Heriot-Watt University

Final Year Dissertation

Ethics of Machine Learning

Author:
Dominic Calina

Supervisor:
Dr. Ron Petrick

A thesis submitted in fulfillment of the requirements for the degree of BSc.

in the

School of Mathematical and Computer Sciences

April 2019

Declaration of Authorship

I, Dominic Calina, declare that this thesis titled, 'Ethics of Machine Learning ' and the work presented in it is my own. I confirm that this work submitted for assessment is my own and is expressed in my own words. Any uses made within it of the works of other authors in any form (e.g., ideas, equations, figures, text, tables, programs) are properly acknowledged at any point of their use. A list of the references employed is included.

Signed:Dominic Calina

Date:22/04/2019

Abstract

As of 2019, Machine Learning is one of the fastest growing areas in Computer Science. The knowledge and trends we can extract from data is ever expanding. The amount of data, computational power and techniques available to build models which can predict present today is unprecedented, but so are the risks. Machine Learning is a field which is not immune to regulation and ethical standards.

This dissertation looked at publicly available data sets with attributes such as age, gender and race. Using Machine Learning algorithms and attribute selection, it researched areas of concern such as discrimination, data misuse and applications of Machine Learning. The experiment used Machine Learning algorithms to investigate whether further considerations should be taken when using data with sensitive attributes. This project will attempt to produce desensitised smaller data sets which can achieve similar accuracy to the full data set. It is important to build models which do not harm individuals. Additionally a public survey was conducted to find the concerns the public had with both Machine Learning and data usage in general.

Acknowledgements

I want to start by thanking my supervisor Professor Ron Petrick for showing his genuine interest in this topic. His support and insight helped give me the guidance needed to refine my project and find interesting topics. Having a supervisor who cares about your topic is an appreciated moral boost.

Dr Frank Broz, my second reader, should also be acknowledged for taking the time to read this project and for the advice given to me in Semester 1 which helped me improve my dissertation.

I would also like to thank my family and friends for the support throughout the year. They actively helped me achieve the number of survey respondents I wanted by spreading my survey in their own circles, especially my father. My friends added the much needed humour to get through this year.

Finally, I must acknowledge the resources online such as Kaggle.com and the UCI Machine Learning repository which provided the data sets which allowed me to conduct my research. Without appropriate data this dissertation would've been a struggle.

Contents

List of Figures

Abbreviations

ML	**M**achine **L**earning
DM	**D**ata **M**ining
NHS	**N**ational **H**ealth **S**ervice
GDPR	**G**eneral **D**ata **P**rotection **R**egulation
AI	**A**rtificial **I**ntelligence
TP	**T**rue **P**ositive
FP	**F**alse **P**ositive

Chapter 1

Introduction

1.1 Overview

The growth of the Internet, improved data storage and rapid advancements in computer hardware has brought a new and highly profitable era of data usage. New data is constantly generated at an exponential rate; by 2020 it's predicted that 1.7MB of data will be created by each person, every second (DOMO, 2018). This data is ready to be explored and with it valuable information gained. Most useful datasets are often far beyond the comprehension of human beings and even computers, without the right technique. Data scientists use statistical techniques to find patterns (Data Mining), confirm hypotheses and learn (Machine Learning) from datasets. The benefit to society is immeasurable; it's helping to improve key fields such as medicine, education and security.

Machine learning is a branch of artificial intelligence in which a computer automates a data driven task. Processed data is inputted and an algorithm applied to model this data (Witten, Frank, and A. Hall, 2011). This can be described as a computer 'learning' from data. Machine Learning is comprised of different techniques such as Neutral Networks, Decision Trees or Nearest Neighbours. This usually means providing a training set to produce a model which can be used for predictions, e.g. oil spill detection. Rules are generated to best represent the data, given the specific algorithm and goal (class).

