Assessing Risk Factors for Treatment Discontinuation in Tuberculosis Patients in Ukraine

REETOM GANGOPADHYAY, BRIAN CAI, CASEY CRARY, KELSON MCBRIDE

ABSTRACT

Treatment adherence is crucial for controlling tuberculosis (TB), especially in regions facing high rates of drug-resistant cases. This study investigates the factors influencing patient dropout from TB treatment in Ukraine from 2018 to 2021. We utilized a combination of statistical and machine learning methodologies to identify and predict key risk factors associated with treatment discontinuation. Our analysis began with multiple logistic regression to establish relationships between potential predictors and dropout. The LASSO regression was then employed to select the most significant variables. Finally, machine learning models, including Classification and Regression Trees (CART) and HistGradientBoostedClassifier, were trained and optimized using GridSearch cross validation to enhance predictive accuracy. Our findings highlight the critical socio-economic and clinical determinants contributing to treatment dropout, providing valuable insights for public health interventions aimed at improving patient retention and TB treatment outcomes in Ukraine.

INTRODUCTION

Tuberculosis (TB) remains a significant public health challenge globally, with treatment adherence being a critical factor in successful disease management. In Ukraine, a country facing a high burden of TB and multi-drug-resistant TB{1,7), patient dropout from treatment programs poses a severe threat to both individual and public health outcomes. The reasons behind this dropout are multifaceted, influenced by socio-economic, demographic, and clinical factors. Identifying these factors is crucial for developing targeted interventions to improve treatment retention and outcomes.

This study aims to explore the determinants of patient dropout from TB treatment in Ukraine, utilizing data from 2018 to 2021. We hypothesize that certain vulnerable groups, such as the homeless and unemployed, are more likely to discontinue treatment prematurely. To investigate this, we employed a combination of biostatistical methods and machine learning techniques. Initially, a multiple logistic regression analysis was conducted to understand the relationship between various predictors and treatment dropout. We then used LASSO (Least Absolute Shrinkage and Selection Operator) to identify the most significant variables influencing dropout. Finally, machine learning models, including Classification and Regression Trees (CART) and HistGradientBoostedClassifier, were trained and optimized using GridSearch to predict dropout risk based on the identified factors. Due to missingness and collection errors, rigorous data cleaning is necessary for any valid analysis.

This integrative approach allows for a robust analysis, combining traditional statistical techniques with advanced machine learning methods to provide a comprehensive understanding of the factors influencing treatment dropout. By identifying key predictors, our findings aim to inform public health strategies and interventions tailored to at-risk populations, ultimately improving treatment adherence and TB control efforts in Ukraine.

METHODS

The data cleaning process was meticulously carried out to ensure the accuracy and consistency of the dataset used for analysis. The details are in the appendix(TABLE 1). Comprehensive data cleaning and preprocessing strategy ensured the dataset's reliability, lay a strong foundation for accurate model training and validation in the subsequent analysis phases.

Our primary research question is regarding dropout rates for tuberculosis (TB) treatment in Ukraine. To address this, we employed a logistic regression model, followed by Least Absolute Shrinkage and Selection Operator (LASSO) regression for variable selection, and machine learning classification models to predict dropout. In the statistics literature, there are multiple ways to carry out variable selection in the context of linear or generalized linear models, however, in our situation – due to the large number of rows and predictors combined with significant missingness, suggest that LASSO is a strong choice for variable selection compared to other methods, such as the commonly utilized stepwise p-value thresholding approach.

Logistic Regression

Logistic regression is a standard statistical method for modeling the relationship between a binary dependent variable Y (dropout: 1 for treatment discontinuation, 0 otherwise) and one or more independent variables x_i (predictors). The logistic regression model is defined as (3):

$$logit\big(P(Y=1)\big) = \beta_0 + \mathrm{\beta}_1 * X_1 + \mathrm{\beta}_2 * X_2 + \cdots \mathrm{\beta}_\mathrm{n} * X_n$$

Where $P(Y = 1)$ is the probability of the event occurring (dropout in this case), β_0 is the intercept and $\beta_1, \beta_2, \dots \beta_n$ are the predictors.

