UNIVERSITEIT•STELLENBOSCH•UNIVERSITY
jou kennisvennoot • your knowledge partner

Research proposal:

Development and optimisation of a model tree forest induction algorithm.

I prefer short, to-the-point, titles :
Model Tree Forests

Werner Van der Merwe
2007223

Supervisor: Prof. A.P. Engelbrecht

MEng Inudstrial Research

16 May 2020

Contents

List of Figures

Chapter 1

Introduction

Machine learning, regarded as a subset of artificial intelligence, is quite an old research field. Many machine learning algorithms used today were first conceptualised in the previous decade. It was not until recent years that computers have been able to produce adequate computing power to allow for the large scale application of these algorithms. Furthermore, the available computing power continues to increase each year. With the increase in computing power comes the increase of interest in machine learning research. More complex models can now be developed that utilise this increased computing power to better capture patterns that are present in data.

Together with the increase of computing power, is the increase of data that is available to build machine learning models from. Both the volume and complexity of gathered data has increased. The ability to properly utilise the available big data, together with complex machine learning models, for problem-solving is to the benefit of the user.

One of the machine learning algorithms often used today is the decision tree model. A decision tree recursively organises information through a series of binary questions. A decision tree can be used to solve both classification and regression problems. Its success is attributed to its ability to handle large amounts of data with faster computation speeds than alternative machine learning models. At first, decision trees were used more often in classification problems than regression as they produced better results in classification tasks. Early research focused on improving decision trees for classification purposes.

Model trees would later be developed as a type of decision tree that yielded increased performance on regression problems. Model trees addressed many of the shortcomings normal decisions trees exhibited in regression applications. The distinguishing factor of a model tree from a decision tree is its ability to predict continuous numerical values. When compared to other machine learning regression models, such as neural networks, model trees produced similar results, but its computation time was significantly less [1].

Ensemble strategies were applied to decision trees to further increase its accuracy and robustness. Research showed that the use of decision trees in

(1) What information? Be careful not to mix terms, e.g.
data vs information
algorithm vs model

(2) What is a binary question? Decision nodes in a decision tree can have more than two outcomes, depending on whether that decision is on a categorical feature.

(3) Not both, but either, depending on the type of decision tree.

(4) For statements like this, you have to provide references to support.

(5) What is a normal decision tree?
You need to very early on make clear that there are the following main types of decision trees: classification, regression, model trees

(6) What are these shortcomings?

(7) You can not refer to decision trees in general here, because classification trees have not been applied to regression problems.

(8) But, a model tree is a decision tree.

(9) This is also done by regression trees.

an ensemble vastly increased its generalisation capabilities [2]. However, little research on applying ensemble strategies to model trees have been published. The use of model trees in an ensemble, referred to as a model tree forest, could ~~allow for the model to produce~~ results better than other regression techniques, whilst being computationally efficient.

Regression models are often used by organisations to gain advantages over competitors. By utilising the large amounts of data available to them, organisations can deduce better insights regarding their work and improve their efficiency. Regression models are used in various industries within South Africa, including manufacturing, banking, retail, transportation and health care. Therefore, a model tree forest has a wide spectrum of applications if it were to be successfully developed within this research.

A short discussion on ensemble approaches for classification trees, regression trees to be included, in order to support your following statement.

(10) You need a stronger motivation for using model trees: problems with many features, huge number of instances; to scale model trees to data rich problems.

You have also not provided sufficient motivation for why looking at model trees instead of other regression algorithms such as neural networks, support vector regression, etc.

(11) Also true for classification problems. Think carefully about what you want to say here

(12) Careful: model tree forests do already exist. What will you offer that existing ones do not.

Chapter 2

Theory and literature analysis

This chapter presents the background on trees to motivate why this research aims to use an ensemble of model trees to produce improved results over alternative machine learning methods.

2.1 The decision tree

The early application of decision trees in machine learning started with simple classification or regression tasks, trained on known datasets, to model the relationships between the input and output variables.

- Classification models were responsible for predicting a categorical output.
- Regression models produced a numerical output.

