Erdös Institute May-2024 Data Science Boot Camp

Exploring the Relationship Between Cancer Occurrence and Contributing Factors in the US Using Machine Learning Algorithms

Priti Singh, Pankaj Singh Dholaniya, Ikenna Nometa, Gbocho Masato Terasaki

THE ERDŐS INSTITUTE

Helping PhDs get and create jobs they love at every stage of their career.

Health Information National Trends Survey

hints®

Introduction

Problem and data

- There is significant literature indicating predictors of cancer, however health information seeking and geography are yet to be fully explored, especially using modern prediction methods.

- Explore this relationship using data collected by the National Cancer Institute, called the Health Information National Trends Survey or HINTS.

- HINTS regularly collects nationally representative data about the American public’s knowledge of, attitudes toward, and use of cancer- and health-related information.

Objectives

- Investigate the relationship between cancer incidence and three key factors: demographics, the utilization of health information technology, and medical history.

- Identify which features best predict the outcome of cancer based on classification models.

- Measure the models' performance(s).

Data preparation

Data Collection

Period: Second cycle of HINTS 4
Duration: Oct. 2012 - Jan. 2013

Total Response: 3630
Regions: Multiple

Total features: 357
Features studied: 20

Data Cleaning

- Data types homogenized
- Missing values taken care of
- Initial features selected

Demographics

- Age
- BMI
- Education
- Census Division
- Gender
- Income

Medical History

- Diabetes
- High Blood Pressure
- Heart Condition
- Lung Disease
- Arthritis
- Depression
- General Health
- Own Ability to Take Care of Health

Utilization of Health Information Technology

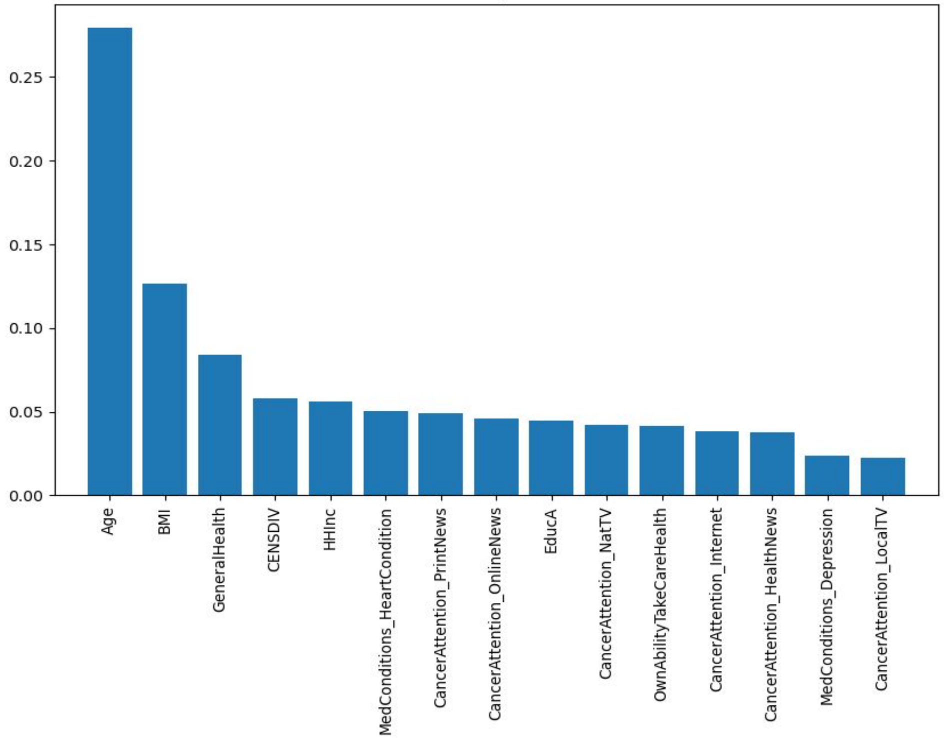
- Health News
- Internet
- Local TV
- National TV
- Online News
- Printed News

Feature Selection Using Recursive Feature Elimination (RFE)

- To identify the most important features for predicting our target variable by leveraging multiple machine learning models and Recursive Feature Elimination (RFE).

