TILBURG UNIVERSITY

PREDICTING INDIVIDUAL PERFORMANCE IN ARTIFICIAL LANGUAGE LEARNING

KIKI PEETERS

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2059705

COMMITTEE

dr. Drew Hendrickson
dr. Murat Kirtay

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

January 13, 2023

ACKNOWLEDGMENTS

Writing this thesis was a real challenge for me, and I couldn't haven done it without the support, guidance and motivation from the people around me. First of all I would like Dr. Drew Hendrickson for his guidance through this process, the weekly meetings were not only useful but also helped to keep me motivated. I would also like to thank my family in general, and in particular my fiancé Eric for always listening to me and supporting me all the way through. Lastly, I would like to thank my peer and friend Rowanne, for always being there when I needed advice or emotional support.

PREDICTING INDIVIDUAL PERFORMANCE IN ARTIFICIAL LANGUAGE LEARNING

KIKI PEETERS

CONTENTS

Abstract

In artificial language learning, it is important to understand how individual performance can be predicted so that effective language learning methods can be designed. This study aimed to address this gap in the literature by comparing the performance of different models, namely the support vector machine, multilayer perceptron, and decision tree ensemble, on a data set collected by Hendrickson and Perfors (2018) in a study on cross-situational learning in a Zipfian environment. The data set consisted of individual word-object pairs, and the models were trained to predict the accuracy of these pairs. The accuracy is either 0 or 1, making the task for this master thesis a binary classification one. The decision tree ensemble was found to be the best performing model, outperforming the baseline model logistic regression, support vector machine and the multilayer perceptron with an accuracy on the test set of 0.686. In addition, no difference in performance was detected between age categories, youth (17 - 24), adults (25 - 64) and seniors (65+) using the decision tree ensemble. Furthermore, the decision tree ensemble was also used to identify the most important features that influence individual performance in artificial language learning. The features were selected based on the mean accuracy decrease. These features were found to be age, the type of experiment, and response time. In conclusion, this study has demonstrated that it is possible to predict individual performance in artificial language learning using a decision tree ensemble. The results of this study can be used to design more effective language learning methods by taking into account the individual factors that influence performance. Total word count thesis: 7423

1 DATA SOURCE, CODE, ETHICS STATEMENT

Work on this thesis did not involve collecting data from human participants or animals. The original owner of the data and code used in this thesis retains ownership of the data and code during and after the completion of this thesis. The author of this thesis acknowledges that they do not have any legal claim to this data or code. A contract between the owner of the data, Andrew Hendrickson, and the author of this thesis, Kiki Peeters, was signed to protect the data rights of the owner and the author. The code used in this thesis is will be made publicly available via https://github.com/kikimgp/Thesis

2 INTRODUCTION

The research goal of this thesis is to find out if it's possible to predict individual performance in artificial language learning accurately. This is done by comparing the performance of different models based on their predictive power.

2.1 *Problem statement*

Artificial languages are languages that are developed for a certain task created up by humans. More often than not, these languages are made to learn more about natural language learning, especially second language learning. Using artificial languages for these types of tasks makes it controllable for the researcher. The more that is learned about artificial language learning, the more that can be learned about natural language learning. Little to no research has been focused on predicting individual performances in artificial language learning. The goal of this master thesis is to close that current gap. Using the knowledge that will be gained for this research will help further understand natural language learning, which is beneficial for several different groups that will be discussed in the next section.

The task of this master thesis is to predict the individual accuracy between object-word pairs. This is a binary classification task because the accuracy, in this case, is either zero or one. Zero if the subject did not select the right word and one if the subject did select the right word. Because this thesis also uses the evaluation metric accuracy, it is important to note that these are not the same.

2.2 *Scientific and societal relevance*

When looking at the language domain in general, it is a broadly researched topic. However, artificial language learning has been less researched. These types of languages are often used in studies about natural language for more control but are less researched on their own. The research conducted in this thesis will help close the gap in the current literature.

Because artificial languages are often used to learn about how natural languages are learned, the knowledge gained from this research can be used to improve natural language learning, in particular second language learning. This can be used by schools to create better learning techniques when teaching languages in class. But also for governments to be able to teach languages more efficiently in integration courses. In this day and age, more people are using apps and online courses to learn second languages. The companies behind these apps and websites, for instance, Duolingo, can also create better learning techniques and plans, which will make it easier for their consumers to learn a language.

2.3 *Research questions*

Based on the sections mentioned above the following main research question is formulated:

To what extent is it possible to predict individual performance in artificial language learning?

RQ1.Which model leads to the best performance, when comparing Support Vector Machine, Decision Tree Ensembles and Multilayer perceptron?

The first research question will model three different models that were found to work in similar predicting tasks, the support vector machine, multi-layer perceptron and decision tree ensemble. Important to mention again that this is a binary classification task, trying to predict the accuracy, which is either 0 or 1, looking at individual word-object pairs. Because the word best can be interpreted in different ways, it is important to note that in this case, the word best means based on the evaluation metric accuracy of the test set, not to confuse with the accuracy mentioned above, the model with the highest accuracy will be the best model.

RQ2.To what extent is there a difference in performance between age groups, for the best performing model?
Prediction can be different for different groups within a data set. The literature about the relationship between age and language, specifically second language learning, showed that the older you are, the harder it can be to learn a new language. To find out if this relationship also existed between the ability to predict language learning and the age of an individual, the best performing model will be used to answer this question. The accuracy will be compared between three different age groups, youth (17 - 24), adults (25 - 64) and seniors (65+). The accuracy of the test set for each age group will be compared to determine the answer.

RQ3. Which features are most important to accurately predict individual performance in artificial language learning, for the best performing model?
Different features can impact the performance of the models in different ways. In this question, the features are compared based on their importance in predicting individual performance in artificial language learning. The answer to these questions will determine which features are most important in language learning and help for the development of (second) language learning techniques. The importance of the features will be based on the mean accuracy decrease. The features with the highest mean accuracy decrease are the most important ones.