A decision tree takes the appearance of a tree to organise information in a recursive manner. It makes decisions through a series of binary questions that resembles branches. In machine learning, decision trees represent an algorithm containing if-else statements to model patterns in data. Decision trees were inspired by the *divide-and-conquer* paradigm, which entails recursively breaking an intricate problem down into smaller sub-problems which each can be solved more simply. These models would not be theoretically refined until computing technologies grew sufficient [3].

Before the conception of decision trees, many of the alternative statistical methods were developed with small datasets in mind, on the assumption of the data being homogeneous, leading to algorithms that focused heavily on only a few parameters to model the influence of a wide range of factors. Applying this to larger datasets further assuming that the data retained its homogeneous structure, would prove to be flawed. datasets are not only subject to being large, but complex as well. This complexity influenced by several factors: high dimensionality; mixtures of data types; non-parametric distribution and non-homogeneity. "The curse of dimensionality", often apparent in these complex

(13) What is a known data set vs an unknown data set?

(14) The itemization does not really flow from the preceeding sentences.

(15) They still are. Careful about your word choices, e.g. responsible.

Classification trees are used to ----

Regression trees are used to ---

(16) Terminology: output vs target.

(17) Biological tree? A tree data structure.

(18) a hierarchical representation of tests/conditions on input features.

(19) Avoid pronouns.

(20) The "questions", i.e. tests are in the nodes (non-terminal nodes); the branches represent the different outcomes of the tests.

(21) The decision tree structure (which is what you refer to here) does not represent an algorithm. In the context of decision trees, there are two algorithms:

- the induction algorithm
- the prediction algorithm.

(22) Not patterns, but to form decision boundaries.

(23) What do you mean by theoretical refinement?

(24) Hereby you imply that decision trees are statistical methods.

(25) Sentence is too long, conveying too many facts.

(26) Avoid -ing words: Application of ... to ...

(27) Why flawed. Carefully think about what you are writing.

(28) Data complexity is influenced by a number of ...: namely ...

(29) I will rather avoid to repetitively refer to data sets as being "complex".

datasets, states that the higher the dimensionality, the higher the variance of the data points are. With the progression of computing technologies, datasets with these attributes are no longer a rare occurrence [3].

To combat this, one could reduce the dimensionality of the data. However, the accompanying drawbacks are unfavourable. This led to the demand for a method in which the most important features of the data are accentuated, the background noise disregarded, and the conclusions interpretable to the analyst. This brought rise to the refinement of decision tree learning. The defining characteristic of decision trees being that it does not only produce satisfactory results, it also gives the user thoughtful information and insight into the data it is being applied on. In classification studies, for example, this helps uncover the predictive nature of the issue and helps the user better understand how specific features contribute to the outcome that is being observed [3].

A simplistic binary tree is portrayed in Figure 2.1, capable of classifying an individual's sex based on two inputs, their weight and height. Nodes X_1 and X_3 each house a split condition. Node X_1 is referred to as a *root* node because it is the starting node of the decision tree. Nodes X_2, X_4 and X_5 contain the label of the output class. These nodes are referred to as *leaf* nodes, due to them terminating all possible *paths* (sequence of nodes) that can be followed within the tree, starting at the root node. It is important to note that decisions trees are always constructed in such a manner that when combining the regions each path denotes, the entirety of the instance space is covered [1]. The *size* of a decision tree refers to the amount of nodes it consists of. Finally, *depth* is the number of nodes in the longest path the tree contains. Therefore a tree is regarded as being *shallow* if its depth is small.

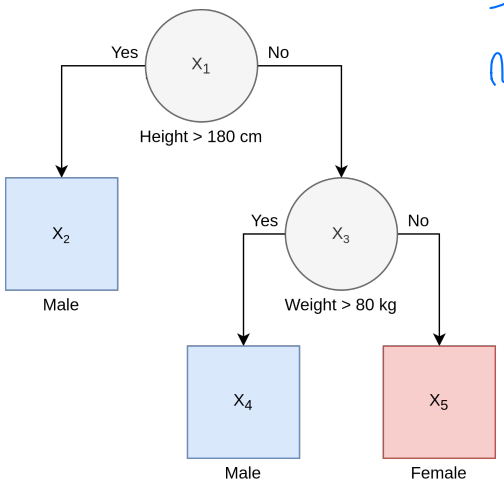


Figure 2.1: A hypothetical two-class tree-structured classifier.

