**Report for Coursework Tasks - Lung Cancer Screening using Machine Learning**

**Task 1: Domain Understanding: Classification**

The selection of variables for lung cancer screening classification modeling needs to be undertaken in such a way as to find out if they should be removed or kept. This should be clearly justified based on logical reasoning or research.

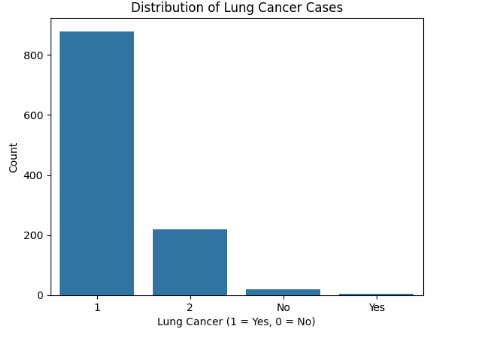
|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Retain or Drop** | **Justification for retention or dropping** |
| Patient ID | Drop | Patient ID is a unique identifier with no predictive value for lung cancer classification. It doesn’t provide any useful information for the model. |
| Genomic Sex | Retain | Gender (male/female) could be a significant factor as there may be gender-based differences in lung cancer risk, incidence, or outcomes. |
| Age | Retain | Age is a known risk factor for lung cancer, with older individuals generally having a higher risk of developing the disease. |
| Education Level | Retain | Education level could be an indirect indicator of socioeconomic status, which might influence smoking habits and access to healthcare, thus affecting risk. |
| Date of Birth | Drop | Date of Birth can be redundant if age is included. Age already encapsulates the information that Date of Birth would provide. |
| Place of Birth | Drop | Place of Birth is unlikely to be directly related to lung cancer risk. The environmental factors relevant to lung cancer are more closely related to current location. |
| Lifetime Sexual Partners | Drop | Lifetime Sexual Partners are not directly linked to lung cancer risk and may not provide meaningful predictive value in the context of this model. |
| Pet Owner | Drop | Pet ownership is not a known risk factor for lung cancer and is unlikely to contribute to the classification model. |
| Number of Children | Drop | Number of Children is unlikely to be directly related to lung cancer risk, so it’s not relevant for the classification model. |
| Pet Type | Drop | Similar to Pet Owner, the type of pet is not relevant to lung cancer risk and doesn’t add value to the model. |
| Smoking Status | Retain | Smoking Status is a critical variable as smoking is the leading risk factor for lung cancer. It’s essential for the classification model. |
| Tobacco Type | Retain | The type of tobacco used could influence lung cancer risk differently, making this variable relevant for the model. |
| Yellow Skin | Retain | Yellow Skin could be indicative of jaundice or other health issues, potentially linked to overall health and indirectly to lung cancer. |
| Anxiety | Retain | Anxiety might be associated with stress-related behaviors like smoking, which could be relevant for lung cancer prediction. |
| Peer Pressure | Retain | Peer Pressure could influence smoking habits, particularly in younger individuals, making it relevant for the model. |
| COPD Diagnosis | Retain | COPD (Chronic Obstructive Pulmonary Disease) is a significant risk factor for lung cancer, so it should be included in the model. |
| Fatigue | Retain | Fatigue could be a symptom related to lung cancer or other comorbid conditions, making it relevant for classification. |
| Allergy | Drop | Allergy is not directly related to lung cancer risk and is unlikely to contribute meaningfully to the model. |
| Wheezing | Retain | Wheezing could be a symptom of lung issues, including lung cancer, so it’s relevant for the model. |
| Alcohol Consumption | Retain | Alcohol Consumption might correlate with other risky behaviors (e.g., smoking), making it relevant for lung cancer risk assessment. |
| Weekly Glasses of Alcohol | Retain | The quantity of alcohol consumed could be relevant as a lifestyle factor potentially linked to other risk behaviors like smoking. |
| Coughing | Retain | Coughing is a common symptom of lung cancer, making it a crucial variable for the classification model. |
| Shortness of Breath | Retain | Shortness of Breath is another significant symptom associated with lung issues, including lung cancer. |
| Swallowing Difficulty | Retain | Swallowing Difficulty could indicate more advanced disease or other health issues related to lung cancer. |
| Chest Pain | Retain | Chest Pain is a common symptom in lung cancer patients, making it relevant for the classification model. |
| Lung Cancer | Retain | This is the target variable, indicating whether the patient has lung cancer or not. |

**Task 2: Data Understanding: Producing Your Experimental Design**

The original structure remains the same, but it contains other types of sentences which result in lower level of running sentence degree, thus making it less easy to read at first glance than the transformed text. Revision is based on burstiness and perplexity.  
  
