Utrecht University

Classifying Sex Based On Eye Tracking Data: A Machine Learning Study

Bachelor Thesis (7.5 ECTS)

Selim Büyük (5959519)

First Supervisor

Christoph Strauch

Second Supervisor

Jakub Dotlacil

Utrecht, July 1st 2021

Word Count: 5195

BSc Artificial Intelligence
Faculty of Humanities

TABLE OF CONTENTS

LIST OF TABLES

Abstract

This paper describes a study into the classification of gender based on viewing behavior. This was done with the data of 1242 visitors of the NEMO museum, to which we had to pick a classification algorithm and decide on what features to use with this algorithm to train and test our given data with. We evaluated the algorithm based on multiple machine learning measures, such as Precision, Recall and F1-score, but the most important measure, which was also the measure we were basing our evaluation on, was the Accuracy measure. Our criteria for a good algorithm was set to 70%, which was based on related work. Our algorithm with the implemented feature set got exactly that as Accuracy, to which we can conclude that it is indeed possible to program an algorithm that can correctly classify sex based on eye tracking data. This has a few implications: by further analysing eye tracking data and successfully furthering algorithms to also correctly classify variables such as age and mood of a person, we can predict the way people are going to behave and make things such as advertisements more effective.

Keywords: machine learning; sex classification; eye tracking; support vector machines

Introduction

Throughout a person's life, there are a few actions they will be doing from the day they are born until the day they are gone. Within these actions, there is an action that is often performed unconsciously and yet one of the most important actions a person can perform. An action which is, on average, performed three times a second (Moss et al., 2012). This action is deciding where to look, which is also a central part of this thesis.

As is common with all the actions a person can perform, two different persons can perform these actions completely differently, even when the same stimulus is considered for both. This is also the case for looking behaviour. For example, when the same picture to look at for people with autism is presented, a difference in gaze can be seen when this is related back to people without autism (Leekam et al., 1998). This is also true for different ages (Gomez et al., 2019; Nikitin and Freund, 2011), differing cultures (Lee et al., 2016) but also between the two sexes (Moss et al., 2012). These studies all use things that can help differentiate where a person will (but also won't) look, which are called features. These features can consist of the shape of objects, colours of objects or even places of the objects that a person may or may not look at.

Reversing this logic, gaze behavior could be conclusive about the person of whose eyes are observed. Since eye tracking became more widely accessible and the technology more accurate, its usage within machine learning algorithms also increased significantly (Kredel et al., 2017). This increase can be directly translated to the advancements in eye tracking technology, as machine learning algorithm rely heavily on the data; having better technology directly reflects on the quality of the data and therefore also on the performance of an algorithm. The aforementioned logic of gaze behaviour being conclusive about the person of whose eyes are observed can therefore be used by machine learning classifiers to try to decode properties of the beholder.

On top of the quality of the devices used, there is another property of good data, which is the amount of data available (Recchia & Jones, 2009). A bigger dataset is almost always better for machine learning purposes, as there is more to train on/learn from (Halevy et al., 2009; Sun et al., 2017). However, bigger datasets are harder to gather for eye tracking purposes as assessing participants in eye tracking is usually time- and cost expensive because of specialised hardware, which results in only a few large scale datasets (Xu et al., 2015).

Since the aforementioned advancement of eye tracking technology, there have been a multitude of articles proposing machine learning techniques to be used on eye tracking data that have yet to be implemented (Al-Rahayfeh & Feazipour, 2013; King et al., 2020; Pierdicca et al., 2020). Additionally, there have been a few researchers that have been successful in creating other algorithms used for eye tracking data, papers from Matsumoto et al. (2017), Bozkir et al. (2020) and David-John et al., (2021) to name a couple. A few of the used algorithms in these articles include the Support Vector Machines and Random Forest Classifiers.

My goal for this thesis was to utilise eye tracking features on the proposed machine learning algorithms named above to correctly distinguish between sexes, as I had been given access to one of the large scale datasets - which was data from 1000+ participants that were tracked for just ten seconds and looked at only one stimulus. Concretely, in this thesis I investigated the following research question: "Is it possible to program an algorithm such that it can correctly classify sex based on eye tracking data?". I answered this question by evaluating my model with standard machine learning metrics, such as Accuracy, Recall, F1-measure and others.