In the application of regression, little changes. The tree-structured predictor is still partitioned into subsets via a series of split decisions. Instead

(30) Many other factors influence "complexity":

- missing values
- noise
- outliers
- skew class distributions for classification problems.
- multiple target features / classes.

(31) ... the dimensionality of ... can be ...

(32) What do you refer to? Number of features and/or number of instances?

(33) Be careful about using one reference.

(34) Does not really flow on the statement in the previous sentence.

(35) Not learning, but induction.

(36) What refinement?

(37) What is a classification study? You mean: for classification problems ...

(38) Please go through the entire document, and rewrite not to use pronouns.

(39) What issue?

(40) Be careful of overstating something.

(41) Gender, please. No sex.

(42) A condition or test on the feature

(43) Within has a too general implication. From root to a leaf

(44) Carefully look at what you are saying!

the tree is partitioned into subsets; nope, the data set at a non-terminal node is partitioned.

(45) split decisions? Need to more clearly define/describe. The non-terminal nodes refer to tests/conditions on input features, which split the current data set into subsets, one for each outcome of the test.

of of predicting which class an observation belongs to, the decision tree now predicts the numerical value associated with a class.

Fundamentally, the regression tree predicts the target numerical value as a dependent variable, based on a multitude of independent variables. The input variables being either discrete, continuous or a even mixture of the two. Through piecewise approximation is the regression tree able to model the relationship between the input and output variables. The leaf nodes of a standard regression tree each have a constant output value [3]. Figure 2.2 illustrates how a regression tree models a sine wave.

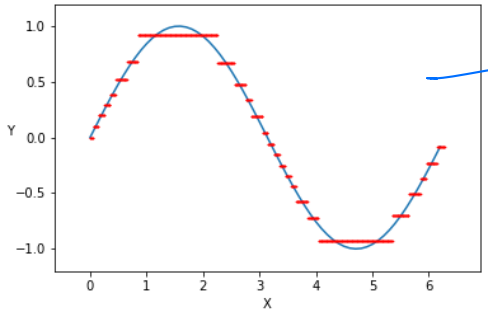


Figure 2.2: A regression tree's piecewise approximation (red) of a continuous input function (blue).

Initial application of decision trees showed that arguably the biggest shortcoming was a product of the high variance that trees exhibited [3]. Trees would often deliver satisfactory results on training sets, but generalised poorly on test sets. During its growing process, trees would continuously optimise the classification boundaries adding additional decision nodes, capturing noise in the training set. This meant that trees were prone to overfitting and would require something to remedy this.

2.2 Decision tree ensembles

One method of minimising variance is that of ensemble learning. The premise of ensemble learning is to develop multiple models in unison to provide a better result than an individual model is capable of. The use of several weak learners together proved to form a strong learner. Taking the result of a decision tree ensemble would prove to lower the variance whilst keeping the bias error low. This meant that tree ensembles would become a very competitive solution in both classification and regression problems, with various differing implementations developed.

One of the simplest tree ensemble methods is bootstrap aggregation, referred to as *bagging* [4]. Several decision trees are induced in parallel, following the aforementioned procedure, each on a subset of the training data. Each subset is determined through a random selection process. The randomness off of which each tree is induced is what brings the variance of the collective model down. For regression applied bagging, the final prediction of the ensemble is derived through averaging the predictions of each individual tree within the ensemble. In the case of classification, the majority vote is taken instead of the average predicted value [5].

Building on the success achieved by bagging, the *random forest* technique was developed as an extension over it; first conceptualised by Breiman [2]. In a random forest, each tree is once again trained on a randomly selected subset of the data. The difference random forest incorporates is a change to the induction of each individual tree. Input features are randomly selected at each node within the tree to determine the best split for that node, evaluated only on those randomly selected features. This decorrelates the trees within the ensemble model, decreasing the model's variance error. Random forest has the benefit of performing well on higher dimensionality data, due to its dimension reduction-like training technique. It is important to note that the individual trees are left unpruned in a random forest. This means that each tree is *fully grown* and retains the low bias trees exhibit. Random forests were shown to outperform many competing classifiers whilst being robust to overfitting [2].

