PREDICTING THE GENDER OF PARTICIPANTS BASED ON COGNITIVE TASK PERFORMANCE

ÁGNES BERNADETT BÁLINT

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

2068267

COMMITTEE

prof. dr. Peter Hendrix
prof. dr. Raquel Garrido Alhama

LOCATION

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

DATE

June 27, 2022

WORD COUNT

8160

ACKNOWLEDGMENTS

The completion of this thesis would not have been possible without the outstanding support of my supervisor, dr. Peter Hendrix. I am also grateful to my partner, for being by my side and cheering me up on the harder days.

PREDICTING THE GENDER OF PARTICIPANTS BASED ON COGNITIVE TASK PERFORMANCE

ÁGNES BERNADETT BÁLINT

Abstract

Many studies have focused on the differences in language processing and producing skills between genders, however, most experimental research was conducted on only a small set of cognitive tasks. This research aims to investigate to what extent gender can be predicted based on cognitive task performance, by finding the optimal machine learning algorithm for this problem and highlighting a subset of tasks with high feature importance. This study uses a dataset containing more than 30 cognitive skill tasks and compares the performance of state-of-the-art methods, such as XGBoost, Random Forests, and Neural Networks. To conduct the research, the Individual Differences in Language Skill database is used, containing the performance of 112 Dutch participants on 33 behavioural measures, across 35 tests. The results of this thesis show, that XGBoost outperforms Linear Regression, Random Forests and the Neural Network in predicting gender based on cognitive task performance. Further analysis shows, that the cognitive tests with the highest predictive power in the best performing model, XGBoost, are the *phrase generation* test, the *visual choice reaction time* test and the *word monitoring in noise* test.

The findings of this study can contribute to a deeper understanding of the relationship of gender and cognitive skills and can be used in author profiling and gender detection.

Data Source, Code and Ethics Statement

Work on this thesis did not involve collecting data from human participants or animals. The original owner of the data used in this thesis retains ownership of the data during and after the completion of this thesis. The data is publicly available at the UK Data Service. The author of this thesis acknowledges that they do not have any legal claim to this data. The code used in this thesis is not publicly available, but can be requested at a.b.balint@tilburguniversity.edu. All images and tables used in this thesis are produced by the author.

1 INTRODUCTION

1.1 *Motivation*

All humans have the ability to process language in some way, but everyone has a different level of that skill. Language producing and processing exists in different forms, such as written, spoken, and nonverbal communication.

In many research, the differences between individuals in cognitive and linguistic skills have been generally overlooked. Most experimental research focuses on the average performance of participants and fails to examine sub-group level- (such as gender) or individual- differences. Given the fact that participants are almost exclusively university students in such experiments, little is known about the general differences in language skills among a wider spectrum of young adults. (Hintz, Dijkhuis, van't Hoff, McQueen, and Meyer (2020))

It is accepted as fact, that there are differences between males and females in psychological domains, but very little is known in which specific domains and skills they differ and why. (Buss, 1995; McCarthy, Arnold, Ball, Blaustein, & De Vries, 2012) In this research, the differences in cognitive skills across genders will be examined. Proving the existence of differences in cognitive task performance across males and females can lead to a better understanding of how different genders process and produce language in different tasks.

Most studies focus on only a few specific problems or a small subset of tasks and examine the results. Results mostly show that men are better at problem solving (Mefoh, Nwoke, Chukwuorji, & Chijioke, 2017; Roberts & Bell, 2003). This approach seems one-sided, as no comprehensive study has been conducted to test a large set of skills and compare tasks that males and females are each better at. Highlighting cognitive tasks or skills in which females outperform males can lead to a more balanced understanding of gender differences. This can contribute to the tendency of recent studies to focus on the individual level, as seen in Jongman, Roelofs, and Meyer (2015); Kidd, Donnelly, and Christiansen (2018).

1.2 *Scientific and Societal Relevance*

The results of this research can contribute to a better understanding of language comprehension and processing differences across genders. It will expand upon results by Alantie, Tyrkkö, Makkonen, and Renvall (2022) and Morgan-Lopez, Kim, Chew, and Ruddle (2017), which show that language processing skills and language usage can reflect an individual's personal characteristics, one of which is gender.

Moreover, results of this study can be used as a base for educational purposes. Knowing which tasks a specific gender performs better at can help teachers to chose the right metrics when assessing a child's cognitive skills. On one hand, at the early stages of development of a child's cognitive skills, professionals could concentrate on tasks that the child's gender group is generally better at, for the purpose of reinforcement and the feeling of success. On the other hand, the tasks that the child's gender group is generally weaker at can be used as a base to identify a set of cognitive skills that can be improved upon and trained from their early years.

In addition, the results can help to understand to a greater extent how different genders perform in specific cognitive tasks, which can be used in different areas of linguistics, such as psycholinguistics and forensic linguistics. In author profiling, they can be used to design a useful approach to identify individuals and get a better understanding of their personality, skills and capabilities.

1.3 *Research Questions*

This thesis aims to answer the following research question:

> *To what extent is it possible to predict the gender of participants based on their performance in cognitive tasks?*

This research question will be answered with the following sub-questions:

> *Which of a set of pre-selected machine learning algorithms provides optimal predictive power for the gender of participants?*

To answer this question, multiple state-of-the-art models are used, such as XGboost, Decision Trees and Deep Neural Networks. These are compared to the baseline performance of Multivariable Logistic regression.

The second sub-question is:

> *Which cognitive tasks offer predictive power for a participant's gender?*

This question aims to determine the importance of each feature and highlight the ones with the highest correlation with, and predictive power for the target variable. Feature ablation is used to identify these features. For a detailed description see Section 3.

In many research papers, the terms *gender* and *sex* are poorly defined. The database used for this study (*Individual Differences in Language Skill*), for instance, claims to include information on the participants' *sex*. The intake questionnaire that collected participant information, however, asked participants to indicate their gender, rather than their sex. As such, there

is an inconsistency in this respect between the collection of the data and the write up in the paper. Consistent with the manner in which the data were collected, I will use the term gender to refer to the two groups of participants in the database throughout this thesis.

1.4 *Findings*

In summary, this study compares Multivariable Logistic Regression, Extreme Gradient Boosting, Random Forests and Artificial Neural Network for gender prediction. Each model is fit to the training set, and after hyperparameter tuning for their optimal performance, each is evaluated on the test set using weighted F_1 and AUC metrics. The analysis shows that XGBoost has the highest F_1 and AUC score (0.732 and 0.706, respectively) in predicting gender based on performance. Phrase generation, followed by Visual choice reaction time test and Word monitoring in noise are the most predictive features, with decrease in F_1 by 0.18, 0.16 and 0.13, respectively.

2 RELATED WORK

This section provides a comprehensive overview of the relevant literature. The *Related Work* section is further divided into subsections: the first subsection highlights studies related to gender prediction, the second subsection gives an overview of previous research done into cognitive tasks and gender performances, the third subsection discusses the machine learning algorithms that are used in this thesis, and the last subsection gives an outline for the current study.

2.1 *Gender prediction*

Predicting the gender of an individual is a fundamental aspect of author profiling. The gender of an author seems to be possible to predict based on lexical and social network features (van der Goot, Ljubešić, Matroos, Nissim, & Plank, 2018). Rangel and Rosso (2013) examined the language usage of males and females, and whether identification of gender is possible in written text. The findings state that females and males differ in their usage or grammatical categories. He concluded that males use more prepositions for hierarchical structuring of their environment, while females tend to use more pronouns and determinants given they are generally more interested in social relationships.