By successfully creating an algorithm that can classify sex based on eye tracking data, this thesis will not only prove that it is possible to correctly classify based on previous works (such as features and recommended machine learning algorithms named earlier), but also layout implications for further work. These implications consist of the knowledge this thesis gains and which can be used for further research so that age can be correctly classified, and if that is possible, classify a person's interests, mood and even mental state. An instance of the above would be for personalised advertisements, as knowledge on a person's sex and age can lead to one advertisement being more successful into persuading the customer into buying the product than another advertisement (Bourreau et al., 2017; Mogaji et al., 2020; Shannon et al., 2009). Even so, if this thesis fails to create such an algorithm, the proposed algorithms from articles that mentioned that these algorithms would be successful in correctly classifying sex, such as the article from Moghaddam and Yang (2000), could be rejected for algorithms that use datasets such as mine (i.e datasets that were obtained outside the lab and with only short viewing times per participant).

In order to answer the research question, two subquestions had to be answered first. These questions pertained the selection of a machine learning algorithm, more specifically;

1. What algorithms can be used for the classification on sex, and on what basis will I pick an algorithm?

2. What are (good) features my algorithm can use to correctly classify sex?

These subquestions were answered in the Background section, followed by the Methods section in which the algorithm, as well as the methods for the experiment that resulted in the big eye tracking dataset, are described. In the Results section, I will go over the results I had obtained and subsequently going over these and discuss among others, the implications of these results in the Discussion section.

Background

To correctly classify sex based on eye tracking data, a classification algorithm has to be picked. Subsequently, features that will be implemented have to be picked. Finally, thresholds for performance have to be set. All of the above will be discussed in this section and will be based on and argued by literature, starting with the classification algorithm.

Algorithm choice

To pick an algorithm means to correctly identify what problem you want to solve with the dataset you have. Because the goal of this thesis is classification, I will pick between classification algorithms. Since I have a labeled dataset with discrete values (datapoints with their correct labels of Male/Female), I will look specifically at the supervised classification algorithms that can handle these values. Thus I will not look at algorithms such as Regression as these methods are most often used for continuous variables, which our dataset labels do not fall under (Altman & Royston, 2006).

The most noteworthy supervised classification algorithms consist of Support Vector Machines, Random Forest Classifiers and Naive Bayes Classifiers (Kotsiantis et al., 2009; Osisanwo et al., 2017; Sen et al., 2020). Each classifier has its pros-and-cons, picking a specific classifier will also rely on these pros-and-cons. By going over each of the algorithms' pros-and-cons we can deliberate over the best choice for our dataset, starting with the Naive Bayes Classifier.

Naive Bayes (NB) Classifiers are statistical in nature, that is to say that they predict the class membership probability of a given input sample. The class membership probability is the probability that the given input sample belongs to a particular class, in our case the Male or Female class. NBs have the advantage that they are fast in training, can be applied to large datasets and are robust but have the downside that they are less accurate compared to other classifiers (Bhavsar & Ganatra, 2012; Friedman et al., 1997).

Second the Random Forest (RF) Classifiers, a classification algorithm that is defined as an algorithm that assembles decision trees while training and results in the majority of the classes of all the trees as output (Chen et al., 2017; Kohestani et al., 2015). As for the advantages of RF's, Gupte et al., (2014) noted that RF's have a high accuracy, high performance, are simple to build and can be used in a variety of applications. Additionally, Zafari et al. (2019) noted that RFs are also known for their reduced sensitivity for overfitting. Yet using RFs on big datasets is not advised, as Genuer et al., (2017) noted that RFs are all computationally expensive when making deep trees and are as such difficult to use for prediction.

Finally the Support Vector Machines, a classification algorithm that it is one of the most powerful training techniques for supervised learning, as stated by Mohamed (2017). SVMs can be defined by Noble (2006) as the following; "a mathematical entity, an algorithm (or recipe) for maximizing a particular mathematical function with respect to a given collection of data". SVM's were evaluated as the best classification algorithm for Accuracy in general by Bhsavar & Ganatra (2012). Furthermore, SVM's can be tuned so that they can solve regression problems, with the pros (and cons) of that of the SVM's as Gunn (1998) and Brereton & Lloyd (2010) have noted. Lastly, SVM's are also excellent tools for binary classification (classification of two classes, in our case Male and Female), as Shao et al. (2014) mention in their article. However, SVM's biggest downside is its interpretability (Martin-Barragan, 2014). Especially in higher dimensional features, interpreting the results and looking at what features are better than others are more difficult (Nalbantov et al., 2006).