In contrast to bagging is the *boosting* technique. Boosting employs a sequential learning strategy; starting with a simplistic, usually shallow tree and building on it with the goal of improving the net error. The idea behind boosting is to develop weak learners, one at a time, with each sequential learner focused on the patterns the previous learner was unable to adequately model. This is achieved through assigning weights to each observation. Instances which proved to have high prediction error were given larger weights so that the next learner would be more likely to correctly model it. Finally, combining these learners results in a model that together outperforms its individual parts. Therefore, in bagging each observed feature has an equal probability of being selected, whereas with boosting this selection probability is influenced by the calculated weights [6].

2.3 Model trees

Model trees were developed as an extension over regression trees to improve its performance. It was clear that the constant output values within the leave nodes were subject to underfitting, as shown in Figure 2.2. A model tree differs from a regression tree as its leave nodes are not limited to having a constant output value. Leaves can incorporate any multivariate model to better capture

(47) For the thesis, a more detailed discussion will be required.

(48) I will start by defining what a model tree is.

the nonlinear sample observations. Therefore, model trees are often used to model a continuous target output. Model trees have the advantage of being much smaller than regression trees, whilst also providing a more accurate fit of the training data [7].

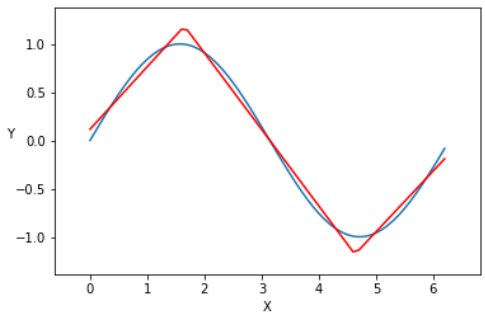


Figure 2.3: A shallow model tree's piecewise approximation (red) of a continuous input function (blue).

To illustrate this, we present the typical fit of a shallow model tree on the same data present in Figure 2.2, now shown in Figure 2.3.

The M5 algorithm, which utilises piecewise linear regression [1], was one of the first conceptualised model trees. During the development of the M5 algorithm, the comparing statistical technique used in performing regression tasks on nonlinear data was MARS (multivariate adaptive regression spline) [7]. MARS is similar to M5 as it is also non-parametric and incorporates piecewise linear approximation. Subsequently, M5 and MARS produced similar accuracy results, but what set M5 apart from MARS was the computational requirement. The computational requirements of MARS grew aggressively with an increase in dimensionality, severely limiting its applicability. M5 was able to handle tasks with up to hundreds of input features, whereas MARS would struggle past no more than twenty [7].

2.4 Shortcomings of M5

The M5 algorithm has been shown to perform inadequately in comparison to other solutions on datasets that exhibit one of either two specific traits. The one documented factor limiting its performance is the model tree's inability to fully capture the patterns present in highly nonlinear data. The other factor is the algorithm's tendency to overfit data, which is an ongoing problem with all decision trees.

2.4.1 Noisy datasets

The M5 model tree's susceptibility to overfitting can be attributed to its complexity. It has been shown that a complex model can be easily disturbed by noise in datasets [8] and that exclusively pruning a model with the intent of increasing its resilience to overfitting could prove to be ineffective [9]. Therefore, additional measures should be taken to effectively decrease the variance of a model.

D. Aleksovski researched ensemble strategies to help his models tolerate noise better, publishing his findings in 2015 [9]. Aleksovski's approach to growing a model tree within an ensemble included three key elements: The split parameters are randomised; non-leaf nodes had their conditions *fuzzified* and trees within the ensemble were pruned individually. Fuzzification refers to Aleksovski's smoothing process to prevent discontinuities each split of the model tree introduced. Afterwards, the ensemble is also pruned as a whole, removing the trees that have the highest contribution to the output error. Aleksovski's results showed that the use of ensembled model trees allows for maintaining accurate results even with the injection of large amounts of noise into the sample data [9].

2.4.2 Nonlinear datasets

A study published in 2016 [10] hypothesised that the reason for the M5 model tree's inferior results on highly nonlinear data is its piecewise linear functions' limited capability to fit the training data. In this study, only six quantitative water quality parameters were used in predicting the monthly chemical oxygen demand of a river. The relationship between these parameters being highly nonlinear. A MARS model was able to achieve, on average, accuracy 19.1% higher than the competing M5 model tree. It is important to note that the M5 model trees were designed with large datasets in mind. Making it the preferred choice given there are a higher number of input parameters [7].

