Project Proposal: Differences in performance between logistic regression and machine learning for clinical risk prediction

Sam Husbands

Candidate Number: 09996

May 2022

Contents

Description of the Problem

Clinical Prediction Models

Clinical risk prediction models are explicit, empirical approaches to estimate the probabilities of disease using person level data (Steyerberg, 2019, p.2). There are two main types of clinical prediction models; clinical prediction models for diagnosis, which seeks to estimate the probability of a patient having a disease; and clinical prediction models for prognosis, which seek to estimate the likely course of a disease.

One widely adopted example of a clinical prediction model is the Framingham risk score (Wilson et al., 1998). This model estimates the ten-year risk of cardiovascular disease using individual level data including age, diabetes, smoking status, blood pressure and cholesterol. Such models have public health benefits, in that prediction models can be used to identify high risk individuals and allow for preventative treatment. The Framingham risk score provides a good example of such benefits, as it was widely adopted in the UK (British Cardiac Society, 2000) and used to recommend preventative treatment, such as statin therapy, and lifestyle modifications, such as quitting smoking in high-risk patients. The focus of this work concerns problems with a binary outcome, representing the occurrence of disease in a given time period, such as being diagnosed with cardiovascular disease or not in the next ten years in the case of the Framingham risk score. However, it is possible to model a time-to-event outcome, modelling how long until a patient experiences a disease or disease outcome.

Due to their applications in clinical practice and public health, it is important that clinical prediction models measure patient risk as accurately as possible. There are two main competing approaches for fitting these personalised clinical risk prediction models for diagnosis: traditional approaches based around logistic regression; and more recent developments utilising supervised machine learning algorithms (Christodoulou et al., 2019).

Logistic Regression

Logistic regression is an extension of linear regression for modelling binary outcomes (Osborne, 2015, p.3). It assumes the dependent variable **Y** follows a Bernoulli distribution. In the clinical risk prediction setting, **Y** is typically the binary outcome of contracting a disease or not within a given time period. Here, p refers to the probability that a dependent variable **Y** is equal to one, and is conditional on a set of explanatory variables or features **X**. In clinical risk prediction models, **X** is a set of individual level characteristics that regularly include age, sex, and body mass index (BMI). When

fitting the model, the probability is transformed into log odds and the features are modelled with respect to the log odds.

$$\ln\left(\frac{p}{1-p}\right) = \beta \boldsymbol{X}$$

The main advantage of modelling the dependent variable in this way is that the bounds of the odds are (0, ∞) and the bounds of the log odds are (-∞, ∞). As a result, we can model the log of the odds (the logit function) as a linear function of a set of predictor variables, and it can never result in impossible predictions for p. The optimal values of the β vector are determined through maximum likelihood estimation. By rearranging this equation, we can model the conditional probability as below:

$$P(Y = 1 \,|\boldsymbol{X}) = \frac{e^{\beta \boldsymbol{X}}}{1 + \ e^{\beta \boldsymbol{X}}}$$

When faced with a new observation, the probability of the patient experiencing the disease in the given time period can be estimated using the optimal values for β and the observed explanatory variables for the new patient. For classification, the patient is deemed to be at high risk if this conditional probability is over a certain threshold, and preventative treatment can be recommended. The commonly used threshold is 0.5, though other thresholds are valid.

The key limitation of logistic regression arises in the model specification stage. The decision of which variables to include, the functional form of the explanatory variables and the interactions between variables is left up to the researcher. As with all parametric methods, the prediction performance of logistic regression is contingent on whether the data follows the assumed model (Couronné et al., 2018) and often, researchers only include the explanatory variables additively and without interactions included. These assumptions are often referred to as linearity and additivity. Linearity is the assumption that there is a linear relationship between the logit transformation of the dependent variable and the predictor variables, and additivity is the assumption that there are no interactions between variables (Couronné et al., 2018). These assumptions are not necessary for logistic regression, but they are frequently the default of researchers and there is no standard practice on how to incorporate non-linearities and interactions.

Machine Learning

Machine learning is a very broad concept that has many competing definitions. In this review, machine learning refers to supervised classification algorithms that learn patterns directly from data (Mitchell, 1997). Related fields such as unsupervised and reinforcement learning are outside the scope of this proposal as they cannot be directly applied for predicting clinical risk. Differences between machine learning and statistical modelling are not clear cut. One distinction between machine learning and statistical modelling is highlighted by Breiman (2001a) who describes two cultures. Firstly, a statistical modelling culture that assumes the outcome variable is generated by a given stochastic data model. In the case of logistic regression, the given stochastic data model is typically that the explanatory variables are linearly and additively associated with log odds of disease. In contrast, the machine learning culture focuses on algorithmic models that treat the underlying data mechanism as unknown. The machine learning culture is therefore more flexible, as it does not directly assume a given distribution or functional form for the data. As a result, machine learning algorithms are therefore generally much more automatic about modelling non-linearities and interactions between variables than logistic regression (Couronné et al., 2018).

Due to the flexible nature of these models overfitting becomes much more likely, as in the case of neural networks (Tu, 1996). Overfitting refers to the problem where an algorithm captures noise in the dataset it was trained on that will not generalise to other datasets (Chandramouli et al., 2018). Preventing overfitting is one of the main challenges of model specification in machine learning, and is generally controlled by hyper-parameter tuning. Hyper-parameters are parameters that cannot be optimised and must be chosen before training starts (Yang and Shami, 2020), and hyper-parameter tuning refers to choosing hyper-parameters that maximise model performance. Performance of a machine learning model depends to a large degree on the ability of the researcher to tune hyper-parameters, and there is no standard practice on how to tune hyper-parameters for machine learning algorithms.

Project Aims and Objectives

The main objective of this project is to compare the performance of logistic regression and multiple machine learning algorithms in the clinical risk prediction setting by applying these methods to clinical risk prediction for modelling the probability of death in a ten-year interval. I then aim to show which factors have a large influence in determining the relative performances of logistic regression and machine learning. The analysis will focus on the number of explanatory variables and the number of data points.

The dataset I will be applying the algorithms to is the National Health and Nutrition Examination Survey (NHANES) (CDC, 2006) merged with the 2015 publicly available linked mortality file (CDC, 2015). NHANES is a United States cross sectional study running every two years, collecting data through a multitude of methods including interviews, questionnaires, health examinations and laboratory analysis. This dataset gives a rich set of variables on an individual level including data on demographics, diets, medical examinations, and laboratory test results. This cross-sectional data is then linked to a mortality file that specifies the individual's mortality status, cause of death and how long they survived after the interview. The aim is to apply logistic regression and supervised machine learning algorithms to estimate the probability of death over a given period. Estimating the probability of death will remove the issue of having to address competing risks. I will then measure performance of these models and see which model shows the highest level of performance. Finally, I plan to run these models again with different levels of training data points and different amounts of explanatory variables to verify if these are important factors in determining the final performance of these classification algorithms.

After the completion of this project, I intend to understand the relative performance of supervised machine learning and logistic regression in the clinical risk prediction space and hope to contribute to understanding what the key drivers behind these differing performance levels are. This should be useful for researchers when determining which model to fit for clinical risk prediction.