To conclude, RFs and NBs are not suited well for this thesis as we have a particularly large dataset, which the RFs don't perform well on, and are measuring with the Accuracy metric, which the NBs don't perform as well as the other algorithms on. Hence, a SVM classification algorithm will be used because SVMs have the highest accuracy of the three classification algorithms and are the best when considering binary classification - what our dataset has.

Feature choice

A good algorithm choice is not the only aspect that matters when creating a good classifier, feature selection is just as important. Features are capable of improving learning performance, lowering computational complexity, building better generalizable models, and decreasing required storage (Tang et al., 2014). We define features as a characteristic or a measure of a phenomenon (Bishop, 2006). For example, features for facial recognition are among other things; eyes, nose, mouth and head outlines (Chen & Wenkins, 2017).

To pick features we can implement, we can look at literature in which there have already been differences noted on males and females. By looking at these differences, some features can be made to underline these in the algorithm and hopefully make a better classification algorithm. One of these differences is the visual processing of images, as Cazzato et al. (2010) noted. Cazatto et al. mention that males have a shorter execution time and a higher path length compared to females. Other articles such as Liu et al. (2020) note that females' gaze patterns are more spatially distributed, meaning that they have more saccades. Saccades are defined as the voluntary rapid eye movement between fixations, while fixations are defined as a comparatively steady state of eye movement - essentially a pause over a region of interest (Rucker et al., 2021; Salvucci & Goldberg, 2000). This difference in saccades between males and females is underlined in countless articles, such as Heisz et al. (2013), Hwang & Lee (2018) and Meyers-Levy & Loken (2015), making it an important aspect of eye tracking research.

On top of the general differences between males and females mentioned above, we can also look at existing literature of classification algorithms that have researched eye tracking/gaze. For example, the article from Sargezeh et al. (2019) uses features such as fixation duration, saccade duration, minimum fixation duration, amplitudes of saccades (length of saccades) and path length. Other research, such as articles from Sammaknejad et al. (2017), Pérez-Moreno et al. (2016), and Emam & Youssef (2012) also underline the differences between fixations and saccades between males and females, making it one of the most important features I can implement in my own algorithm.

By using both the literature on differences between males and females and analysing the features already present in gaze classifiers, I can answer my second subquestion. To conclude, I will focus on the differences in fixation, saccades, path length, and amplitudes as features to classify sex as these are the most recurring features in eye tracking research for sex classification.

Threshold choices

Finally, we have to decide on thresholds on when the classification algorithm can be evaluated as "good". Since thresholds for machine learning metrics differ in certain applications, as for example, threshold for Accuracy in classifying a cancer cell wrongly versus threshold for classifying a colour wrongly are different from each other, I decided to make the threshold to be "correct/good" if the accuracy metric reached 70% or more, based on Sargezeh et al. (2019) investigating a related task.

We are specifically interested in the Accuracy metric because Accuracy gauges the overall ability of a classifier in correct classification of samples. Other metrics such as Recall and Specifity on the other hand, illustrate the ability of a classifier in correct classification of positive and negative samples (True Positives and True Negatives), respectively (Baratloo et al., 2015). We will however also calculate these measures as it gives us more room to interpret the results.

On top of determining the threshold for evaluating the algorithm, we also have to set a threshold for deciding a train and test set. This threshold will be based on the research of Rácz et al. (2021) in which they researched the different Train/Test split ratios, especially when considering differing dataset sizes. The research suggested that the best Train/Test split ratio for larger datasets was the 80%/20% split ratio. This ratio will also be used in this thesis.

Methods

Participants

Participants were 1242 visitors of the NEMO Science Museum, Amsterdam, The Netherlands, 549 male, 614 female, ages from 11 to 59 years old. Furthermore, there were 79 participants that did not want to disclose their sex, these participants were excluded from the data. Lastly, the age groups 10, 20 and 60 were excluded as there were bugs with the implementation. The visitors of the NEMO Science museum participated voluntarily at an exhibit in one of the museum rooms and gave informed consent to the usage of their data for scientific purposes. An overview of the participants can be seen in Table 1.

Table 1: Demographics of participants.