Logistic regression was chosen for this analysis due to its suitability for modeling binary outcomes, which aligns perfectly with our research question about patient dropout from tuberculosis treatment. This method is ideal because it handles cases where the response variable has two possible outcomes—dropout or continuation—allowing us to estimate the probability of dropout effectively. Unlike linear regression, logistic regression transforms the linear combination of predictors into a probability score between 0 and 1, providing insights into the likelihood of dropout and allowing for probabilistic interpretations. Additionally, logistic regression offers clear interpretability, as its coefficients can be converted into odds ratios, which help in understanding the impact of each predictor on dropout odds. The method's flexibility in handling both continuous and categorical predictors make it well-suited for incorporating a range of demographic, socio-economic, and clinical factors. Furthermore, logistic regression serves as a valuable baseline model, enabling us to compare its performance with more complex models, such as LASSO regression and machine learning classifiers. Its capacity to manage imbalanced datasets, where dropout events may be less frequent, further reinforces its appropriateness for our analysis.

LASSO

To refine our model and select the most relevant predictors, we applied LASSO regression, which enhances the logistic regression by adding a penalty to the regression coefficients. The LASSO regression aims to minimize the following objective function(5) :

$$-2\log\,(\mathrm{L}(\beta)) + \lambda * \sum_{j=1}^{p} |\beta_\mathrm{j}|$$

where ß is $(\beta_1, \beta_2 \ldots \beta_p)$ and L(ß) is the likelihood function for logistic regression model, λ is the regularization parameter controlling the L1 penalty given by the sum of the absolute value of the coefficients $(|\beta_j|)$.

The LASSO procedure shrinks some coefficients to exactly zero, effectively performing variable selection and simplifying the model. This results in retaining only the variables with the strongest predictive power—those that predict patient dropout with the highest efficacy. LASSO balances model parsimony against predictive power, ensuring that the final model is both interpretable and robust.

Model Implementation for Logistic Regression with LASSO

1. **Defining Variables:**
   - Response variable (Y): dropout (1 if the patient's final outcome was "Treatment discontinuation", 0 otherwise).
   - Predictor variables (X): Demographic factors (age, sex, region), socio-economic factors (homelessness, unemployment), clinical characteristics (HIV status, drug resistance profiles), and treatment-related factors.
2. **Model Fitting:**
   - We first fit a logistic regression model using all predictors.
   - Next, we applied LASSO regression using the glmnet package in R. We defined X as a model matrix including the selected predictors and Y as the dropout variable.
   - Cross-validation was performed to select the best λ value. The final LASSO model was fitted using this optimal λ

3. **Coefficients and Variable Selection:**
    - The coefficients of the LASSO model were examined to identify significant predictors. Non-zero coefficients indicate variables that contribute to predicting treatment dropout.

By employing logistic regression and LASSO for variable selection, we aimed to build a robust model to identify key factors associated with TB treatment dropout. This approach not only improves model interpretability but also ensures that only the most relevant predictors are retained, reducing the risk of overfitting. The LASSO operation, in particular, provides a method for achieving an optimal balance between model complexity and predictive accuracy, making it a strong method for our analysis.

CART Analysis

CART (Classification and Regression Trees) is a non-parametric decision tree algorithm that is widely used for classification tasks. The primary advantage of CART lies in its ability to model complex, non-linear relationships between the predictors and the response variable. CART works by recursively partitioning the dataset into subsets that are increasingly homogeneous with respect to the target variable. The splitting criterion is based on measures like Gini impurity or entropy, with the goal of maximizing the purity of the nodes.

The decision tree in CART is represented as the following:

$$f(x) = \sum_{m=1}^{M} \mathrm{c}_{\mathrm{m}} 1(x \in R_m)$$

Where:

- x represents the input features,
- M is the number of terminal nodes (leaves) in the tree,
- c_m is the prediction associated with region R_m
- $1(x \in R_m)$ is an indicator function that equals 1 if x belongs to region R_m and 0 otherwise.

In this analysis, the CART model was optimized using GridSearchCV, where a range of hyperparameters—such as max_depth, min_samples_split, min_samples_leaf, and max_features—were tuned to achieve the best possible performance. The optimal parameters were then used to fit the model on the training data, and the model's performance was evaluated using accuracy, confusion matrices, and classification reports.

Histogram-based Gradient Boosting Classification Tree

HistGradientBoostedClassifier is an advanced ensemble learning method that builds upon the principles of gradient boosting. It generates an ensemble of decision trees, where each tree is trained to correct the errors of the preceding ones. The "Hist" in HistGradientBoostedClassifier refers to the histogram-based method it uses to speed up the training process by discretizing continuous features. This is a similar model to models such as LightGBM or XGboost which are highly utilized models in statistics and machine learning. Previous work has been performed by Bui et al. (6) which utilizes several machine learning models – including gradient boosted models and a decision tree – in regards to drug resistance.