The next sections show my literature search and review of the applications of logistic regression and supervised machine learning in the field of clinical risk prediction. The final selection gives a more complete description of the project's objectives and outlines a clear work plan for the project's completion over the summer.

Literature Search

Prior to conducting the literature search, I first reviewed several textbooks relevant to the field. Especially helpful was Osborne's (2015) overview of logistic regression, which gave a good break-down and understanding of the key assumptions of logistic regression. Chandramouli et al's (2018) overview of modern machine learning classification algorithms was especially helpful in overviewing random forests, support vector machines and neural networks, and gave a good breakdown of the validation procedure in all machine learning algorithms. Steyerberg's (2019) application of statistical modelling was also rich in examples of classification models in the clinical risk prediction setting and gave a good grounding in typical clinical risk prediction models.

Following on from this, I looked at modern reviews of the literature and applications of algorithms to datasets. Christodoulou et al. (2019) gave a strong overview of 71 different studies that compared logistic regression and a variety of machine learning techniques in the clinical risk prediction field and highlighted the shortcomings of previous research. Overall, this paper concluded there was no systemic advantage of machine learning techniques over logistic regression, though limitations in these studies prevented a straightforward comparison. Kotsiantis et al. (2007) gave a comprehensive review of regularly used supervised machine learning algorithms generally, as well as their limitations. I then looked at key papers that were cited by the overview papers. This helped me identify key limitations of the current methods for measuring clinical risk and the seminal papers in the comparison of these methods. Boulesteix and Schmid (2014) highlighted the issues of model misspecification in logistic regression and how machine learning models avoid misspecification by allowing for greater flexibility. Van Calster et al. (2016) highlight how calibration is a pivotal aspect of model performance that is often unreported in machine learning literature. Saeb et al. (2017) highlights the issues of biased model validation, where often hyper-parameters are tuned on the entire dataset, which results in the model overfitting and not generalising to other datasets, overstating the model's accuracy.

These overview papers gave me a starting point from which to conduct my literature review, which I separated into four main topics. The first main topic to research was focussed on the practical implementations of logistic regression for clinical risk prediction, particularly with respect to model building and feature selection, as the issue of which features to include in the model has clear implications on model performance. The next topic was applications of supervised machine learning algorithms for clinical risk prediction, focussing on the main algorithms of random forests, support vector machines and neural networks as these are the most regularly used and typically report the best performance. The last main topics were evaluations of model performance and model validation, which is often done poorly in medical literature.

Once I understood logistic regression and machine learning algorithms in the clinical risk prediction setting, I conducted a thorough literature search of the clinical risk prediction literature using a systematic search of multiple databases. The three databases used were MathSciNet, Web of Science and Google Scholar. The results of this literature search are outlined below. In general, I found MathSciNet less useful, as it was largely focussed on the mathematics of optimisation of the algorithms rather than implementation or performance. Web of Science and Google Scholar were both very useful, with Web of Science containing a narrower range of results that made it easier to find the key papers and Google Scholar having a wider range of results that ensured no papers useful to the analysis were left out.

Database	Search	Results	Date	Comments
MathSciNet	Clinical* AND Prediction* *Anywhere*	238	15/04/2022	Too broad, 238 results
MathSciNet	Clinical* AND Prediction* *Title*	45	15/04/2022	Narrow list of results. Van Calster et al. (2020) useful in discussing shrinkage.
MathSciNet	Clinical* AND Prediction* AND Logistic	13	15/04/2022	Too narrow a list, not relevant.
MathSciNet	Clinical* AND Prediction* AND ("Machine Learning") *Anywhere*	13	15/04/2022	Too narrow list of results.
MathSciNet	“Logistic” AND ("Machine Learning")	146	15/04/2022	Broad and largely about optimisation. However, Fernández-Delgado et al. (2014) gave a comprehensive application of supervised ML.
MathSciNet	“Logistic” AND ("Machine Learning") *Anywhere* Valid* *Title*	2	15/04/2022	Too narrow, both largely irrelevant.
Web of Science	“Clinical*" AND "Prediction*”	131,825	15/04/2022	Too Broad
Web of Science	“Clinical* Risk*” AND "logistic"	13,387	15/04/2022	Very broad, but Steyerberg and Vergouwe (2014) gave good review on feature selection in logistic regression.
Web of Science	“Clinical*" AND "Prediction*” AND "Discrimination" AND "Calibration"	2877	15/04/2022	Alba et al. (2017) and Steyerberg et al. (2010) most useful, among plenty of useful results.
Web of Science	“Clinical*" AND "Prediction*” AND "Principal Component Analysis" AND "logistic"	47	15/04/2022	Thottakkara et al. (2016) most relevant, major example of dimensionality reduction being used for logistic regression. Otherwise not common.
Web of Science	“Clinical*" AND "Prediction*” AND "Machine Learning"	8,003	15/04/2022	Too broad.
Web of Science	“Clinical*" AND "Prediction*” AND "Random Forest*"	2,246	15/04/2022	Broad, but Strobl et al. (2009) gave useful breakdown of random forests.

Web of Science	“Clinical*" AND "Prediction*” AND "Support Vector Machine*"	2,343	15/04/2022	Orru et al. (2012) gave a useful breakdown of SVM, but mostly too broad.
Web of Science	“Clinical*" AND "Prediction*” AND "Neural Network*"	4,776	15/04/2022	Amato et al. (2013) gave strong review of neural networks in medical diagnosis
Google Scholar	"clinical" "prediction"	4,570,000	15/04/2022	Far too broad
Google Scholar	"clinical" "prediction" "logistic"	650,000	15/04/2022	Still too broad
Google Scholar	"clinical" "prediction" "logistic"	611,000	15/04/2022	Shipe et al. (2019) a useful discussion paper on building logistic regression models for clinical use.
Google Scholar	"clinical" and "prediction" and "random forest*" and "diagnosis"	51,500	20/04/2022	Too broad, but many good papers on random forests for diagnosis
Google Scholar	"clinical" and "prediction" and "support vector machine" and "diagnosis"	72,600	20/04/2022	Again Orru et al. (2012) a useful paper.
Google Scholar	"clinical" and "prediction" and "neural network" and "diagnosis"	137,000	20/04/2022	Largely too broad.
Google Scholar	"clinical" and "prediction" and "logistic regression" and "diagnosis" and “principal component analysis”	15,300	20/04/2022	Reddy et al. (2021) gave another example of PCA being used for feature selection in logistic regression. Most papers about imaging or treatment effects.
Google Scholar	“Hyper-parameter*” and “Tunability” and “Machine Learning”	354	20/04/2022	Probst et al. (2019a) very useful
Google Scholar	“Automatic” and “Selection” and “Hyper-parameter*”	24,000	20/02/2022	Luo (2016) useful in discussing hyper-parameter tuning.

My review of the literature revealed that logistic regression models have been used to predict clinical risk since at least the late 1980s, with seminal papers typically focussing on breast cancer and cardiovascular disease. Logistic regression seems to be the most widely used approach for

estimating clinical risk, having yielded the most results on all three databases, and they seem to be favoured by clinicians due to their relative simplicity and interpretability (Boulesteix and Schmid, 2014). Applications of machine learning to clinical risk are relatively newer, as these algorithms are more recent than logistic regression (Cortes and Vapnik, 1995) (Breiman, 2001b). However, there are examples of applications of recent supervised machine learning algorithms in clinical risk prediction since the mid-2000s (Lisboa and Taktak, 2006). There seems to be a growing interest in machine learning algorithms in the clinical risk prediction space, as there is a general feeling that the emergence of larger, more feature rich datasets may favour these flexible machine learning algorithms (Saeb et al., 2017). However, the general conclusion of the current literature seems to be that this hypothetical improvement in performance is not being seen in contemporary clinical risk prediction models.