3 LITERATURE REVIEW

The motivation behind this thesis is to contribute to the improvement of natural language learning techniques. This literature review will explore related topics, such as predicting individual learning performance and predicting performance in second language learning. Next, this literature review will cover the topic of feature importance, looking into the important features from studies performed in the past and how they could relate to this master thesis. Lastly, the difference in language learning for ages is discussed in relation to the second research question of this master thesis.

3.1 *Model comparison*

The different models will first be individually discussed, and after that the conclusion will summarise the differences between the studies and this master thesis.

3.1.1 *Decision tree ensemble*

A decision tree ensemble is an algorithm that combines the predictions of multiple decision trees to make a more accurate prediction. Decision trees are a type of tree-like model that uses a series of binary splits to classify data points into different classes.
In 2018 the language learning app Duolingo launched a machine learning challenge (*2018 Duolingo Shared Task on Second Language Acquisition Modeling (SLAM)*, 2018). The goal of this challenge was to predict future mistakes that app users would make when learning English, Spanish and French based on their history of mistakes. In total, 15 different teams participated in this challenge which led to very important findings that can be used for this master thesis. The data used to perform this task consisted of 2 million words (tokens) from answers from more than 6,000 Duolingo users over a 20-day course. Different machine and deep learning models were used to predict future mistakes. The two best performing models were the recurrent neural network and decision tree ensemble (Osika, Nilsson, Sydorchuk, Sahin, & Huss, 2018); (Xu, Chen, & Qin, 2018) ;(Rich, Osborn Popp, Halpern, Rothe, & Gureckis, 2018).The decision tree ensemble that performed best was the gradient boosted decision tree ensemble. As well as model comparison, the teams also investigated the effect of complex feature engineering and feature importance. Complex feature engineering turned out to be less important than the choice of model (Settles, Brust, Gustafson, Hagiwara, & Madnani, 2018), teams that used psychologically motivated features combined with less complex models did not perform as well as the more complex models with less complex features (Settles et al., 2018); (Xu et al., 2018). A big weakness of this research lies in the fact that it's possible that the data set was used to prevent the features from being useful when training the models. The question is if the models will perform better if the data set is more diverse and contested data collected over a long time period.

3.1.2 *Multilayer perceptron*

Another common model that is used in predicting performance in second language learning is a multilayer perceptron. This type of neural network

consists of multiple layers of interconnected nodes, allowing them to model complex relationships between input and output variables.
A study by Babić and Benčina (2017) used a multilayer perceptron for a slightly different but not less interesting task. They used this model to predict the reading comprehension ability in English for people for whom English is a foreign, meaning non-native, language. They were able to achieve an accuracy of 72%. However, the researchers stated that a big weakness and problem of their research was the lack of generalizability. In order to achieve further research should focus on a more varied and larger sample (Babić & Benčina, 2017). Another study by Widyahastuti and Tjhin (2017) focused on predicting the performance of students by comparing a linear regression model and a multilayer perceptron. They concluded that the multilayer perceptron outperformed the linear regression based on the accuracy of the test set.

3.1.3 *Support vector machine*

One of the most widely used algorithms for predicting individual performance in artificial language learning is support vector machine. Support vector machines are a type of linear classifier that uses a linear decision boundary to separate data points into different classes. Support vector machine models have been trained on a variety of features, including language aptitude, motivation, and prior knowledge of related languages. In 2022 a study by Arashpour et al. (2022) compared an artificial neural network and a support vector machine in predicting individual learning performance. This research consisted of two different tasks, a regression task and a classification task. To keep in line with the task of this master thesis, which is a binary classification, the focus is on the findings of the classification task. The goal of this research was to predict whether a student would pass or fail an exam. The support vector machine outperformed the artificial network in all measures for all data. The performance of both models was measured using the precision, recall, accuracy, F1-score, Fowlkes–Mallows index and Matthews Correlation Coefficient (Arashpour et al., 2022). One limitation of the study was the method used to assess engagement, which relied on analyzing the students' click stream data, given that the majority of instruction was conducted in an online setting. To improve the applicability of this measure in non-online contexts, alternative methods for determining engagement could be explored. Another study by Al-Shehri et al. (2017) also used a support vector machine to predict students' performance in the final examination. They based their finding on measuring the correlation coefficient. The support vector machine outperformed the K-Nearest neighbours algorithm with a correlation coefficient of 0.96. However, looking at the accuracy of the performance, the model performed less well, with accuracy varying between 52% and 82%, for the three different final score categories, which were bad, good and excellent. Realizing the difference between accuracy for these categories is the biggest weakness of this study.

3.1.4 *Conclusion: model comparison*

The models described above all predict well when predicting performance in a second language and other topics related to language learning. These models, support vector machine and decision tree ensemble, will be compared in this thesis to see if they are also able to predict individual performance when working with a data set consisting of information about artificial language learning. As mentioned before, the benefit of using artificial language when researching language is that there is more control for the researchers. All subjects have no prior knowledge about the language, which is often not the case in second language learning. Even the researchers mention multiple times that prior knowledge is an important predictor in their models. This thesis will investigate what happens when you use a data set containing artificial language learning information when using the models mentioned above. And what the effect is on their performance.

3.2 *Feature importance*

Studies that are mentioned previously all focus on different features that are important when conducting their studies, especially the ones which had high predictive power. These vary between studies. The Duolingo study mentioned that psychological features had less importance according to Settles et al. (2018) and Xu et al. (2018). Whilst other studies like Arashpour et al. (2022) and Babić and Benčina (2017) did focus on more psychological motivation features, like motivation and engagement, which did result in a strong predictive model (Arashpour et al., 2022);Al-Shehri et al. (2017). The data set of this master thesis has different types of features available, but all of these are less psychologically related and more focused on personal characteristics, the way of learning and word-related features. How these different features are related to the ability of the performance of the models will be discussed later on and will be answered in research question three. These results will be compared further to the studies from this literature review in the discussion.