- We chose 4 models (Logistic Regression (base model), LinearSVC, Decision Tree, 'very shallow' Random Forest), and for each model, we use RFE to recursively eliminate the least important features.

- For each model, we use RFE to recursively eliminate the least important features.
  - We collect the feature rankings from each model.
  - Then, accumulate the rankings and compute the average ranking for each feature across all models.
  - Finally, we selected the top 15 features based on these average rankings.

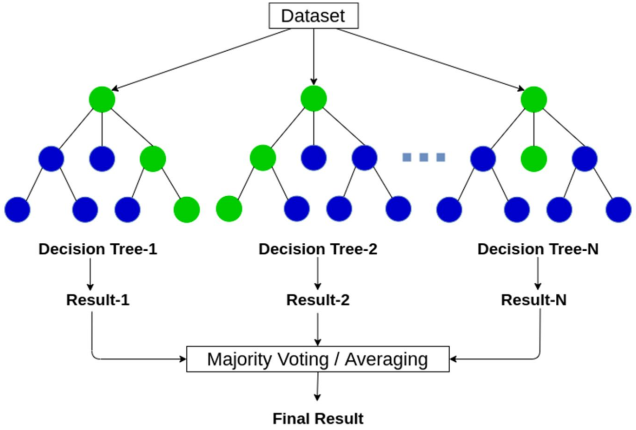


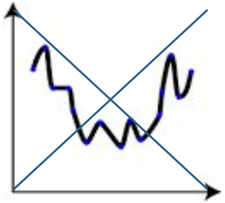
Feature Importance plot of the top 15 features based on average ranking are

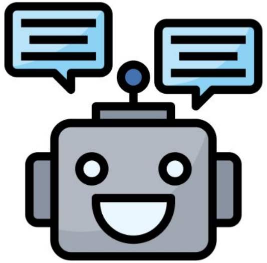
Model Choice for Our Problem: Random Forest

While there exist tree-ensemble models such as Gradient Boosting Trees which generally excel in predictive accuracy, particularly in datasets with imbalance classes, we choose to build our pipeline around Random Forest model.