My search of the literature also leads me to the conclusion there is a lack of papers implementing research into what factors drive the differing performance of logistic regression and machine learning in this setting. This is something I intend to explore further, particularly with relation to which algorithms handle a greater number of features or more data points. There is also very limited application of ensemble learning techniques and feature engineering, which I also intend to explore if there is sufficient time.

Literature Review

Logistic Regression approaches

The traditional approach of estimating a clinical risk prediction model is logistic regression. A recent example of this is North et al.'s (2011) model of pre-eclampsia (high blood pressure) in pregnant women. This model estimates the probability of developing pre-eclampsia using individual level data including age, BMI, alcohol intake and family history of pre-eclampsia.

The main challenge of this logistic regression approach is model development, which broadly refers to the identification and modelling of the predictor variables (Shipe et al., 2019). In logistic regression there is no agreed upon method for determining the best multivariate diagnostic model from a set of candidate predictors. It is regarded as impractical and poor practice to include all the candidate predictors when the number of candidate predictors is large relative to the number of data points, as this can result in model overfitting (Pavlou et al., 2015) as the fitted model then captures random noise present in the development dataset. Though a rule of thumb, it has been commonly found that there should be at least ten events per feature to mitigate the effect of overfitting (Steyerberg et al., 1999), as reliability of the model estimates greatly decreases when this is not the case. If the number of features exceeds the number of data points it becomes impossible to estimate a logistic regression (Strobl et al., 2009) and therefore including interaction effects and high order terms often leads to difficulties with reliable parameter estimation.

Feature Selection

There are two main approaches to deciding which features to use for clinical risk prediction, which are often used in conjunction with one another. The first is using specialist knowledge to refine the variables to be used in model fitting, whilst the other is to use software-controlled methods. Software-controlled methods are typically iterative methods, such as stepwise regression that either use backward elimination or forward stepwise to iteratively drop the worst predictor or include the next best predictor, based on p-values or the Akaike Information Criterion (Royston et al., 2009). Univariate analysis can also often be used, where each candidate explanatory variable is measured for its association with the outcome and only statistically significant candidate variables are used

going forwards. All these methods have limitations. Univariate analysis can result in useful candidate predictors being rejected, as can using specialist knowledge. Stepwise regression is unstable, and the resultant estimated coefficients tend to be too high, resulting in overfitting and overestimation of model performance (Steyerberg and Vergouwe, 2014). An alternative, much more rarely used approach in the clinical risk prediction space is to use dimensionality reducing algorithms such as principal component analysis (Thottakkara et al., 2016). PCA reduces the dimensionality of the data and extracts the most important information from the data (Abdi and Williams, 2010) and is widely used across many scientific disciplines. Bizarrely, it is very rarely used for clinical risk prediction, despite Thottakkara et al. (2016) finding that this means of feature selection improved the performance of logistic regression.

Modelling Continuous Variables and Interactions

A related issue to feature selection is how to model continuous predictors and interactive terms. The literature is clear that continuous variables should not be dichotomised as this results in information loss (Steyerberg and Vergouwe, 2014) and implicitly assumes there is no effect within levels of the variable. Nevertheless, dichotomisation is relatively common in the clinical risk prediction literature.

The other way that researchers typically model continuous variables is as a linear term. This necessarily assumes that the effect is the same at each part of the range of the predictor (Steyerberg, 2019), for instance, growing a year older has the same effect on the log odds of developing pre-eclampsia whether the patient is 20 or 35. A more flexible approach would be to use restricted cubic splines. Restricted cubic splines are a flexible way of modelling non-linearities and can fit a wide range of curves (Harrell, 2015, p. 23). Despite this, non-linearities are rarely modelled in logistic regressions for clinical risk prediction, and the decision of which non-linearities to model was often unclear (Christodoulou et al., 2019). Similarly, interaction terms were rarely modelled. Interaction terms capture the idea that the value of one predictor can have an impact on the effect of another predictor. As most logistic regressions in the clinical risk prediction space do not model non-linearities or interactions, the models are likely to oversimplify the true relationship between features and the outcome variable and thus underfit, making model performance lower than it would have been had interactions and non-linearities been properly addressed.

In the motivating example of pre-eclampsia, a multitude of methods were used for feature selection. The researchers excluded variables based on domain knowledge or where there was a high frequency of missing data. They then performed univariate regression analysis on each of the explanatory variables and excluded those variables which were statistically insignificant at the 10% level. After doing this, they performed a forward stepwise regression. The handling of non-linearities and interactions was unclear, though as neither appeared in the final equation, it is probable the researchers tacitly assumed linearity and additivity. This example highlights a general trend, in that the implementation of feature selection and model building is very inconsistent (and often unclear) in clinical risk literature.

After deciding on the number of features and their functional form, it can be recommended to apply shrinkage to the regression coefficients, shrinking them toward zero (Van Calster et al., 2020). Shrinkage is used to prevent overfitting in logistic regression and aims to prevent predicted risks that are too extreme. However, it has been shown that shrinkage and other penalised logistic regression methods often work poorly in individual clinical datasets and do not solve the problems associated with either small sample size or small numbers of events per variable.

Logistic Regression Takeaways

In summary, the main challenge of fitting logistic regression to predict disease occurrence is the issue of model development. Typically, in the literature, the methodology for doing this either involves directly assuming only linear direct effects, or discretization, transforming the variable into a categorical variable and assuming the variable to have a relation only across levels. It is rare in this field for non-linearities and interactions to be considered, despite the fact these relationships in nature are unlikely to be linear. There are other challenges in logistic regression, including model estimation, the encoding of variables, the handling of missing data and the measurement of performance, although these issues are often ones that supervised machine learning algorithms share.

Machine Learning

Hypothetical Advantages of Machine Learning

Overall, the literature around logistic regression shows that there are clear limitations when it comes to applying this method to real world data. Methods of feature selection and the typical implicit assumption that the explanatory variable is linearly related to the log odds of the outcome variable in modern literature suggest that there are ways to improve model performance with machine learning techniques.

In the clinical risk prediction setting, machine learning generally describes a set of computationally intense and highly flexible algorithms that identify patterns in complex data structures (Jiang et al., 2021). Unsupervised and reinforcement learning are not within the scope of this project, as these machine learning techniques are not used for prediction when the outcome is known. Machine learning contrasts with logistic regression and other more rigid statistical modelling techniques in that it does not require the same statistical assumptions around the functional form of the relationship between the explanatory variables and the outcome. There are at least seventeen different regularly used machine learning classifiers (Fernández Delgado et al., 2014) and in the clinical risk prediction space, these are most often random forests, support vector machines and artificial neural networks, which are widely used machine learning algorithms across many domains. It is plausible they confer an advantage in this area due to the flexible way they model non-linearities and interactions between variables. It is regularly claimed in the literature that machine learning algorithms are better able to handle both a larger number of predictors and their interactions, and the flexible nature of these algorithms prevents researchers from having to pre-specify the important predictor variables as these directly learn the important predictors from data (Rajkomar et al., 2018).