Mathematically, we denote this model by the following(4) :

$$\mathrm{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \widehat{y_i}^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)$$

Where l(x) is some CART learner. And y_i is some label from the training data.

The HistGradientBoostedClassifier was also optimized using GridSearchCV, focusing on parameters such as learning_rate, max_iter, max_leaf_nodes, and min_samples_leaf. By fine-tuning these hyperparameters, the model was able to balance the trade-off between bias and variance, leading to an accurate and efficient classifier. The model's performance was evaluated using standard metrics like accuracy, confusion matrices, and classification reports.

RESULTS

LASSO

The variable list provided for the LASSO operation was the following:

FORMULA:

```
dropout ~ { Sex, Region, imputed_weight, Age, DST_R, Localization,
hiv_def, hiv, Cot, Alcohol.abuse, Injecting.drug.user, Homeless,
Unemployed, healthcare_worker, Prisoner, migrant_refugee,
prev_treatment, Bactec, LJ, GeneXpert, DST_E, DST_Z, DST_S, DST_H,
DST_Am, DST_Cm, DST_LFX, DST_MFX, DST_PAS, DST_Km, DST_Ofx, DST_Et,
DST_Lzd, DST_Cs} (See TABLE 1 for full)
```

The LASSO procedure on the Logistic Regression returned the following results:

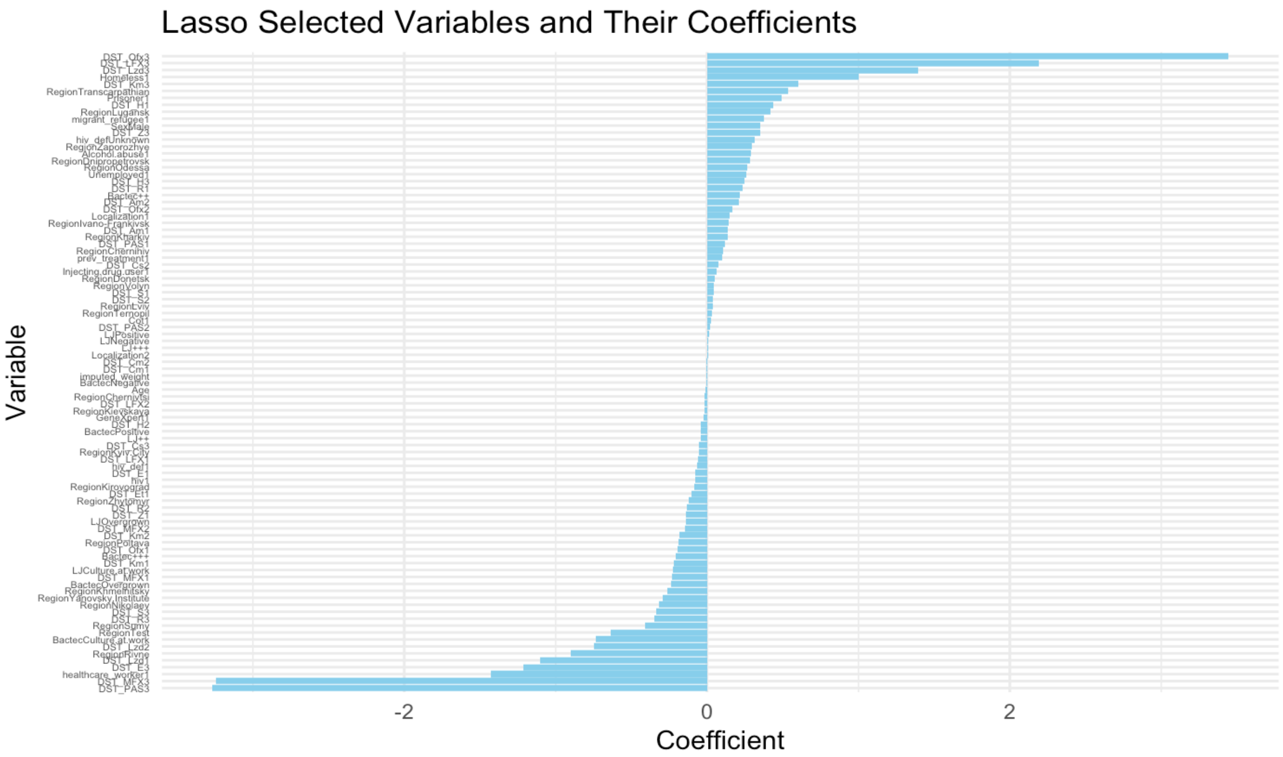


Fig1: LASSO variable Selection plot with coefficients

From the original list of variables there are several that were dropped (see appendix) and there are also several variables which have a coefficient very close to zero. These variables can be considered to be unimportant.

