Analysis of COVID-19 mortality among old aged population using machine learning algorithms

Marjan Qazvini

September, 2022

Outline

Introduction

- ELSA is a longitudinal survey study that interviews people aged 50+.
- The participants are selected from those who participated in the Health Survey for England.
- New cohorts have been introduced in waves $3, 4, 6, 7$ and 9.
- ELSA COVID-19 Substudy, collects individual-level health, behavioural and social data to study the effects of COVID-19 pandemic on old aged population.
- The first wave was from June 3, 2020 to July 16, 2020.
- The second wave was from November 4, 2020 to December 20, 2020.
- Survey was conducted online and on the phone.

Interview mode in 2 waves and the number of core members

	Online	Phone	Core members
Wave 1	5,791	1,249	5,825
Wave 2	5,652	1,142	5,338

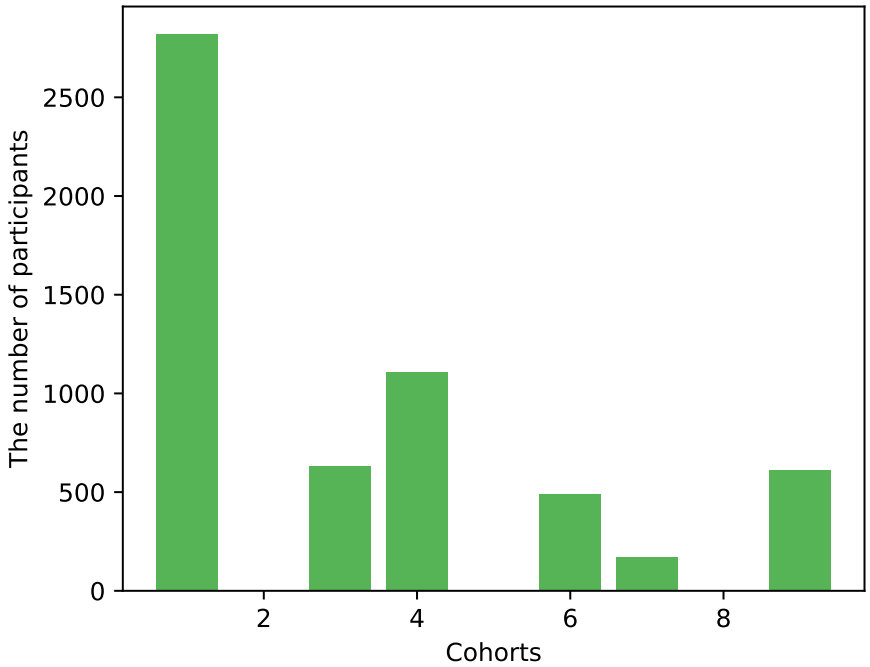
Source: User guide for ELSA COVID-19 Substudy

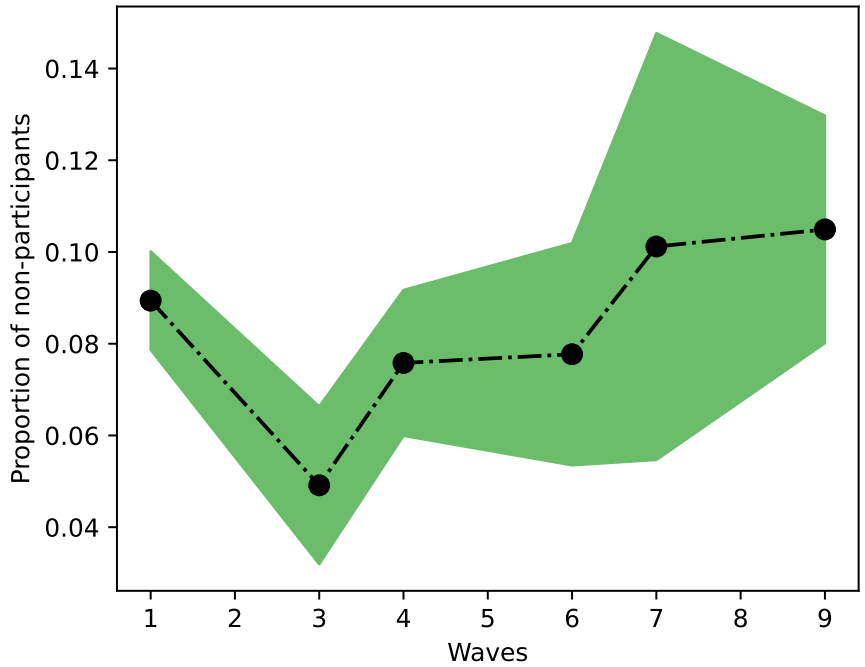
ELSA core members meet three criteria:

- Fitted the age eligibility criteria of a given ELSA cohort.
- Participated in the sample-origin HSE (Health Survey for England) survey.
- Participated in the first wave of ELSA when invited to join the study.

Features

- Questions are selected from the following topics: demographics, mental health, COVID-19-related health, employment and work, physical health and health behaviours from wave 1.
- 486 individuals did not participate in wave 2.
- Participants are from Cohorts $1, 3, 4, 6, 7$ and 9.

