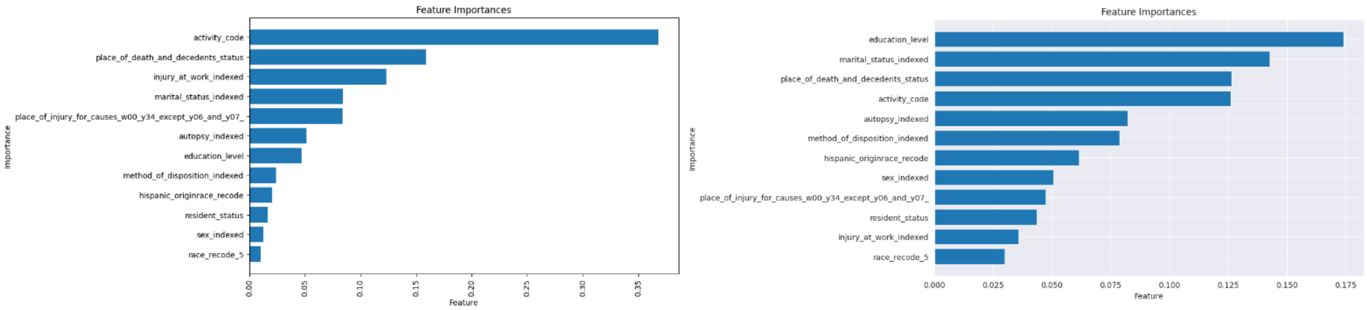


Figure 5: Feature Importance plots for 2005 unweighted RF (left) and weighted (right).

The change in the feature importance of all variables was significant along the board, with the ranking above showcasing that weighting addressed the imbalance by amplifying the importance of features relevant to minority classes predictions, such as education level, that is otherwise overshadowed in the imbalanced setting. The same can be observed in the feature importance plots for 2015 (Fig.6), with education level in particular exhibiting the same behaviour as the 2005 column.

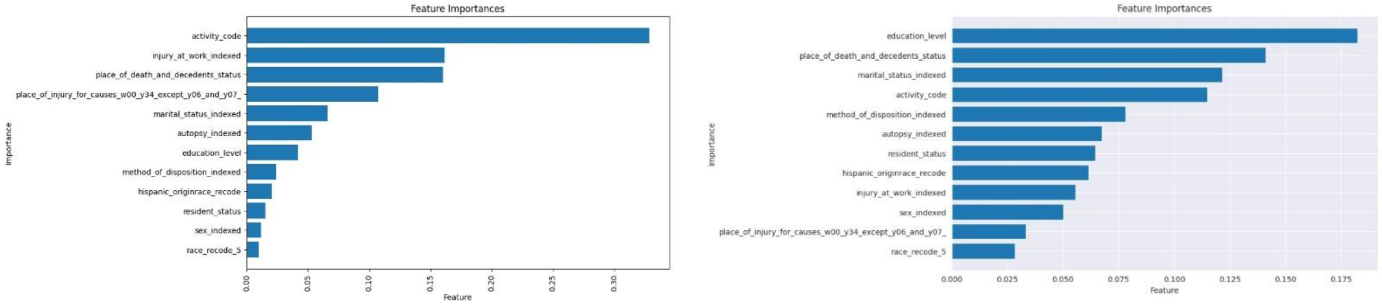


Figure 6: Feature Importance plots for 2015 unweighted RF (left) and weighted (right).

Reduced Version

The results of the 22 labels classification tasks were overall insightful yet the models lacked the ability to predict the rarest diseases, thus we reshaped the task by grouping diseases semantically, to assess whether the underlying values that the features assumed were in a way swaying the prediction dramatically towards the majority classes; to do this we have identified the following groups of diseases:

Table 3: Mapping labels into 7 classes.

Index	Group	Labels included
0	Circulatory system	Kept the same to relate to the other models
1	Neoplasms	
2	External causes of Morbidity and Mortality	
3	Respiratory, digestive and genitourinary disorders	Respiratory system, Digestive system, Genitourinary disorders, Pregnancy/childbirth and puerperium
4	Neurological, behavioural and sensory disorders	Mental and behavioural disorders, Nervous system diseases, Eye and adnexa, Ear and mastoid process
5	Endocrine, metabolic and structural disorders	Endocrine, Nutritional and metabolic diseases, Skin and subcutaneous tissue, Musculoskeletal system and connective tissue

6	Infectious, immune and developmental disorders	Infectious and parasitic diseases, Blood diseases, CMDCA, Perinatal period-originating conditions, Abnormal symptoms

After defining the new labels by mapping their values into the new column grouped_diseases, we proceeded the analysis by replicating the same steps of the previous task, while also including the weighted versions of the models to test whether the weights can be handled better on a slightly less imbalanced setting.