Research by Alantie et al. (2022) shows significant correlation between a set of linguistic performance skills of elderly Finnish participants and

their personal characteristics, such as gender. In their results, the age of the participants was the most powerful predictor; the influence of gender was presented, but not further investigated. Following these studies, this thesis aims to investigate the extend to which gender can be predicted based on a broad set of cognitive tasks.

2.2 *Cognitive tasks*

A correlation between different cognitive task results and an individual's personal characteristics has been proven in several studies. Some of those studies have attempted to find significant differences between the cognitive performance of different genders. No comprehensive answer was found, as most studies mainly focus on one, or a small set of cognitive tasks.

Weiss et al. (2003) reported, that while solving visuospatial cognitive tasks, different brain activation is seen between males and females. Similarly, Brañas-Garza, Kujal, and Lenkei (2019) reported that in cognitive reflection tests, there is a significant negative correlation between being female and the number of correct answers given. A study by Mefoh et al. (2017) suggests that in problem solving types of cognitive tasks males outperform females. Mefoh et al. (2017) reports a statistically significant difference between males and females when measuring the amount of puzzle problems solved. His research was conducted on only one type of task, namely problem solving. Additionally, two- and three-dimensional mental orientation tasks were performed and analysed by Roberts and Bell (2003). In that study, they concluded that males performed better than females in the three-dimensional tasks, but there was no difference in performance for the two-dimensions tasks. There was no further research done into why there is difference in three-dimensions but not in two-dimensions, and no other tasks were performed.

During a study done back in 1994, Gallagher examined whether females use different solution strategies for mathematical problem solving than males, and whether that choice of approach is correlated with performance. His participants were students with high mathematical ability. In his paper, Gallagher and De Lisi (1994), he reported results that confirm his theory; female students had generally used conventional strategies as problem solving methods. He reported correlation with the choice of conventional strategies and worse performance. In his experiment, males outperformed females, alike to findings on problem solving by Mefoh et al. (2017). He showed, that the results were due to the fact that males had an unconventional strategy approach rather that conventional.

A paper by Feingold (1988), titled *'Cognitive gender differences are disappearing'* tested different kinds of cognitive tasks. Based on his experiments

with teenagers, conducted over a period of more than 20 years, he reported that girls outperformed boys on spelling, grammar tasks and perceptual speed, while boys scored higher than girls on tasks measuring spatial visualisation, mechanical aptitude and mathematics. He reported no difference in verbal and figural reasoning.

Alike Gallagher, several researchers in the 90' have focused on differences between males and females in mathematical problem solving performance, but have found contradicting results. Duffy, Gunther, and Walters (1997) examined gender differences in performance on the *GAUSS* test and the *Canadian Test of Basic Skills* (CTBS). Interestingly, for the *GAUSS* test, there was no gender difference among twelve year old individuals, thus generalising for mathematical tests may lead to false assumptions about gender and performance. Baker and Jones (1993) conducted experimental research over the duration of almost twenty years, with participants from several European countries. His results show that during that 20 years of his experiment, the effect of gender on mathematical performance decreased over the years. His further findings show that the gender-performance correlation was based on the educational opportunities, which were less accessible for women.

More recent studies also show contradicting results with previous findings. Lindberg, Hyde, Petersen, and Linn (2010) did not find significant differences in the mathematical performance of males and females. They used meta-analysis on data from over 200 studies, with publishing dates ranging from 1990 to 2007. In the full database, they had data on more than a million individuals in total. They concluded that males and females perform similarly in mathematical tasks.

These gender stereotypes about females performing worse than males at mathematical tasks may not just not be true, it might even have its impact on the performance of females as well. Schmader (2002) conducted a quasi-experimental study which showed that females with high levels of gender identification performed worse than males. On the other hand, females with lower levels of gender identification had similar results to males. Kiefer and Sekaquaptewa (2007) conducted a similar study on female university students participating in calculus courses. Both studies show, that female performance is affected by the stereotype of women having lower aptitude for math.

Other than an individual's own gender, the gender of additional people in a test environment can also influence the performance. As seen in work done by Inzlicht and Ben-Zeev (2000), performance of females can decline when male participants are present during mathematical problem solving tasks. There was no performance decrease in case of a verbal test. For males, the presence of females did not affect the results in either

task. This further proves, that performance is not solely dependent on the mathematical ability of an individual, but on other factors as well.

Some research aimed to find the reason behind the gender difference phenomenon. Research conducted by Weber, Skirbekk, Freund, and Herlitz (2014) attempted to find the cause of the difference in performance between males and females in cognitive tasks. The results suggest, that cognitive performance differences across genders can be due to the living conditions and the educational levels. This research provided more insight into the cause of the gender difference.

Current thesis aims to build upon these previous studies by highlighting cognitive tasks that females or males perform better at, and thus revealing correlation between gender and cognitive skills. Knowing the difference in cognitive skill performances can help understand the thinking processes of genders.

2.3 *Models and Algorithms*

Machine learning algorithms are commonly used for gender prediction. Current study has tested XGBoost, Random Forests, and a Neural Network, along with Logistic Regression for baseline.

XGBoost has been proven to be a computationally efficient machine learning model for the purpose of classification. It was used by Dehzangi et al. (2021) to predict the well-being of opioid patients based on cognitive test results and other predictors. They reported a 96.12 % accuracy, which was the best performing model in their research.

Random Forests have proven to give a good performance in predicting gender as well. In Raghunadha Reddy, Vishnu Vardhan, GopiChand, and Karunakar (2018), they report a 91.17% accuracy of gender prediction for author profiling using Random Forests as a classifier.

Both Deitrick et al. (2012), Rafique et al. (2019) and Cichy and Kaiser (2019) are studies using Neural Networks for gender related research. Deitrick et al. (2012) predicted the gender of the author based on email texts. They reported an accuracy of 95% on word based features, and thus displaying a difference between males and females. Rafique et al. (2019) used Convolutional Neural Networks (CNN) for age and gender prediction on images of individuals with 79% of accuracy. This thesis will use Multilayer Perceptrons for age prediction, anticipating high F_1 scores.

In this study, the set of algorithms listed above will be compared to find the best performing machine learning algorithm. Further details on these algorithms and their performances can be seen in later sections.

2.4 *Current study*

In this study, the aim is to give a comprehensive overview of the predictive power of 33 cognitive measures across 35 tasks in gender prediction, and to highlight a set of tests that have the highest predictive power.

For this research, a dataset based on the Individual Differences in Language Skill (IDLaS) is used. It contains 33 different cognitive measures divided into 35 tasks, and each task is repeated twice for each participant, approximately a month apart. The dataset includes the performance of each person for each task and test occasion, resulting in 70 performance indicators per participant. The experiment was conducted with the help of 112 Dutch participants between the age of 18-29. This age range shows improvement upon previous research on the topic, and gives better insight into the general performance of young adults. (Hintz et al., 2020)

While most research is focused on only a small subset of tasks, current thesis aims to study a broader range of correlations between genders and cognitive task performance using all 35 tasks as predictors. A set of state-of-the-art algorithms is used for the the predictions. A comparison of machine learning models for gender identification on cognitive tasks is a unique approach that has not been done previously. Feature ablation is used to find tasks that have a significant role in predicting gender. Further analysis is conducted to identify which gender performs better at those specific cognitive tasks.