3.3 *Difference in language learning for different age categories*

Generally speaking, different people learn in different ways, and humans tend to want to generalize that to certain groups. It is commonly observed that older people are less efficient when performing a certain task than younger people. When looking at second language learning, many researchers have noticed that age can affect the ability to learn a second language (Birdsong, 2018). Birdsong (2018) States that the ease with which a second language can be learned slows down older people get. Another study by Zhang (2022) analyzed the influence of age on second language acquisition. They also concluded that age is an importance factor with a high influence on second language learning. In this master thesis, the best performing model will be compared based on performance between three different age categories, youth (17 - 24), adults (25 - 64) and seniors (65+). This is to find out if there is a difference between the groups, which is expected based on the literature. When looking for feature importance, the age feature will also be analyzed to see if this confirms the study's findings that were presented previously.

3.4 *Literature review: conclusion*

In this Master Thesis, the models that were used in a different field and within natural language learning will be tested in predicting the performance of artificial language learning. We will learn whether or not the models also perform well when predicting the individual performance of data from an artificial language experiment. Next to that, feature importance seems to be a topic of discussion, which is why this master thesis will go further into investigating which features are important to predict the accuracy in individual learning of an artificial language. Lastly, the literature shows that age is an important factor of influence in language learning. The findings from the disparate group analysis and feature importance research will be used to compare if the findings from the literature review agrees with the findings from this master thesis.

4 METHOD

4.1 *Models*

In this chapter, the three models, support vector machine, multilayer perceptron and decision tree ensemble, will be discussed. The baseline model, logistic regression, will also be explained.

4.1.1 *Support vector machine*

Support vector machines are powerful and versatile machine learning algorithms that can be used for binary classification problems by finding the optimal hyperplane that separates data points into two classes. This is done based on a range of parameters such as the choice of kernel, gamma, and c (Cervantes, Garcia-Lamont, Rodríguez-Mazahua, & Lopez, 2020). One of the major strengths of the Support vector machine is the ability for the user to choose the kernel, which can help to prevent overfitting when combined with the correct c and gamma value (Han & Jiang, 2014). Additionally, Support vector machines are known to be memory efficient (Pisner & M.Schnyer, 2020).

The choice to use SVMs for this study is justified by their effectiveness in binary classification problems and their demonstrated ability to predict individual performance in related research (Arashpour et al., 2022). Support vector machines have been proven to work well in a wide range of applications. Additionally, the flexibility of the algorithm allows it to be adapted to different types of data and problems, making it a suitable choice for this study.

4.1.2 *Multilayer perceptron*

In the Duolingo challenge, one of the best performing models was the recurrent neural network (Osika et al., 2018) (Xu et al., 2018) (Rich et al., 2018). However, for the data set used in this master thesis, using the recurrent neural network is not totally impossible but is not the right fit for data set, due to missing recurrent features. However, there are many different other neural networks that can be used that work well on the data set that will be used. One of them is a multilayer perceptron, an addition of a forward neural network. The multilayer perceptron has three layers; input, hidden and output layers. This model is often used for prediction, for example, in the study about predicting reading comprehension ability (Babić & Benčina, 2017) and another study by Widyahastuti and Tjhin (2017) focused on predicting the performance of students by comparing a linear regression model and a multilayer perceptron. For binary classification, the model defines a linear decision boundary. Similar to the support vector machine, it finds the best hyperplane that minimizes the distance between misclassified points and the decision boundary. It uses stochastic gradient descent to do that. Lastly, it uses an activation function which is initially the sigmoid function, to decide whether a neuron will fire or not. A multilayer perceptron has different hidden layers that have different neurons stacked together. Backpropagation allows the model to iteratively adjust the weights in the network, to minimize the cost function.

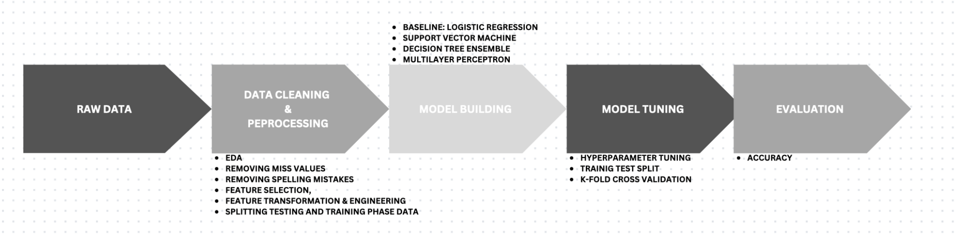

Figure 1: Pipeline. Source: the author's illustration.

4.1.3 *Decision tree ensemble*

A decision tree ensemble is a model which combines several decision trees to create a better performing model. There are different techniques that can perform an ensemble of decision trees, bagging and boosting (González, García, Del Ser, Rokach, & Herrera, 2020). Bagging is used when it's necessary to reduce the variance of decision trees. Boosting is fitting trees existing of random samples, and at every step, the goal of the tree is to solve the net error from the tree before it (Hancock & Khoshgoftaar, 2020). In this master thesis, the boosting method will be used by using the XGBClassifier algorithm from scikit-learn. In order to work, all numeric features should be scalded, and categorical features have to be encoded. Generally speaking, it is one of the best and most used machine learning models out there.

4.2 *Baseline model: logistic regression*

For the baseline model, to which the other models will be compared too, logistic regression will be used. This is a very basic model, often used as a baseline for binary classification (King, 2008). It has the capability to predict and train fast, and there is no need to scale features. It does assume a linear relationship between features, but that is not a problem for the data set in this master thesis. For this model, the Sigmoid function is used to predict the outcome. The model first computes a weighted sum of the different input features and adds a bias term. Then it passes through the Sigmoid function. The parameter theta is determined by the cost function. With this, positive instances will have a higher probability, and negative instances will have a lower probability (King, 2008).