Quinlan recommends researching the application of nonlinear models at the leave nodes [7], improving the ability of the model to capture these highly nonlinear patterns present in complex datasets. This would also allow for a model tree to be grown smaller, without negatively influencing its predictive capabilities. Growing a smaller sized tree is likely to result in an increased bias in the model, making it less sensitive to noise and less prone to overfitting.

However, careful consideration should be given to the increase in computation that comes with the implementation of complex nonlinear models. A model has to justify its increased computation time with a clear and substantial improvement in its accuracy. It is also worth noting that exclusively increasing the complexity of a model does not guarantee an improvement in its performance. We hypothesise these as contributing factors to the lack of abundant research on model trees incorporating nonlinear models.

2.5 Nonlinear solutions

Modelling the relationship between independent variables and a dependent variable as an n^{th} degree polynomial equation allows a model to better capture the nonlinear patterns present in sample observations. Although the regression function is nonlinear, estimating the parameters is a linear problem [11]. The value of the dependent variable y can be expressed through the n^{th} order polynomial equation of the independent variable x, as:

This is only univariate.

$$y = \alpha_0 + \alpha_1 x + ... + \alpha_{n-1} x^{n-1} + \alpha_n x^n = \sum_{j=0}^{n} \alpha_i x^j \tag{2.1}$$

The parameters of a regression problem are the coefficients $\alpha_0, ..., \alpha_n$ of an arbitrary polynomial function, such as Equation (2.1).

For multi-variate polynomials.

Through the use of a genetic algorithm in 2007, Potgieter and Engelbrecht estimated the parameters resulting in an optimal polynomial structure on two criteria, the shortest polynomial with the best possible approximation. The genetic algorithm proposed in this study produced comparable results to that of a neural network, but at a significantly faster approach [11], attributing its success to the incorporation of specialised mutation and crossover operators in the algorithm.

2.6 Proposed model tree forest

As mentioned, Aleksovski's results showed that the use of ensembled model trees improved the robustness of his models [9]. Aleksovski employed the M5 algorithm to achieve this, limiting the capability of his model to adequately capture highly nonlinear patterns within data. It would be expected that using model trees, that incorporate polynomial functions, in an ensemble would be impractical due to the increased computation time. However, the research of Potgieter and Engelbrecht showed that model trees with polynomial functions can be induced with decreased computation times using a genetic algorithm. [1].

Successfully ensembling model trees that incorporate polynomial functions could mitigate the aforementioned shortcomings of trees and provide a machine learning regression solution capable of outperforming alternative methods on large and complex datasets. As is the case with decision trees, there are various strategies one could follow to ensemble model trees, such as bagging and boosting. Extensive research that evaluates these strategies is needed.

Ok, but this section does not provide a proposed model tree forest.

Chapter 3

Research problem statement and research objectives

3.1 Problem Statement

Review of previous literature shows that model trees produce comparable accuracy to other regression techniques. Decision trees have also shown to be more robust and accurate when used in ensemble methods. Little to no research has been done on the use of models trees in an ensemble. Can the use of an ensemble of model trees produce improved accuracy results when compared to a single model tree? Furthermore, what are the optimal parameters for the ensemble?

3.2 Research Objectives

The primary objective of this research is to develop an algorithm capable of inducing a model tree forest from large training sets where the target value to be predicted is a numerical value. This study aims to investigate the following aspects of the model tree forest:

1. The number of model trees in the ensemble.
2. Approaches to constructing model trees on a subset of the input features.
3. Different methods of fusing the decisions of the individual model trees.
4. How the ensemble as a whole can be optimised post-induction.
5. The process in which data is subsampled for each induced model tree.

To properly evaluate the outcome of these objectives, the model tree forest will first be trained on benchmark databases. These databases are openly available from various machine learning repositories.

Once the aforementioned objectives have been satisfied, the final objective of this research is to find a suitable problem to which the model tree forest can be applied and its performance compared to alternative machine learning solutions.