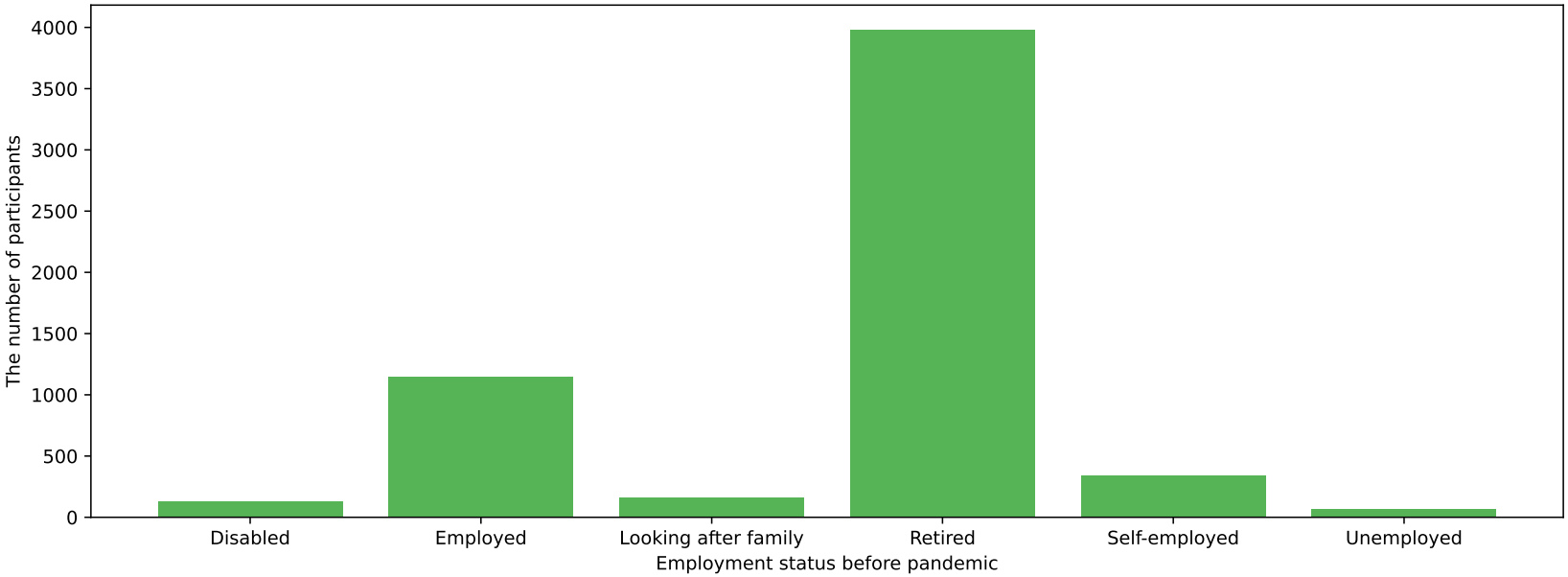


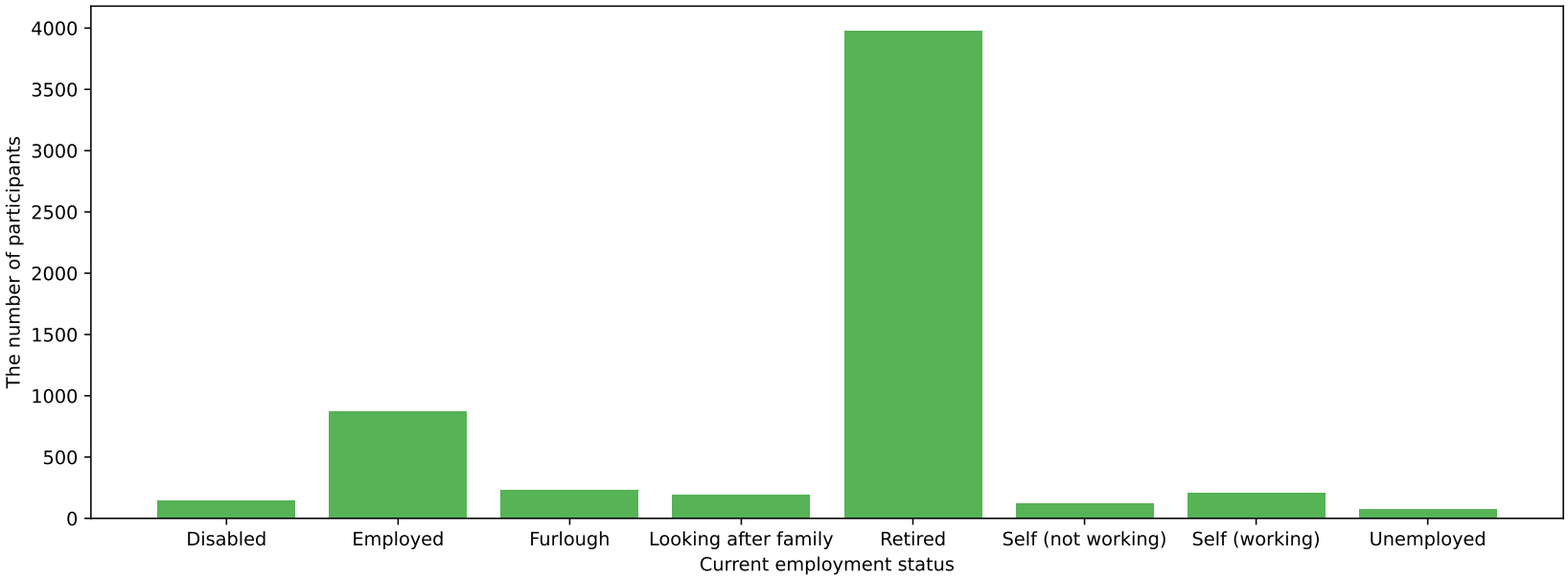


Place of living of participants and non-participants in wave 1

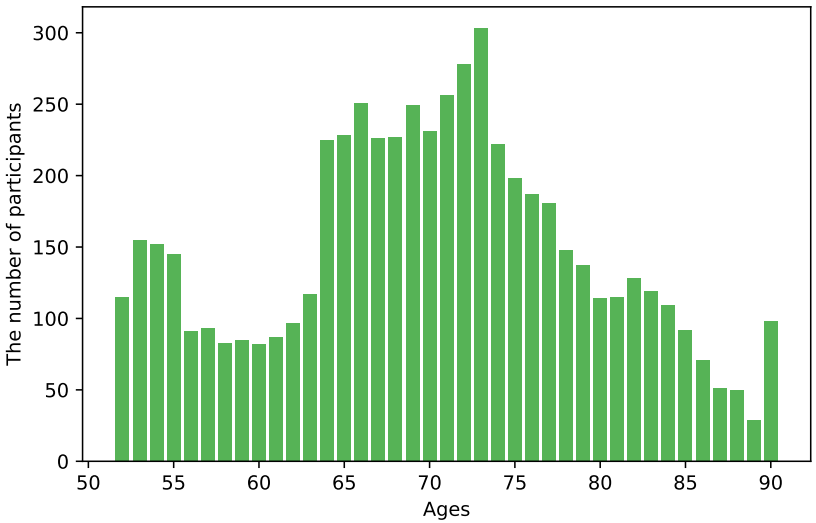
	At home	Care home	Hospital	Other's home	Else
0	5264	5	2	45	23
1	469	1	0	9	7

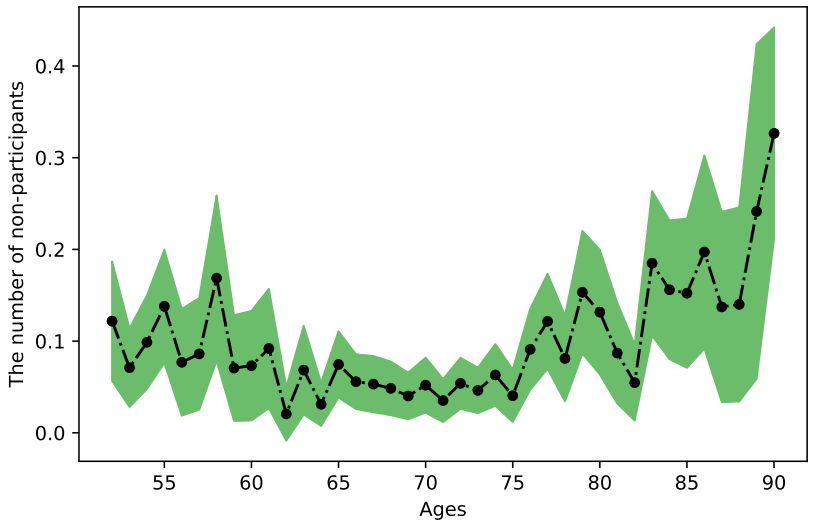
- We have 51 features.
- High temperature, cough, shortness of breath, fatigue, loss of smell or taste, diarrhoea, abdominal pain, loss of appetite, admitted to hospital, death of friends and/or families, smoking, high blood pressure, heart attack, heart failure, diabetes, stroke, chronic lung disease, asthma, arthritis, cancer, dementia, blood disorder, infectious disease, neoplasma, immune system, malnutrition, mental disorder, sleep-wake disorder, nervous system, visual system, urban/rural. diseases of the ear, circulatory system, respiratory system, digestive system, skin, connective tissue, geniratory system, sexual health, disability.
- All these features (variables) are binary.


The number of participants
0
500
1000
1500
2000
2500
3000
3500
4000
Disabled
Employed
Looking after family
Retired
Self-employed
Unemployed
Employment status before pandemic

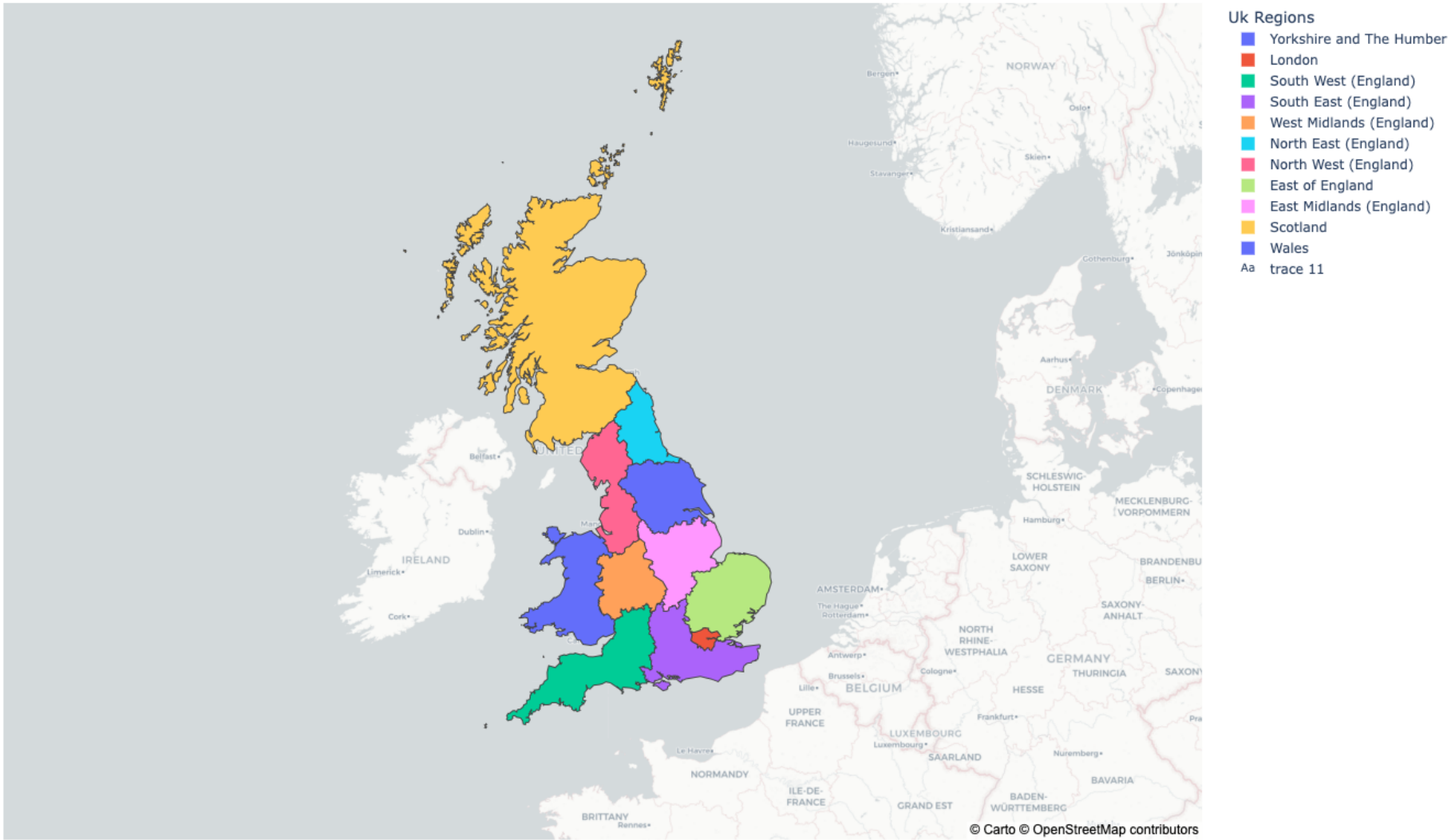

The number of participants
0
500
1000
1500
2000
2500
3000
3500
4000
Disabled
Employed
Furlough
Looking after family
Retired
Self (not working)
Self (working)
Unemployed
Current employment status

- Self-reported weight (average = 76.71).
- Physical activities.
- Region: South East, South West, North West, East of England, West Midlands, East Midlands, Yorkshire and The Humber, London, Wales, Scotland.
- Missing values are imputed based on the most frequent response.