However, despite the theoretical improved performance, the real-world studies have resulted in mixed conclusions. A general survey on all classifiers by Fernández-Delgado et al. (2014) on a wide range of datasets found that random forests tended to yield the highest accuracy in classification. Accuracy here is the percentage of correctly classified units, with accuracy assessed at a 50% cut-off. However, in the clinical risk setting, the theoretical improved performance of machine learning does not seem to materialise (Christodoulou et al., 2019).

Typical Supervised Machine Learning Algorithms

Fernández Delgado et al. (2014) highlighted seventeen different families of machine learning algorithms for classification, and all could feasibly be used to model clinical prediction risk. However, the scope of this project is not to decide upon the optimal classifier in clinical risk prediction, but to

see whether there are gains from these machine learning algorithms and why these improvements in performance are likely to arise. As such, in the project I will be exclusively focussing my attention on three algorithms: random forests, support vector machines and neural networks. This is because these algorithms are some of the most regularly used, and often report the highest performance, though the optimal classifier does appear to be linked to the characteristics of the individual dataset (Khan et al., 2020).

Random Forests

There are many variants of the random forest algorithm, however, the one I will be implementing is the original introduced by Breiman (2001b). The random forest is an ensemble learning technique, aggregating the estimate of multiple classification decision trees (Couronné et al., 2018). Each tree in the forest is built based on a recursive partitioning, where the features are iterated over and split such that the resultant splits of data contain observations that are more like one another. The original method to choose the optimal split was to minimise the Gini impurity. An example of a constituent tree is shown below, which seeks to predict myocardial infarctions (heart attacks) over the next ten years based on age and smoking status. To evaluate a constituent tree, the data starts at the root node and at each split, goes left if the data satisfies the criteria. In this tree, only over 50s who smoke would be classified as at risk of a heart attack.

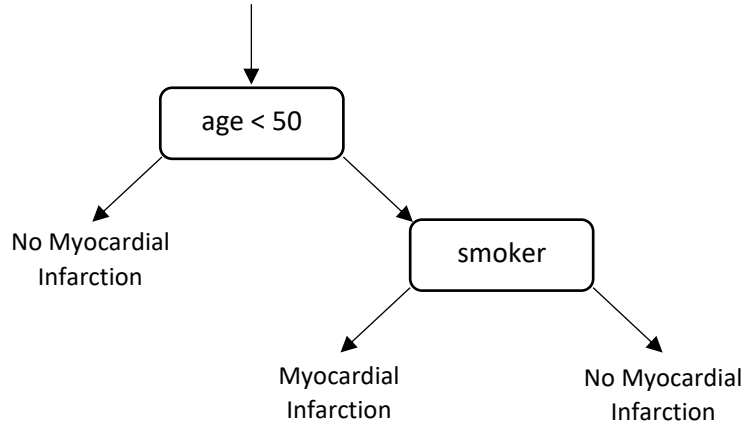

Figure 1: Example constituent decision tree. Loosely based on Karaolis et al. (2010)

To prevent each tree being the same, each tree experiences randomisation during training. The first way to force diversity in the random forest is by bagging, where each tree is built on a bootstrap sample of the original dataset (Chandramouli et al., 2018). Additionally, at each split, only a given number of features are given to the tree as candidates to split upon.

This method has potential advantages over logistic regression in that it captures dependency and non-linearity in its features by design, whereas any non-linearities and interaction in logistic regression must be explicitly included by the researcher. Any subsequent split is contingent on the previous splits, and so in classification the features are not modelled additively. It has been shown in simulated data that random forests are able to better classify dependent, non-linear structures in simulated data (Strobl et al., 2009).

A potential downside of random forests is that the performance is impacted by hyper-parameters which are set before the algorithm starts. Examples of these are the number of candidate predictors at each split and the minimum number of datapoints at terminal nodes (Couronné et al., 2018). The

performance of random forest is contingent upon these hyper-parameters to some degree, and methods to identify the best hyper-parameters are not consistent in literature (Probst et al., 2019b). A more detailed look at hyper-parameter tuning is reserved for the next section.

A recent example of random forests applied to the clinical risk prediction space is implemented by Su et al. (2020). They develop both a logistic regression and random forest for predicting whether the patient has cardiovascular disease based on individual level data such as age and BMI. Despite seemingly assumingly linearity and additivity in their logistic regression, the random forest had worse discrimination performance. This is a result that broadly generalises across clinical risk prediction literature, despite evidence of random forest generally being one of the top classifiers. Ultimately, the best classifier is contingent on the innate characteristics of the dataset such as the number of features, the class imbalance, the number of datapoints and the distribution of the data (Khan et al., 2020).

Support Vector Machines

Support Vector Machines (SVMs) estimate a high dimensional surface called a hyperplane, defined over the features that best separate the groups in the data (Orru et al., 2012). Such groups could be patients with Alzheimer's disease and those without, and the features could include age and gender. The hyperplane is estimated by maximising the margin of separation between the two distinct groups.

A simplified, hypothetical model of this is an algorithm that predicts ten-year risk of myocardial infarction using only weight and hours of exercise per week as features.

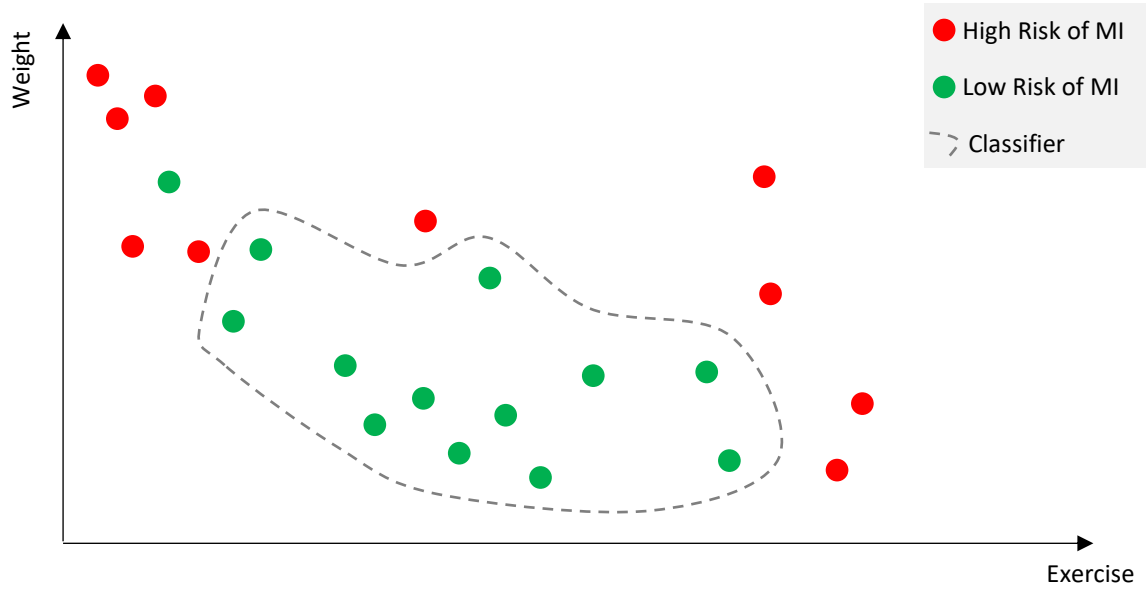

Figure 2: Hypothetical Support Vector Machine classifier

Figure 2 represents synthetic data. In two dimensions, this data is not linearly separable, however the support vector machine is able to create a non-linear decision boundary for this data. A 1-dimensional hyperplane that separates this data can be expressed by:

$$w_1 exercise + w_2 age + b = 0.$$

The objective of a support vector machine is to choose values of w_1, w_2 and b such that two hyperplanes (margins) can be estimated:

$$w_1 exercise + w_2 age + b \geq +1$$

$$w_1 exercise + w_2 age + b \leq -1$$

This ensures data above the hyperplane is classified as +1 and therefore has a high risk of myocardial infarction and data below the hyperplane is classified as -1 and thus has a low risk of myocardial infarction. To maximise the distance between the hyperplanes (maximise the soft margin), it can be shown that this problem can be represented as (Jakkula, 2006):

$$\min_{w_1,w_2,b} \frac{\sqrt{{w_1}^2 + {w_2}^2}}{2} - \sum_{i=1}^{n} \lambda_i (y_i * (w_1 exercise_i + w_2 age_i + b) + s_i - 1) + \alpha \sum_{i=1}^{n} s_i \ \forall \ i \ \in (1, \dots, N)$$

Where n is the total number of datapoints, s_i represents a slack variable that allows for points to be the wrong side of the hyperplane, and α represents a Lagrange multiplier that penalises large slacks. α is a hyper-parameter that is optimised via hyper-parameter tuning. Mathematically, this is the way that SVM determines the optimal boundary of the classifier whilst allowing for misclassifications. At present however, the SVM is unable to create the non-linear classifier above and would only be able to create linear classifiers (straight lines) that would be a poor fit for the data.

The last element of a support vector machine is a kernel that facilitates non-linear boundaries. This entails passing the data through a non-linear map resulting in higher dimensional data. A common kernel choice is the polynomial kernel (Jakkula, 2006). as shown below:

$$K(\overrightarrow{age}, \overrightarrow{exercise}) = (< \overrightarrow{age}. \overrightarrow{exercise}' > +1)^d$$

Where d represents the order of the polynomial kernel. Passing the features through this kernel results in features that are polynomials and interactions of the original two features of age and exercise, and in this higher feature space, there is a soft margin maximum classifier or linear decision boundary that fits the data well. The choice of kernel is another hyperparameter that can be tuned.

SVMs typically are estimated with more than two features, and thus cannot be easily shown visually. However, the principle of a soft-margin maximum classifier using features passed through a kernel function to make a non-linear classifying boundary is the same regardless of the number of features used. As with random forests, the theoretical advantage of this method is that the algorithm can identify non-linearities and interactions present in the data. However, the model again requires the selection of hyper-parameters, this time including the regularisation parameter that determines the soft margin and the choice of kernel function (Duan et al., 2003).

A recent example of SVM applied to clinical risk prediction is implemented by Cui et al. (2018). They develop a support vector machine to measure risk of readmission for patients with diabetes, alongside a logistic regression model and many other supervised machine learning algorithms. This paper engages in multiple issues with data preparation, such as class imbalance, which distorts risk predictions and can worsen model performance (Goorbergh et al., 2022). However, this paper does find considerable gains for support vector machines modelling risk of readmission for patients with diabetes.

Neural Networks

Artificial Neural Networks (ANNs) constitute a multitude of approaches that are inspired by human neural architecture (Amato et al., 2013). The main advantage of this algorithm compared to

parametric models such as logistic regression is in its ability to capture non-linearities and interactions. The most widely used neural network model is the multilayer perceptron. In this network, the neurons are arranged in layers. There is an input layer, that takes in individual level patient data and represents individual patient characteristics, and multiplies them by weights and connections, a hidden layer, that further multiplies these inputs, and an output layer, that typically reflects patient risk of having a particular disease. It is the hidden layer that distinguishes the neural network from a straightforward logistic regression, and the hidden layer or layers allow for neural networks to model many different non-linear and interactive relationships between variables more automatically than linear regression (Steyerberg, 2019, p. 71).

A simple, hypothetical example of this is adapted from Fei et al. (2018) and predicts acute lung injury in patients already diagnosed with pancreatitis, using the variables age and BMI.

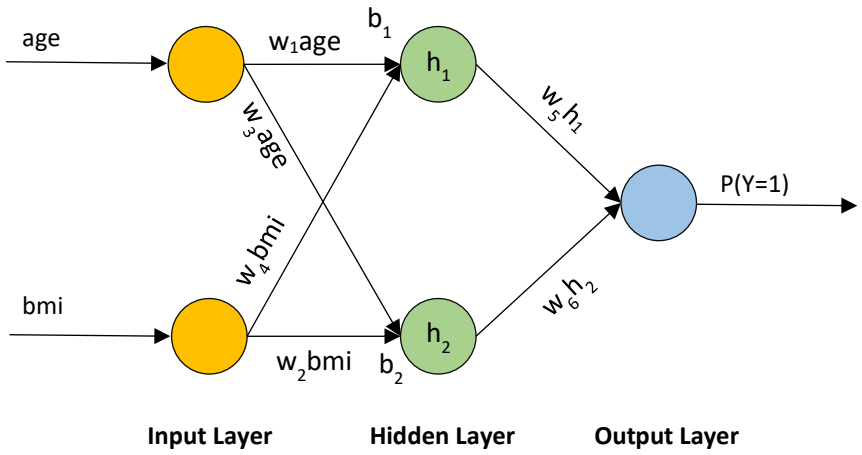

Figure 3: Hypothetical Artificial Neural Network (loosely based on Fei et al., 2018)

The neural network consists of nodes and connections between the nodes (Chandromouli et al., 2018). The input layer simply acts as a means of inputting data into the algorithm. The algorithm then multiplies the data by weights (w) in each connection and adds a bias term (b) before passing the data to the nodes in the hidden layer. The data is therefore fed through into the first hidden layer through a weighted link function. For instance, the input (x) for the first hidden node in the above network is given by:

$$x_1 = w_1 age + w_4 bmi + b_1$$

In general, the input to a node in a neural network is the sum of all its connections and a bias. In the hidden layer, the input is then passed through an activation function. A common choice for this is the sigmoid activation function given below:

$$h_1 = f(x_1) = \frac{1}{1 + e^{-kx}}$$

Where k is a constant that can be tuned. The choice of activation function is a hyper-parameter that can be tuned. After the inputs have been passed through the activation function, they are again

passed through connections and biases reaching an output layer. The final output is an estimated probability resulting from the output layer, which again applies a sigmoid activation function.

Whilst this is a simple example, typical fitted neural networks contain the same constituent parts, though they regularly contain many more input variables, connections, and hidden layers. The values of the weights and biases of the neural network are parameters to be estimated. These parameters are determined by backpropagation, which trains the model sequentially and updates the weights and biases using gradient descent to determine in which direction the weights should be adjusted, and a learning rule that determines the size of the step. The typical objective function used in backpropagation is the sum of squared residuals. Each training cycle is referred to as an epoch, and results in the weights and biases being adjusted (Basheer and Hajmeer, 2000). Like with random forests and support vector machines, the neural network has hyper-parameters that dictate its final performance, including the number of hidden layers, the number of neurons in each layer, the activation function, the learning rate, and the number of epochs.