5 EXPERIMENTAL SETUP

The experimental setup will be discussed in depth, following the flow of the data pipeline illustrated in figure 1. First, the raw data set will be described, how the data was collected and what the available features are. After that, the data cleaning and pre-processing steps will be discussed, focusing on missing values, feature engineering and feature selection. After that, the process of hyperparameter tuning and modelling will be discussed. As well as the cross validation technique and the evaluation metric. Finishing the chapter off with the software and hardware that will be used for this master thesis.

5.1 *Dataset*

The data set was collected by Hendrickson and Perfors (2018), in a research about cross-situational learning in a Zipfian environment. The research study examines how the statistics of word usage across different scenes (cross-situational word learning) relates to real-life language, specifically in terms of the difficulty of learning a lexicon (vocabulary) in a Zipfian distribution. The Zipfian distribution is a distribution in which a small number of words are used very frequently, and a large number of words are used infrequently, as is common in natural language (Hendrickson & Perfors, 2018). The study found that when the distribution of words and meanings follows a Zipfian distribution, learning is not impaired and is often improved. Through a series of experiments, the study provides evidence that this is because Zipfian distributions help people to disambiguate the meanings of other words in a given situation. The experiments were conducted with an artificial language rather than a natural language. This gives the researcher more control over the experiment and removes any chance of difference between the prior knowledge of the language between participants.

It consists of data from 924 participants who participated in different types of experiments. For each of the experiments, there is data available from the training phase and testing phase. To clarify here, if this thesis mentions the training phase and testing phase, it means the phases of the experiments. If it mentions training data and testing data, it means the data sets that are used to train and test the different models.

The complete data set contained 281,135 observations. The variables without any data cleaning or pre-processing are as follows:

Label	Explanation
Language	spoken language
Age	age
Gender	gender
Country	country
Completion code	random generated code to complete experiment
Distribution	condition of the experiment
Subject ID	random generated ID
Break frequency	the number of training screens between each break
Type of data	type of data being recorded
Experiment	ID of experiment code
Trail within screen	number (0 of N) of previous words displayed
Screen words	screen words
Screen objects	screen objects
Phase	current phase of experiment, training or testing
Screen	current screen number (0 to N)
Correct word	correct word
Correct object	ID of correct word (1 to N)
Response time	measured in milliseconds
Location selected	ID of selected objects location (1 to N)
Object selected	ID of current selected object (1 to N)
Word of object selected	word of object selected
Accuracy	accuracy current selection
Data base key	ID of the current entry

Table 1: Data labels and explanation

5.2 *Data cleaning and missing data*

The data cleaning processes started with looking for illegal values and missing values. No illegal values or outliers were detected.
In total, there were 78,708 missing values within the data set which is 38,89% of the observations in the data set. When investigating further, it was shown that these missing values occurred over the same rows, all missing the value for language, gender, country, distribution, break frequency, trial within screen, phase, screen, correct word, correct object, response time, location selected, object selected, objected selected frequency, word of object selected and accuracy. Even though a lot of observations were missing in this case, the choice was made to remove all rows consisting of missing values, for the reason that changing the missing values would mean that, for example, 38,89% of the observations in the age columns

by the mean of the other 62% age observations. This would lead to a less robust training data set. However, removing these observations would lead to a data set consisting of a total of 202,427 observations, which is still enough to train a reliable and robust model. After the removal of all of the missing values, all duplicates were removed. Ending in a data set of 202,427 observations of 21 variables.

The language column contained some spelling mistakes and different writing of the same words. These were cleaned up so that all writing was consistent throughout. All other character and string columns were also checked for spelling mistakes, and no inconsistencies were found. It consists of data from 924 participants who participated in different types of experiments. For each of the experiments, there is data available from the training phase and testing phase. To clarify here, if this thesis mentions the training phase and testing phase, it means the phases of the experiments. If it mentions training data and testing data, it means the data sets that are used to train and test the different models.

The complete data set contained 281,135 observations. The variables without any data cleaning or pre-processing are as follows:

5.2.1 *Data pre-proccesing: age categories*

The age variables was grouped into youth (17 - 24), adults (25 - 64) and seniors (65+) to answer research question 2.

5.3 *Feature engineering & transformation*

To make sure that the data is in the right format for the models, all variables containing characters and/or strings were transformed into categories. The variable screenWords, which contained a list of all words that appeared on screen, was removed because this was not transformable into a usable format. Because the data set contained words from an artificial language, the word itself is less useful, but the characteristics of the words can be more useful. The effect of word length on the ability to remember the word has been studied by different researchers. It has been concluded that the longer the word is, the harder it is to remember (Nichelli, 2016)

Because of that, the correctWord variable and selectedWord are engineered into variables containing the length of both words. For both the new variables correctWord_length and selectWord_lenght are created, and the "old" variables containing the words are removed.

5.4 *Feature selection*

Before feature selection, the dataset was split into two sets based on the phase of the experiment, which was either training or testing. The training phase data set contained 169,075 observations, and the testing phase data set contained 33,352 observations. The goal of feature selection is to remove the variables that add no value. Meaning has no predictive powers, and the ones that are too highly correlated. In other to do this, a correlation matrix was created. The variables that had a correlation of zero were: subjectID, breakFrequency, typeOfData, phase and databasekey. Looking at this variables this seems logicaly, the variables breakFrequency, typeOfData and phase had the same value for every observation. The subjectID and databasekey were randomly generated, and were not expected to have any predictive powers, so this is not surprising either. These features will be removed from the testing phase data set. The high correlation variables, a threshold of >0.80, are correctObject and ObjectSelected, correctObjectFreq and objectSelectedFreq, selectWord_lenght and correctWord_lenght. This also is not a surprise. It is obvious that if the correctObjectFreq and the objectSelectedFreq are not the same numbers, the accuracy would be 0. These features will be removed from the testing data set. Only the correctWord_length feature will not be removed. This feature will be tested for feature importance later on in this master thesis.