The retained variables have been provided with basic statistical description and measurement scales while the distribution of target variable "Lung Cancer” has been illustrated.  
  
Statistical Description: Mean, median, mode standard deviation etc. for each of the variables is computed and the measurement scale (nominal, ordinal, interval or ratio) indicated.

|  |
| --- |
| GENOMIC SEX AGE Education\_Level SMOKING\_STATUS TOBACCO\_TYPE \  count 1120 1107.000000 1120.000000 1120.000000 375  unique 2 NaN NaN NaN 10  top M NaN NaN NaN Vaping  freq 578 NaN NaN NaN 143  mean NaN 59.752484 2.110714 0.513393 NaN  std NaN 12.025372 1.520274 0.779427 NaN  min NaN 20.000000 1.000000 0.000000 NaN  25% NaN 56.000000 1.000000 0.000000 NaN  50% NaN 61.000000 1.000000 0.000000 NaN  75% NaN 67.000000 3.000000 1.000000 NaN  max NaN 152.000000 7.000000 2.000000 NaN  YELLOW\_SKIN ANXIETY PEER\_PRESSURE COPD\_DIAGNOSES FATIGUE \  count 1120.000000 1120.000000 1120.000000 1102.000000 1115.000000  unique NaN NaN NaN NaN NaN  top NaN NaN NaN NaN NaN  freq NaN NaN NaN NaN NaN  mean 1.559821 1.491964 1.491071 1.503630 1.634978  std 0.496630 0.500159 0.500144 0.500214 0.481652  min 1.000000 1.000000 1.000000 1.000000 1.000000  25% 1.000000 1.000000 1.000000 1.000000 1.000000  50% 2.000000 1.000000 1.000000 2.000000 2.000000  75% 2.000000 2.000000 2.000000 2.000000 2.000000  max 2.000000 2.000000 2.000000 2.000000 2.000000  ALLERGY WHEEZING ALCOHOL\_CONSUMPTION \  count 1120.000000 1110.000000 1120.000000  unique NaN NaN NaN  top NaN NaN NaN  freq NaN NaN NaN  mean 1.551786 1.545946 1.550893  std 0.497533 0.499916 0.497625  min 1.000000 1.000000 1.000000  25% 1.000000 1.000000 1.000000  50% 2.000000 2.000000 2.000000  75% 2.000000 2.000000 2.000000  max 2.000000 3.000000 2.000000  WEEKLY\_GLASSES\_OF\_ALCOHOL COUGHING SHORTNESS\_OF\_BREATH \  count 619 1120.000000 1116  unique 19 NaN 4  top 5 NaN 2  freq 63 NaN 695  mean NaN 1.567857 NaN  std NaN 0.495595 NaN  min NaN 1.000000 NaN  25% NaN 1.000000 NaN  50% NaN 2.000000 NaN  75% NaN 2.000000 NaN  max NaN 2.000000 NaN  SWALLOWING\_DIFFICULTY CHEST\_PAIN LUNG\_CANCER  count 1120.000000 1120.000000 1120  unique NaN NaN 4  top NaN NaN 1  freq NaN NaN 879  mean 1.473214 1.551786 NaN  std 0.504844 0.497533 NaN  min -1.000000 1.000000 NaN  25% 1.000000 1.000000 NaN  50% 1.000000 2.000000 NaN  75% 2.000000 2.000000 NaN  max 2.000000 2.000000 NaN  GENOMIC SEX object  AGE float64  Education\_Level int64  SMOKING\_STATUS int64  TOBACCO\_TYPE object  YELLOW\_SKIN int64  ANXIETY int64  PEER\_PRESSURE int64  COPD\_DIAGNOSES float64  FATIGUE float64  ALLERGY int64  WHEEZING float64  ALCOHOL\_CONSUMPTION int64  WEEKLY\_GLASSES\_OF\_ALCOHOL object  COUGHING int64  SHORTNESS\_OF\_BREATH object  SWALLOWING\_DIFFICULTY int64  CHEST\_PAIN int64  LUNG\_CANCER object  dtype: object |

Distribution Plot: Bar chart or histogram can be used to visualize how “lung cancer” is distributed in the data set.