Random Forest



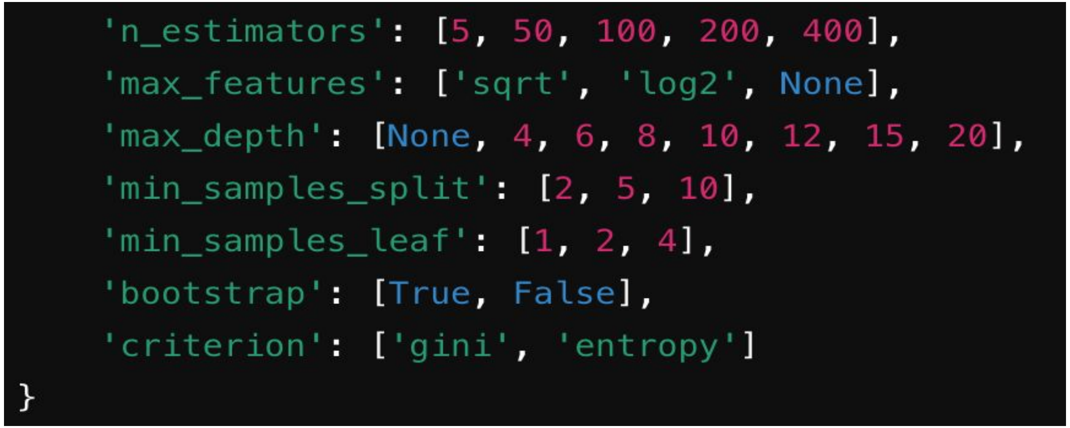




Hyperparameter Tuning - Random Forest Model

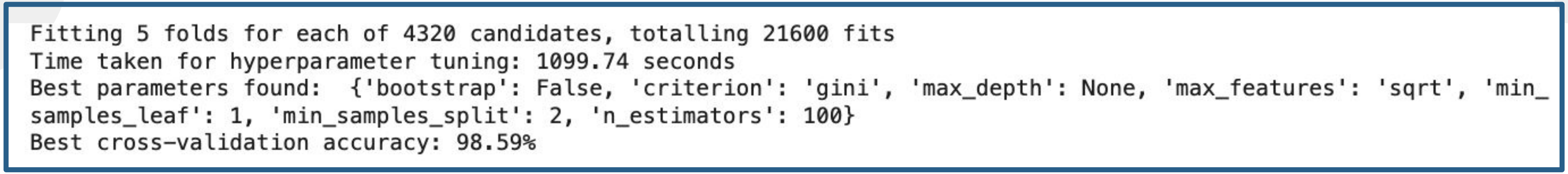
In order to optimize the performance of our machine learning model, we systematically search for the best combination of hyperparameters. This will enhance its predictive accuracy and generalization capability.

We define the following comprehensive set of hyperparameters for tuning:



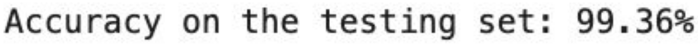
```
    'n_estimators': [5, 50, 100, 200, 400],
    'max_features': ['sqrt', 'log2', None],
    'max_depth': [None, 4, 6, 8, 10, 12, 15, 20],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
    'bootstrap': [True, False],
    'criterion': ['gini', 'entropy']
}
```

Results: Hyperparameter Tuning

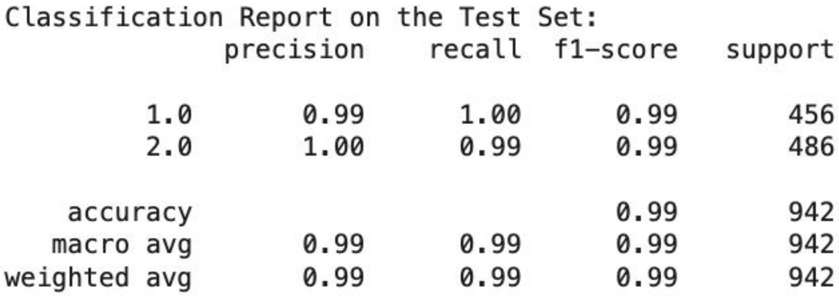


```
Fitting 5 folds for each of 4320 candidates, totalling 21600 fits
Time taken for hyperparameter tuning: 1099.74 seconds
Best parameters found:  {'bootstrap': False, 'criterion': 'gini', 'max_depth': None, 'max_features': 'sqrt', 'min_
samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100}
Best cross-validation accuracy: 98.59%
```

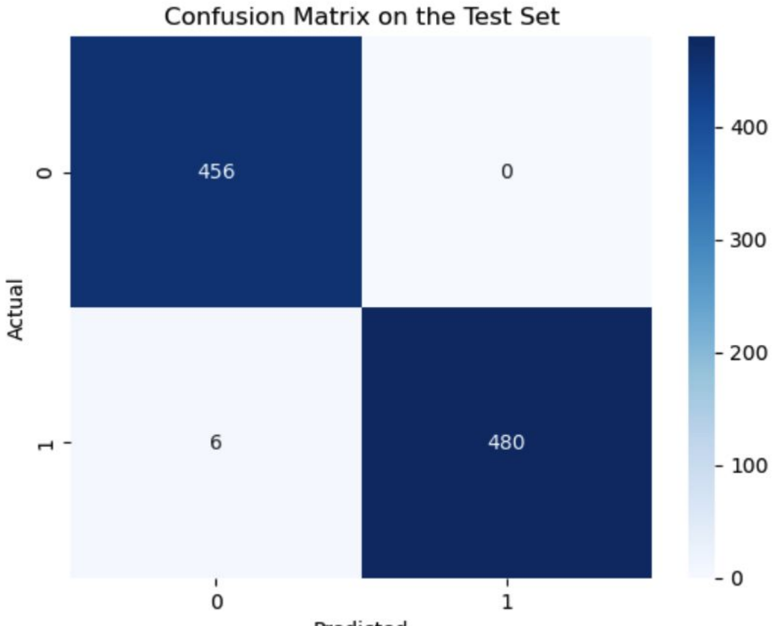
Performance on Test Set



```
Accuracy on the testing set: 99.36%
```