Age (years)		*M*	*SD*	*n*	*%*
	Male	28.6	13.1	549	47.2
	Female	28.4	12.3	614	52.8
	Total	29.5	12.7	1163	100.0

Materials

In order to record participant's eye movements, a Tobii 4C eye tracker whose sampling rate was set to 60 Frames Per Second was used. There was a distance of 80 centimeters from the eyes to the screen for every participant. The screen itself was a 27 inch, 1920 x 1080 pixel and 16:9 ratio screen where the brightness was constant and controlled. The experiment, which was interactable software created by the NEMO museum in Amsterdam, was a single task. The stimuli used in this experiment was the same for all participants, which can be seen in Appendix A.

Procedure

All visitors of the NEMO museum could voluntarily participate in this experiment as multiple eye trackers were situated in one of the exhibition halls of the museum. Visitors that were interested had to answer two questions before they could start with the experiment, namely their sex (options consisting of Male - Neutral - Female) and their age. The visitors could look at the left side of the screen for the Male option, the middle for Neutral option and the right side for the Female option. As for their age, visitors looked at the value in the middle of the screen and added a year by looking to the left of the screen or subtracted a year to ultimately get the year of birth. After answering these questions, the visitors were asked to look at the screen where they could look at the picture in Appendix A for 10 seconds. After the ten seconds were over, the visitors would see visualizations of their gaze behavior and thereafter which areas they looked at more than other participants. These areas were highlighted by brightening the areas that were looked at more while the rest was dimmed. Subsequently, in video-form, a researcher asked participants whether they would agree in giving their data to science. Upon agreement, data gathered by the experiment including the demographics were saved and the visitors were deemed participants, otherwise the data from the experiment was absolved. Finally the participant was thanked and the next participant could start.

Measures

In this experiment, gaze position over time had been recorded along with the sex of the participant, the year of birth of the participant and the three most viewed areas of interest. To be more concrete, for every second,

we got 60 x- and 60 y-coordinates with their timestamp given by (yyyy:mm:dd:hh:mm:ss:SSS), which reflected the change of gaze position over time (because of the sampling rate of 60 Frames Per Second). For the whole experiment this resulted in 600 points for one participant.

Algorithm description

To answer the research question, we had to create an algorithm that could achieve an accuracy of 70% or more and thereby classify sex correctly. In the Background section we discussed among others the algorithm and features that we were going to use to achieve this. In this section I will discuss how the final algorithm had been implemented, as well as give an overview of the features used by the algorithm and the steps undertaken to do so.

The algorithm I had chosen for this classification was the Support Vector Machine algorithm. SMVs have a parameter that can be modified, which is the kernel to use. Kernels are defined as mathematical functions that can modify the data given as input to the desired form. There are a few kernels to chose from, among which are the linear-, polynomial-, sigmoid- and rbfkernels. Each of these kernels has a use case, but for this thesis and algorithm, I had chosen the rbfkernel. The rbfkernel is often the preferred type of kernel among many applications, as it is localized and has a finite response along the complete x-axis (Mezghani et al., 2010).

For evaluation purposes, in which I originally was planning to look at the Accuracy, as it gauges the overall ability of a classifier in correct classification of samples, I also added the Balanced Accuracy measure. This measure is used to make the algorithm more robust, specifically for imbalanced datasets. In training sets where there were more males than females, the normal Accuracy measure gives a skewed view while the Balanced Accuracy stabilises this (Broderson et al., 2010).

Furthermore, I added another step for the preprocessing, which was the scaling of the train and test data (specifically for the feature sets, as the labels did not have to get scaled). This was also done to make the algorithm more robust against imbalanced datasets; normalising - a variant of scaling - was used to achieve this. To be more precise, I used the min-max scaling as noted by Géron (2020), in which values are shifted and rescaled so that they end up ranging from 0 to 1. This is done by subtracting the min value from a feature and dividing by the max of the features minus the min of the features, which can be rewritten in equation form like:

$$x_{scaled} = \frac{x_i - min(x)}{max(x) - min(x)}.$$

As for the features, in the same matter noted in the Background section, I laid the attention to features concerning fixations and saccades. Analogous to these were the amount of saccades and fixations for each participant for example, but also the statistical measures - such as mean and standard deviation - for fixations and saccades. The fixations and saccades were not saved by the eye tracker but had to have been calculated and identified separately with the x and y points. The calculation of fixations and saccades was done with the Identification by two-means clustering (I2MC) algorithm as noted by Hessels et al (2017). This algorithm was chosen because the I2MC algorithm's output was the most robust against high noise, is automatic, works offline, and is suitable for eye tracking data recorded with remote or tower-mounted eye-trackers using static stimuli, which this experiment deals with.