5.5 *Content of the data set used for models*

The features with a high or low correlation are removed from the testing phase data. The following variables remained: language, age, gender, country, distribution, experiment, trialWithinScreen, screen, responseTime, accuracy and correctWord_length. The language feature varies between English, Vietnamese, Hindi, French, Albanian, Russian, Spanish, Swedish, Tagalog, Malayalam, Ukrainian, Telugu, Mandarin, Polish, Italian and Portuguese. The age feature ranges from 18 to 71. The countries vary between United States of America, India, Ukraine, Portugal, Sweden, Philippines, Sri Lanka, United Kingdom, Egypt, Canada, Poland and Italy. The distribution is either normal or Zipfian. There are 13 different experiments from which the data was collected. The trailWithScreen varies between 0 and 39. The screen number varies between 0 and 70. The response time varies between 0 and 193130 seconds. The correct word length varies between 3 and 9. And accuracy is either 0 or 1.

5.6 *Data split and cross validation*

From now on, when this master thesis mentions "the data set" it refers to the data set containing the testing phase data. This part of the data will be used for training and testing. Before using the scikit-learn function RandomizedSearchCV, further explained in the hyper parameter tuning section, the testing data set is split into a training set and testing. 70% of the data will be for training, and 30% will be for testing. The split was done completely at random. For the training data, a 5-fold cross-validation method will be used, the default for the RandomizedSearchCV function.

5.6.1 *5-fold cross validation*

5-fold cross-validation is a widely used method for evaluating the performance of a machine learning model. The main idea behind this method is to divide the data set into five smaller sets or "folds", and then train the model on four of these folds while using the remaining fold as a validation set. This process is repeated five times, with each fold being used as the validation set once. After all five iterations, the performance metric of the model in this master thesis, which will be accuracy, is calculated and averaged across all the iterations.

One of the main advantages of this method is that it helps to reduce the risk of overfitting by providing an estimate of performance that is less dependent on a specific training and validation set. This is because, by using different subsets of the data as the validation set, the model is exposed to a variety of input data, which reduces the risk of it becoming too closely adapted to the specific training set. Additionally, averaging the performance metric across all five iterations provides a more robust estimate of the model's performance compared to using a single validation set. 5-fold cross-validation is a simple yet powerful method to evaluate the performance of a machine learning model. Providing an estimate of performance that is less dependent on a specific training and validation set helps to reduce the risk of overfitting and provides a more robust estimate of the model's performance.

5.7 *Hyper parameter tuning*

For hyper parameter tuning, the RandomizedSearchCV function from scikit-learn was used. It takes an estimator, a parameter distribution, and a number of iterations as inputs and performs a randomized search on the parameter space. It randomly samples the parameters from the given distribution for a fixed number of iterations and fits the estimator to the data with those parameters. It then finds the best parameters by comparing the cross-validated performance of each set of parameters to the others. In this case, a 5-fold cross-validation. This function is useful when dealing with a large search space of potential parameters, and the goal is to find the best set of parameters quickly. It can save a lot of time compared to a grid search, which tries every possible combination of parameters. But works more efficiently when in comparison to manual search. 5-fold cross-validation is a simple yet powerful method to evaluate the performance of a machine learning model. Providing an estimate of performance that is less dependent on a specific training and validation set helps to reduce the risk of overfitting and provides a more robust estimate of the model's performance.

5.8 *Evaluation metric*

The evaluation metric accuracy will be used to compare all models. This metric was chosen because it falls in line with the studies from the literature review. In the literature, accuracy was the most commonly used evaluation metric.

5.9 *Software and hardware*

Data cleaning and processing was performed using the programming language R 4.1.1 in R studio. The R Base package was used. For all other steps the programming language Python 3.0 in Juypter Notebook was used, with the following packages: Pandas (pandas development team, 2020) , NumPy (Harris et al., 2020), Sklearn (Pedregosa et al., 2011), Seaborn (Waskom, 2021) and Matplotlib (Hunter, 2007).

6 RESULTS

In this chapter, the results are discussed. First, look at the results of hyperparameter tuning, then compare the accuracy of the test set between the baseline model, support vector machine, multilayer perceptron and decision tree ensemble. The results of the error analysis will also be discussed, as well as the results of the disparate group analysis. Lastly, the topic of feature importance is discussed based on the results.

6.1 *Scaling*

Before running every model, the data set is scaled using the StandardScaler function from scikit-learn. This function starts with calculating the mean and standard deviation of the features in the data set and then standardizing the features by subtracting the mean and dividing by the standard deviation. This to ensures that each feature has a mean of 0 and a standard deviation of 1, which can help some algorithms converge faster and perform better.

6.2 *Hyperparameter tuning*

To train the models and to find the best hyperparameters in the training set, 70% of the testing phase was used. To perform this randomized search was used, using the RandomizedSearchCV function from scikit-learn. The ranges and options of the hyperparameters were based on studies from the literature review and prior knowledge.

6.2.1 *Hyperparameter tuning support vector machine*

For the support vector machine, there are three parameters that can be tuned, C, gamma and kernel. In this case, the kernel was set to RBF, the most commonly used for classification problems. The range for C was set to range from 0.1 up to 100, and the options for gamma were set to vary between 1, 0.1, 0.01, 0.001, and 0.001.

The cross-validation method was set to the default 5-fold cross-validation, and the scoring was accuracy.

The best performing model had the following hyperparameter settings: C is 10, and gamma is 1.

6.2.2 *Hyperparameter tuning multilayer perceptron*

For the multilayer perceptron, there are multiple parameters that can be tuned. The hidden layer sizes determines the number of hidden layers was set to vary between (150, 100, 50), (120, 80, 40) and (100, 50, 30). The maximum iteration was set to 2500. The solver option was sgd, stochastic gradient descent, or adam, another type of stochastic gradient descent. The activation, the activation function for the hidden layers, was set to tahn, hyperbolic tan function or relu, rectified linear unit function. The learning rate, the learning rate schedule for weights updates, was set to adaptive, keeping the learning rate constant to 'learning_rate_init' as long as training loss keeps decreasing or constant, constant learning rate given by 'learning_rate_init'.

The cross-validation method was set to the default 5-fold cross-validation, and scoring was accuracy.