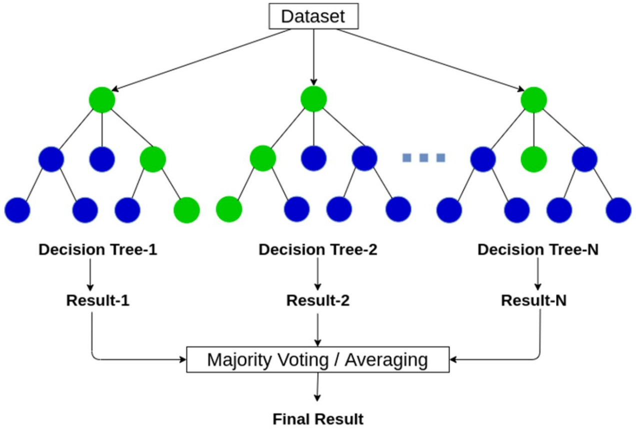
```
Classification Report on the Test Set:
              precision    recall  f1-score   support

         1.0       0.99      1.00      0.99       456
         2.0       1.00      0.99      0.99       486

    accuracy                           0.99       942
   macro avg       0.99      0.99      0.99       942
weighted avg       0.99      0.99      0.99       942
```



Future Direction 1: Evaluate Model Performance

On new data from subsequent years from HINTS

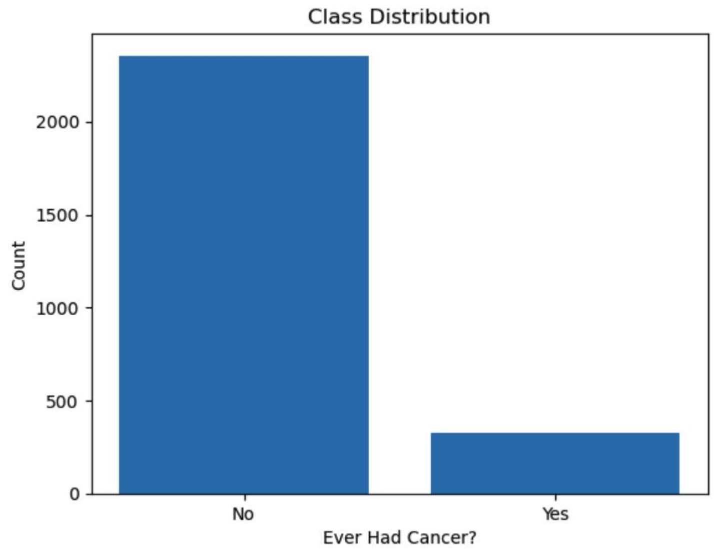
Random Forest



?

Need some new
HINTS?
Social Media/ Misinformation
Telehealth
Social Determinants of Health
and More!
HINTS 6 (2022) is now available.
hints
Download the data set at hints.cancer.gov today!

Future Direction 2: Explore more the effects of oversampling vs undersampling on the model accuracy.



Undersampling Techniques

- Random Undersampling

Oversampling Techniques

- Random Oversampling
- SMOTE (**S**ynthetic **M**inority **O**ver-sampling **TE**chnique)

Future Direction 3:

Develop an app that will automatically predict whether a person has had cancer based on the model we develop.

Acknowledgements:

- Roman Holowinsky, Alec Cott, Steven Gubkin and the entire Erdös Institute Summer-May-2024 team.
- Greg Edwards (our project mentor).
- NIH's National Cancer Institute's (for HINTS).

NIH NATIONAL CANCER INSTITUTE
Health Information National Trends Survey®

Health Information National Trends Survey hints®

THE ERDŐS INSTITUT
Helping PhDs get and create jobs t
love at every stage of their career.