Having a complete overview of different tasks can help breaking down the general belief that males outperform females in overall cognitive tasks.

3 METHODOLOGY

In the following section, the methodology used for this thesis is introduced. The first subsection contains the description of the machine learning algorithms used in this study. The second subsection explains the process used for the feature importance.

3.1 *Machine Learning Models*

To answer the first research sub-question, the following machine learning algorithms were compared: Logistic Regression, Random Forests, XGBoost and Neural Network. In this section, the methodology behind these algorithms are explained.

3.1.1 *Logistic Regression*

Logistic Regression is one of the base algorithms in machine learning. A simple logistic regression predicts a binary outcome Y as a function on an independent variable X. Multivariable Logistic Regression is used when there is a single outcome (Y) and multiple independent variables (X1....Xn), using the following formula:

$$Pr(Y_i = 1|X_i) = \frac{exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n)}{1 + exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n)} \quad (1)$$

where $Pr(Y_i = 1|X_i)$ is the probability of outcome 1 for observation i, given the vector of predictor variables X_i. Y is the dependent variable, whereas $X_1, X_2, \ldots, X_n$ are the independent variables. (Ebrahimi Kalan, Jebai, Zarafshan, & Bursac, 2021) In this study, Multivariable Logistic Regression is used as the baseline to compare the other models to.

3.1.2 *Random Forests*

Random Forests is an ensemble method that combines the output of multiple decision trees, and thus performs more robust than a single tree. Being an extension of bagging algorithms, Random Forests is taking bootstrap samples from the observations, using a subset of features for each tree. The final output, in case of classification, is the majority vote of the trees. In this method, feature bagging ensures that the individual decision trees have low correlation by having only a subset of the features. (Zhou, 2012)

For reaching optimal performance, several hyperparameters can be tuned, such as: number of trees, number of features at every split, maximum depth (levels) of each tree, minimum number of samples to split a node, minimum number of samples to be stored in a leaf node and whether bootstrapping is used for sampling data points.

The key benefits of Random Forests include smaller risk of overfitting, usability for both regression and classification, and, given the tree-structure, it makes determining feature importance relatively easy. One downside of Random Forests is that training is time-consuming when using a large dataset.

3.1.3 *XGBoost*

Extreme gradient boosting, or commonly referred to as XGBoost, is a fast and high-performance implementation of gradient boosted decision trees. XGBoost is a state-of-the-art algorithm and software system, developed by Tianqi Chen. (Chen & Guestrin, 2016) Whereas Random Forests fit trees independently, trees are fit in a sequential manner in XGBoost, with each

tree being an expert at the shortcomings of the previous trees. In gradient boosting, a sequence of tree-based models is trained, and after each tree, the error is calculated, and the parameters are updated for the next tree. (Friedman, 2002) Similar to Random Forests, numerous parameters of the model can be tuned. Such hyperparameters include maximum depth (levels) of each tree, learning rate, subsample ratio of the training examples, number of estimators and subsample ratio of columns by tree, level, or node.

The advantage of XGBoost lies in its efficiency (performance and speed) and wide range of applications. It has proven its usability in several applications not only among researchers, as well as in numerous competitions on the data science platform Kaggle. (Nielsen, 2016)

3.1.4 *Neural Networks*

Artificial Neural Networks are designed based on inspiration from the neural network system in the human brain. In this study, a Multilayer perceptron (MLP) is used as a neural network approach. A Multilayer perceptron, similar to other feed-forward artificial neural networks, consists of an input layer of nodes (neurons), one or more hidden layers of nodes, and an output layer of nodes. Each node combines information from the data that was given as an input and weights, that is summed and passed to the activation function to get the significance of the node and decide whether the information is passed on or not. This method is trying to mimic the way the human brain works with activating specific neurons based on the importance of the information it carries. (Wang, 2003)

Neural Networks are a good tool for pattern recognition, and can be used on all kinds of real-life data, giving good performance. (Abiodun et al., 2018) A downside of Neural Networks, however, is that it tend not to be optimal for relatively small datasets, and thus can perform worse than traditional machine learning algorithms. (Feng, Zhou, & Dong, 2019)

3.2 *Feature ablation*

To answer the second research sub-question, feature ablation is conducted on the best performing model to find features with the highest predictive power for the target variable. In several studies, such as Fraser et al. (2014), it has been proven to be a powerful tool for feature importance. During feature ablation, one feature is left out of the model to test its contribution to the performance. Since the current dataset includes tests repeated twice for each individual, the model fitting is repeated 35 times, leaving out two features at a time, being the same cognitive task repeated over two

occasions. The performance of those 35 individual models are compared, with the decrease in model performance indicating the importance of the feature being left out.

4 EXPERIMENTAL SETUP

In this section, the experimental setup of this study is explained, including detailed information about the dataset, pre-processing steps, hyperparameter optimization of each model and the metrics used for the final evaluation.

4.1 *Dataset and Pre-processing*

The Individual Differences in Language Skill is a battery of tests and was completed by 112 native Dutch individuals (18–29 years old) in 2019, at Max Planck Institute for Psycholinguistics in Nijmegen. The data is freely available on the UK Data Service website [1] . It includes 33 behavioural measures across 35 tests, evaluating language skills and domain-general cognitive skills that are likely to be involved in language tasks. Tests evaluating language experience (e.g. vocabulary size), linguistic processing skills (e.g. word comprehension) and general cognitive abilities (e.g. memory) were included in the battery. The test was conducted in a tightly supervised laboratory setup. Each participant completed the battery twice, in two different days (approximately a month apart), with one occasion lasting for about four hours. (Hintz et al., 2020)

The final dataset contains the performance of the 112 participants (39 males and 73 females). 87 of the participants attended university, 24 attended vocational college and one individual was a high-school graduate. The tests in the battery included linguistic experience tests, namely: *Stairs4Words*, *Peabody picture vocabulary test*, *spelling test*, *author recognition test*, *idiom recognition test*, *prescriptive grammar test* and *sentence-picture verification test*. For testing non-verbal processing speed, the *auditory simple reaction time test*, *auditory choice reaction time test* and *letter comparison test* were used. As visual reaction time tests, the *visual simple reaction time test* and the *visual choice reaction time test* were used. Working memory was examined by the *digit span test* and the *Corsi block clicking test* (both forwards and backwards). Measuring inhibition was done by the *Eriksen Flanker and antisaccade test*, while the non-verbal intelligence was done by the *Raven's advanced progressive matrices test*. Word production tests included the *picture naming test*, *rapid automatized naming*, *antonym production*, *verbal*

[1] https://reshare.ukdataservice.ac.uk/854399/

fluency, *maximal speech rate*, *one-minute-test* and the *Klepel test*. For word comprehension, they used the *monitoring in noise in lists*, *rhyme judgment*, *auditory lexical decision* and *semantic categorization* tests. *Phrase and sentence generation*, and *spontaneous speech* measured sentence production. Lastly, for sentence comprehension, they tested on *gender cue activation during sentence comprehension*, *verb semantics activation during sentence comprehension* and *monitoring in noise in sentences*. Having forward and backward variation for two of the tests, the total number of tests in the dataset is 35, resulting in a total of 71 feature columns, including gender.