	2005			*2015*		
	Precision	*Recall*	*F1*	*Precision*	*Recall*	*F1*
OVR no weights	0.31	0.42	0.28	0.26	0.39	0.25
OVR with weights	0.38	0.26	0.23	0.33	0.24	0.22
RF no weights	0.19	0.30	0.21	0.19	0.27	0.19
RF with weights	0.23	0.13	0.13	0.19	0.27	0.18

Table 4: Classification results for 7-label task

The results obtained by grouping labels were generally better, and although the Weighted Random Forest Classifier was capable of identifying and classifying correctly minority classes instances to produce a non-negative per class f1 score, the overall F1 score of the models for both years did not warrant the over penalization being applied to the majority class, Circulatory System; this can especially be observed on the Confusion Matrices[3] (Fig. 7) which clearly highlight that the model is confusing class 0 (Circulatory System) with classes 1, 3 and 4 equally for year 2005; while for year 2015 we observed that instances from class 0 get misclassified into class 1, like its model for the 2005 counterpart, class 5 and class 6.

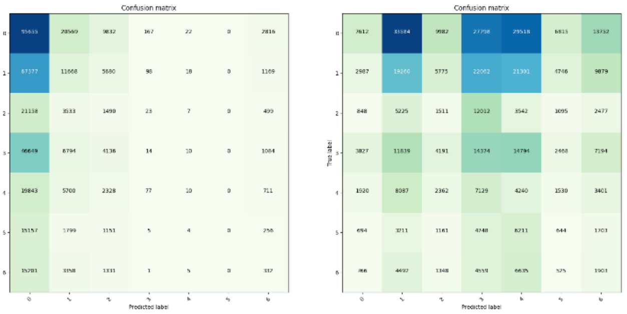


Figure 7 : Matrices for RF, unweighted(left) and weighted (right).

Even though the premises of a reduced task should have helped making a case for the usage of weights to counter the imbalance, we conclude the classification portion of our work by declaring the unweighted versions of the Random Forest to be the better methods for both the extended and the reduced tasks, both by examining the overall evaluation scores and by finding the insights they provide in terms of class distinguishability and feature explanations to be a more truthful and significant representation of the context of our datasets: hence the next part of our study covers Unweighted Random Forests only.

[3] Here included are the Confusion matrices for year 2005.

EXPLAINABILITY

Moving onto the next section of our work, we decided to delve deeply into the Random Forest black box and try to build upon the insights both the OVR models, the Confusion Matrices and especially the Feature Importance plots left us with.