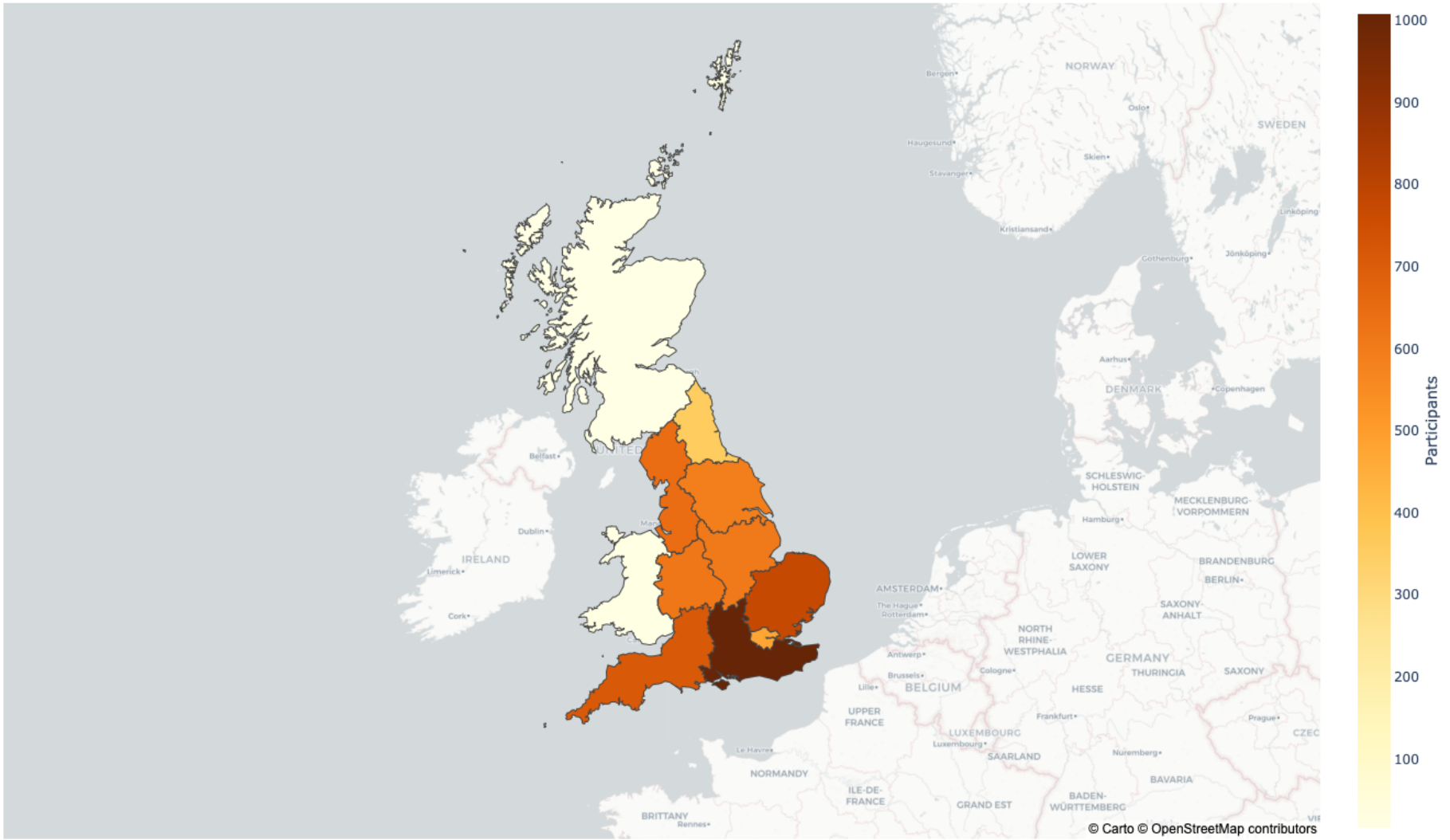


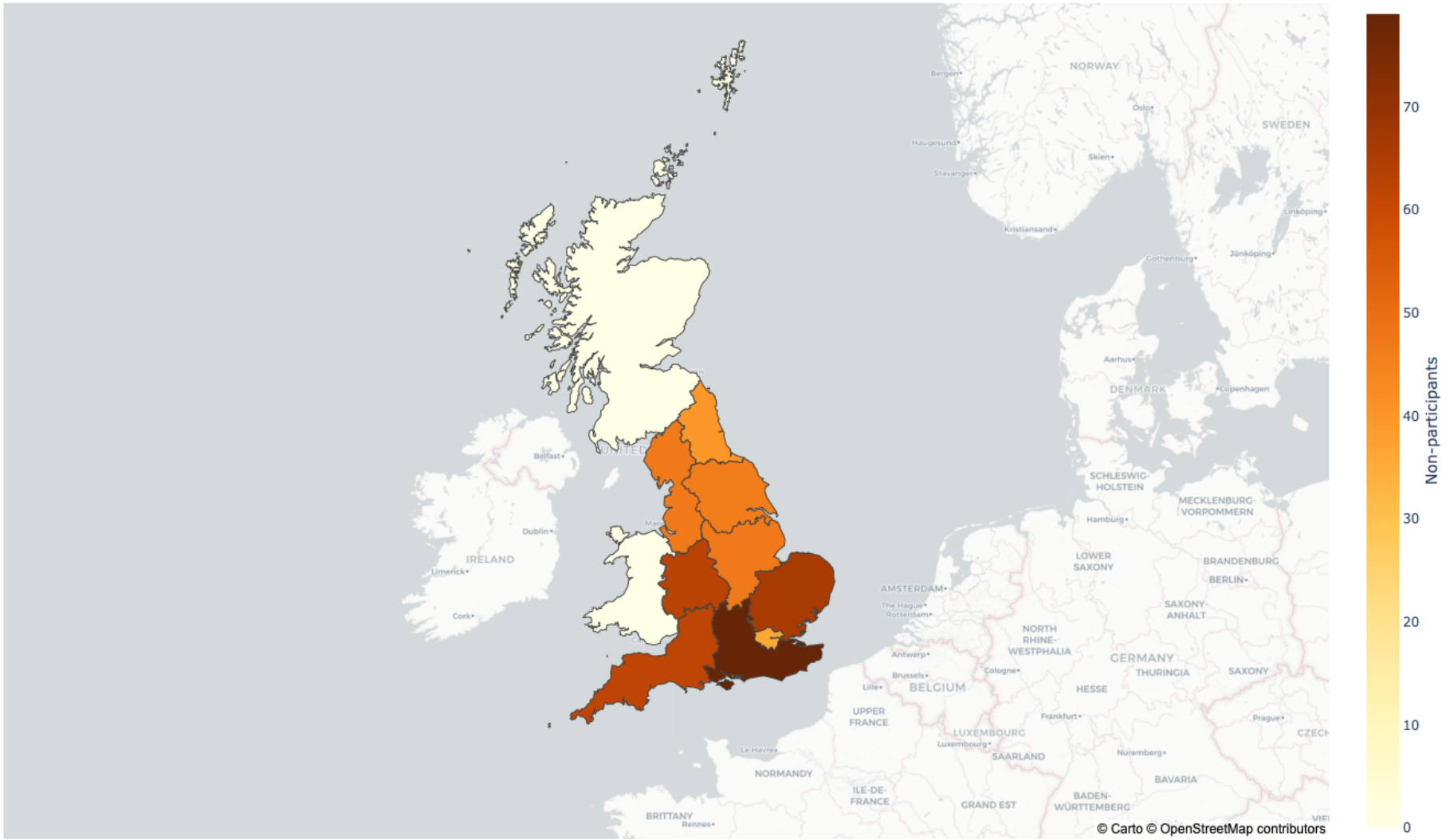


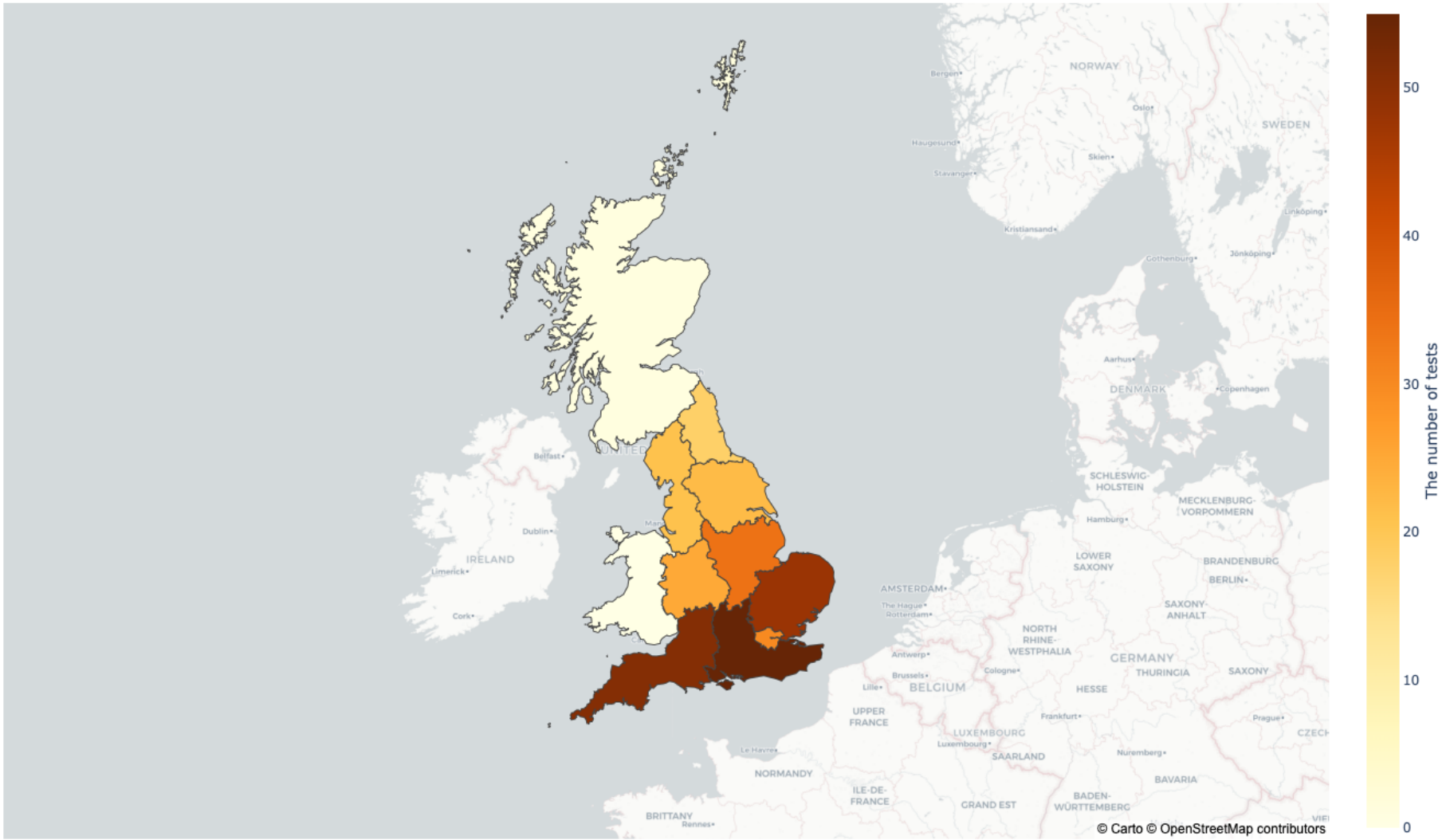
NUTS Level 1 (January 2018)