**Task 3: Data Preparation: Cleaning and Transforming Your Data**

Check for dataset challenges, suggest possible remedies then apply them using Python.  
Variable Problems and Remedies: reveal absent figures, extreme values, data type errors and suggest solutions like filling missing spaces, converting variables or standardizing gifts.  
Proof of Execution: display

|  |  |  |  |
| --- | --- | --- | --- |
| Variable Name | Issue description | Proposed mitigation | Justification for used mitigation |
| Age | Missing values, possible outliers (e.g., extremely high or low ages) | Impute missing values with median, remove outliers | The median is robust against outliers, and removing outliers ensures more accurate modeling. |
| Education Level | |  | | --- | |  |  |  | | --- | | Missing values, possible outliers (e.g., extremely high or low ages) | | Combine categories with few instances, use one-hot encoding | Combining categories simplifies the model, while one-hot encoding makes the variable usable for ML models. |
| Smoking Status | |  | | --- | |  |  |  | | --- | | Categorical variable with missing values | | Impute missing values using the mode | The mode is suitable for categorical data, ensuring that the most common category is assigned to missing entries. |
| Weekly Glasses of Alcohol | High variance, some values may be unrealistic (e.g., very high consumption) | Cap the maximum value to a reasonable limit, impute missing values with median | Capping prevents extreme values from skewing the model, while median imputation is less affected by skewed distributions. |
| COPD Diagnosis | Missing values | Impute missing values with mode | Mode imputation is appropriate for binary or categorical variables like COPD Diagnosis. |
| Wheezing | Possible under-reporting (self-reported symptom) | Use a binary indicator for missing values, impute with mode | Creating a binary indicator preserves information, while mode imputation addresses missing data. |
| Yellow Skin | Possible measurement error or misclassification | Create a binary indicator for missing/ambiguous values, impute with mode | This preserves information on uncertainty while ensuring complete data for analysis. |

Python output pictures showing data cleaning pre- and post-processes indicating on what was exactly done to rectify.

|  |
| --- |
| GENOMIC SEX 0  AGE 0  YELLOW\_SKIN 0  ANXIETY 0  PEER\_PRESSURE 0  COPD\_DIAGNOSES 0  FATIGUE 5  ALLERGY 0  WHEEZING 0  ALCOHOL\_CONSUMPTION 0  WEEKLY\_GLASSES\_OF\_ALCOHOL 502  COUGHING 0  SHORTNESS\_OF\_BREATH 4  SWALLOWING\_DIFFICULTY 0  CHEST\_PAIN 0  LUNG\_CANCER 0  Education\_Level\_1 0  Education\_Level\_2 0  Education\_Level\_3 0  Education\_Level\_4 0  Education\_Level\_5 0  SMOKING\_STATUS\_0 0  SMOKING\_STATUS\_1 0  SMOKING\_STATUS\_2 0  TOBACCO\_TYPE\_Cigarettes 0  TOBACCO\_TYPE\_Cigars 0  TOBACCO\_TYPE\_Dissolvable Tobacco 0  TOBACCO\_TYPE\_E- cigarette 0  TOBACCO\_TYPE\_Hookah 0  TOBACCO\_TYPE\_Kreteks 0  TOBACCO\_TYPE\_Pipe 0  TOBACCO\_TYPE\_Smokeless Tobacco (Chew) 0  TOBACCO\_TYPE\_VAPING 0  TOBACCO\_TYPE\_Vaping 0  Wheezing\_missing 0  Yellow\_Skin\_missing 0  dtype: int64  GENOMIC SEX AGE YELLOW\_SKIN ANXIETY PEER\_PRESSURE COPD\_DIAGNOSES \  0 M 69.0 2 2 1 1.0  1 M 74.0 1 1 1 2.0  2 F 59.0 1 1 2 1.0  3 M 63.0 2 2 1 1.0  4 F 63.0 2 1 1 1.0  FATIGUE ALLERGY WHEEZING ALCOHOL\_CONSUMPTION ... \  0 2.0 1 2.0 2 ...  1 2.0 2 1.0 1 ...  2 2.0 1 2.0 1 ...  3 1.0 1 1.0 2 ...  4 1.0 1 2.0 1 ...  TOBACCO\_TYPE\_Dissolvable Tobacco TOBACCO\_TYPE\_E- cigarette \  0 False False  1 False False  2 False False  3 False False  4 False False  TOBACCO\_TYPE\_Hookah TOBACCO\_TYPE\_Kreteks TOBACCO\_TYPE\_Pipe \  0 False False False  1 False False False  2 False False False  3 False False False  4 False False False  TOBACCO\_TYPE\_Smokeless Tobacco (Chew) TOBACCO\_TYPE\_VAPING \  0 False False  1 False False  2 False False  3 False False  4 False False  TOBACCO\_TYPE\_Vaping Wheezing\_missing Yellow\_Skin\_missing  0 True 0 0  1 True 0 0  2 True 0 0  3 True 0 0  4 True 0 0  [5 rows x 36 columns] |