Exploratory data analysis of the features can be seen in Appendix A (page 28). The table shows the feature name, the mean of the scores, the standard deviation of the scores and number of missing values. For the reaction time based tests, the absolute value of the score is the response time in milliseconds.

All individuals have their gender registered as either male or female. 73 females and 39 males participated in the experiment, resulting in a moderately imbalanced dataset. This imbalance is being addressed when choosing evaluation metrics.

The final dataset used in this study was compiled and pre-processed based on the guidelines recommended in the paper of the researchers; Hintz et al. (2020). In the same publication, a detailed description can be found about the pre-processing method they used to remove outliers, and invalid entries and participants from the original dataset.

As some outliers or invalid entries have been changed to missing values, during the pre-processing of the dataset for this thesis, the missing values have been imputed with the median value of the feature column. Using the median is a proper choice for imputing missing values, as it is an accurate representation of the majority value of each feature. (Berkelmans et al., 2022) The target variable has been one-hot encoded, with 0 and 1, representing females and males, respectively.

After the necessary pre-processing steps, the dataset has been split randomly into training and test data, with the test set being 20% of the whole dataset. In order to get the same sets of data for each model and at any later time, random seeds have been used for splitting the data. After the split, it was confirmed that the test data has similar distribution to the training data regarding target variables. The test data has been set aside and only used for the final evaluation for each model.

Table 1: Hyperparameter values tested for Random Forests and XGBoost, and their best performing value

Model	Hyperparameter	Hyperparameter Values	Best value
Random Forests	n_estimators	5, 20, 50, 100	50
	max_features	'auto', 'sqrt'	'sqrt'
	max_depth	10, 20, 40, 60, 80	40
	min_samples_split	4, 6, 10, 12, 14	12
	min_samples_leaf	1, 3, 5	1
	bootstrap	True, False	False
XGBoost	n_estimators	40, 50, 60	50
	learning_rate	0.01, 0.1, 0.3	0.3
	max_depth	5, 10, 20	10
	subsample	0.1, 0.5, 1	0.5
	colsample_bylevel	0.4, 0.6, 1	0.6
	colsample_bytree	0.3, 0.7, 1	1

4.2 *Hyperparameter tuning*

The models have been trained and evaluated on the training data using cross-validation, with weighted F_1 score as the scoring metric. During this phase, hyperparameters have been tuned to reach optimal performance.

In order to establish the baseline, all 70 features have been used for the Linear Regression model. During the training, ten-fold cross-validation was performed.

For Random Forests and XGBoost, randomized grid search with cross-validation was used. Given the large number of potential parameter combinations, the randomized grid search was chosen in order to reduce the computational time and resources. Cross-validation was used to provide better generalization.

For XGBoost, six hyperparameters were tuned, each with three potential values. The hyperparameters are: the *number of estimators*, *learning rate*, *maximum depth*, *subsample ratio*, *column sample per level-* are *column sample per tree ratio*. The potential values for the hyperparameters were chosen based on research by Putatunda and Rama (2018).

For Random Forests, twenty-one potential values have been given across six hyperparameters. The hyperparameters are: the *number of trees*, *maximum number of features per split*, *maximum depth of each tree*, *minimum number of samples to split node*, *minimum number of samples for a leaf node*, and whether *bootstrapping* was used for sampling the data.

The hyperaprameter values per model can be seen in Table 1. The table shows the name of the model, the name of the hyperparameter in the python model (this name may vary per library or programming language),

the values that have been tested in the randomized grid search method, and finally the value that have been used in the final model in combination with the other best performing values. The optimal model for Random Forests consists of 50 trees without bootstrapping, with 12 being the minimum sample number per split, 1 as minimum number of samples to split a node, 40 as the depth of each tree, and the square root of all features as the maximum number of features. XGBoost was optimally built with 50 trees, 0.3 the learning rate, 10 as the maximum depth for each tree, 0.5 as subsample ratio, and 1 and 0.6 for the column sample per tree and the column sample per level ratio, respectively.

It can be interesting to note, that the range of the potential values for the number of trees used in the models (*n_estimators*) was different for Random Forests and XGBoost, but both models gave the best performance with the same value, being 50.

The total number of models fitted for Random Forests and XGBoost is 500 for each, both having 100 iterations over 5 folds.

For the Multilayer Perceptron, the training set was further divided into a training and a validation set. In the final division, the ratio of the training, the validation and the test set to the whole dataset is 0.6, 0.2 and 0.2, respectively. The input variables have been standardized before being fed into the network.

The Neural Network has an input layer, three hidden dense layers and an output layer. The input layer has 50 nodes (with input size being 70), the first hidden layer has 20 nodes, the second hidden layer has 10 nodes, and the output layer gives a single output. For the input and hidden layers the *ReLu* activation function was used, and for the output layer *sigmoid* activation function. *Adam* was used an the optimiser and binary-cross entropy as the loss function. Early stopping was used to monitor the validation loss and avoid overfitting. The Multilayer Perceptron has a total of 4791 trainable parameters. The number of layers and nodes in the Neural Network was determined based on domain knowledge and trial-and-error processes. During the trial-and-error processes, architectures with two to five hidden layers have been tested, with the number of nodes ranging from 10 to 70. Out of these architectures, the current model gave the best performance.

4.3 *Evaluation*

Given that the dataset is moderately imbalanced, weighted F_1 and *Area Under the Curve (AUC)* were chosen to evaluate the performance of each model. Both the weighted F_1 score and the *Area Under the Curve* score are appropriate measures for imbalanced sets, measuring the performance of

the model for both positive (female) and negative (male) classes. (Huang & Ling, 2005; Tharwat, 2020)

4.3.1 F_1

The weighted F_1 score combines two important measures; *precision* and *recall*. The equation for both can be seen in Equation 2 and Equation 3, respectively, with TP being the number of true positives, FP being the number of false positives and FN being the number of false negatives.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

The base formula for F_1 is shown in Equation 4. It is the ratio of two times the multiplication of precision and recall, and the sum of precision and recall. The lowest possible F_1 score is 0, and occurs when either *precision* or *recall* is zero. The highest possible score is 1.0, indicating perfect *precision* and *recall* scores. (Tharwat, 2020)

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

The weighted F_1 score calculates the average of F_1 scores weighted by their supports. Here, the supports are the number of true occurrences in each class. This alteration to the base F_1 score accounts for class imbalance.

4.3.2 *Area Under the Curve*

AUC combines *sensitivity* and $specificity$, and treats negative and positive classes equally. (Huang & Ling, 2005) In current study, females represent the positive class, while males represent the negative class. The equation for *sensitivity* and $specificity$ can be seen in Equation 5 and Equation 6 respectively, where TN is the number of true negatives.

$$Sensitivity = Recall = \frac{TP}{TP + FN} \tag{5}$$

$$Specificity = \frac{TN}{FP + TN} \tag{6}$$

AUC is calculated as the area under the curve of *sensitivity* (the true positive rate), and *(1-specificity)* (the false positive rate). *AUC* therefore ranges between 0 and 1.0, with 0 indicating zero correct predictions and 1.0 indicating only correct predictions.