Fei et al. (2018) implemented a neural network to predict acute lung injury in patients, alongside a logistic regression model. This paper is reflective of many machine learning papers in that the methodology reporting was unclear, particularly with respect to the choice of activation function. However, this paper found evidence that artificial neural networks do improve performance over logistic regression in predicting acute lung injury risk. The AUC of the neural network was 0.86, and the AUC of the logistic regression was 0.70. This was deemed to be a statistically significantly higher AUC value (see section on model performance), which is a measure of discrimination.

Ensemble Approaches

An ensemble approach is the approach of combining multiple estimators for improved performance. An example of this would be combining the estimators of a random forest, a support vector machine, and a neural network together and letting them vote on a classification. This tends to improve performance, especially when diversity is high in the estimators and the correlation between errors is small (Sewell, 2008).

One example of a hybrid ensemble being applied in clinical risk prediction is Nikookar and Naderi (2018), who used an ensemble of five different learners including a random forest and a support vector machine to estimate individual patient risk of heart disease and showed that ensembles of many heterogeneous estimators outperform individual ML algorithms.

Model Performance

There are a wide range of ways that clinical risk prediction models report performance. There are two main ways through which clinical prediction models should be assessed: discrimination and calibration (Alba et al., 2017). Discrimination refers to the ability of a model to differentiate high risk patients from low-risk patients, and calibration refers to the accuracy of absolute risk estimates.

Discrimination

The most common measure of discrimination reported is the C statistic, which measures the area under the receiver operator curve (Steyerberg et al., 2010). The receiver operator curve plots the true positive rate against the false positive rate for every threshold. A threshold is the value above which an estimated probability must be to classify as positive. The C statistic can be interpreted as the probability that given two patients, one who has the disease and one who does not, the model successfully prescribes a higher probability to the individual who had the disease.

Calibration

Another important measure of model performance is calibration. Calibration captures the reliability of model estimates, and a model is said to be properly calibrated if, for those individuals with an estimated probability of $\hat{p}$, a proportion $\hat{p}$ develop the predicted outcome. This is typically explored with a calibration plot, which plots the predicted probability from the patients against the observed proportions in the data, typically within subgroups of participants ranked by increasing predicted probability as below:

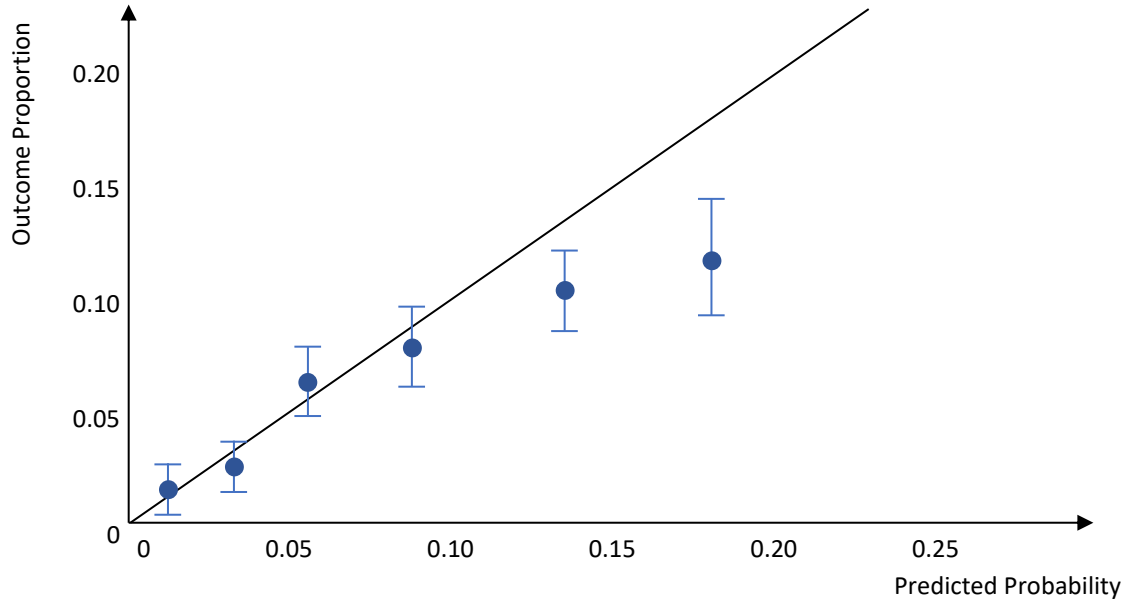

Figure 4: Example calibration plot (loosely based on North et al., 2011)

For instance, in the above calibration plot, it can be observed that the model overestimates the clinical risk in high-risk categories. It is also possible to apply smoothing techniques such as Loess as an alternative to using subgroups (Steyerberg et al., 2010). However, the specific smoothing algorithm used can affect the graphical impression. Calibration is very rarely discussed in recent literature (Van Calster et al., 2019) despite the fact a poorly calibrated model can make predictions misleading.

Decision Curve Analysis

The final, rarer, means of comparing model performance is decision curve analysis, first introduced by Vickers and Elkin (2006). The main limitation of AUC metrics is that they focus solely on accuracy, and do not factor in information on the consequences of a false negative or false positive classification. Decision curve analysis incorporates the consequences of false positive and false negative outcomes using the consequence of threshold probabilities. A threshold probability is the probability of disease at which a patient is indifferent between treatment and no treatment. For each threshold probability, the net benefit of the prediction model can be calculated as below, and it is possible to estimate the net benefit of the prediction model for all threshold probabilities. Including decision curves is useful as they demonstrate the benefit of a clinical risk model for clinical practice, comparing the net benefit of treating based on threshold probabilities to the net benefit of treating all patients.

Model Validation

Cross Validation and Bootstrapping

Model validation refers to ensuring the model predictions generalise to new observations. This is typically done by partitioning the dataset into a training set, to fit the model, and a testing set to validate performance. The main reason to do this is to ensure model predictions generalise to data the model was not trained on (Chandramouli et al., 2018). Machine learning algorithms are highly flexible and can overfit, capturing noise in the dataset they were trained on. The most common method of model validation is k-fold cross validation, where the data is randomly split into k-equally sized portions (Saeb et al., 2017). The model is trained on the training set (k-2 portions), the model's hyper-parameters are tuned on a subsection (the validation set), and the last subsection is used to evaluate performance (the test set). This is repeated k-times, and each time a different subsection is used as test data. The average AUC and calibration across the k test sets are then reported as model performance (Chandramouli et al., 2018). This avoids overfitting as the validation set is used to tune hyper-parameters to maximise model performance on data the model was not trained on, and overfitting is prevented as the reported test set accuracy is independent of the validation and training sets.

In addition to k-fold cross-validation, bootstrap sampling is another common way to partition the data into test-validate-train subsections by using simple random sampling with replacement for each section. This allows for the same data points to be in both the training and test sections but is useful for small datasets as the number of training-validation-test data samples are unlimited.