The best performing model had the following hyper parameter settings: hidden layer size (120,80,40), adaptive learning rate and max iteration 2500.

6.2.3 *Hyper parameter tuning decision tree ensemble*

For the decision tree ensemble, the parameters that were n_estimators, the number of boosting stages to perform, which was set to vary between 5, 50, 250 and 500. The max depth estimator, which limits the number of nodes in the tree, was set to vary between 1, 3, 5, 7, and 9. The learning rate, which shrinks the contribution of each tree by learning rate, was set to vary between 0.01, 0.1, 1, 10, and 100.

The cross-validation method was set to the default 5-fold cross-validation, and the scoring was accuracy.

The best performing model had the following hyper parameter settings: learning rate of 1, max depth of 5, n estimators of 500.

6.3 *Performance of baseline model, support vector machine and multilayer perceptron*

-	Accuracy: train set	Accuracy: test set
Baseline	0.608	0.610
SVM	0.652	0.663
MLP	0.606	0.626
DTE	0.705	0.686

Table 1: Performace of the baseline model (logistic regression), support vector machine (SVM), multilayer perceptron (MLP), decision tree ensemble (DTE) based on accuracy

In table 1, the performance of the baseline model, support vector machine and multilayer perceptron is shown. All models outperformed the baseline model on the test set, with the decision tree ensemble being the best performing model. The multilayer perceptron had the lowest performance and seems to be the weakest model.

6.4 *Error analysis*

The error analysis is done by comparing the confusion matrix of each model. Overall they don't show a difference in terms of errors when examining each confusion matrix for each model in figure 3. Generally

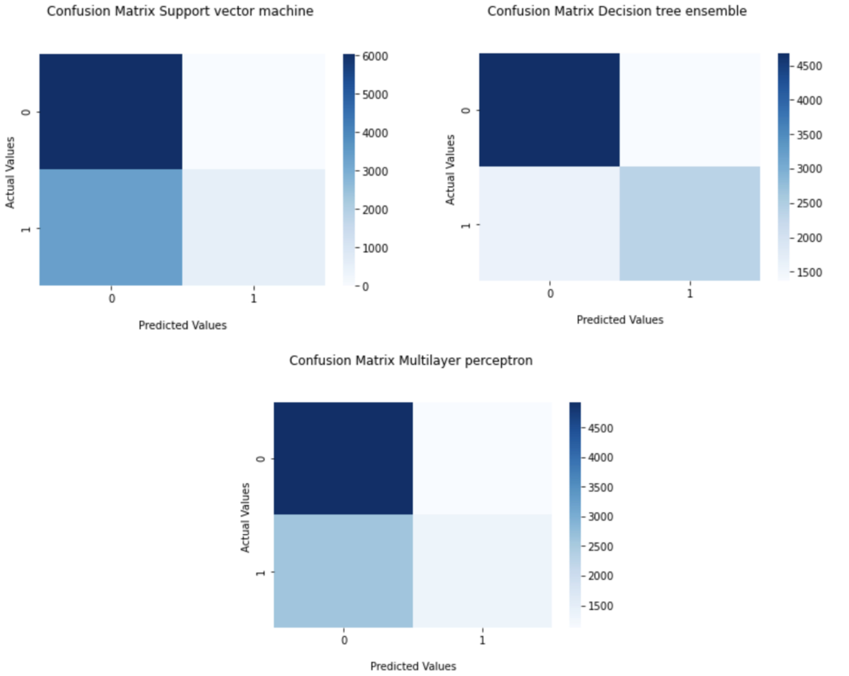

Figure 2: Confusion matrix of the support vector mahcine and multilayer perceptron. Source: the author's illustration.

speaking, all models did especially well in predicting the true values for an accuracy of 0 and showed low errors in falsely predicting a 1 for an accuracy of 0. The decision tree ensemble did show to be slightly better a predicting a true accuracy for 1 than the other models.

6.5 *Performance between age categories of decision tree ensemble.*

-	Accuracy: test set
Youth	0.693
Adult	0.0.703
Seniro	0.715

Table 2: Performace of the three different age categories: youth (17-24), adults (25-64), and seniors (65+) for the decision tree ensemble.

The best performing model, the decision tree ensemble was used to test if there is a performance difference between age categories. The different age categories were youth (17 - 24), adults (25 - 64) and seniors (65+). Examining the results of the evaluation metric accuracy and the test set shows that there was no difference in performance for different age categories. The results are shown in Table 2.

6.6 *Feature importance*

Feature importance was measured using the permutation feature importance technique. Permutation feature importance is a common method for measuring feature importance in machine learning models, especially in the field of feature selection and model interpretation ability. It is a model-agnostic method, meaning it can be used with any learning algorithm, including a decision tree ensemble. The method works by evaluating the change in model performance when the values of a single feature are randomly shuffled. The feature that causes the largest decrease in performance is considered the most important.

The permutation feature importance is a simple yet powerful method for measuring feature importance in machine learning models. It can be implemented easily with any machine learning algorithm and provides an interpretative way to understand the importance of each feature in a data set.

Concluded from the results displayed in figure 3, the most important features are age, the experiment type and response time. Different things can be concluded from this. Age seems to contradict when comparing