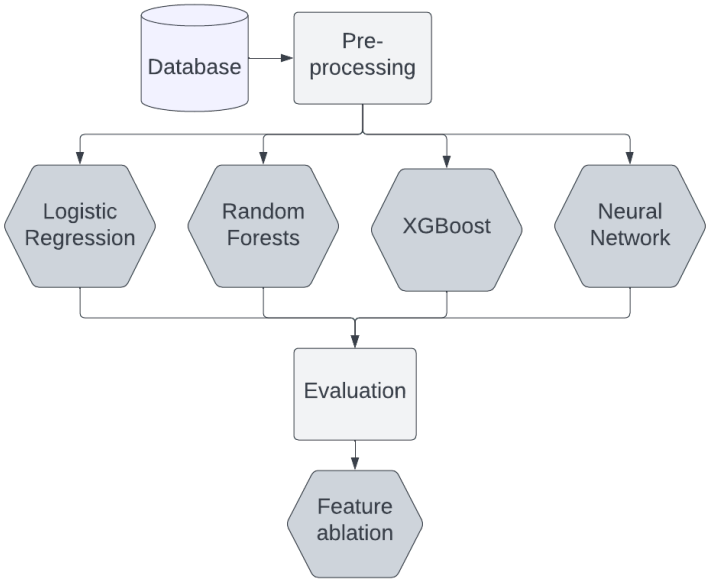


Figure 1: Pipeline of the current research

4.4 *Workflow*

Figure 1 illustrates the steps that have been followed throughout this research. As seen in the flowchart, the pipeline starts with the obtained, processed data. After the pre-processing steps described in Subsection 4.1, four models are built, and hyperparameter optimization have been performed on them. The results of these models are evaluated based on the metrics mentioned above. Feature ablation is then performed on the model with the highest evaluation scores.

4.5 *Sotfware and Algorithms*

This research was conducted in Python (version 3.6) using Google Colab notebooks and its basic 12GB RAM. For basic exploratory data analysis, *pandas*, *numpy*, and *statistics* were used. For model implementation, the following libraries were used: *sklearn*, *tensorflow*, *keras* and *xgboost*. For visualization and creating plots, *matplotlib*, *seaborn* and *ann_visualizer* were used.

Table 2: Model performance with respect to F_1 and AUC scores

	Performance	
Model	F_1	AUC
Logistic Regression	.544	.562
Random Forests	.632	.623
XGBoost	**.732**	**.706**
Neural Network	.698	.69

5 RESULTS

In the *Results* section, the findings of this thesis will be presented. This section is divided into *Model comparison* (further divided into *Performance* and *Error analysis* subsections) and *Feature importance*.

5.1 *Model comparison*

The following subsections give an overview of the performance of each model, with the error analysis giving an insight into the limitations of the models.

5.1.1 *Performance*

In this subsection, the performance of the models described in Section 3 and Subsection 4.2 are presented. The results of the gender prediction can be seen in Table 2. For the F_1 score and the *Area Under the Curve*, the Multivariable Logistic Regression achieved a score of 0.544 and 0.562, Random Forests achieved 0.632 and 0.623, XGBoost achieved 0.732 and 0.706 and the Neural Network achieved 0.698 and 0.69, respectively.

All models performed over chance level for predicting gender, as can be seen in the scores in Figure 2. For both evaluation metrics, Random Forests, XGBoost and the Neural Network all outperformed the baseline, which was set by Logistic Regression (F_1 being 0.544 and AUC being 0.562). XGBoost gave the best performance in terms of F_1 (0.732), as well as AUC (0.706) for predicting the gender of participants based on cognitive task performance.

It can be seen, that while XGBoost showed best overall performance, it has a relatively big difference between its two evaluation metric scores. For other models, the average difference between F_1 and AUC is 0.012, while for XGBoost it is 0.026. This can result from fact that the model is slightly biased towards the majority class. Further discussion on this can be found in Subsection 5.1.2.

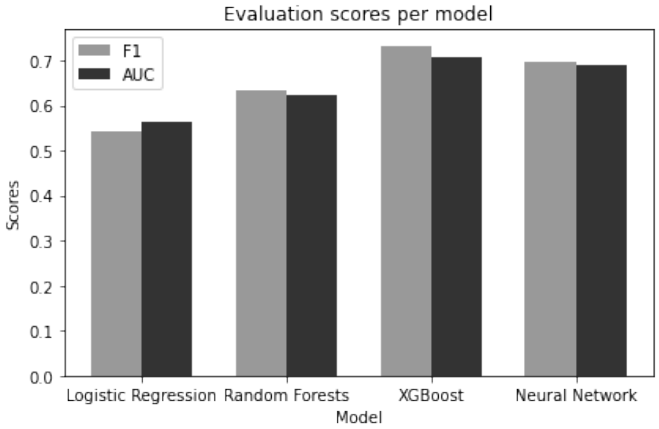


Figure 2: Performance per model

The Neural Network and Random Forests both showed good results as expected (for the F_1 score they are 0.698 and 0.632, and for the AUC score they are 0.69 and 0.623, respectively), with the former performing moderately better.

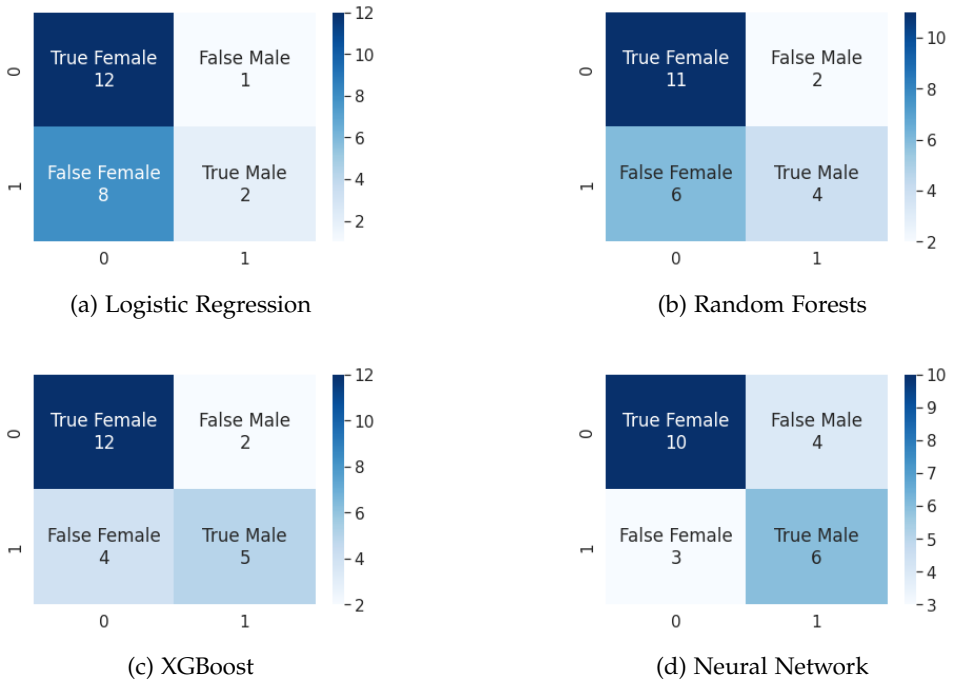
5.1.2 *Error analysis*

Analysing the confusion matrices for each model gave a better insight into the classes where the models performed better or worse. Given the class imbalance, it is important to assess each class and the distribution of the predictions.