1.2 The Problem

With the rapidly expanding field Machine Learning comes key areas of concern. Regulation and laws can often struggle to keep up with the rate of technology. The ability to now do so much with data means that it also comes with greater risk. There is no general consensus on why and how data should be used; even international data laws often greatly vary (DLA Piper, 2018). Development in this area has led to its widespread usage throughout businesses and scientific research. Machine Learning is now used to automate difficult but mundane tasks. The adoption of this technology has led to both increased profits and improved knowledge. It's widespread usage has unfortunately also brought concern and caution. The algorithms by themselves are usually not the issue, but rather how they are used (the data). However there is a separate issue of algorithm bias. This leads on to the overall scope of this dissertation; the ethics of Machine Learning. The background research covers key areas of concern such as: Anonymisation, Discrimination, Consent, Data Abuse and Data Transparency.

Chapter 2

Aim and Objectives

2.1 Aim

The aim of this project is to address the risks of Machine Learning and in turn contribute towards safer data usage. The risks focused on are those that affect the public. We shall discuss and analyse chosen data sets, with ethics in mind. The data set experiments will try to be combined with a public survey to form guidance. This research aims to find the technical problems with Machine Learning and the concerns of the public to conclude how Machine Learning can achieve a standard practise.

2.2 Objectives

2.2.1 Conduct Public Opinion Survey

A public survey to gain the thoughts and concerns about data usage, anonymisation and consent, in both the private and public sector. This will investigate the trust the public have in how their data is used. The survey will attempt to find the public consensus on the topics raised and discover where organisations (such as Netflix, Facebook and Amazon) can improve their relationship with the public. The relationship refers to the public's use of their services and data they give away.

2.2.2 Run ML experiments on datasets

Using Machine Learning techniques, attribute selection and preprocessing we will look at how to appropriately use chosen data sets. This will cover points covered in the background research: including anonymisation, data abuse and discrimination etc. Appropriate usage mean trying to avoid the reasons for concern and attempting to help prevent it. The observations found in the datasets will be used to provide guidelines based on firsthand research in this project. It will look at the different algorithms, accuracy of classification, other metrics such as kappa statistic, test options and any other areas of interest.

These data sets include:

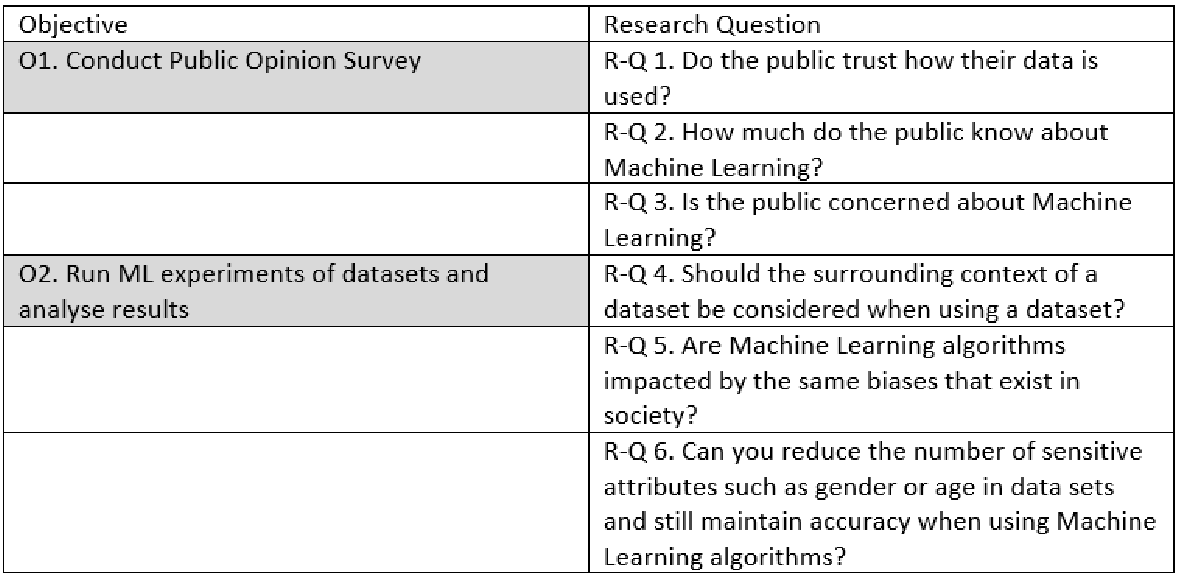
- Student Alcohol Consumption (UCI Machine Learning, 2018): A survey of student's Math and Portuguese language courses in secondary school. It contains social, gender and school study data.
- Diabetes (UCI Machine Learning, 2019b): Data originally created by the National Institute of Diabetes and Digestive and Kidney Diseases. This data set can be used to predict if a patient has diabetes.
- Credit Risk (Hofmann, 1994): This dataset can be used to decide if a customer has good or bad credit risk.
- Adult Income (UCI Machine Learning, 2019a): A 1994 US Census bureau database was used to form this dataset which can be used to predict if an individual earns more or less than $50K a year.
- Absenteeism (Martiniano et al., 2012): Data from a Brazilian courier which can used to predict the Absenteeism time in hours.
- Stop and Search Data (Home Office, 2019a): British Police force stop and search data from 2016-2018.