Other features included age and their age category (Young - Old), whether or not the participant looked more to the left side of the screen rather than the right side of the screen and the top three most looked at areas per participant. For explanation, the top three most looked at areas had to be normalised too, which can not be done with discrete values. Therefore these features had to be transformed with help of a Label Encoder. This was done by numbering every discrete value with a specific unique number (for example, the Turtle area had been given the unique number 1, the Zeppelin area with 2 etc). By doing so, the algorithm could normalise and used these features as well.

Finally, to make research easier, I had made an application in which the SVM (and the other two named algorithms in the Background section, NB and RF) could be modified to work with the features I had implemented. All the features can be seen in the listbox as options and can even be turned off to see its impact on the overall (Balanced) Accuracy and other metrics, such as the Precision. In this application, the demographics can be shown too. Lastly, the application has a "Help" button to support the user in case of confusion. The Graphical User Interface (GUI) of this application can be seen in Appendix B.

Results

Before we could answer the research question, we answered two subquestions first. We deliberated on what algorithm we could use to classify sex and on what basis that decision would be made. Based on literature and earlier relatable work, we came to a conclusion that Support Vector Machines were the best algorithm for this specific dataset as they have the highest accuracy of the three classification algorithms and are the best when considering binary classification. Furthermore picked features that the algorithm could use, which was also done on the basis of literature. To be more precise, we looked at sex differences in gaze research and related those differences to the features already present with eye tracking research. By looking at both of these aspects, I laid my focus on the differences in fixation, saccades, path length, and amplitudes as features to classify sex.

We posed if it was possible to classify someones sex based on eye tracking data on just ten seconds of tracking with only one stimulus as our main research question. By answering our two subquestions, we made it possible to test this. The results can be seen in Table 2 and Table 3. Table 2 gives a rough idea on how many data points in the test set were classified correctly, while Table 3 gives more details about these classifications. For example, the Recall measure specifically tells us how many females had been correctly classified, divided by the number of classifications.

Finally, the measure we were most interested in, as it was the measure we were going to use to evaluate whether or not our algorithm was good, was the Accuracy measure. The Accuracy for this algorithm with the implemented feature set was 70%, exactly the minimum value of our criteria.

Table 2: Confusion matrix of test data.

n = 233	Predicted: Female	Predicted: Male
Actual: Female	97	29
Actual: Male	42	65

Table 3: Classification report of test data.

	Precision	Recall	F1-score
Female	0.70	0.77	0.73
Male	0.69	0.61	0.65

Discussion

In this paper, it was investigated whether sex could be classified from eye-tracking data using a machine learning algorithm. The data used in this paper was the eye tracking data of 1242 visitors of a museum that were tracked for just ten seconds and looked at only one stimulus. In this single task experiment, every gaze, along with the sex of the participant, the year of birth of the participant and the three most viewed areas of interest had been saved. Fixations and saccades, which were used to create features such as fixation amplitude, fixation length and more, were detected by using the I2MC algorithm. By doing so, these features could be used in the Support Vector Machine classifying algorithm, which resulted in a 70% Accuracy measure.

These results imply a few things, as achieving 70% on the Accuracy metric - by only having ten seconds of eye tracking data (600 datapoints per participant) with just one stimulus - means that it is possible to make an algorithm that can classify sex based on eye tracking data. The possibility of automatically classifying sex based on eye tracking data makes way for different aspects of person - such as age - to be classified too. By also accomplishing that, more information on a person just based on their gaze behaviour can be gathered, in which case things like personalised advertisements can be introduced, as knowledge on a person's sex and age can lead to one advertisement being more successful into persuading the customer into buying the product than another advertisement (Bourreau et al., 2017; Mogaji et al., 2020; Shannon et al., 2009). It also underlines previous work, such as different features used that have an impact on the classification and recommended machine learning algorithms, as previous research suggested that Support Vector Machines would be a good choice for classifying sex (Moghaddam & Yang 2000).