The confusion matrices shown in Figure 3 provide an overview of the counts for true positive (*True Female*), false positive (*False Female*), true negative (*True Male*) and false negative (*False Male*) predictions. On both axes, 0 is representing females, while 1 is representing males. Along the X axis are the predicted labels, with the ground truth labels being on the Y axis.

Based on Figure 3 we can state, that the Neural Network gave the best prediction for males, identifying six out of nine of them. It also predicted the highest amount of males of all models, having a total number of 10 male predictions. Logistic Regression got the most true positives (correctly identified females), but did a poor job at identifying males. Random Forests also predicted females better than males, having 11 true positives and 4 true negatives, yet giving better performance than Logistic Regression. XGBoost classified 12 females and 5 males correctly.

For all models, *sensitivity* (also referred to as *precision* or true positive rate (TPR)) and $specificity$ (also referred to as false positive rate (FPR)) have been calculated. The equation for *sensitivity* and $specificity$ can



(a) Logistic Regression

(b) Random Forests

(c) XGBoost

(d) Neural Network

Figure 3: Confusion matrix of each model's performance

be seen in Equation 5 and Equation 6, respectively. Table 3 shows that Logistic Regression had the highest true positive rate, in combination with the lowest true negative rate. It has an outstanding 0.92 for *sensitivity*, and a very poor 0.2 for *specificity*. As seen from both Figure 3 and Table 3, *sensitivity* is relatively high for all four models, meaning that they all achieved good results in predicting the positive class (females). The big difference in *sensitivity* and *specificity* shows that the models are moderately biased towards the majority class, being females, and do poorer job at predicting the minority class, being males. The Neural Network showed the most balanced *specificity* and *sensitivity* with 0.67 and 0.71, respectively, thus showing the least bias towards class imbalance. The best performing model, XBoost, showed relatively good *specificity* (0.56) compared to Logistic Regression and Random Forests, and moderately high *sensitivity* (0.86). It is important to note, that given the small size of the dataset, and hence the limited number of observations in the test data, misclassification of even one instance can lead to big changes in the evaluation metrics.

Table 3: Sensitivity and Specificity scores for each model

Model	Sensitivity	Specificity
Logistic Regression	0.92	0.2
Random Forests	0.85	0.4
XGBoost	0.86	0.56
Neural Network	0.71	0.67

Table 4: Top 5 average F_1 score decrease per test

Test name	F_1 decrease
Phrase generation	**0.18**
Visual choice reaction time test	0.16
Word monitoring in noise	0.13
Auditory simple reaction time test	0.12
Corsi block clicking test	0.11

5.2 *Feature importance*

In this subsection, the results of the feature ablation will be presented. The feature ablation process (as described in Section 3) was conducted on the best performing model, XGBoost. The top five tests that had the highest predictive power over gender in XGBoost can be seen in Table 4.

Table 4 also shows the corresponding decrease in weighted F_1 score when the test-pair is excluded from the model. The highest decrease, and thus the test with the highest predictive power, was the *phrase generation* test (decrease of 0.18), followed by the *visual choice reaction time* test (decrease of 0.16) and the *word monitoring in noise* test (decrease of 0.13). The *phrase generation* test is part of the sentence production measure, the *visual choice reaction time* test is part of the visual reaction time measure, and the *word monitoring in noise* test is part of the word comprehension measure.

Further analysis was conducted into the relationship between the tests with the highest predictive power, and the genders. It was found that among the five most predictive measures, averaged across the two test days, females performed better at the *visual choice reaction time* test and the *word monitoring in noise* test, while males performed better at the *phrase generation* test, the *auditory simple reaction time* test and the *Corsi block clicking* test. The description of these tests can be seen in Table 5, together with *Gender* indicating the gender that performed better on the corresponding test.

Table 5: Description of the five tests with the highest predictive power

Test name	Test description	Gender
Phrase generation	Participants had to name the object(s) on pictures. Performance is measured by the average reaction time of correct answers.	Male
Visual choice reaction time	Participants had to press a button corresponding to the shape displayed on the screen. Performance is measured by the average reaction time of correct answers.	Female
Word monitoring in noise	Participants had to recognise words that were identical or related to a cue word given prior. Performance is measured by the proportion of correct answers to false positives.	Female
Auditory simple reaction time	Participants had to press a button as soon as they have heard the tone. Performance is measured by the average reaction time.	Male
Corsi block clicking test	Participants had to repeat the sequence of blocks lighting up, in a forwards or backwards order. Performance is measured by the sum of correct answers.	Male

6 DISCUSSION

6.1 *Interpretation of the results*

This thesis aims to answer the following question: *"To what extent is it possible to predict the gender of participants based on their performance in cognitive tasks?"*. It expands on previous studies by comparing a set of machine learning algorithms on the problem of gender prediction based on cognitive task performance. Findings show, that by using state-of-the-art machine learning algorithms it is possible to predict a participant being either male or female (weighted F_1 score of 0.732). These findings can contribute to research into profiling and general gender prediction tasks, alike Alantie et al. (2022); Rangel and Rosso (2013) and van der Goot et al. (2018).

Regarding the first research sub-question, *"Which of a set of pre-selected machine learning algorithms provides optimal predictive power for the gender of participants?"*, XGBoost has proven to perform well on gender prediction, in line with findings by Dehzangi et al. (2021). It has outperformed the

Multivariable Linear Regression, Random Forests and the Neural Network. Comparison of state-of-the-art models on cognitive skill performance based gender prediction has not been performed before. Comparing state-of-the-art methods in such a way can highlight the strength of new and existing models for specific tasks, and give an overview and base for future research.

As seen in previous studies, researchers tend to focus on only a small subset of cognitive skills to compare the performance of females and males, as seen in Mefoh et al. (2017) and Roberts and Bell (2003). As stated in the second research sub-question, *"Which cognitive tasks offer predictive power for a participant's gender?"*, this current study aims to identify a subset of cognitive tasks which have predictive power for gender. The results show, that the *phrase generation* test, the *visual choice reaction time* test and the *word monitoring in noise* test are the three most important features in the best performing XGBoost model. Results further show, that out of the five most important features in the model, females performed better at the *visual choice reaction time* test and the *word monitoring in noise test*, while males performed better at *phrase generation*, the *auditory simple reaction time* test and the *Corsi block clicking* test. The fact that males performed better at the *Corsi block clicking* test task is in line with findings by Mefoh et al. (2017) and Roberts and Bell (2003) that males perform generally better at pattern based tasks.

The novelty of the current study lies in the fact that, given the wide range of tasks tested in the dataset, these findings give insight into a more comprehensive overview of the different cognitive skill performance of different genders.

The societal relevance of this study includes its contribution to gender prediction and to the understanding of language comprehension and processing. As seen in the literature, there is a general idea of males outperforming females in specific cognitive tasks. (Brañas-Garza et al., 2019; Mefoh et al., 2017) This stereotype not only may not be true, but can also carry harmful effect for females' performance, as seen in Inzlicht and Ben-Zeev (2000). Highlighting the sets of tasks that each gender is generally better at can help break down such stereotypes. These sets of tasks can be used for educational purposes and for helping the development of cognitive skills for children. Professionals can use specific tasks that the child's gender is generally better at to boost their confidence. On the other hand, they can refer to tasks at which the child's gender performs worse on average, in order to further develop those specific skills.