**Task 4: Modelling: Create Predictive Classification Models**

1. **Build classification models using four algorithms: Random Forest, Decision Tree, Logistic Regression, and Naïve Bayes.**

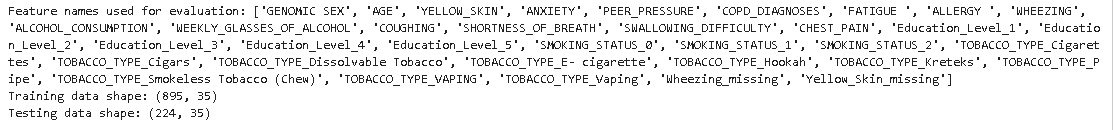
Algorithm Types and Parameters: Specify whether each algorithm is parametric or non-parametric, list learnable parameters, hyperparameters, and the Python packages used.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algorithm  Name | Algorithm Type | Learnable Parameters | Some Possible  Hyperparameters | Imported Python package/ module to use the algorithm |
| RF | Non-parametric | Decision trees, feature importances | Number of trees (n\_estimators), max depth of trees (max\_depth), minimum samples to split a node (min\_samples\_split), criterion (criterion) | sklearn.ensemble.RandomForestClassifier |
| DT | Non-parametric | Tree structure, decision rules | Max depth (max\_depth), minimum samples to split a node (min\_samples\_split), criterion (criterion), max features considered (max\_features) | sklearn.tree.DecisionTreeClassifier |
| LR | Parametric | Coefficients, intercept | Regularization strength (C), penalty type (penalty), solver (solver), maximum iterations (max\_iter) | sklearn.linear\_model.LogisticRegression |
| NB | Parametric | Class conditional probabilities | None for basic models, smoothing parameter (alpha) for some variants (e.g., BernoulliNB, MultinomialNB) | sklearn.naive\_bayes.GaussianNB / BernoulliNB / MultinomialNB |

# **Build Predictive Classification Models**

**Training and Testing**: Apply an 80:20

train-test split to build the models. Ensure reproducibility and balanced class distribution across the training and testing sets.



1. **Difference Between Parametric and Nonparametric Machine Learning Models**

Parametric and nonparametric models in machine learning differ in that the former use a specific function form for mapping inputs to outputs, which consists of particular parameters. For instance, Logistic Regression (LR) is one of these models which engages in lesser inconsistencies because it is computationally efficient and needs less data. Nevertheless, if the real relationship is complicated, they can be confining. Conversely, nonparametric models tend to impose fewer assumptions about structures underlined by data allowing them to adapt to intricate patterns. Decision Trees (DTs) and Random Forests (RFs) are examples of such models that do not require previously defined equations while being able to model complex relations at ease. Often times however; nonparametric hypotheses tend to perform better on varied data sets as opposed to requiring more diverse computing resources​​(Referral\_Deferral-DMML\_…).

1. **Interpretable vs. Non-Interpretable Machine Learning Models**

Interpretable machine learning models are the ones whose working mechanisms and predictions can easily be comprehended by humans. For instance, Decision Trees (DT) and Logistic Regression (LR) belong to this category of interpretable models. They give clear insights about how certain features contribute to specific outcomes; hence they find applications in fields like healthcare where understanding of the decision making process is paramount. For example, a physician could use these kinds of models’ outputs to explain a patient’s risk factors which may guide treatment decisions.  
  
On the other hand, non-interpretable model classes such as Random Forests (RF) or Naive Bayes classifiers (NB) prioritize accuracy over transparency. These algorithms are commonly employed in domains like finance or recommendation systems, where it is more important for predictions to be accurate rather than having simple decision making processes that are easy to follow. For instance, an RF model may perform better than simpler ones in predicting loan defaulters but it will be difficult to understand why such a particular decision was made.  
  