However there are also a few remarks to be made on the experiment and the algorithm. Firstly the experiment; Because the eye tracker had an unadjustable height, some visitors could not participate in the experiment as the height of the eyes needed to be about 114 cm. In other words, children that were smaller than 114 cm were having a hard time to look into the eye tracker. This might limit the lower and upper boundary regarding age due to body length. To add on top of that, because the eye tracker was situated in one of the exhibition halls in a museum without supervision, there could have been visitors that participated multiple times which would result in duplicate data. The reason that this is problematic is that participants that have seen the picture once (or more times) already, can look at totally different areas thenceforth. By doing so, the data of most looked areas (in the training- but also in the testset) would not be accurate, as the most looked areas don't accumulate over different trials.

As for the algorithm itself, there are also a few remarks. Generally your algorithm wants to have enough features so that the classification gets better but not too much that it overfits the data, the term used for this concept is the complexity of a model. The more features you have, the more complex your model is and therefore more likely to overfit the data (Yu et al., 2015). Because of this overfitting, the algorithm scores high on the machine learning metrics in the train data but are performing worse when seeing new data. To counteract this, an algorithm needs to have a healthy balance of the complexity of its features. Something can be said about the amount of features used in the algorithm in this thesis, as a few features (more specifically, the amplitude based features) result in the same Accuracy as without the features - this can also be shown with the application. To go with machine learning standards, these features could be removed, however, I used these features because these features resulted in a higher performance in classifying males. Because this algorithm had the task to classify sex (so females and males equally well), I decided to leave these features in, which resulted in the classification of females being slightly worse but the classification of males to be slightly better and therefore making them more balanced.

Lastly, we propose a couple of improvements for further research, including improvements of our own experiment and its shortcomings. The first proposition would be to have participants look longer and at more (varying) stimuli and then use the algorithm described in this thesis, which could result in an Accuracy higher than we have achieved here, as the data from this study had participants look at just ten seconds at one stimuli without supervision. In other words, this algorithm could achieve a better Accuracy metric in a controlled setting with more stimuli and time. The second proposition would be the addition of an adjustable stand for the eye tracker, so that children and people that can not get to the eye tracker for whatever reason, to also be included in the experiment. The third proposition is to adjust the algorithm by adding a variant of a *best subset selection* method so that the most predictive feature(s) from the chosen specific feature set can be depicted. Even though the user can go over each feature and click it on and or off to check its impact on the Accuracy, there are a lot of combinations of features that work better if they are both

on/off. I have 24 features in my features set, which gives $2^{24} = 16777216$ possible combinations of features - which would be too much for a person to go over manually. The fourth proposition would be to add a supervisor to the experiment/eye tracker for a small subset of participants (e.g 100 participants) to understand how often people would participate twice. Another advantage of having a supervisor would be to aid people that have difficulties with adjusting the stand of the aforementioned adjustable stand. Finally, I used the rbfkernel and based my features around this specific kernel. However, there may also be a kernel and feature set that gets a higher Accuracy and scores higher on the other machine learning metrics if paired together. The last proposition would be to look at these different kernels and the feature sets already present.

References

Al-Rahayfeh, A., & Faezipour, M. (2013). Eye tracking and head movement detection: A state-of-art survey. *IEEE journal of translational engineering in health and medicine*, *1*, 2100212-2100212.

Altman, D. G., & Royston, P. (2006). The cost of dichotomising continuous variables. *Bmj*, *332*(7549), 1080.

Baratloo, A., Hosseini, M., Negida, A., & El Ashal, G. (2015). Part 1: simple definition and calculation of accuracy, sensitivity and specificity.

Bhavsar, H., & Ganatra, A. (2012). A comparative study of training algorithms for supervised machine learning. *International Journal of Soft Computing and Engineering (IJSCE)*, *2*(4), 2231-2307.

Bishop, C. M. (2006). Pattern recognition and machine learning. springer.

Bourreau, M., De Streel, A., & Graef, I. (2017). Big Data and Competition Policy: Market power, personalised pricing and advertising. *Personalised Pricing and Advertising (February 16, 2017)*.

Bozkir, E., Günlü, O., Fuhl, W., Schaefer, R. F., & Kasneci, E. (2020). Differential privacy for eye tracking with temporal correlations. *arXiv preprint arXiv:2002.08972*.