2.2.3 Combine research from experiments and public survey

The background research, insights from the public opinion survey and observations from the analysis of the data sets will be used to create guildlines to take away from the project. The guideline will be inspired by already existing non-legal documents used to help companies fall inline with GDPR. The public viewpoint will shape the most important points to consider when concluding from the data set experiments .

2.3 Research Questions

These research questions were decided by the areas found in the background research. The Machine algorithms refer to Decision Trees, Random Forest, K-Nearest Neighbour and Naive Bayes.



Objective	Research Question
O1. Conduct Public Opinion Survey	R-Q 1. Do the public trust how their data is used?
	R-Q 2. How much do the public know about Machine Learning?
	R-Q 3. Is the public concerned about Machine Learning?
O2. Run ML experiments of datasets and analyse results	R-Q 4. Should the surrounding context of a dataset be considered when using a dataset?
	R-Q 5. Are Machine Learning algorithms impacted by the same biases that exist in society?
	R-Q 6. Can you reduce the number of sensitive attributes such as gender or age in data sets and still maintain accuracy when using Machine Learning algorithms?

Chapter 3

Background

3.1 Anonymisation

Data Anonymisation is the process of removing sensitive or identifiable attributes from data sets. This is vital for both data mining and data storage, in general. The General Data Protection Regulation has set a good precedent for how stakeholders should be handling sensitive data (European Parliament, 2016). Anonymisation helps regulate businesses and researchers in a way that helps to protect individual rights, without removing the ability to find useful trends in data. Identifiers should be kept out of data sets with human participants.

3.1.1 Techniques

Data anonymisation can be performed by an assortment of techniques such as: Attribute Suppression, Record Suppression, Character Masking, Pseudonymisation, Generalisation, Swapping, Data Perturbation, Synthetic Data and Data Aggregation. Each data set has its own procedure for anonymisation. It is often a case of combining multiple techniques. Below is a summary of these techniques (Personal Data Protection Commission of Singapore, 2018).

1. Attribute Suppression: Removal of an entire part of a dataset (column), which could for example be your National Insurance Number.
2. Record Suppression: Removal of an entire record in a dataset. This applies to outlier record which are unique, such as the only professor of a nationality, in a University department.
3. Character Masking: Changing of the characters of a data value, e.g. by using a constant symbol like “*”.
4. Pseudonymisation: Replacement of identifying data with made up values, e.g. changing company names to fictional characters.
5. Generalisation: Reduction in the precision of data, e.g. grouping by Town rather than postcode.
6. Swapping: Rearrange the dataset, e.g. swap the age rows in the dataset.
7. Data Perturbation: Data is modified to be slightly different, e.g. add or subtract 1 from each age in a dataset.
8. Synthetic Data: Generate synthetic data based off a real dataset, e.g. generating a synthetic dataset of sprint times from the 100m Final at the Olympics, based off real historic data from the last 50 years.