Hyper-parameter Tuning

Hyper-parameters are model parameters that are not optimised during training (Probst et al., 2019a) and are set by the researcher before model training. They contrast with parameters, which are learned by the model during model training. An example hyper-parameter would be the minimum data points per leaf node in the constituent decision trees of a random forest, whereas an example parameter would be which feature to split upon in the constituent decision tree (Luo, 2016). When building ML algorithms, it is possible to use the default values of these parameters from software packages, however, manual configuration of hyper-parameters using a validation set typically leads to improved performance (Luo, 2016). Once the dataset is split into train-validate-test subsections, the training set is used to train the model with different values of the hyper-parameters. The validation set is used to measure the model's performance with each hyper-parameter configuration, and the test set is used to measure performance with the optimal hyper-parameter configuration. There are a multitude of ways of determining how to choose which subset of hyper-parameters to use in model training, including random search and gradient based methods (Luo, 2016).

Hyper-parameter tuning is not something that is done well in recent literature. Christodoulou et al. (2019) show that the model validation procedure was biased or unclear in 68% of studies. Typically, this arises when hyper-parameters are tuned on the whole dataset. The results in overfitting as it uses the test data to train the model, therefore allowing the model to capture noise in the dataset the model was trained on that will not generalise to other datasets**.**

Work Plan

The aim of this project is to investigate the different performances of logistic regression and supervised machine learning for clinical risk prediction, and to identify and explain what drives the differences in performance. This will be done by applying the methods to a specific dataset, the NHANES cohort survey, merged with the linked mortality file. As this survey is publicly available and anonymised, there are no issues with accessing and using this data for this purpose.

Part 1 – Implementation of methods on whole dataset

Task Number	Task	Duration	Dependencies
1	Review implementations of logistic regression, random forests, SVM and neural networks in python 3.6.	1 week	None
2	Clean and merge the 2005-06 NHANES cohort survey with the linked mortality file.	½ a week	None
3	Apply multiple imputations to address missing values.	½ a week	Task 2
4	Apply logistic regression with cubic splines and interaction terms to model performance to the NHANES dataset on a restricted set of known strong predictors of mortality from medical literature.	1 week	Tasks 2 and 3
5	Apply supervised ML algorithms (including an overall ensemble) to the NHANES dataset on a restricted set of known powerful predictors.	1 week	Tasks 2 and 3
6	Measure performance using AUC, calibration, and decision curves.	½ a week	Tasks 4 and 5
7	Write up overall results.	1 week	All previous tasks

Total Duration: 5½ - 6 Weeks

Part 2 – Implementing methods on differing amounts of data and features

Task Number	Task	Duration	Dependencies
1	Divide the dataset randomly into arbitrarily sized chunks of individuals (e.g., 500, 1000, 2000, whole data).	1 day	None
2	Reperform analysis with each new sub-dataset and evaluate performance.	½ a week	Task 1
3	Write up overall results on how the performance of different algorithms changes with different amounts of data.	½ a week	Tasks 1 and 2
4	Perform cleaning of additional likely candidate features with low missing variable count. Perform multiple imputations to address missing values	½ a week	All previous tasks
5	Reperform analysis with different levels of features and evaluate performance.	½ a week	All previous tasks
6	Apply principal component analysis and reperform analysis with the new, smaller principal components.	1 week	Tasks 1 through 4
7	Write up the final results.	2 weeks	All previous results

Total Duration: 5 – 6 Weeks

This outline of the tasks required to complete this project will result in my achieving the overall objectives of the project, in comparing the performance of logistic regression and supervised machine learning as well as investigating some of the driving factors behind the differences in performance in these widely applied classification techniques. My literature search indicates that exploring model performance differs when exposed to different methods of feature selection, different number of features and different amounts of data is lacking in recent studies, and this project should demonstrate when these different models are appropriate given how many features and how much data is available to the researcher.

References

Alba, A.C., Agoritsas, T., Walsh, M., Hanna, S., Iorio, A., Devereaux, P.J., McGinn, T. and Guyatt, G., 2017. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. *Jama*, *318*(14), pp.1377-1384.

Amato, F., López, A., Peña-Méndez, E.M., Vaňhara, P., Hampl, A. and Havel, J., 2013. Artificial neural networks in medical diagnosis. *Journal of Applied Biomedicine*, *11*(2), pp.47-58.

Basheer, I.A. and Hajmeer, M., 2000. Artificial neural networks: fundamentals, computing, design, and application. *Journal of Microbiological Methods*, *43*(1), pp.3-31.

Boulesteix, A.L. and Schmid, M., 2014. Machine learning versus statistical modeling. *Biometrical Journal*, *56*(4), pp.588-593.

Breiman, L., 2001a. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), pp.199-231.

Breiman, L., 2001b. Random forests. *Machine learning*, *45*(1), pp.5-32.

British Cardiac Society, British Hypertension Society, British Hyperlipidaemia Association and British Diabetic Association, 2000. Joint British recommendations on prevention of coronary heart disease in clinical practice: summary. *BMJ: British Medical Journal*, *320*(7236), p.705.

Centers for Disease Control and Prevention (CDC). and National Center for Health Statistics (NCHS)., 1999-2021. *National Health and Nutrition Examination Survey Data* [Online]. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention,. Available from: https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2005. [Accessed 16th April 2022]

Centers for Disease Control and Prevention (CDC)., National Center for Health Statistics (NCHS). And Office of Analysis and Epidemiology., 1999-2014. *2015 Public-Use Linked Mortality Files* [Online]. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention,. Available from: https://www.cdc.gov/nchs/data-linkage/mortality-public.htm. [Accessed 16th April 2022]

Chandramouli, Dutt, Das, Dutt, Saikat, and Das, Amit. Machine Learning. Pearson Education India, 2018. Web.

Christodoulou, E., Ma, J., Collins, G.S., Steyerberg, E.W., Verbakel, J.Y. and Van Calster, B., 2019. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, *110*, pp.12-22.

Cortes, C. and Vapnik, V., 1995. Support-vector networks. *Machine Learning*, *20*(3), pp.273-297.

Couronné, R., Probst, P. and Boulesteix, A.L., 2018. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics*, *19*(1), pp.1-14.

Cui, S., Wang, D., Wang, Y., Yu, P.W. and Jin, Y., 2018. An improved support vector machine-based diabetic readmission prediction. *Computer Methods and Programs in Biomedicine*, *166*, pp.123-135.

Duan, K., Keerthi, S.S. and Poo, A.N., 2003. Evaluation of simple performance measures for tuning SVM hyperparameters. *Neurocomputing*, *51*, pp.41-59.

Fei, Y., Gao, K. and Li, W.Q., 2018. Artificial neural network algorithm model as powerful tool to predict acute lung injury following to severe acute pancreatitis. *Pancreatology*, *18*(8), pp.892-899.

Fernández-Delgado, M., Cernadas, E., Barro, S. and Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems?. *The Journal of Machine Learning Research*, *15*(1), pp.3133-3181.

Goorbergh, R.V.D., van Smeden, M., Timmerman, D. and Van Calster, B., 2022. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *arXiv:2202.09101*. Preprint. Available from: https://doi.org/10.48550/arXiv.2202.09101 [Accessed 22nd April 2022]

Harrell, F.E., 2015. *Regression Modeling Strategies*. New York: Springer.

