Variance threshold – 0.9892

**Chi Squared Test - 0.9852**

**Mutual Information - 0.9847**

**Info Gain – 0.9815**

**Backward Elimination - 0.9849**

**FS - 0.9839**

**RFE - 0.9836**

**Lasso - 0.9782**

**Ridge - 0.9881**

**RF - 0.9892**

**Let me clarify how the Variance Threshold method works and suggest more appropriate feature selection techniques for your binary COVID-19 dataset.**

1. **Variance Threshold Explanation: Variance Threshold is actually one of the simplest feature selection techniques. It works by:**
   * **Calculating the variance of each feature in your dataset**
   * **Removing features whose variance falls below a specified threshold**
   * **It does NOT consider the target variable at all - it's a completely unsupervised method**
   * **It simply removes features that don't change much across samples**

**So no, it doesn't give you the "top 10 based on the target" - it just removes low-variance features regardless of their relationship with your prediction target.**

1. **Suitability for Binary COVID-19 Prediction: Variance Threshold has limited usefulness for binary data because:**
   * **For binary features (0s and 1s only), the variance is p\*(1-p) where p is the proportion of 1s**
   * **A binary feature with 99% 1s would have variance 0.99\*0.01 = 0.0099, which might be filtered out despite potentially being predictive**
   * **It doesn't consider the relationship between features and your target (COVID-19 status)**
2. **Better Feature Selection Techniques for Binary Classification:**
3. **Chi-Square Test - Tests statistical dependence between categorical variables, ideal for binary data as it directly measures the relationship between each feature and your binary target.**
4. **Mutual Information - Measures how much information one binary variable provides about another, capturing non-linear relationships between features and target.**
5. **Recursive Feature Elimination (RFE) - Recursively removes the weakest features based on classifier importance weights, which works well for binary prediction tasks.**
6. **L1 Regularization (Lasso) - Adds a penalty that forces some feature coefficients to zero, effectively selecting only the most important features for prediction.**
7. **Random Forest Feature Importance - Uses decision trees to rank features based on how much they improve prediction accuracy, which works well with binary data.**
8. **Boruta Algorithm - Compares feature importance against random noise features, ensuring only truly relevant features are selected for your COVID prediction.**
9. **Correlation-based Feature Selection - Identifies subsets of features that are highly correlated with the COVID-19 outcome but have low correlation with each other.**
10. **Sequential Feature Selection - Iteratively adds or removes features to find the subset that maximizes prediction performance, directly optimizing for your classification task.**
11. **Genetic Algorithm Feature Selection - Uses an evolutionary approach to find optimal feature subsets, particularly useful for complex relationships in medical data.**
12. **Feature Importance from XGBoost/LightGBM - These gradient boosting algorithms provide importance scores that work exceptionally well for imbalanced binary medical data like COVID-19 diagnosis.**

**For your specific use case (binary COVID-19 prediction with binary features), I'd particularly recommend Chi-Square, Mutual Information, and ensemble-based methods (Random Forest, XGBoost) as they're specifically designed to handle categorical relationships and can identify subtle patterns in medical diagnostic data.**