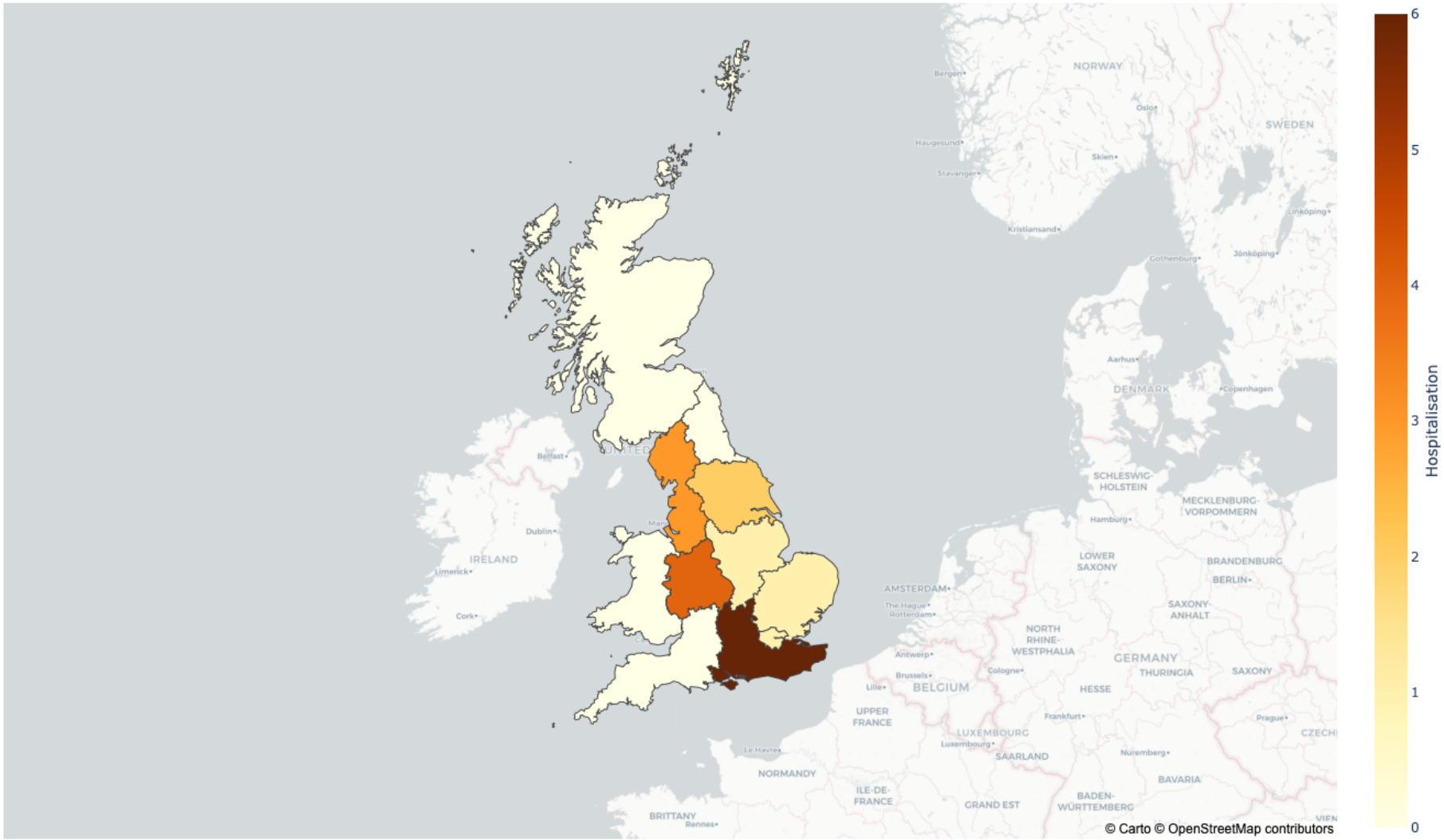


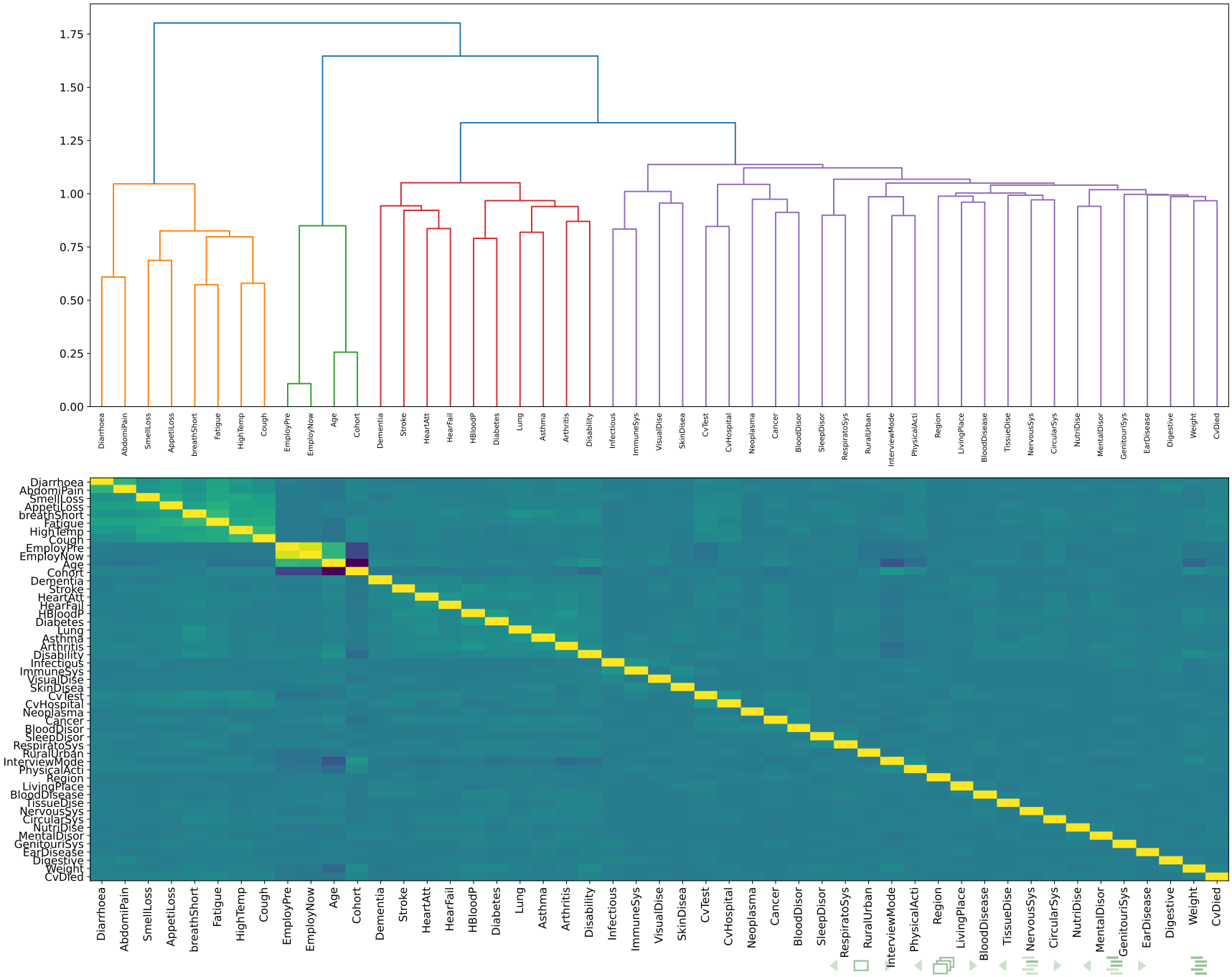
Source: `https://geoportal.statistics.gov.uk/datasets/ons::nuts-level-1-january-2018-ultra-generalised-clipped-boundaries-in-the-united-kingdom/about`


1000
900
800
700
600
500
400
300
200
100
Participants
NORWAY
SWEDEN
DENMARK
IRELAND
BELGIUM
GERMANY
LUXEMBOURG
AMSTERDAM
NORMANDY
BRITTANY
BAVARIA
SAARLAND
SAXONY
THURINGIA
HESSE
BRANDENBURG
BERLIN
LOWER SAXONY
SCHLESWIG-HOLSTEIN
MECKLENBURG-VORPOMMERN
SAXONY-ANHALT
NORTH RHINE-WESTPHALIA
UPPER FRANCE
ILE-DE-FRANCE
GRAND EST
BADEN-WÜRTTEMBERG
Bergen
Oslo
Haugesund
Stavanger
Skien
Kristiansand
Linköping
Jönköping
Gothenburg
Aarhus
Copenhagen
Hamburg
Belfast
Dublin
Limerick
Cork
The Hague
Rotterdam
Antwerp
Brussels
Lille
Cologne
Frankfurt
Luxembourg
Prague
Nuremberg
Le Havre
Rennes
© Carto © OpenStreetMap contributors