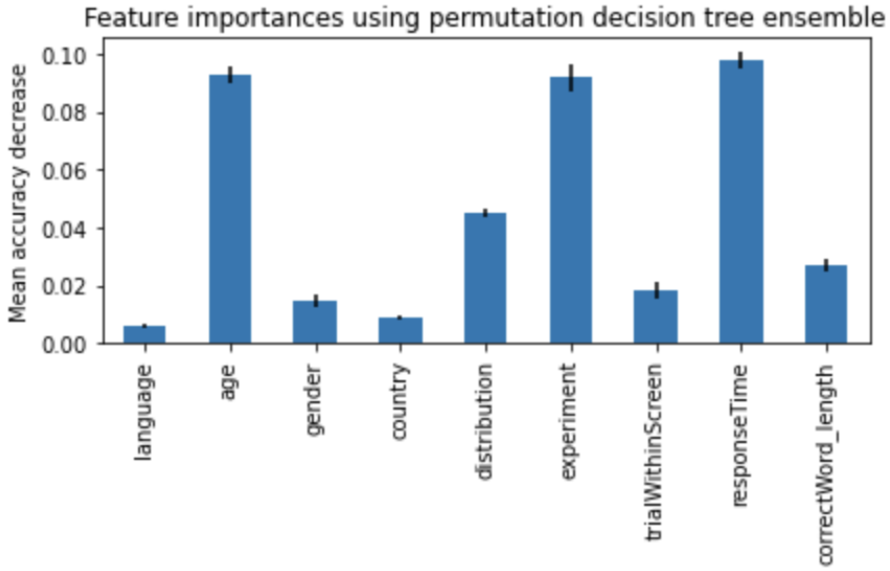

Figure 3: . Feature importance using the decision tree ensemble based on mean accuracy decrease. Source: the author's illustration.

the answer to the previous research question because no difference was detected performance of the models between different age categories. However, this does not mean that, in general, age can't be an important predictor. The type of experiment that was conducted also seems of high importance, meaning that the learning technique that was used when conducting the experiments from which the data was collected is of high importance. Lastly, response time seems to be an important feature as well. It seems that the time that a subject takes to respond is a resourceful bit of information. What all of this means for developing learning techniques will be discussed in the discussion.

6.7 *General discussion*

This master thesis has proven that the models that are used are able to predict individual performance in artificial language learning well. All models outperformed the baseline model based on the accuracy of the test set. However, every research has its shortcomings, and this master thesis also has some points which are needed to be addressed.

In general, the goal of this master thesis was achieved. The findings from this master thesis are useful and can be a stepping stone for further research. With that being said, this master thesis does lack some generalizability and depth. The chosen models and their parameters could have been expanded, and more complex models with more available parameters to tune could further improve the generalizability of the findings. Next to that, the overall finding of this master thesis is useful but could use more depth to have a bigger societal impact. This can be done by using resources and combining studies between a data scientist and cognitive science researchers to further investigate models and techniques and the findings of the research relating to cognitive science.

6.8 *Models performance*

As expected by past researchers, all models were able to predict individual performance in artificial language learning fairly well. The three models, support vector machine, decision tree ensemble and multilayer perceptron, outperformed the logistic regression baseline model based on the accuracy of the test set. But the studies in the literature did mostly achieve a higher accuracy than the models in this master thesis. Even though you can not directly compare the scores because different data sets were used, it is still a point that stands out. This could be because of the less complex features in this master thesis. There was no use of any psychologically motivated features, which could have been useful. However, there is no way of knowing this without testing this in feature research.

6.9 *Difference between age categories*

The studies in the literature suggested that age was an important factor in language learning. That is the reason behind looking into the difference in performance between the different age categories. This master thesis concluded that there was no difference between the performance of different age categories. Meaning that the decision tree ensemble had no problem in predicting the performance of an individual if there the individual fell into a different age category. However, it did conclude that age is an important feature in predicting the accuracy of individual performance, which falls in line with the findings from the literature review (Birdsong, 2018);Zhang (2022).

6.10 *Feature importance*

Feature importance might be the most important question to be able to contribute to the societal impact, helping develop better learning techniques and a better understanding of language learning. Age was one of the important features in predicting the accuracy, which falls in line with the findings in the literature review (Birdsong, 2018);Zhang (2022). Another important feature was the experiment, the data set that was used contained data from different types of experiments. This is especially meaning full because this means that the way people learn a language is important for their success in learning a language. The last important feature was the response time. Whether a participant responded faster or slower, this feature was of high importance to be able to predict the individual's performance. This is something that has not come up when reviewing the literature but is nevertheless useful for further research. However, this might be a topic that can be researched within the cognitive science field rather than data science on its own. Lastly, during the pre-processing of the data, the correct word feature was engineered into the correct word length feature. Here the value of the feature was the number of characters of the correct word. This was done because studies in the past have shown that word length is related to the ability to remember a word (Nichelli, 2016). However, this master thesis shows that the word length does not help more, compared to other features, in predicting individual performance.

6.11 *Limitations*

This master thesis dealt with one big limitation that the researcher experienced, which was the lack of knowledge which was needed to increase the depth of the findings. Even though the internet is full of papers on language learning from different perspectives, the researcher still experienced a lack of understanding of what some findings could mean in different settings and if these are out of the ordinary or not.

6.12 *The next step*

For the next step, researchers can do multiple things. Even though the models in this master thesis performed well, even higher accuracy could be obtained by using different, more complex models. These complex models, with more hyper parameters available to tune, can also be used to generalize the findings better. Another thing could be to further investigate the findings of the feature importance; age, experiment and response time. What the exact importance of these features are? For example, does a higher response time lead to better accuracy or even why is response time such an important feature? Investigating this further can even help better improve learning techniques. The code from this master thesis will be made available via `https://github.com/kikimgp/Thesis`.

6.13 *Societal impact*

The conclusions of this thesis can be used to create better learning techniques when teaching languages in class. But also for governments to be able to teach languages more efficiently in integration courses. In this day and age, more people are using apps and online courses to learn second languages. The companies behind these apps and websites, for instance, Duolingo, can also create better learning techniques and plans, which will make it easier for their consumers to learn languages. Especially the feature importance question can contribute to the societal impact. The features that seem to have a higher degree of importance can be further investigated and used when developing language learning techniques. This study has shown that age and the way of teaching a language are important indicators when predicting individual performance, meaning that these are also important factors to keep in mind when developing learning techniques.

7 CONCLUSION

To conclude the research, the main research question: **To what extent is it possible to predict performance in artificial language learning?** ,has to be answered by using the three sub questions.

1.Which model leads to the best performance, when comparing Support Vector Machine, Decision Tree Ensembles and Multilayer perceptron?
The accuracy of the test set was used to determine the best performing model. The decision tree ensemble outperformed the baseline and the other two models, the support vector machine and multilayer perceptron. The accuracy of the decision tree ensemble was 0.686 on the test set..