Brereton, R. G., & Lloyd, G. R. (2010). Support vector machines for classification and regression. *Analyst*, *135*(2), 230-267.

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010, August). The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition* (pp. 3121-3124). IEEE.

Cazzato, V, Basso, D, Cutini, S and Bisiacchi, P (2010) Gender differences in visuospatial planning: an eye movements study. BEHAVIOURAL BRAIN RESEARCH, 206 (2). pp. 177-183. ISSN 0166-4328

Chen, J., & Jenkins, W. K. (2017, August). Facial recognition with PCA and machine learning methods. In *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)* (pp. 973-976). IEEE.

Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Bui, D. T., ... & Ma, J. (2017). A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. *Catena*, *151*, 147-160.

David-John, B., Hosfelt, D., Butler, K., & Jain, E. (2021). A privacy-preserving approach to streaming eye-tracking data. *IEEE Transactions on Visualization and Computer Graphics*, *27*(5), 2555-2565.

Friedman, N., Geiger, D., Goldazmidt, ―Bayesian Network Classifiers,‖ Machine Learning, vol. 29, pp. 131-163, 1997.

Géron, A. (2020). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems. O'Reilly.

Gomez, P., von Gunten, A., & Danuser, B. (2019). Eye gaze behavior during affective picture viewing: Effects of motivational significance, gender, age, and repeated exposure. Biological psychology, 146, 107713.

Gunn, S. R. (1998). Support vector machines for classification and regression. *ISIS technical report*, *14*(1), 5-16.

Gupte, A., Joshi, S., Gadgul, P., Kadam, A., & Gupte, A. (2014). Comparative study of classification algorithms used in sentiment analysis. *International Journal of Computer Science and Information Technologies*, *5*(5), 6261-6264.

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*, *24*(2), 8-12.

Heisz, J. J., Pottruff, M. M., & Shore, D. I. (2013). Females scan more than males a potential mechanism for sex differences in recognition memory. Psychological Science, 24 (7), 1157–1163.

Hessels, R. S., Niehorster, D. C., Kemner, C., & Hooge, I. T. (2017). Noise-robust fixation detection in eye movement data: Identification by two-means clustering (I2MC). *Behavior research methods*, *49*(5), 1802-1823.

Hwang, Y. M., & Lee, K. C. (2018). Using an eye-tracking approach to explore gender differences in visual attention and shopping attitudes in an online shopping environment. *International Journal of Human–Computer Interaction*, *34*(1), 15-24.

King, A. J., Cooper, G. F., Clermont, G., Hochheiser, H., Hauskrecht, M., Sittig, D. F., & Visweswaran, S. (2020). Leveraging eye tracking to prioritize relevant medical record data: Comparative machine learning study. *Journal of medical Internet research*, *22*(4), e15876.

Kohestani, V.R., Hassanlourad, M., Ardakani, A., 2015. Evaluation of liquefaction potential based on CPT data using random forest. Natural Hazards, 79(2), 1079-1089.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*(1), 3-24.

Kredel, R., Vater, C., Klostermann, A., & Hossner, E. J. (2017). Eye-tracking technology and the dynamics of natural gaze behavior in sports: A systematic review of 40 years of research. Frontiers in psychology, 8, 1845.

Lee, Y. J., Greene, H. H., Tsai, C. W., & Chou, Y. J. (2016). Differences in sequential eye movement behavior between Taiwanese and American viewers. Frontiers in psychology, 7, 697.

Leekam, S. R., Hunnisett, E., & Moore, C. (1998). Targets and cues: Gaze-following in children with autism. Journal of child psychology and psychiatry, 39(7), 951-962.

Martin-Barragan, B., Lillo, R., & Romo, J. (2014). Interpretable support vector machines for functional data. *European Journal of Operational Research*, *232*(1), 146-155.

Matsumoto, R., Yoshimura, H., Nishiyama, M., & Iwai, Y. (2017, September). Feature extraction using gaze of participants for classifying gender of pedestrians in images. In *2017 IEEE International Conference on Image Processing (ICIP)* (pp. 3545-3549). IEEE.

Meyers-Levy, J., & Loken, B. (2015). Revisiting gender differences: What we know and what lies ahead. *Journal of Consumer Psychology*, *25*(1), 129-149.

Mezghani, D. B. A., Boujelbene, S. Z., & Ellouze, N. (2010). Evaluation of SVM kernels and conventional machine learning algorithms for speaker identification. *International journal of Hybrid information technology*, *3*(3), 23-34.