The results from the LASSO procedure provided key insights into the predictors of patient dropout in tuberculosis treatment. LASSO, by design, shrinks the coefficients of less important variables towards zero, effectively performing variable selection. In this analysis, several variables were either dropped entirely or had their coefficients reduced to values close to zero, indicating that these variables contribute minimally to predicting patient dropout.

For instance, variables such as `RegionKherson` and `RegionVinnitsa` were excluded from the model, while others like `RegionPoltava` and `DST_Lzd1` exhibited coefficients close to zero, implying limited predictive power. On the other hand, certain variables retained larger coefficients, highlighting their significance. For example, `Homeless1` and `DST_Ofx3` had coefficients of 0.999 and 3.44, respectively, indicating a stronger relationship with the dropout outcome.

The LASSO results, therefore, help refine the model by focusing on the most impactful predictors, enhancing both the interpretability and predictive accuracy of the logistic regression model that follows. The final model, now more parsimonious, is well-suited for subsequent analyses and predictive tasks, offering a clearer understanding of the factors most closely associated with treatment discontinuation.

In our analysis, the selection of the regularization parameter, lambda, plays a crucial role in balancing the trade-off between model complexity and prediction accuracy. The binomial deviance is a key metric used to assess the model's fit to the data, with lower values indicating better fit.

As we adjusted the lambda parameter, we observed the following:

- When log(λ) is approximately -6.2, the binomial deviance is around 0.51.
- The absolute minimum binomial deviance occurs at log(λ) ≈ −8.8, where the deviance reaches about 0.507.

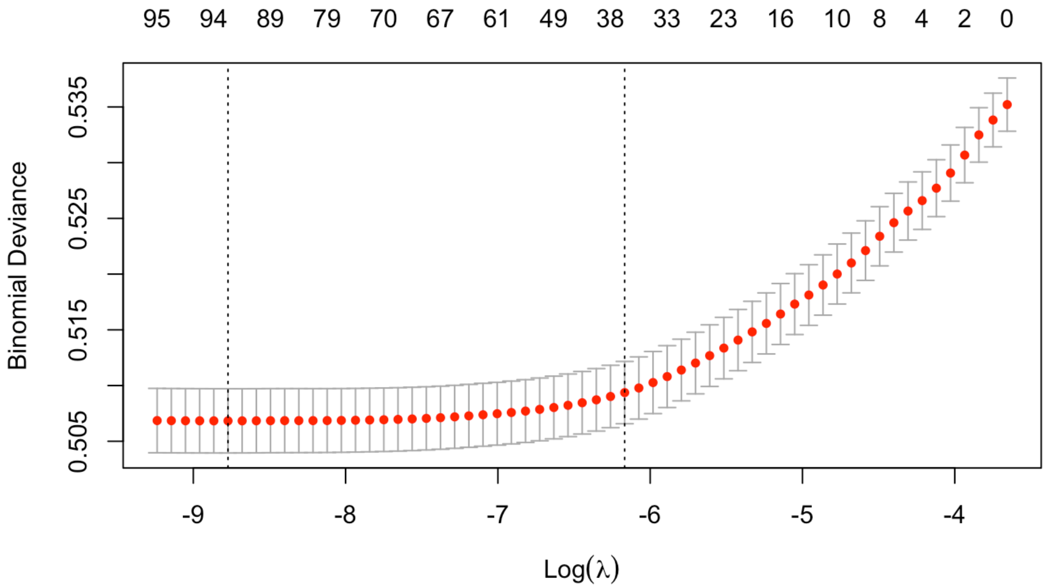


Fig2: Binomial Deviance from Lambda parameter of LASSO

This indicates that as we decrease the regularization strength (i.e., reduce λ), the model becomes more flexible, allowing for a better fit to the data. However, this comes with the risk of overfitting, which must be balanced against the benefits of a lower deviance. The slight reduction in binomial deviance as log(λ) decreases suggests that while the model's fit improves, the gain in predictive accuracy may be marginal. In our case the optimal log(λ) would be the value near -6.2, due to the relatively marginal increase in deviance.

Based on the coefficients from the LASSO model, several key variables emerge as significantly important in predicting patient dropout from tuberculosis treatment.