NORWAY
SWEDEN
DENMARK
IRELAND
BELGIUM
GERMANY
NETHERLANDS
AMSTERDAM
Bergen
Oslo
Haugesund
Stavanger
Skien
Kristiansand
Linköping
Jönköping
Gothenburg
Aarhus
Copenhagen
Belfast
Dublin
Limerick
Cork
Hamburg
SCHLESWIG-HOLSTEIN
MECKLENBURG-VORPOMMERN
LOWER SAXONY
BRANDENBURG
BERLIN
SAXONY-ANHALT
SAXONY
THURINGIA
HESSE
NORTH RHINE-WESTPHALIA
Antwerp
Brussels
Cologne
Lille
Frankfurt
Prague
Nuremberg
Luxembourg
LUXEMBOURG
SAARLAND
BAVARIA
BADEN-WÜRTTEMBERG
GRAND EST
ILE-DE-FRANCE
UPPER FRANCE
NORMANDY
BRITTANY
Le Havre
Rennes
The Hague
Rotterdam
70
60
50
40
30
20
10
0
Non-participants
© Carto © OpenStreetMap contributors


NORWAY
SWEDEN
DENMARK
IRELAND
BELGIUM
NETHERLANDS
GERMANY
LUXEMBOURG
NORMANDY
BRITTANY
Bergen
Oslo
Stavanger
Haugesund
Skien
Kristiansand
Gothenburg
Jönköping
Linköping
Aarhus
Copenhagen
Belfast
Dublin
Limerick
Cork
Hamburg
SCHLESWIG-HOLSTEIN
MECKLENBURG-VORPOMMERN
LOWER SAXONY
BRANDENBURG
BERLIN
SAXONY-ANHALT
AMSTERDAM
The Hague
Rotterdam
Antwerp
Brussels
Lille
Cologne
NORTH RHINE-WESTPHALIA
THURINGIA
SAXONY
HESSE
Frankfurt
Prague
Luxembourg
SAARLAND
Nuremberg
BAVARIA
UPPER FRANCE
ILE-DE-FRANCE
GRAND EST
BADEN-WÜRTTEMBERG
Le Havre
Rennes
© Carto © OpenStreetMap contributors
The number of tests
50
40
30
20
10
0


Hospitalisation
6
5
4
3
2
1
0
NORWAY
SWEDEN
DENMARK
IRELAND
BELGIUM
GERMANY
AMSTERDAM
Bergen
Oslo
Haugesund
Skien
Stavanger
Kristiansand
Linköping
Jönköping
Gothenburg
Aarhus
Copenhagen
Belfast
Dublin
Limerick
Cork
SCHLESWIG-HOLSTEIN
MECKLENBURG-VORPOMMERN
Hamburg
LOWER SAXONY
BRANDENBURG
BERLIN
SAXONY-ANHALT
The Hague
Rotterdam
NORTH RHINE-WESTPHALIA
Antwerp
Brussels
Cologne
Lille
THURINGIA
SAXONY
HESSE
UPPER FRANCE
Frankfurt
Prague
LUXEMBOURG
Luxembourg
SAARLAND
Nuremberg
Le Havre
NORMANDY
BAVARIA
ILE-DE-FRANCE
GRAND EST
BADEN-WÜRTTEMBERG
BRITTANY
Rennes
© Carto © OpenStreetMap contributors


1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Diarrhoea
AbdomiPain
SmellLoss
AppetiLoss
breathShort
Fatigue
HighTemp
Cough
EmployPre
EmployNow
Age
Cohort
Dementia
Stroke
HeartAtt
HearFail
HBloodP
Diabetes
Lung
Asthma
Arthritis
Disability
Infectious
ImmuneSys
VisualDise
SkinDisea
CvTest
CvHospital
Neoplasma
Cancer
BloodDisor
SleepDisor
RespiratoSys
RuralUrban
InterviewMode
PhysicalActi
Region
LivingPlace
BloodDisease
TissueDise
NervousSys
CircularSys
NutriDise
MentalDisor
GenitouriSys
EarDisease
Digestive
Weight
CvDied

Algorithms

Let $\mathbf{x} \in R^d$, where d represents the number of features and $\mathbf{y} \in \{0, 1\}$ be the features and target (label), respectively. Then, the training set is $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$.

- Preprocessing
  - Encoding
  - Scaling
    - Standard scaler
    - Min-Max scaler
- Cross validation to avoid overfitting
- Validation curves to determine hyper-parameters
- Feature importance by permutation
- Scikit-learn in Python

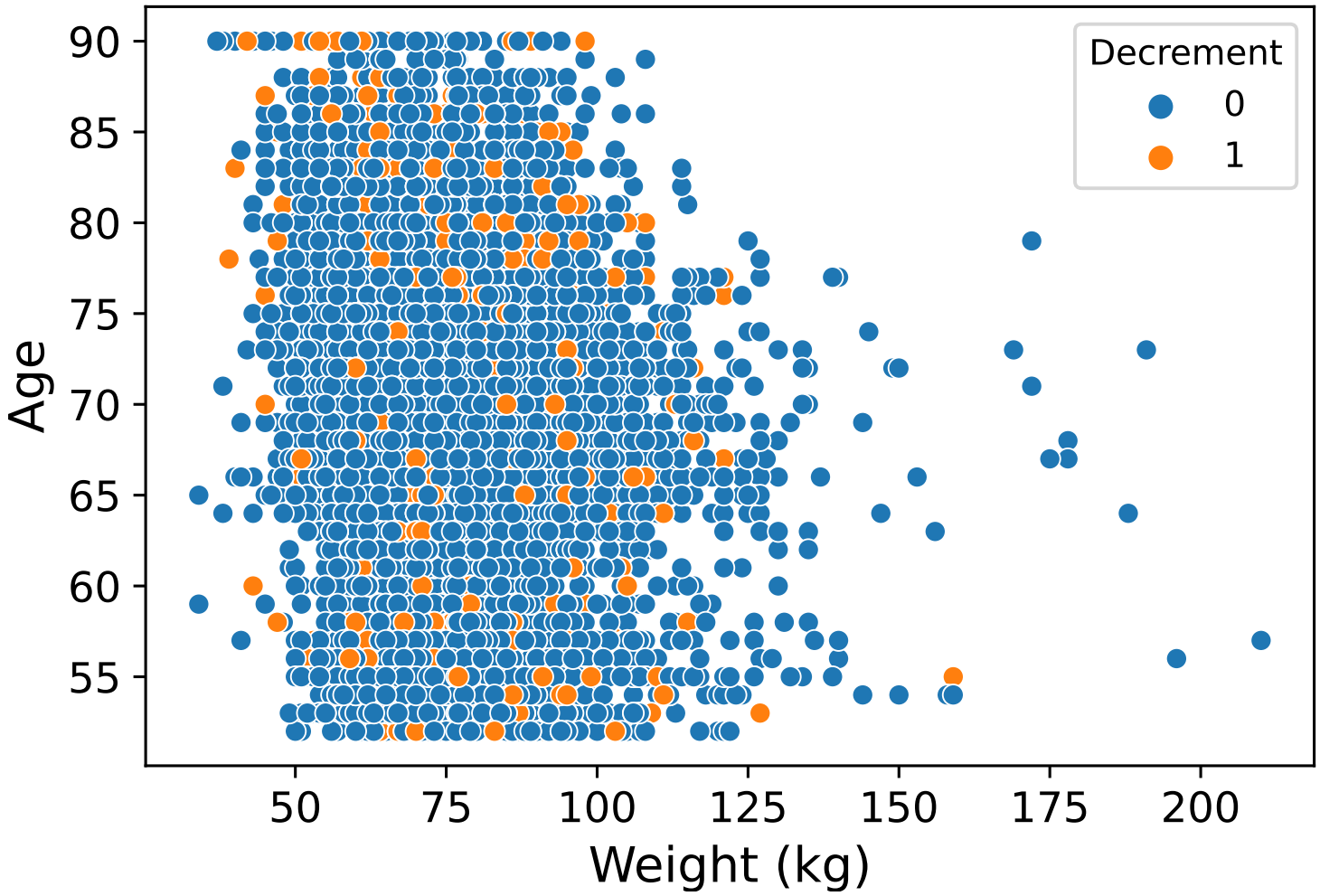
K-nearest neighbours algorithm

Suppose $\mathbf{x}$ is the new data point and $(\mathbf{x}^*, y^*)$ is the closest point to $\mathbf{x}$. Then