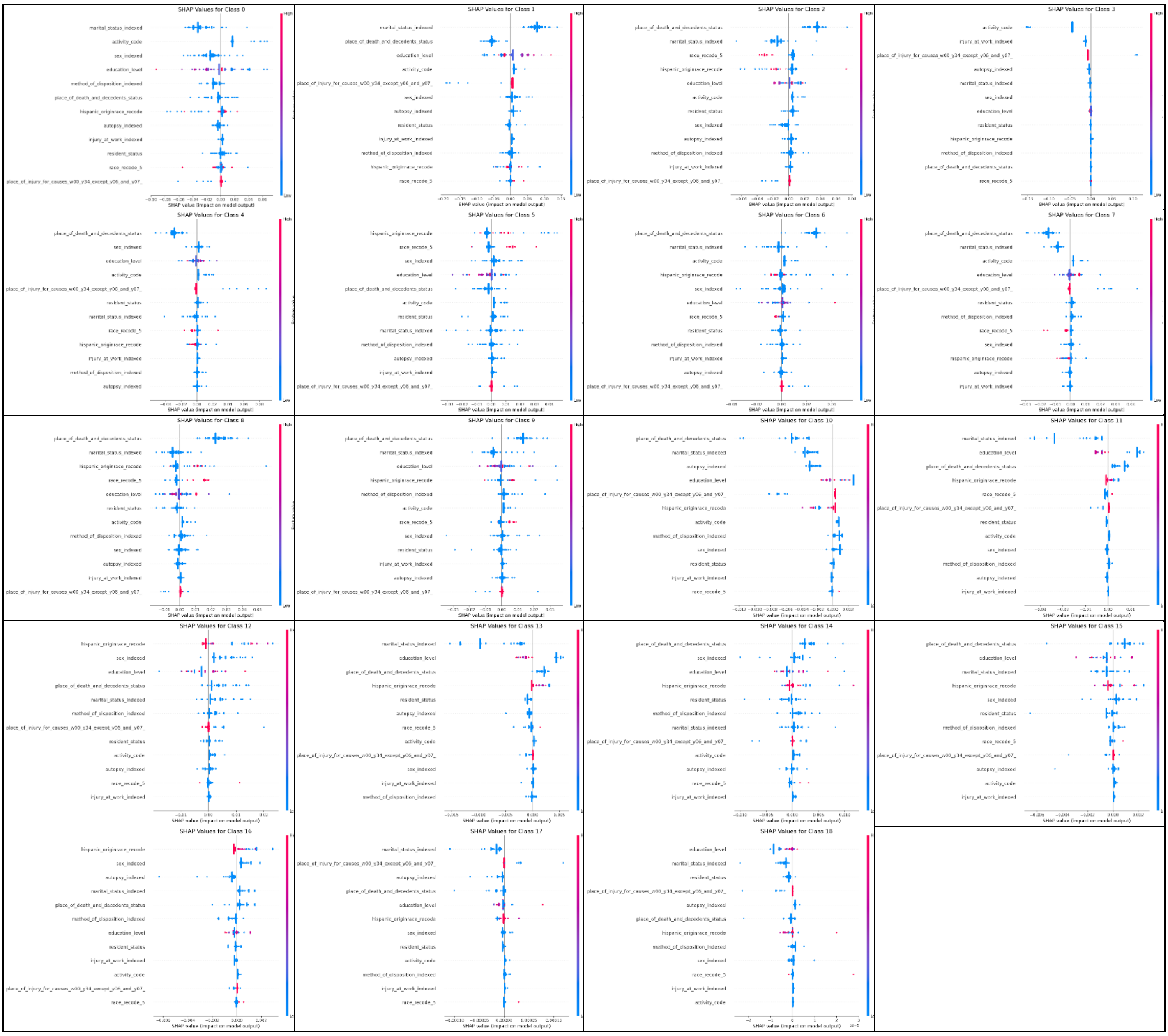
We chose SHAP, particularly its summary plots, to assess the impact of the features in 1,000 misclassified instances[4], and applied it to both the Unweighted Random Forest for the 22-class setting and the 7-class one. Note that the SHAP implementation available on PySpark was not compatible with an indexed label, hence we resorted to training a Random Forest Classifier using the sklearn implementation[5] and then was passed as the black box model for SHAP to explain. Once the SHAP explainer extracted the SHAPley values we then proceeded to plot it; since SHAP was ran on misclassified instances the values on the right side of the line are the ones that swayed the prediction into a misclassification for the specific feature. Upon further inspection of SHAP for the 22-labes task with it became evident that some features were considerably influential into leading to the misclassification of specific types of disease:

- **Race_recode = Asian or Pacific Islander** is a highly discriminative pattern for class 2 (Respiratory System) in misclassifications, suggesting this demographic group being somewhat less exposed to this type of death, while also implying that the other demographic groups, especially **Race_recode = White**, are more vulnerable to it; **Place of Death and Decedents Status= Hospital's Inpatient** is also very influential in misclassifying instances for class 2 (Respiratory system), pointing to this disease not being the leading major cause of death in hospitals, with the swift progression of respiratory system conditions resulting in death before the person is even admitted as an inpatient.
- **Place of Death and Decedents Status= Hospital's Inpatient** is relevant in misclassifying several diseases, including 6 (Digestive System), 8 (Infectious and parasitic), 9 (Genitourinary system), 14 (Blood) and 15 (Skin and subcutaneous tissue), likely for the reason mentioned beforehand with Respiratory System diseases.
- **Education level** showed a generally balanced behaviour over almost the entirety of the classes, spare for class 10 (Abnormal symptoms, signs and findings not elsewhere specified) where its low values, for **Education Level = 8th grade or less,** lead it to be the most important feature for misclassification; and class 1 (Neoplasms) for the opposite, since high values for Education Level, which refer to higher degrees of education, lead to the misclassification of the instances for the class.

[4] Due to technical issues the number of instances chosen was reduced from an initial 10,000 to 1,000.

[5] Also note that the sklearn model was used only after verifying that the underlying trend in (mis)classifying labels was consistent with respect to the RF Classifier trained natively on Spark; the fact that the sklearn implementation produced better results was not considered in the analysis and reasoning.

Table 5: SHAP summary plots for RF 2005.



CONCLUSIONS

Although the results of the classification attempts were not extensive when relating the labels to each other, SHAP's summary plot allowed us to uncover deeper insights into the most underrepresented labels in the dataset by analysing the heterogeneity in the distribution of the SHAP values; some classes for instance, (e.g. class 3, External Causes of Morbidity and Mortality) are strongly influenced by a single features, making them more susceptible to being swayed, while others require a combination of multiple factors to be misclassified or correctly classified. Combined with results of the Confusion Matrices and the Feature Importance plots, our study provided a nuanced understanding of the interaction between features and labels, particularly in the context of underrepresented classes.