The variable `Homeless1` stands out with the highest coefficient of 0.999, indicating that being homeless is a strong predictor of dropout. Similarly, the variable `healthcare_worker1` has a substantial negative coefficient of -1.431, suggesting that healthcare workers are significantly less likely to drop out of treatment. The coefficient for `Prisoner1` is 0.495, reflecting a notable association with an increased risk of dropout. Additionally, the `migrant_refugee1` variable, with a coefficient of 0.378, also points to a higher likelihood of dropout among migrants and refugees.

Drug susceptibility test results play a critical role as well. Variables such as `DST_MFX3` and `DST_PAS3` have very large negative coefficients of -3.242 and -3.267, respectively, indicating a strong association with a reduced risk of dropout. Conversely, `DST_Ofx3` and `DST_LFX3` have significant positive coefficients of 3.442 and 2.189, suggesting these results are strongly associated with an increased risk of dropout.

CART and HistGradientBoostedClassifier

The results (see appendix$_{\text{(TABLE 3)}}$) from the CART and Histogram-Based Gradient Boosting Classifier (HistGradientBoosting) models reveal distinct differences in their performance, particularly in how they manage the imbalanced dataset, where the majority class (0) vastly outnumbers the minority class (1).

The initial CART model achieved an accuracy of 91.6%. However, it struggled significantly with predicting the minority class, yielding a very low recall of 0.04 and a precision of 0.24, resulting in an F1-score of just 0.08. This indicates that the model was heavily biased towards predicting the majority class, as reflected in the confusion matrix where 2,185 instances of the minority class were misclassified. After tuning with GridSearchCV, the CART model's accuracy slightly improved to 92.4%. Despite this, the recall for the minority class dropped to zero, meaning the model failed entirely to detect any instances of the minority class. This result underscores the limitation of the CART model in handling severe class imbalance, even when optimized through hyperparameter tuning.

The initial HistGradientBoosting model performed better than CART, with an accuracy of 92.4% and a slightly higher recall of 0.06 for the minority class. Although the recall was still low, the model managed to detect more instances of the minority class, resulting in an F1-score of 0.10. After tuning, the HistGradientBoosting model's accuracy improved marginally to 92.5%. The precision for the minority class increased to 0.63, but the recall remained low at 0.05, showing that the model, while better than CART, still struggled with the minority class.

When both models were evaluated again using a reduced number of variables, based on the most important features identified through the LASSO procedure, the results remained consistent. The CART model maintained an accuracy of 92.4% but continued to fail in detecting the minority class, with a recall of zero. The HistGradientBoosting model also showed similar performance with a recall of 0.0004 for the minority class, indicating that the reduction in parameters did little to improve its ability to predict the minority class.

Overall, while both models achieve high overall accuracy, they are not well-suited for handling class imbalance in this dataset, particularly for predicting the minority class. These results suggest that alternative approaches, such as resampling techniques or adjusting class weights, might be necessary to improve the detection of the minority class and create a more balanced model performance. Likely this issue arose due to the high number of binary or ordinal variables and very few continuous variables as a result of the structure of the data.

CONCLUSION

This study focused on identifying factors predicting patient dropout from tuberculosis treatment in Ukraine from 2018 to 2021. We utilized a variety of statistical and machine learning techniques, including multiple logistic regression, LASSO for variable selection, and advanced classification models such as CART and HistGradientBoostedClassifier. The results highlight several significant predictors of dropout, including socio-demographic factors like homelessness and unemployment, as well as clinical factors such as previous treatment history and the presence of drug resistance.

Our analysis revealed that LASSO regularization effectively reduced the number of predictors, with some variables having coefficients close to zero, suggesting they are less important in predicting dropout. The optimal lambda value for LASSO was found to be around $\log(\lambda)\approx 6.2$, resulting in a binomial deviance of approximately 0.51. The absolute minimum deviance observed was about 0.507 at $\log(\lambda)\approx -8.8$, indicating a trade-off between model fit and complexity.

The variables found to be highly important (Homeless1, healthcare_worker1, Prisoner1, migrant_refugee1, DST_MFX3, DST_PAS3, DST_Ofx3, DST_LFX3) highlight the significant impact of socio-demographic factors and drug resistance profiles on treatment adherence. Understanding these key predictors can guide targeted interventions to reduce dropout rates and improve treatment outcomes.