Mogaji, E., Olaleye, S., & Ukpabi, D. (2020). Using AI to personalise emotionally appealing advertisement. In *Digital and Social Media Marketing* (pp. 137-150). Springer, Cham.

Moghaddam, B., & Yang, M. H. (2000, March). Gender classification with support vector machines. In Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580) (pp. 306-311). IEEE.

Mohamed, A. E. (2017). Comparative study of four supervised machine learning techniques for classification. *Information Journal of applied science and technology*, *7*(2).

Moss, F. J. M., Baddeley, R., & Canagarajah, N. (2012). Eye movements to natural images as a function of sex and personality. PloS one, 7(11), e47870.

Nalbantov, G., Bioch, J. C., & Groenen, P. J. (2006). Solving and interpreting binary classification problems in marketing with SVMs. In *From Data and Information Analysis to Knowledge Engineering* (pp. 566-573). Springer, Berlin, Heidelberg.

Nikitin, J., & Freund, A. M. (2011). Age and motivation predict gaze behavior for facial expressions. Psychology and Aging, 26(3), 695.

Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, *48*(3), 128-138.

Pérez-Moreno, E., Romero-Ferreiro, V., & García-Gutiérrez, A. (2016). Where to look when looking at faces: Visual scanning is determined by gender, expression and tasks demands. *Psicológica*, *37*(2), 127-150.

Pierdicca, R., Paolanti, M., Quattrini, R., Mameli, M., & Frontoni, E. (2020). A Visual Attentive Model for Discovering Patterns in Eye-Tracking Data—A Proposal in Cultural Heritage. *Sensors*, *20*(7), 2101.

Rácz, A., Bajusz, D., & Héberger, K. (2021). Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification. *Molecules*, *26*(4), 1111.

Recchia, G., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior research methods*, *41*(3), 647-656.

Rucker, J. C., Rizzo, J. R., Hudson, T. E., Horn, A. K., Buettner-Ennever, J. A., Leigh, R. J., & Optican, L. M. (2021). Dysfunctional mode switching between fixation and saccades: collaborative insights into two unusual clinical disorders. *Journal of Computational Neuroscience*, 1-11.

Salvucci, D. D., & Goldberg, J. H. (2000, November). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications* (pp. 71-78).

Sammaknejad, N., Pouretemad, H., Eslahchi, C., Salahirad, A., & Alinejad, A. (2017). Gender classification based on eye movements: A processing effect during passive face viewing. *Advances in cognitive psychology*, *13*(3), 232.

Sargezeh, B. A., Tavakoli, N., & Daliri, M. R. (2019). Gender-based eye movement differences in passive indoor picture viewing: An eye-tracking study. *Physiology & behavior*, *206*, 43-50.

Sen, P. C., Hajra, M., & Ghosh, M. (2020). Supervised classification algorithms in machine learning: A survey and review. In *Emerging technology in modelling and graphics* (pp. 99-111). Springer, Singapore.

Shannon, R., Stabeler, M., Quigley, A., & Nixon, P. (2009). Profiling and targeting opportunities in pervasive advertising. In *1st Workshop on Pervasive Advertising@ Pervasive*.

Shao, Y. H., Chen, W. J., & Deng, N. Y. (2014). Nonparallel hyperplane support vector machine for binary classification problems. *Information Sciences*, *263*, 22-35.

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision* (pp. 843-852).

Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. *Data classification: Algorithms and applications*, 37.

Xu, P., Ehinger, K. A., Zhang, Y., Finkelstein, A., Kulkarni, S. R., & Xiao, J. (2015). Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*.

Yu, B., Pan, D. Z., Matsunawa, T., & Zeng, X. (2015, January). Machine learning and pattern matching in physical design. In *The 20th Asia and South Pacific Design Automation Conference* (pp. 286-293). IEEE.

Zafari, A., Zurita-Milla, R., & Izquierdo-Verdiguier, E. (2019). Evaluating the performance of a random forest kernel for land cover classification. *Remote sensing*, *11*(5), 575.

Appendices

Appendix A: Experiment Picture

The picture that the participants got to look at for ten seconds.

Appendix B: App GUI

Layout of the GUI of the app so that you can choose the features you want to use and see the impact it has on the model