Jakkula, V., 2006. Tutorial on support vector machine (SVM). *School of EECS, Washington State University, 37*(2), p.3.

Jiang, T., Gradus, J.L. and Rosellini, A.J., 2020. Supervised machine learning: a brief primer. *Behavior Therapy*, *51*(5), pp.675-687.

Khan, I., Zhang, X., Rehman, M. and Ali, R., 2020. A literature survey and empirical study of meta-learning for classifier selection. *IEEE Access*, *8*, pp.10262-10281.

Karaolis, M.A., Moutiris, J.A., Hadjipanayi, D. and Pattichis, C.S., 2010. Assessment of the risk factors of coronary heart events based on data mining with decision trees. *IEEE Transactions on Information Technology in Biomedicine*, *14*(3), pp.559-566.

Kotsiantis, S.B., Zaharakis, I. and Pintelas, P., 2007. Supervised machine learning: A review of classification techniques. *Emerging Artificial Intelligence Applications in Computer Engineering*, *160*(1), pp.3-24.

Lisboa, P.J. and Taktak, A.F., 2006. The use of artificial neural networks in decision support in cancer: a systematic review. *Neural networks*, *19*(4), pp.408-415.

Luo, G., 2016. A review of automatic selection methods for machine learning algorithms and hyper-parameter values. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *5*(1), pp.1-16.

Mitchell, T.M., 1997. *Machine Learning.* New York: McGraw Hill.

Nikookar, E. and Naderi, E., 2018. Hybrid ensemble framework for heart disease detection and prediction. *IJACSA: International Journal of Advanced Computer Science and Applications*, *9*(5), pp.243-248.

North, R.A., McCowan, L.M., Dekker, G.A., Poston, L., Chan, E.H., Stewart, A.W., Black, M.A., Taylor, R.S., Walker, J.J., Baker, P.N. and Kenny, L.C., 2011. Clinical risk prediction for pre-eclampsia in nulliparous women: development of model in international prospective cohort. *BMJ: British Medical Journal* [Online], *342*. Available from: https://doi.org/10.1136/bmj.d1875 [Accessed 20th April 2022]

Orru, G., Pettersson-Yeo, W., Marquand, A.F., Sartori, G. and Mechelli, A., 2012. Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. *Neuroscience & Biobehavioral Reviews*, *36*(4), pp.1140-1152.

Osborne, J.W., 2015. *Best Practices in Logistic Regression.* London: Sage Publications.

Pavlou, M., Ambler, G., Seaman, S.R., Guttmann, O., Elliott, P., King, M. and Omar, R.Z., 2015. How to develop a more accurate risk prediction model when there are few events. *BMJ: British Medical Journal* [Online], *351*. Available from: https://doi.org/10.1136/bmj.h3868 [Accessed 20th April 2022]

Probst, P., Boulesteix, A.L. and Bischl, B., 2019a. Tunability: importance of hyperparameters of machine learning algorithms. *The Journal of Machine Learning Research*, *20*(1), pp.1934-1965.

Probst, P., Wright, M.N. and Boulesteix, A.L., 2019b. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery, 9*(3), p.1301.

Rajkomar, A., Oren, E., Chen, K., Dai, A.M., Hajaj, N., Hardt, M., Liu, P.J., Liu, X., Marcus, J., Sun, M. and Sundberg, P., 2018. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, *1*(1), pp.1-10.

Reddy, K.V.V., Elamvazuthi, I., Abd Aziz, A., Paramasivam, S. and Chua, H.N., 2021, July. Heart Disease Risk Prediction using Machine Learning with Principal Component Analysis. In *2020 8th International Conference on Intelligent and Advanced Systems (ICIAS)* (pp. 1-6). IEEE.

Royston, P., Moons, K.G., Altman, D.G. and Vergouwe, Y., 2009. Prognosis and prognostic research: developing a prognostic model. *BMJ: British Medical Journal* [Online], *338*. Available from: https://doi.org/10.1136/bmj.b604 [Accessed 20th April 2022]

Saeb, S., Lonini, L., Jayaraman, A., Mohr, D.C. and Kording, K.P., 2017. The need to approximate the use-case in clinical machine learning. *Gigascience*, *6*(5), pp.1-9.

Sewell, M., 2008. Ensemble learning. *RN*[Online], *11*(02). Available from: http://www.machine-learning.martinsewell.com/ensembles/ensemble-learning.pdf [Accessed 22nd April 2022]

Shipe, M.E., Deppen, S.A., Farjah, F. and Grogan, E.L., 2019. Developing prediction models for clinical use using logistic regression: an overview. *Journal of Thoracic Disease* [Online], *11*(Suppl 4). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6465431/ [Accessed 20th April 2022]

Steyerberg, E.W. and Vergouwe, Y., 2014. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European Heart Journal*, *35*(29), pp.1925-1931.

Steyerberg, E.W., 2019. *Clinical prediction models*. Cham: Springer International Publishing.

Steyerberg, E.W., Eijkemans, M.J. and Habbema, J.D.F., 1999. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *Journal of Clinical Epidemiology*, *52*(10), pp.935-942.

Steyerberg, E.W., Vickers, A.J., Cook, N.R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M.J. and Kattan, M.W., 2010. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, *21*(1), p.128.

Strobl, C., Malley, J. and Tutz, G., 2009. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*(4), p.323.

Su, X., Xu, Y., Tan, Z., Wang, X., Yang, P., Su, Y., Jiang, Y., Qin, S. and Shang, L., 2020. Prediction for cardiovascular diseases based on laboratory data: an analysis of random forest model. *Journal of Clinical Laboratory Analysis* [Online], *34*(9). Available from: https://doi.org/10.1002/jcla.23421 [Accessed 22nd April 2022]

Thottakkara, P., Ozrazgat-Baslanti, T., Hupf, B.B., Rashidi, P., Pardalos, P., Momcilovic, P. and Bihorac, A., 2016. Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications. *PloS one* [Online], *11*(5). Available from: https://doi.org/10.1371/journal.pone.0155705 [Accessed 20th April 2022]

Van Calster, B., McLernon, D.J., Van Smeden, M., Wynants, L. and Steyerberg, E.W., 2019. Calibration: the Achilles heel of predictive analytics. *BMC Medicine*, *17*(1), pp.1-7.

Van Calster, B., Nieboer, D., Vergouwe, Y., De Cock, B., Pencina, M.J. and Steyerberg, E.W., 2016. A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of Clinical Epidemiology*, *74*, pp.167-176.

Van Calster, B., van Smeden, M., De Cock, B. and Steyerberg, E.W., 2020. Regression shrinkage methods for clinical prediction models do not guarantee improved performance: simulation study. *Statistical Methods in Medical Research*, *29*(11), pp.3166-3178.

Vickers, A.J. and Elkin, E.B., 2006. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, *26*(6), pp.565-574.

Wilson, P.W., D'Agostino, R.B., Levy, D., Belanger, A.M., Silbershatz, H. and Kannel, W.B., 1998. Prediction of coronary heart disease using risk factor categories. *Circulation*, *97*(18), pp.1837-1847.

Yang, L. and Shami, A., 2020. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, pp.295-316.