6.2 *Limitations and Future research*

The generalisability of the results of this thesis is limited by the number of observations in the dataset, as well as the distribution of the target variable. With limited number of participants, the gender imbalance could influence interpretability of the results. To address the imbalance, appropriate evaluation metrics have been chosen in this study, providing an accurate interpretation of the robustness of the results for both genders. Due to the limited number of observations in the test set, one single misclassification can have a big impact on the evaluation metrics, hence the results are less reliable. To overcome this issue, future studies are advised to experiment with upsampling techniques to gain a larger, more balanced dataset.

Regarding generalisation of the results based on the demographics of the participants, the dataset includes individuals with both university and vocational college background, with one participant being a high-school graduate. This distribution of educational level gives better general results for young adults, however, it is still a limited range for age. The results hence may not generalise well to the middle age adults. For a follow-up study, it could lead to interesting findings to include participants from lower educational level, as well as individuals of older age. Another limitation based on the demographics of the participants is that evidently all participants are native Dutch speakers. Accordingly, the generalisation is limited to Dutch young adults and may not be extended to natives of other languages.

Future research on gender identity can also expand upon results found in this study. Current study was conducted on a dataset containing a binary input for gender: male and female. Researchers could consider moving towards non-binary labels for gender and conduct experiments with participants from a wider spectrum.

Additionally, while this dataset excluded personal information of the participants other than gender, the educational level of an individual could show importance in the analysis. Conducting research with more demographic information on the individuals, such as their education level or income, and an extended variety of cognitive skills could potentially yield new and interesting findings.

7 CONCLUSION

This research aimed to identify the optimal machine learning algorithm for gender prediction based on cognitive task performances and highlighted a subset of tasks with highest predictive power. By comparing a set of machine learning models, the results of this thesis had shown that XGBoost

outperformed Linear Regression, Random Forests and the Neural Network in cognitive task performance based gender prediction.

Conducting further analysis, it was concluded that in the best performing model, being XGBoost, the cognitive tasks that had the highest feature importance and thus the most predictive power on gender were the *phrase generation* task, the *visual choice reaction time* test and the *word monitoring in noise* test. Results further indicated, that from the five most predictive tasks, females outperformed males in the *visual choice reaction time* test and the *word monitoring in noise* test, while males performed better at the *phrase generation* test, *auditory simple reaction time* test and the *Corsi block clicking* test.

These findings can contribute to the general understanding of gender differences in language producing and processing, and it can help develop better methods for improving the cognitive development of young individuals.

REFERENCES

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., & Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. *Heliyon*, *4*(11), e00938.

Alantie, S., Tyrkkö, J., Makkonen, T., & Renvall, K. (2022). Is old age just a number in language skills? language performance and its relation to age, education, gender, cognitive screening, and dentition in very old finnish speakers. *Journal of Speech, Language, and Hearing Research*, *65*(1), 274-291.

Baker, D. P., & Jones, D. P. (1993). Creating gender equality: Cross-national gender stratification and mathematical performance. *Sociology of education*, 91–103.

Berkelmans, G. F., Read, S. H., Gudbjörnsdottir, S., Wild, S. H., Franzen, S., van der Graaf, Y., ... Dorresteijn, J. A. (2022). Population median imputation was noninferior to complex approaches for imputing missing values in cardiovascular prediction models in clinical practice. *Journal of Clinical Epidemiology*, *145*, 70–80.

Brañas-Garza, P., Kujal, P., & Lenkei, B. (2019). Cognitive reflection test: Whom, how, when. *Journal of Behavioral and Experimental Economics*, *82*, 101455.

Buss, D. M. (1995). Psychological sex differences: Origins through sexual selection.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in cognitive sciences*, *23*(4), 305–317.

Dehzangi, O., Shokouhmand, A., Jeihouni, P., Ramadan, J., Finomore, V., Nasrabadi, N. M., & Rezai, A. (2021). Xgboost to interpret the opioid patients' state based on cognitive and physiological measures. In *2020 25th international conference on pattern recognition (icpr)* (pp. 6391–6395).

Deitrick, W., Miller, Z., Valyou, B., Dickinson, B., Munson, T., & Hu, W. (2012). Author gender prediction in an email stream using neural networks.

Duffy, J., Gunther, G., & Walters, L. (1997). Gender and mathematical problem solving. *Sex roles*, *37*(7), 477–494.

Ebrahimi Kalan, M., Jebai, R., Zarafshan, E., & Bursac, Z. (2021). Distinction between two statistical terms: multivariable and multivariate logistic regression. *Nicotine and Tobacco Research*, *23*(8), 1446–1447.

Feingold, A. (1988). Cognitive gender differences are disappearing. *Ameri-*

can Psychologist, *43*(2), 95.

Feng, S., Zhou, H., & Dong, H. (2019). Using deep neural network with small dataset to predict material defects. *Materials & Design*, *162*, 300–310.

Fraser, K. C., Hirst, G., Graham, N., Meltzer, J. A., Black, S. E., & Rochon, E. (2014). Comparison of different feature sets for identification of variants in progressive aphasia. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality* (pp. 17–26).

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, *38*(4), 367–378.

Gallagher, A. M., & De Lisi, R. (1994). Gender differences in scholastic aptitude test: Mathematics problem solving among high-ability students. *Journal of Educational Psychology*, *86*(2), 204.

Hintz, F., Dijkhuis, M., van't Hoff, V., McQueen, J. M., & Meyer, A. S. (2020). A behavioural dataset for studying individual differences in language skills. *Scientific data*, *7*(1), 1–18.

Huang, J., & Ling, C. X. (2005). Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, *17*(3), 299–310.

Inzlicht, M., & Ben-Zeev, T. (2000). A threatening intellectual environment: Why females are susceptible to experiencing problem-solving deficits in the presence of males. *Psychological science*, *11*(5), 365–371.

Jongman, S. R., Roelofs, A., & Meyer, A. S. (2015). Sustained attention in language production: An individual differences investigation. *Quarterly Journal of Experimental Psychology*, *68*(4), 710–730.

Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual differences in language acquisition and processing. *Trends in cognitive sciences*, *22*(2), 154–169.

Kiefer, A. K., & Sekaquaptewa, D. (2007). Implicit stereotypes, gender identification, and math-related outcomes: A prospective study of female college students. *Psychological Science*, *18*(1), 13–18.

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: a meta-analysis. *Psychological bulletin*, *136*(6), 1123.

McCarthy, M. M., Arnold, A. P., Ball, G. F., Blaustein, J. D., & De Vries, G. J. (2012). Sex differences in the brain: the not so inconvenient truth. *Journal of Neuroscience*, *32*(7), 2241–2247.

Mefoh, P. C., Nwoke, M. B., Chukwuorji, J. C., & Chijioke, A. O. (2017). Effect of cognitive style and gender on adolescents' problem solving ability. *Thinking Skills and Creativity*, *25*, 47–52.

Morgan-Lopez, A. A., Kim, A. E., Chew, R. F., & Ruddle, P. (2017). Pre-

dicting age groups of twitter users based on language and metadata features. *PloS one*, *12*(8), e0183537.

Nielsen, D. (2016). *Tree boosting with xgboost-why does xgboost win" every" machine learning competition?* (Unpublished master's thesis). NTNU.