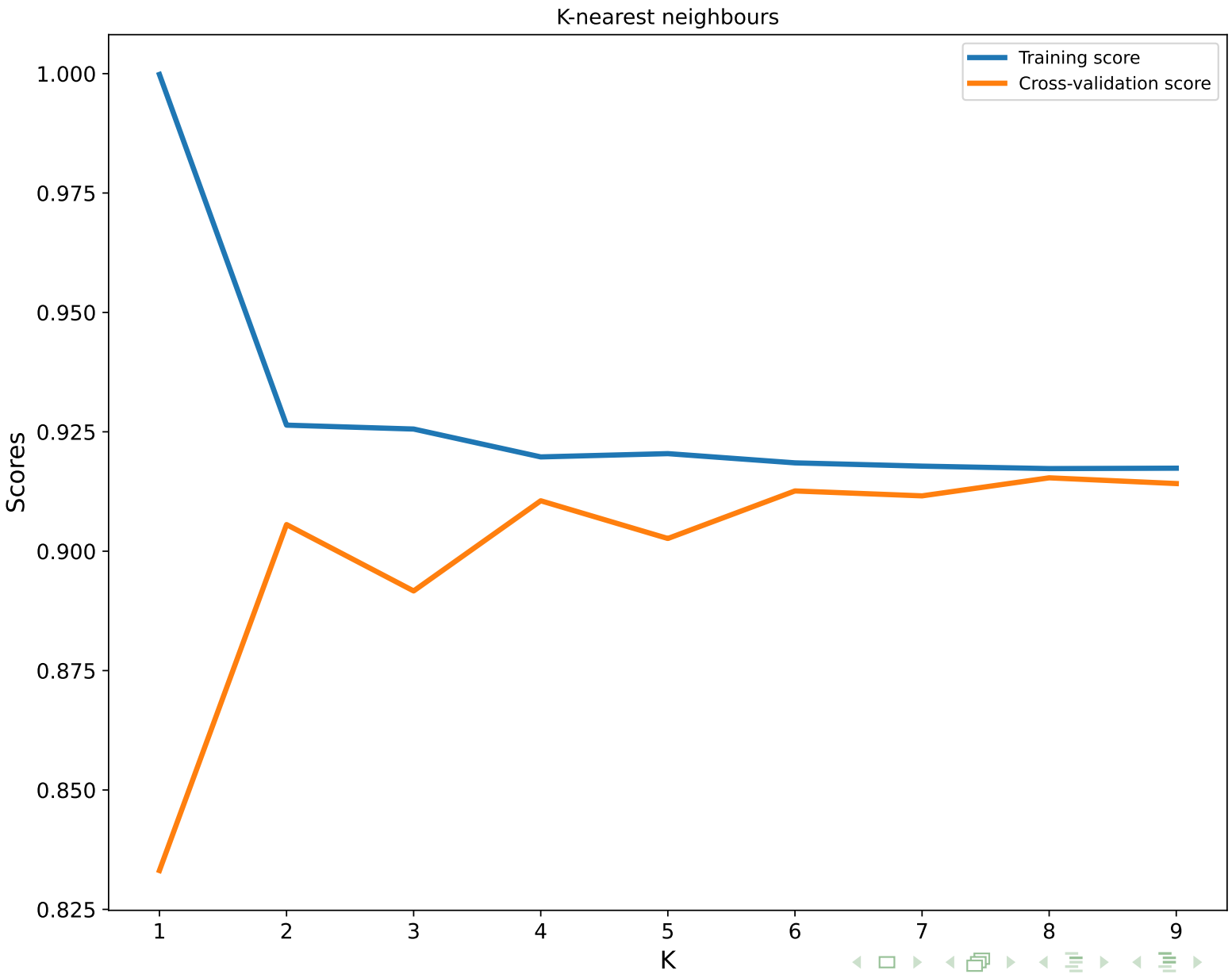
$$\mathbf{x}^* = \arg \min_{\mathbf{x}_i \in \text{train set}} \text{distance}(\mathbf{x}_i, \mathbf{x}).$$

Then the label of the new data point is y^*. Different distance measures in scikit-learn are: Euclidean, Chebyshev, Manhattan, Minkowski distances.

- K is the hyper-parameter that can be tuned through Cross Validation. Too small may overfit, too large K may underfit.
- Suffers from the curse of dimensionality: in high dimensions, most points are far apart.
- Scaling is needed.


Decrement
0
1
Age
90
85
80
75
70
65
60
55
50
75
100
125
150
175
200
Weight (kg)

Validation curves



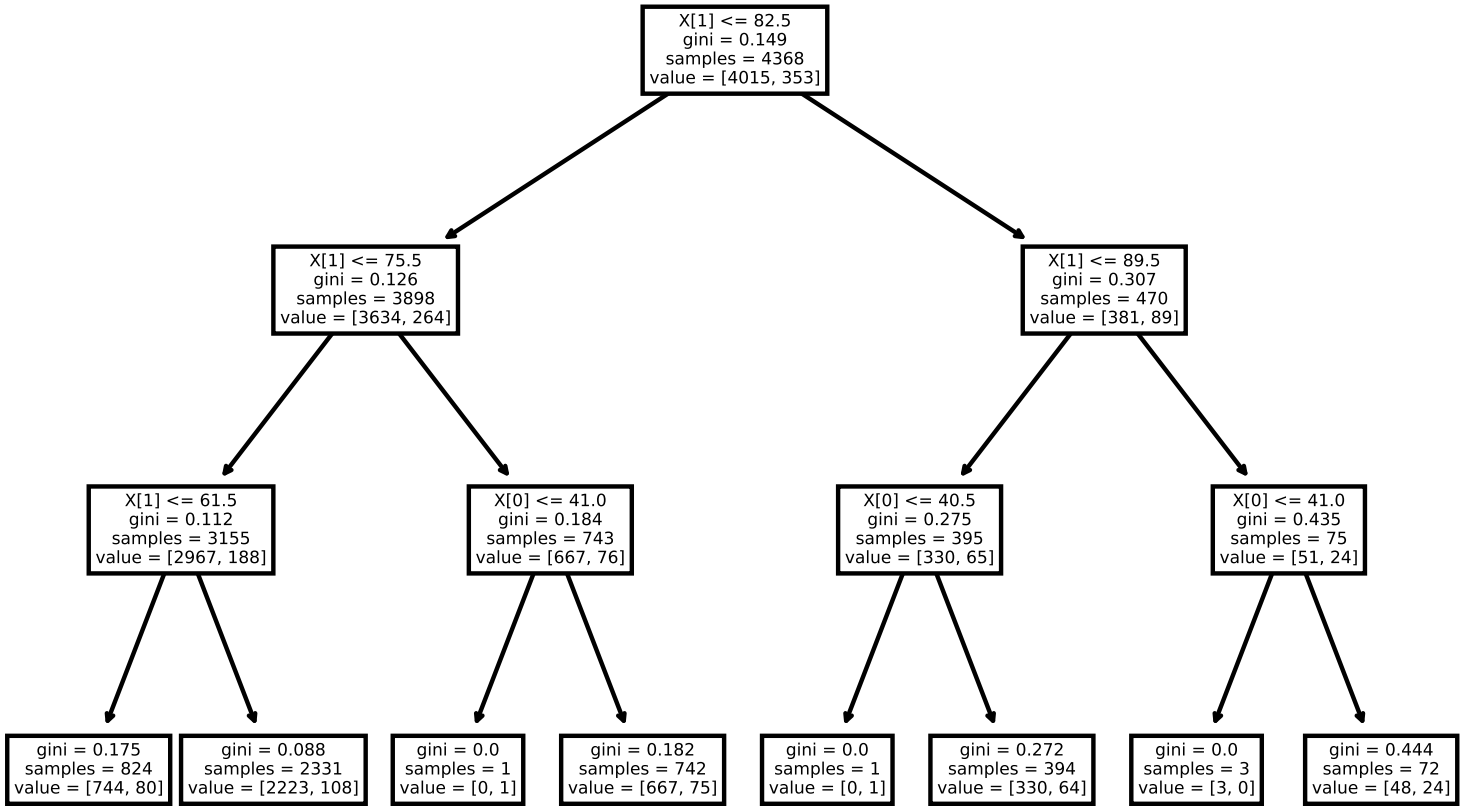
Classification trees

- Non-parametric supervised learning. A set of if/then decision rules.
- It partitions the space by considering a single feature at a time.
- Different algorithms for decision trees:
  - ID3 (Iterative Dichotomiser) (Quinlan, 1986). It only works with discrete and nominal data.
  - C4.5 (Quinlan, 1993). It works with both discrete and continuous data.
  - CART (Classification and regression tress) (Breiman et al., 1984). It is used in scikit-learn. It is based on numerical splitting.
- To decide which features to split on we can use

$$\text{cost}(D) - \left(\frac{|D_L|}{|D|} \text{cost}(D_L) + \frac{|D_R|}{|D|} \text{cost}(D_R) \right).$$

Cost: the impurity criteria such as Gini index, entropy or log loss.
D, D_L, D_R: sample size of the parent, left and right nodes.

- $X[1]$: Age, $X[2]$: Weight
- Value$[y = 0, y = 1]$.
- Depth: 3
- Gini: $p(0|t)[1 - p(0|t)] + p(1|t)[1 - p(1|t)]$, where $p(i|t)$ represents the relative frequency of each class in node t.