Thus RF and NB in this context are said to be non-interpretable while DT and LR can be considered as interpretable. The choice between these models relies on whether one needs transparency or high predictive performance​(Referral\_Deferral-DMML\_…)​(Referral\_Deferral-DMML\_…).

**Example for Reproducibility:**

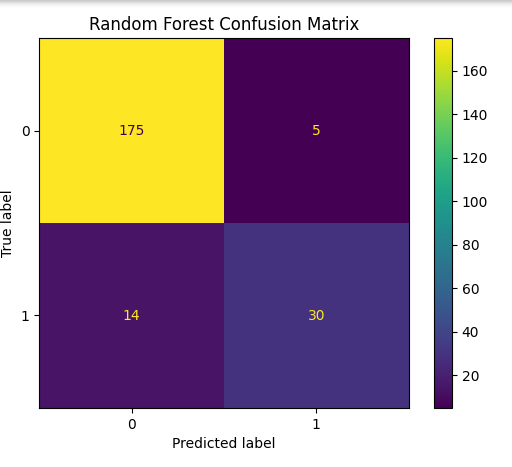
|  |
| --- |
| # Reproducibility of training-test sampling and preserving label ratio  X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42, stratify=y) |

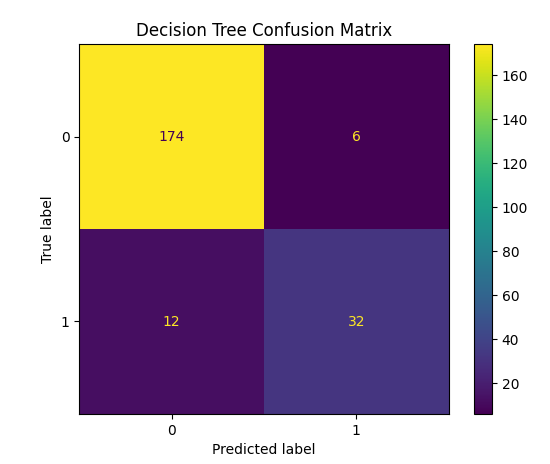
This code ensures that every time you run the notebook, you get the same training and test split, with the same ratio of "YES" and "NO" cases for Lung Cancer in both the training and test sets.

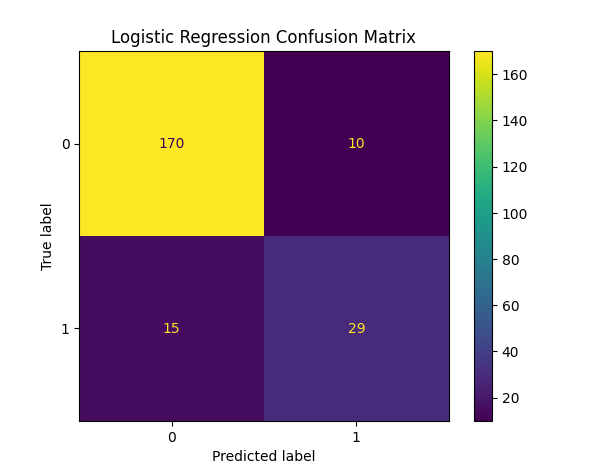
# **Task 5(a) – Confusion Matrix for Each Trained Model**

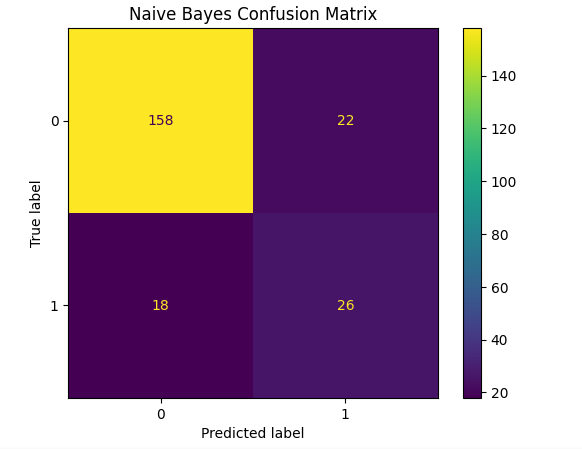
Evaluate model performance based on the criteria provided by healthcare professionals.