3.3 Importance of the research problem

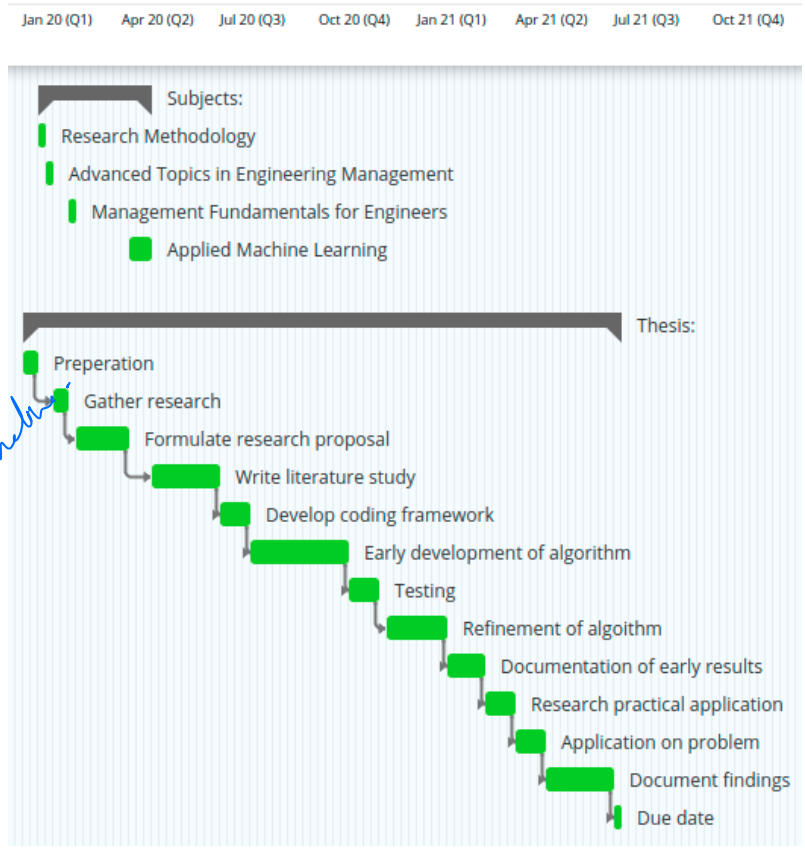
Organisations gain competitiveness through, among other things, efficiency. Therefore it is to the benefit of the organisation to ensure that their decisions are made with this in mind. To help an organisation make the decisions that will ensure efficiency is optimised, regression analysis is employed on various aspects of the business. It helps the organisation interpret data better. This data often being large and complex.

The proposed model tree forest specialises in regression application where the target variable takes on a continuous numerical value. If it were to satisfy the research objectives, the model tree forest can outperform other regression techniques to produce better predictions on the regression problems organisations have. As a result, the organisation can improve its efficiency and consequently its competitiveness.

Chapter 4

Proposed research strategy and schedule

The proposed schedule is shown in Figure 4.1. The research approach starts with adequate research on existing machine learning solutions, through both self-research and academic modules. This is followed by the development of the algorithm itself and the optimisation of it, based on the research objectives.



This does not really discuss a research strategy, mainly gives a timeline

Figure 4.1: A Gantt chart of the proposed research schedule.

Finally, the focus will be directed at solving/optimising a real-world problem through the application of the proposed model tree forest and comparison to alternative machine learning methods. Figure 4.2 breaks down the strategy followed to successfully formulate and apply a model tree forest induction algorithm.