2.To what extent is there a difference in performance of the models, when comparing age categories, for the best performing model?
To answer this question, the decision tree ensemble was used to see if there was a difference in performance between the three age categories, youth (17-24), adults (25-64), and seniors (65+). No difference was measured when looking at the accuracy score of the test set.

3.Which features are most important to accurately predict individual performance in artificial language learning, for the best performing model?
To answer this question, the decision tree ensemble was used with the permutation feature importance technique. The three most important

predictors based on the mean accuracy decrease are age, experiment and response time.

REFERENCES

2018 Duolingo Shared Task on Second Language Acquisition Modeling (SLAM). (2018). Retrieved from https://sharedtask.duolingo.com/2018.html

Al-Shehri, H., Al-Qarni, A., Al-Saati, L., Batoaq, A., Badukhen, H., Alrashed, S., ... Olatunji, S. O. (2017, 4). Student performance prediction using Support Vector Machine and K-Nearest Neighbor. *2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE).* Retrieved from http://dx.doi.org/10.1109/ccece.2017.7946847 doi: 10.1109/ccece.2017.7946847

Arashpour, M., Golafshani, E. M., Parthiban, R., Lamborn, J., Kashani, A., Li, H., & Farzanehfar, P. (2022). Predicting individual learning performance using machine-learning hybridized with the teaching-learning-based optimization. *Computer Applications in Engineering Education*, *n/a*(n/a). Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/cae.22572 doi: https://doi.org/10.1002/cae.22572

Babić, I. Đ., & Benčina, K. (2017). Prediction of reading comprehension ability in english as a foreign language. In *42nd atee annual conference 2017 conference proceedings* (p. 152).

Birdsong, D. (2018, 3). Plasticity, Variability and Age in Second Language Acquisition and Bilingualism. *Frontiers in Psychology*, *9*. Retrieved from http://dx.doi.org/10.3389/fpsyg.2018.00081 doi: 10.3389/fpsyg.2018.00081

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020, 9). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*, 189–215. Retrieved from http://dx.doi.org/10.1016/j.neucom.2019.10.118 doi: 10.1016/j.neucom.2019.10.118

González, S., García, S., Del Ser, J., Rokach, L., & Herrera, F. (2020). A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Information Fusion*, *64*, 205–237.

Han, H., & Jiang, X. (2014, 1). Overcome Support Vector Machine Diagnosis Overfitting. *Cancer Informatics*, *13s1*, CIN.S13875. Retrieved from http://dx.doi.org/10.4137/cin.s13875 doi: 10.4137/cin.s13875

Hancock, J., & Khoshgoftaar, T. M. (2020). Performance of catboost and xgboost in medicare fraud detection. In *2020 19th ieee international conference on machine learning and applications (icmla)* (pp. 572–579).

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., . . . Oliphant, T. E. (2020, September). Array programming with NumPy. *Nature*, *585*(7825), 357–362. Retrieved from https://doi.org/10.1038/s41586-020-2649-2 doi: 10.1038/s41586-020-2649-2

Hendrickson, A. T., & Perfors, A. (2018, Nov). *Cross-situational learning in a zipfian environment.* PsyArXiv. Retrieved from psyarxiv.com/6jumv doi: 10.31234/osf.io/6jumv

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. doi: 10.1109/MCSE.2007.55

King, J. E. (2008). Binary logistic regression. *Best practices in quantitative methods*, 358–384.

Nichelli, P. (2016). Consciousness and Aphasia. *The Neurology of Conciousness*, 379–391. Retrieved from http://dx.doi.org/10.1016/b978-0-12-800948-2.00023-6 doi: 10.1016/b978-0-12-800948-2.00023-6

Osika, A., Nilsson, S., Sydorchuk, A., Sahin, F., & Huss, A. (2018). Second language acquisition modeling: An ensemble approach. Retrieved from https://arxiv.org/abs/1806.04525 doi: 10.48550/ARXIV.1806.04525

pandas development team, T. (2020, February). *pandas-dev/pandas: Pandas.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.3509134 doi: 10.5281/zenodo.3509134

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pisner, D. A., & M.Schnyer, D. (2020). Support vector machine. *Machine Learning*, 101–121. Retrieved from http://dx.doi.org/10.1016/b978-0-12-815739-8.00006-7 doi: 10.1016/b978-0-12-815739-8.00006-7

Rich, A., Osborn Popp, P., Halpern, D., Rothe, A., & Gureckis, T. (2018, June). Modeling second-language learning from a psychological perspective. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications* (pp. 223–230). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W18-0526 doi: 10.18653/v1/W18-0526

Settles, B., Brust, C., Gustafson, E., Hagiwara, M., & Madnani, N. (2018, June). Second language acquisition modeling. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications* (pp. 56–65). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://aclanthology

.org/W18-0506 doi: 10.18653/v1/W18-0506

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. Retrieved from https://doi.org/10.21105/joss.03021 doi: 10.21105/joss.03021

Widyahastuti, F., & Tjhin, V. U. (2017, 7). Predicting students performance in final examination using linear regression and multilayer perceptron. *2017 10th International Conference on Human System Interactions (HSI)*. Retrieved from http://dx.doi.org/10.1109/hsi.2017.8005026 doi: 10.1109/hsi.2017.8005026

Xu, S., Chen, J., & Qin, L. (2018, June). CLUF: a neural model for second language acquisition modeling. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications* (pp. 374–380). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W18-0546 doi: 10.18653/v1/W18-0546

Zhang, H. (2022). Analysis on the Influence of the Age Factor on Second Language Acquisition: Rethink the Meaning of Critical Period Hypothesis and Create New Teaching Methods. *Advances in Social Science, Education and Humanities Research*. Retrieved from http://dx.doi.org/10.2991/assehr.k.220131.151 doi: 10.2991/assehr.k.220131.151