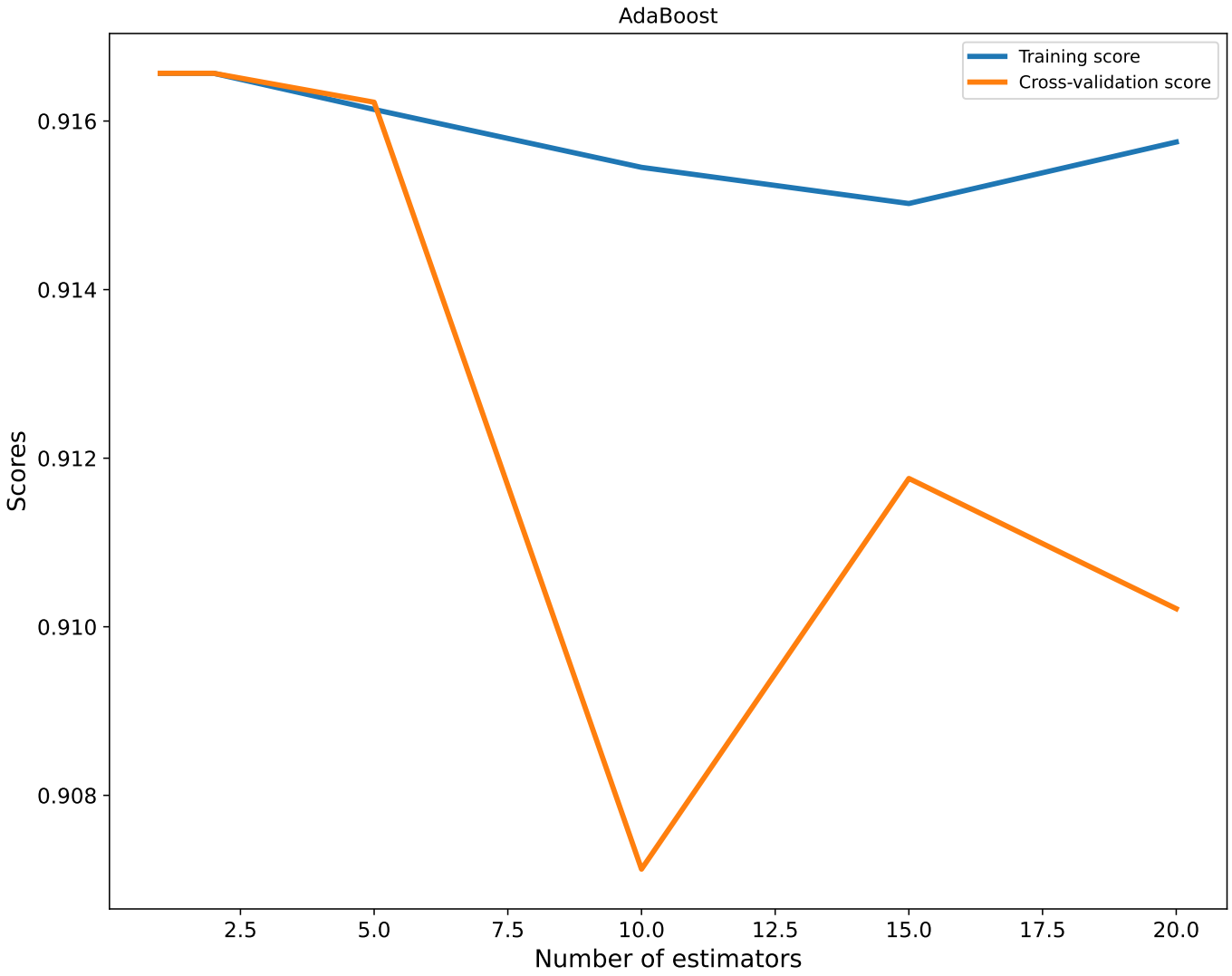


Ensemble methods

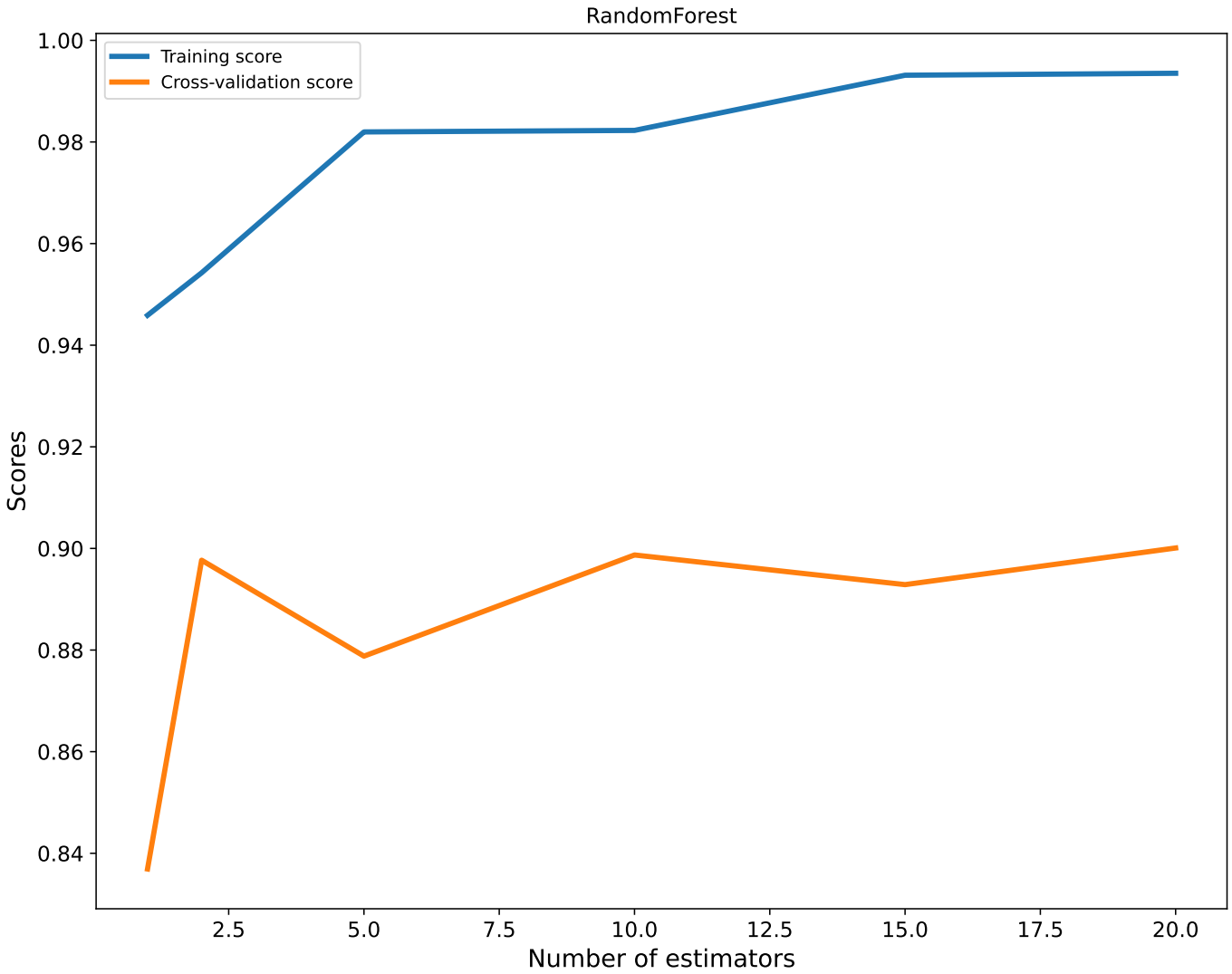
Ensemble methods combine multiple learners (models).

- Boosting (Shapire and Freund, 2012): combining weak learners that perform only slightly better than chance to reduce bias.
- AdaBoost (Adaptive Boosting) is a combination of weak trees (one level trees) that are trained sequentially by re-weighting training data points that were previously misclassified.
- Bagging (Bootstrap Aggregating) (Breiman, 1996): combining different learners on subsets of data chosen randomly with replacement to reduce variance.
- Random forest is a combination of different trees on randomly chosen subsets of data and randomly chosen subsets of features.

Validation curves



Validation curves



Logistic regression

- Binary logistic regression for the positive class is given by

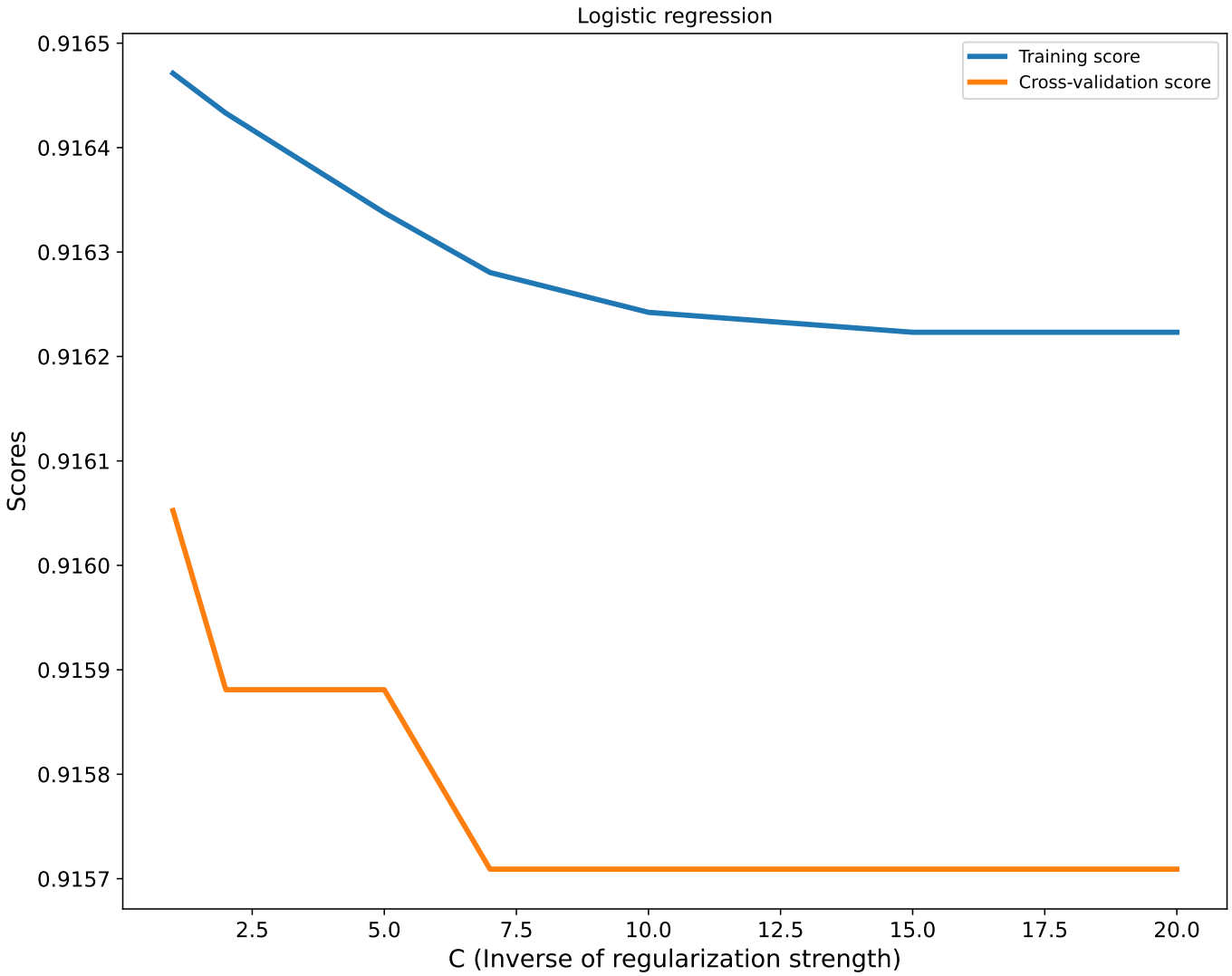
$$\Pr(y_i = 1|\mathbf{X}_i) = p(\mathbf{x}_i; \mathbf{w}) = \frac{1}{1 + \exp(-\mathbf{w}^T \mathbf{x}_i + b)}.$$

  To find the optimum values of w, we have

$$\min_w \sum_{i=1}^{n} (-y_i \log(p(x_i; w)) - (1 - y_i) \log(1 - p(x_i; w))) + \alpha r(w),$$

- $r(w)$ is the regularisation term.
- $\alpha = 1/C$ is the hyper-parameter.