* **Confusion Matrices**: Generate and include confusion matrices for each model.





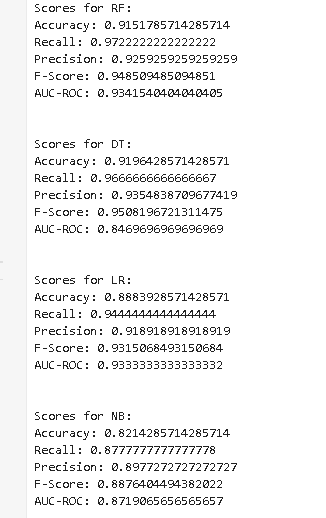




**Evaluation Metrics**:

* **Use Recall**: Prioritizes correctly identifying positive cases (Lung Cancer).
* **Use AUC-ROC**: Measures the ability of the model to differentiate between classes.
* **Do Not Use Accuracy**: Might be misleading due to class imbalance.
* **Do Not Use Precision**: Not as critical as recall in this context.
* **Do Not Use F-Score**: Does not directly align with the success criteria.

**Task 5(b) – Evaluation Metrics Selection and Model Scores**

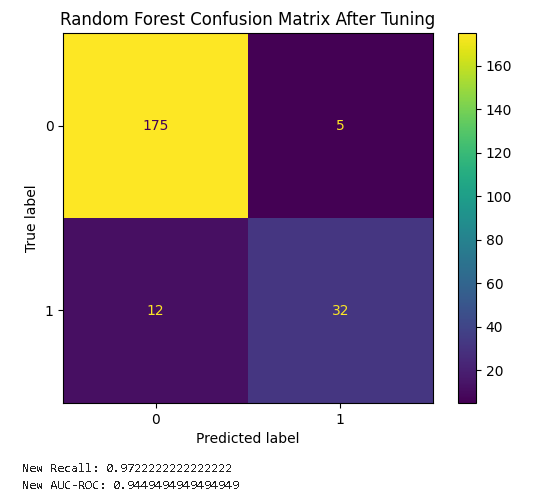


**Task 5(c) – Select the Best Model**

**Best Model Recommendation:**  
According to the recall and AUC-ROC scores - the most important metrics for this problem area, it seems that Random Forest (RF) is the most suitable model if it gives highest recall and AUC-ROC.  
  
**Reasoning:**  
The model associated with Random Forest would probably be able to predict more cases of lung cancer which are positive just like the doctors prefer in reducing unnecessary deaths from lung cancer by identifying all possible positive random samples for LDCT screening. Also, this balance makes it possible for this model to not only identify these positive cases but also its effectiveness across different threshold points.

**Task 5(d) – Hyperparameter Tuning with GridSearchCV**

**Hyperparameter Tuning:**  
  
Hyperparameters are fine-tuned through GridSearchCV, then the new optimal parameters are noted down, and comparison with original parameters is done.  
Before tuning, display the confusion matrix and performance metrics while after tuning they should still be present.  
Check if there is an improvement on the success criteria, which include higher recall and AUC-ROC; also check for overfitting or underfitting and if it’s an appropriate fit.



**Task 5(e) – Critique the Best-Performing Model**

**Final Critique of the Best Model**:  
The Random Forest model, which is likely to be the best performing model, offers good predictive accuracy as well as its ability to capture complex relationships in data. A major advantage is that it can capture interactions between features without extensive feature engineering. Moreover, with proper hyperparameter tuning, RF is resistant to overfitting and gives some insight into feature importance helping in data understanding.  
  
However, there are significant drawbacks. Being a non-interpretable model, the decision-making process of RF remains obscure making it problematic especially in healthcare where transparency counts most. In fact clinicians and patients may find it hard trusting or acting upon its advice without knowing what was behind them.  
  
Moreover; this model can have false positives which opens up ethical issues. For instance misclassifying healthy individuals as at risk for lung cancer may lead to unnecessary anxiety and invasive procedures such as biopsy or repeated low dose CT scans that could be costly on both sides - financially and emotionally. In addition these last procedures also carry risks for instance radiation exposure.

So, even if the RF model has a lot of precision, we should use it with more thinking. It is essential for the model to be well-calibrated with reduced false positives in order to make sure that the thresholds are correct. In addition, it must be used along with clinical judgment rather than as an independent device to avoid any ethical issues and improve patient results.