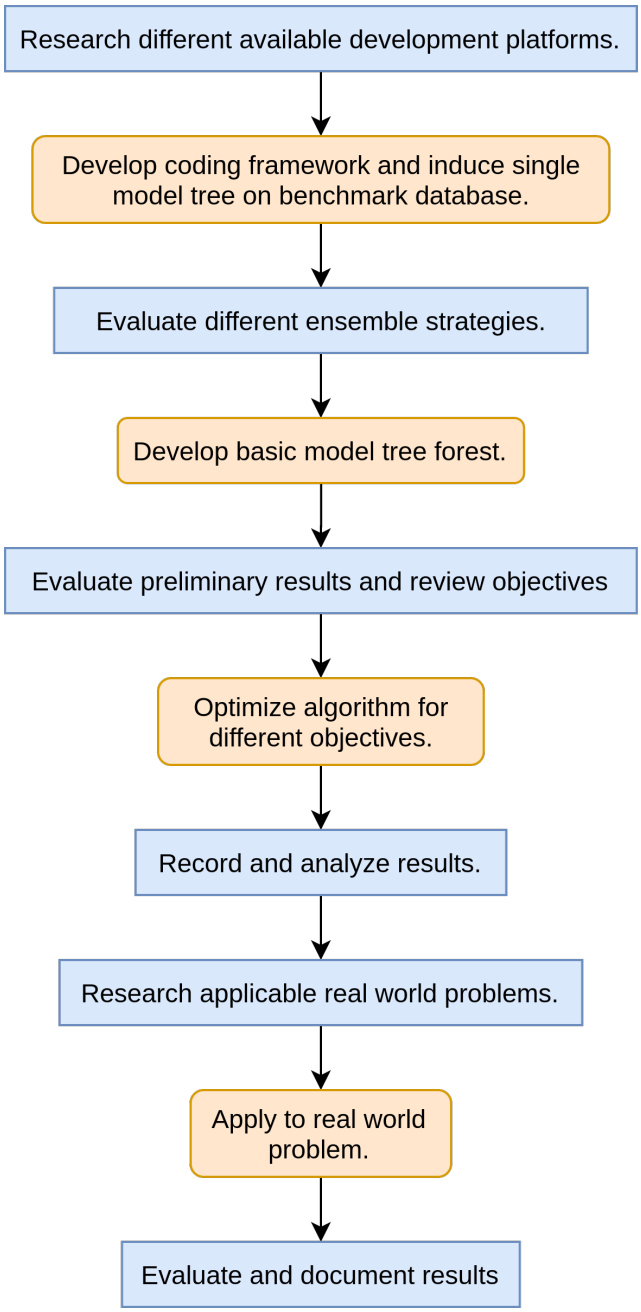


Figure 4.2: Research strategy presented in a flow diagram.

Chapter 5

Conclusion

The literature reviewed in this research proposal shows that ensemble strategies can be applied to trees to increase both its robustness and accuracy, two aspects that are essential to the success of a machine learning model. However, there is room for further improvement as the M5 model tree used in these ensemble models does exhibit certain flaws.

By building on the research of Engelbrecht and Potgieter that showed how single model trees could be induced with decreased computation time [1], this research aims to produce a model tree forest capable of modeling a continuous target value using polynomial functions at the leaf nodes of the trees in the ensemble, distinguishing this research from Aleksovski [9]. Furthermore, this research will conclude which of several ensemble strategies produce the best results.

Bibliography

[1] G. Potgieter, "Mining continuous classes using evolutionary computing," Ph.D. dissertation, University of Pretoria, 2006.

[2] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[3] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Monterey, CA: Wadsworth and Brooks, 1984.

[4] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[5] A. Nagpal, "Decision Tree Ensebmles- Bagging and Boosting," 2017. [Online]. Available: https://towardsdatascience.com/decision-tree-ensembles-bagging-and-boosting-266a8ba60fd9. [Accessed: 23- Mar- 2020].

[6] G. Moisen, "Classification and regression trees," *In: Jørgensen, Sven Erik; Fath, Brian D.(Editor-in-Chief). Encyclopedia of Ecology, volume 1. Oxford, UK: Elsevier. p. 582-588.*, pp. 582–588, 2008.

[7] J. R. Quinlan *et al.*, "Learning with continuous classes," in *5th Australian joint conference on artificial intelligence*, vol. 92. World Scientific, 1992, pp. 343–348.

[8] S. Singh, "Understanding the Bias-Variance Tradeoff," 2018. [Online]. Available: https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229. [Accessed: 22- Mar- 2020].

[9] D. Aleksovski, J. Kocijan, and S. Džeroski, "Model-tree ensembles for noise-tolerant system identification," *Advanced Engineering Informatics*, vol. 29, no. 1, pp. 1–15, 2015.

[10] O. Kisi and K. S. Parmar, "Application of least square support vector machine and multivariate adaptive regression spline models in long term prediction of river water pollution," *Journal of Hydrology*, vol. 534, pp. 104–112, 2016.

[11] G. Potgieter and A. P. Engelbrecht, "Genetic algorithms for the structural optimisation of learned polynomial expressions," *Applied Mathematics and Computation*, vol. 186, no. 2, pp. 1441–1466, 2007.