Drawbacks

1. **Missing Data Handling**: While multiple imputation was used to address missing data, the approach to treating coefficients close to zero (≤ 0.01) as missing could potentially mask useful information and affect the model's performance. This method of imputation might not capture all nuances in the data. Missing data also forced the usage of ordinal or binary variables - which may have potentially removed important information from the data if they were continuous.
2. **Model Overfitting**: The flexibility of the model increases with lower lambda values, which can lead to overfitting. Although the binomial deviance decreased with lower lambda, it is crucial to ensure that the model remains generalizable and does not capture noise as significant predictors.
3. **Limited Scope of Data**: The analysis is constrained to data from Ukraine and may not be generalizable to other regions or settings. Differences in healthcare infrastructure, patient demographics, and disease prevalence could affect the applicability of these findings elsewhere.
4. **Unobserved Variables**: Certain potentially important predictors may not have been included in the dataset, leading to omitted variable bias. Factors such as patient adherence behavior, social support systems, and local healthcare practices could also influence dropout rates.

Future Work

1. **Enhanced Data Collection**: Future research should aim to collect more comprehensive data, including additional socio-economic, behavioral, and clinical variables that could impact dropout rates. Longitudinal data could provide deeper insights into patient behavior over time.
2. **Model Improvement**: Exploring alternative regularization techniques and models, such as Elastic Net or deep learning approaches, could improve predictive performance and handle complex interactions between variables more effectively.
3. **Cross-Regional Studies**: Extending the analysis to other regions or countries would help determine the generalizability of the findings and identify region-specific factors influencing patient dropout.
4. **Incorporation of Qualitative Data**: Qualitative research methods, such as patient interviews or focus groups, could provide additional context and insights into the reasons behind dropout, complementing quantitative findings.
5. **Real-Time Data Analysis**: Implementing real-time data collection and analysis could facilitate more immediate interventions and support for patients at risk of dropping out, potentially improving treatment adherence and outcomes.

In addition to LASSO, another variable selection method that could potentially improve the model's performance is **Elastic Net**. Elastic Net combines both LASSO and Ridge regression penalties (L1 and L2 respectively), allowing for the selection of groups of correlated variables while also shrinking some coefficients to zero. This hybrid approach balances the strengths of both techniques, making it particularly useful when the dataset contains many correlated variables, as might be the case with the geographic regions or DST results in this study.

Elastic Net performs variable selection like LASSO, but its inclusion of Ridge regularization helps prevent some of the over-penalization seen in LASSO, where certain variables might be prematurely removed. This dual regularization might help retain more informative variables while reducing model variance, improving overall predictive power.

Tuning the mixing parameter, which controls the trade-off between LASSO and Ridge, allows for more flexibility compared to using LASSO alone. This could result in a more robust model, especially when dealing with multicollinearity among the predictors. Thus, implementing Elastic Net could refine feature selection and improve generalization.By addressing these areas, future research can build on these findings to develop more effective strategies for improving tuberculosis treatment adherence and reducing dropout rates.

APPENDIX

TABLE 1) Data Cleaning

1. **Variable Standardization and Renaming**:
   - **Age**: Corrected two entries showing values ">100" by taking the modulo 100, converting 134 to 34 – for example.
   - **Treatment Dates**: Renamed Treatment.start.date and Treatment.end.date to start_date and end_date respectively, and converted them to date/time format.
   - **HIV Testing**: Renamed HIV.testing to hiv_testing and converted it to date/time format.
   - **Has.started.to.take.ART.**: Renamed to takes_art. The start date for ART was set in m/d/y format. Missing values were considered for special handling but left commented out.
   - **Cotrimoxazole Treatment**: Changed to m/d/y format, with missing values considered for potential imputation.
2. **Categorical Variable Conversion**:
   - **Localization**: Re-encoded as 1 for Pulmonary, 2 for Extra-pulmonary, and 0 for Both.
   - **Cavitation**: Recoded as 1 for Yes and 0 for No.
   - **HIV Definition**: Recoded as 0 for negative and 1 for positive.
   - **Alcohol Abuse** and **Injecting Drug User**: Set to 0 if '-' or NA, and 1 otherwise.
   - **Homeless, Unemployed, Healthcare Worker, Prisoner**: Recoded as 0 for No and 1 for Yes, with '-' considered as No.
   - **GeneXpert**: Recoded as 0 for Negative and 1 for Positive, with NA considered as Negative.
3. **New Variables**:
   - **Imputed Weights**: A new variable, imputed_weights, was created following specific rules:
     1. Converted weights from pounds to kilograms for values between 140 and 240.
     2. Added a decimal point for weights between 240 and 1000.
     3. Set values greater than 1000 to NA for imputation.
     4. Multiple imputation was conducted using the MICE package in R, based on Age, Sex, and Weight, ensuring the distribution was plausible.
4. **DST Variables**:
   - Resistance testing results were uniformly transformed: missing values were set to 0, "Resistant" to 1, "Sensitive" to 2, and "Contaminated" to 3. The variables were renamed with the convention DST_XYZ (e.g., DST.results.R became DST_R).
5. **Additional Derived Variables**:

- **HIV Status (hiv)**: Created as a binary indicator where 1 indicates the patient has started ART and 0 otherwise.
- **Cotrimoxazole Treatment Indicator (Cot)**: Created as a binary indicator where 1 indicates treatment and 0 otherwise.
- **Previous Treatment (prev_treatment)**: Derived from the new_prev variable, where "Previously treated" was set to 1 and other values to 0.