Putatunda, S., & Rama, K. (2018). A comparative analysis of hyperopt as against other approaches for hyper-parameter optimization of xgboost. In *Proceedings of the 2018 international conference on signal processing and machine learning* (pp. 6–10).

Rafique, I., Hamid, A., Naseer, S., Asad, M., Awais, M., & Yasir, T. (2019). Age and gender prediction using deep convolutional neural networks. In *2019 international conference on innovative computing (icic)* (pp. 1–6).

Raghunadha Reddy, T., Vishnu Vardhan, B., GopiChand, M., & Karunakar, K. (2018). Gender prediction in author profiling using relieff feature selection algorithm. In *Intelligent engineering informatics* (pp. 169–176). Springer.

Rangel, F., & Rosso, P. (2013). Use of language and author profiling: Identification of gender and age. *Natural Language Processing and Cognitive Science*, *177*.

Roberts, J. E., & Bell, M. A. (2003). Two-and three-dimensional mental rotation tasks lead to different parietal laterality for men and women. *International Journal of Psychophysiology*, *50*(3), 235–246.

Schmader, T. (2002). Gender identification moderates stereotype threat effects on women's math performance. *Journal of experimental social psychology*, *38*(2), 194–201.

Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*.

van der Goot, R., Ljubešić, N., Matroos, I., Nissim, M., & Plank, B. (2018). Bleaching text: Abstract features for cross-lingual gender prediction. *arXiv preprint arXiv:1805.03122*.

Wang, S.-C. (2003). Artificial neural network. In *Interdisciplinary computing in java programming* (pp. 81–100). Springer.

Weber, D., Skirbekk, V., Freund, I., & Herlitz, A. (2014). The changing face of cognitive gender differences in europe. *Proceedings of the National Academy of Sciences*, *111*(32), 11673–11678.

Weiss, E., Siedentopf, C., Hofer, A., Deisenhammer, E., Hoptman, M., Kremser, C., ... Delazer, M. (2003). Sex differences in brain activation pattern during a visuospatial cognitive task: a functional magnetic resonance imaging study in healthy volunteers. *Neuroscience letters*, *344*(3), 169–172.

Zhou, Z.-H. (2012). *Ensemble methods: foundations and algorithms*. CRC press.

APPENDIX A

Feature name	Mean	SD	Missing values
Gender	-	-	0
X2_Peabody_picture_vocabulary_test	56.11	24.56	0
X2_Peabody_picture_vocabulary_test_Session2	60.35	23.28	0
X3_Spelling_test	0.56	0.18	0
X3_Spelling_test_Session2	0.55	0.17	0
X4_Author_recognition_test	0.2	0.12	0
X4_Author_recognition_test_Session2	0.2	0.12	0
X5_Idiom_recognition_test	0.76	0.13	0
X5_Idiom_recognition_test_Session2	0.75	0.13	0
X6_Prescriptive_grammar_test	0.69	0.13	0
X6_Prescriptive_grammar_test_Session2	0.7	0.13	0
X8_Auditory_simple_reaction_time_test	-2.35	0.08	0
X8_Auditory_simple_reaction_time_test_Session2	-2.36	0.09	0
X9_Auditory_choice_reaction_time_test	-2.6	0.09	0
X9_Auditory_choice_reaction_time_test_Session2	-2.6	0.11	0
X10_Letter_comparison_test	-3.02	0.08	5
X10_Letter_comparison_test_Session2	-2.88	0.63	0
X11_Visual_simple_reaction_time_test	-2.37	0.05	0
X11_Visual_simple_reaction_time_test_Session2	-2.39	0.07	0
X12_Visual_choice_reaction_time_test	-2.62	0.07	0
X12_Visual_choice_reaction_time_test_Session2	-2.63	0.09	0
X13_Digit_span_test_Forward	7.95	2.06	0
X13_Digit_span_test_Forward_Session2	8.53	1.95	0
X13_Digit_span_test_Backward	6.75	2.12	2
X13_Digit_span_test_Backward_Session2	7.17	2.56	2
X14_Corsi_block_clicking_test_Forward	8.01	1.59	1
X14_Corsi_block_clicking_test_Forward_Session2	8.61	1.95	1
X14_Corsi_block_clicking_test_Backward	7.27	2.02	4
X14_Corsi_block_clicking_test_Backward_Session2	7.48	1.94	4
X15_Eriksen_Flanker_test	0.09	0.03	6
X15_Eriksen_Flanker_test_Session2	0.09	0.03	6
X16_Antisaccade_test	0.89	0.1	1
X16_Antisaccade_test_Session2	0.9	0.09	1
X17_Raven.s_advanced_progressive_matrices_test	0.55	0.17	0
X17_Raven.s_advanced_progressive_matrices_test_Session2	0.6	0.2	0
X18_Picture_naming_test	-2.95	0.06	1
X18_Picture_naming_test_Session2	-2.92	0.05	1

Appendix A continued from previous page

Feature name	Mean	SD	Missing values
X20_Antonym_production	0.72	0.1	1
X20_Antonym_production_Session2	0.75	0.1	1
X21_Verbal_fluency_Categories	24.37	4.94	6
X21_Verbal_fluency_Categories_Session2	26.58	5.2	6
X21_Verbal_fluency_Phonology	15.55	4.43	0
X21_Verbal_fluency_Phonology_Session2	17.38	4.9	0
X22_Maximal_speech_rate	-3.6	0.09	6
X22_Maximal_speech_rate_Session2	-3.58	0.09	4
X23_One.minute.test	89.8	14.34	1
X23_One.minute.test_Session2	100.05	12.23	1
X24_Klepel_test	62.8	12.08	1
X24_Klepel_test_Session2	66.5	12.92	1
X25_Non.word_monitoring_in_noise	0.58	0.17	0
X25_Non.word_monitoring_in_noise_Session2	0.61	0.17	0
X25_Word_monitoring_in_noise	0.83	0.12	0
X25_Word_monitoring_in_noise_Session2	0.82	0.15	0
X25_Meaning_monitoring_in_noise	0.3	0.19	3
X25_Meaning_monitoring_in_noise_Session2	0.31	0.21	3
X26_Rhyme_judgment	-2.89	0.08	3
X26_Rhyme_judgment_Session2	-2.89	0.09	3
X27_Auditory_lexical_decision	-2.93	0.05	0
X27_Auditory_lexical_decision_Session2	-2.93	0.05	0
X28_Semantic_categorization	-2.92	0.06	3
X28_Semantic_categorization_Session2	-2.92	0.06	3
X29_Phrase_generation	-2.86	0.07	0
X29_Phrase_generation_Session2	-2.85	0.07	0
X29_Sentence_generation	0.79	0.19	0
X29_Sentence_generation_Session2	0.86	0.17	0
X31_Gender_cue_activation_during_ sentence_comprehension	587.81	654.53	7
X31_Gender_cue_activation_during_ sentence_comprehension_Session2	773.16	668.7	7
X32_Verb_semantics_activation_during_ sentence_comprehension	742.13	672.59	0
X32_Verb_semantics_activation_during_ sentence_comprehension_Session2	1028.54	653.01	0
X33_Monitoring_in_noise_in_sentences	0.09	0.13	0
X33_Monitoring_in_noise_in_sentences_Session2	0.11	0.14	0