Validation curves



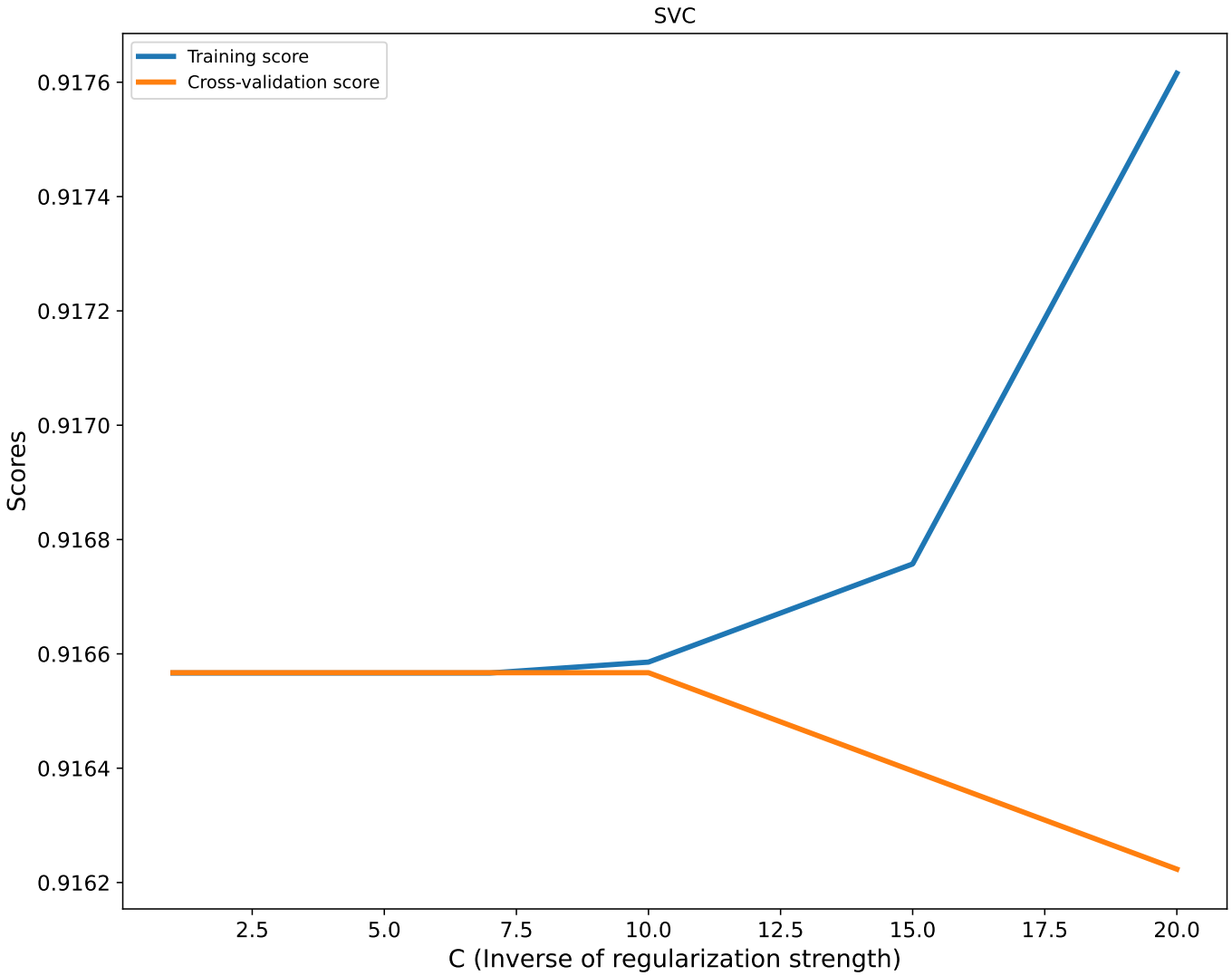
Support Vector Classification

- Margin: the distance from the decision boundary (hyperplane) to the closest point across both classes.
- Support vectors: the closest points to the decision boundary.
- Slack variables ψ_i: points within the margin, even misclassified.

$$\min_{w,b,\xi} \mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i$$
$$s.t. \quad y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i,$$
$$\xi_i \geq 0$$

- C is a hyper-parameter that trades off the margin with the amount of slack.
- If C is large, the SVC becomes strict and if C is small, the SVC becomes loose.

Validation curves



Confusion matrix

	$y = 1$	$y = 0$
$\hat{y} = 1$	True positive (TP)	False positive (FP)
$\hat{y} = 0$	False negative (FN)	True negative (TN)

- Recall = TP / (TP + FN)
- Precision = TP / (TP + FP)
- Accuracy = (TP + TN) / (TP + TN + FP + FN)
- True positive rate = TP / (TP + FN)
- False positive rate = FP / (FP + TN)

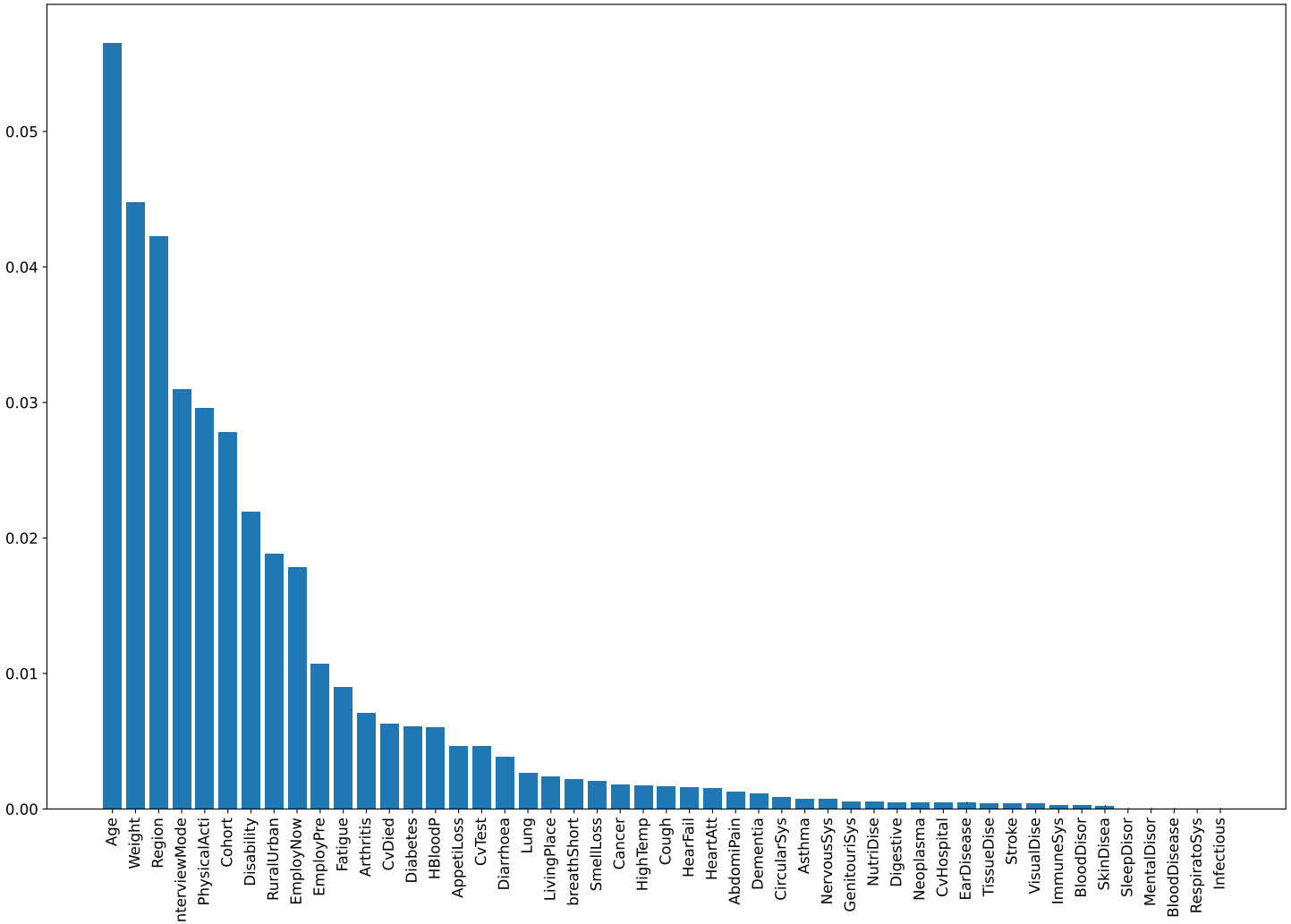
Comparison of models and algorithms

	KNN	RF	AdaBoost	Logistic	SVC
Train	0.916	0.985	0.917	0.915	0.915
Test	0.922	0.911	0.914	0.921	0.922
Precision	0.667	0.105	0.176	0.333	0.000
Recall	0.018	0.018	0.026	0.009	0.000
AUC	0.600	0.640	0.710	0.740	0.580

Feature importances by permutation

KNN	Random forest	AdaBoost	Logistic
Region	Age	Age	Cohort
Fatigue	Weight	IntMode	Age
CVTest	Region	PhysiActi	BreathShort
HeartFail	IntMode	EmployNow	SkinDisea
Cancer	PhysiActi		
	Cohort		
	Disability		
	RuralUrban		
	EmployNow		
	EmployPre		

Feature importances for random forest model based on feature permutation



Conclusions

- We have an imbalanced dataset.
- AdaBoost can discriminate between positive and negative labels better than other models and algorithms in terms of AUC and Recall.
- K-nearest neighbours performs better in terms of Precision.
- K-nearest neighbours and SVC provide the highest test accuracy.
- The important features are not the same for all algorithms and models.
- Almost all features are considered important by Random Forest.
- Age and region are selected as important feature by most models and algorithms.