6. **Handling of Missing Values and Data Imputation**:
   - For key variables, missing data were addressed through a combination of recoding and multiple imputation techniques, ensuring that the dataset was suitable for subsequent statistical analyses and machine learning modeling.

TABLE 2) Results From LASSO

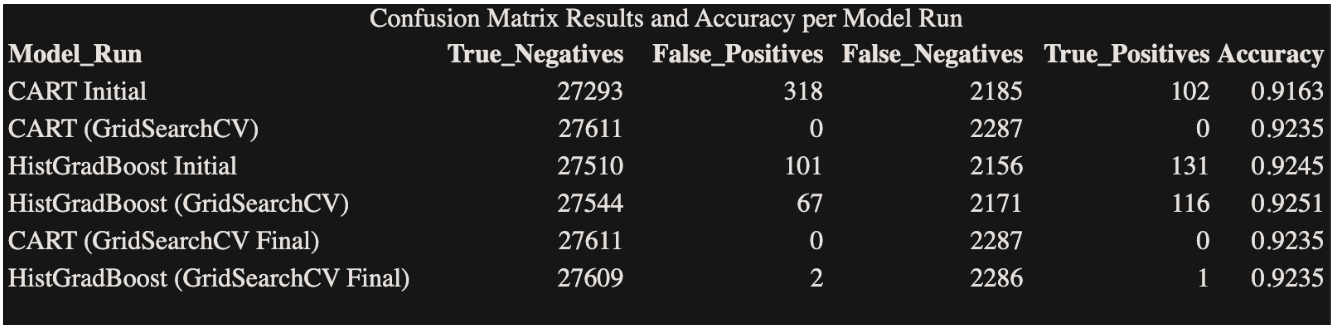
Variable	Coefficient
SexMale	0.3516979
RegionChernihiv	0.1066500
RegionChernivtsi	-0.0137311
RegionDnipropetrovsk	0.2853924
RegionDonetsk	0.0491183
RegionIvano-Frankivsk	0.1449775
RegionKharkiv	0.1350622
RegionKherson	NA
RegionKhmelnitsky	-0.2615948
RegionKievskaya	-0.0144628
RegionKirovograd	-0.0820276
RegionKyiv.City	-0.0532337
RegionLugansk	0.4201572
RegionLviv	0.0395706
RegionNikolaev	-0.3178960
RegionOdessa	0.2624691
RegionPoltava	-0.1900173
RegionRivne	-0.8999509
RegionSumy	-0.4109054
RegionTernopil	0.0334789
RegionTest	-0.6375044
RegionTranscarpathian	0.5363307
RegionVinnitsa	NA
RegionVolyn	0.0469573
RegionYanovsky.Institute	-0.2957390
RegionZaporozhye	0.2973721

Variable	Coefficient
RegionZhytomyr	-0.1227272
imputed_weight	-0.0067873
Age	-0.0082956
DST_R1	0.2373843
DST_R2	-0.1326427
DST_R3	-0.3491663
Localization1	0.1483004
Localization2	0.0071788
hiv_def1	-0.0669804
hiv_defUnknown	0.3158505
hiv1	-0.0802313
Cot1	0.0258594
Alcohol.abuse1	0.2910178
Injecting.drug.user1	0.0643145
Homeless1	0.9994783
Unemployed1	0.2576963
healthcare_worker1	-1.4305063
Prisoner1	0.4950078
migrant_refugee1	0.3784666
prev_treatment1	0.0996938
Bactec++	0.2191241
Bactec+++	-0.2071847
BactecCulture.at.work	-0.7374457
BactecNegative	-0.0069354
BactecOvergrown	-0.2404565
BactecPositive	-0.0405997
LJ++	-0.0416882
LJ+++	0.0086706
LJCulture.at.work	-0.2253568
LJNegative	0.0090783
LJOvergrown	-0.1414491
LJPositive	0.0165675
GeneXpert1	-0.0239987
DST_E1	-0.0751946
DST_E2	NA
DST_E3	-1.2142835
DST_Z1	-0.1395251

Variable	Coefficient
DST_Z2	NA
DST_Z3	0.3512278
DST_S1	0.0437481
DST_S2	0.0411886
DST_S3	-0.3385024
DST_H1	0.4357974
DST_H2	-0.0384911
DST_H3	0.2446603
DST_Am1	0.1359413
DST_Am2	0.2118227
DST_Am3	NA
DST_Cm1	-0.0043929
DST_Cm2	-0.0039722
DST_Cm3	NA
DST_LFX1	-0.0614302
DST_LFX2	-0.0138170
DST_LFX3	2.1891223
DST_MFX1	-0.2339815
DST_MFX2	-0.1431959
DST_MFX3	-3.2423270
DST_PAS1	0.1199452
DST_PAS2	0.0182629
DST_PAS3	-3.2665255
DST_Km1	-0.2203190
DST_Km2	-0.1827297
DST_Km3	0.6052821
DST_Ofx1	-0.1938571
DST_Ofx2	0.1702000
DST_Ofx3	3.4415141
DST_Et1	-0.1025485
DST_Et2	NA
DST_Et3	NA
DST_Lzd1	-1.1049551
DST_Lzd2	-0.7440722
DST_Lzd3	1.3926213
DST_Cs1	NA
DST_Cs2	0.0754622

Variable	Coefficient
DST_Cs3	-0.0528177

TABLE 3) Machine Learning Results



Confusion Matrix Results and Accuracy per Model Run

Model_Run	True_Negatives	False_Positives	False_Negatives	True_Positives	Accuracy
CART Initial	27293	318	2185	102	0.9163
CART (GridSearchCV)	27611	0	2287	0	0.9235
HistGradBoost Initial	27510	101	2156	131	0.9245
HistGradBoost (GridSearchCV)	27544	67	2171	116	0.9251
CART (GridSearchCV Final)	27611	0	2287	0	0.9235
HistGradBoost (GridSearchCV Final)	27609	2	2286	1	0.9235

BIBLIOGRAPHY

1) UKRAINE TUBERCULOSIS ROADMAP OVERVIEW, FISCAL YEAR 2023. (2023).

https://www.usaid.gov/sites/default/files/2024-02/Ukraine_TB_roadmap_narrative-22_508.pdf

2) Hauer B, Kröger S, Haas W, Brodhun B. Tuberculosis in times of war and crisis: Epidemiological trends and characteristics of patients born in Ukraine, Germany, 2022. Euro Surveill. 2023 Jun;28(24):2300284. doi: 10.2807/1560-7917.ES.2023.28.24.2300284. PMID: 37318760; PMCID: PMC10318937.

3) Multiple Logistic Regression Analysis. (n.d.). https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704

ep713_multivariablemethods/BS704-EP713_MultivariableMethods4.html

4) Leventis, D. (2022, January 2). XGBoost Mathematics Explained - Dimitris Leventis - Medium. Medium.

https://dimleve.medium.com/xgboost-mathematics-explained-58262530904a

5) A gentle introduction to logistic regression and lasso regularisation using R. (2017, October 6). Eight to Late. https://eight2late.wordpress.com/2017/07/11/a-gentle-introduction-to-logistic-regression-and-lasso-regularisation-using-r/

6) Bui VCB, Yaniv Z, Harris M, Yang F, Kantipudi K, Hurt D, Rosenthal A, Jaeger S. Combining Radiological and Genomic TB Portals Data for Drug Resistance Analysis. IEEE Access. 2023;11:84228-84240. doi: 10.1109/access.2023.3298750. Epub 2023 Jul 25. PMID: 37663145; PMCID: PMC10473876.

7) Kuzyk PV, Padilla R, Rybak NR, Hoshovska II, Kitov VO, Savchyna MO, Jenkins HE, Chiang SS, Horsburgh CR, Dolynska M, Petrenko V, Gychka SG. Missed Tuberculosis Diagnoses: Analysis of Pediatric Autopsy Data From General Hospitals in Lviv, Ukraine. J Pediatric Infect Dis Soc. 2022 Jun 22;11(6):300-302. doi: 10.1093/jpids/piac016. PMID: 35395086; PMCID: PMC9214781.

8) Uniting for Ukraine – Drug Resistance Background and Recommendations. (2024, March 11). Tuberculosis (TB). https://www.cdc.gov/tb/php/dear-colleague-letters/2022-uniting-for-ukraine-drug-resistance.html