9. Data Aggregation: Converting a dataset from a list of records to summarised values, e.g. group average alcohol consumption by each nation rather than individuals.

3.1.2 Re-identification

It cannot be assumed that datasets are ethically sound based on anonymisation alone. It has been shown with academic research that with the addition of other data, seemingly anonymised data sets can become de-identifiable (Porter, 2008). Further research has been conducted in this area and the evidence found a high reidentification rate in the 14 data sets used (El Emam et al., 2011). An example of this is publicly available data on New York taxi journeys. Anyone in the public can combine timestamped pictures of celebrities using New York taxis with the 'anonymised' dataset to find how much they tip for a journey (Tockar, 2014). This is a clear invasion of privacy which was from the result of a dataset that had been anonymised. Surge in publicly available information online and improved computer hardware has made it easier to reidentify so called 'anonymized data'. This is often done by combining two or more data sets to find links between attributes in each e.g. time, in the case of the taxi dataset (Altman et al., 2013). Re-indentification is a well known amongst those that handle data but the public may be unaware. This project will find out whether re-identification impacts public trust.

A study in 2015 of credit card data found that it only took four credit card transactions to identify 90% of the individuals in the dataset (Montjoye et al., 2015). The dataset consisted of 3 months of credit card transactions, recording the spending of 1.1 million people in 10,000 shops in a single country. Every individual has a unique spending pattern which makes a dataset such as this very vulnerable to reidentification. You just need to correlate the dataset with outside information about an individual. The bank providing the metadata anonymised the dataset by removing credit card numbers, shop addresses and time of the transactions (Bohannon, 2015). Despite this anonymisation the individual's right to anonymity was not protected. Regulating the way even anonymised data is made publicly available may be a good step towards reducing what de Montjoye calls a 'correlation attack'.

3.1.3 Regulation vs Research

The NHS is a good example of the battle between anonymisation, regulation and impactful research. There has been a debate between different interest groups on how patient data should be used. Privacy campaigners have focused on confidentiality and consent, whereas epidemiologists are pushing the research benefit of using data which contains the entire population

(I. Brown, L. Brown, and Korff, 2010). The research letter (I. Brown, L. Brown, and Korff, 2011) discusses the results of an editorial on regulation of clinical research in the UK (Smyth, 2011). The research letter notes that the editorial pays insufficient attention to anonymisation. Participants are often misled about the chance of re-identification. Some data is even 'partially' pseudonymised. There is no such thing as partially anonymising data. It is simply a measure in place to prevent immediate identification of patients e.g. First and last name. The data users specifically kept the ability to re-identify patients. Re-identifying individuals is desirable in medical research especially when researching rare disease (Hansson et al., 2016). Consent was not clearly presented regarding this. As of 2018, it is very unlikely that this would still comply with current data laws in the UK. Under GDPR (European Parliament, 2016), personal data cannot be identifiable, there is a greater push for anonymisation both in the private and public sector. This poses a moral dilemma. Smyth claims that regulation is interfering with medical research due to excessive bureaucracy. It is irresponsible however to not value the individual rights of participants, regardless of data research. Her point of view is still one of great important to the discussion of ethics in Data Mining and Machine learning however. Effective anonymisation can often reduce the value of a data set and the insight that can be made (Smart, 2016). The goal of using medical data sets is often to improve healthcare for individuals. The risk of de-anonymising individuals can lead to important advancements in medical research, due to more valuable data. The research letter highlights the gap between clear consent to participants and data use by researchers. Both Brown and Smyth agreed that engagement with the public is a useful step towards improving public trust in how their data is used.

3.1.4 Summary

Anonymisation is not enough of a measure of how ethical a dataset and its subsequent analysis is. Anonymisation is often relative to each individual example. On the surface it may appear to be a good way of protecting individual rights, but it cannot be relied on, on its own. There is no consensus on what anonymisation entails, regardless of the laws, as it is often defined by the individual data user. Anonymisation can therefore be viewed as a good starting point when preparing data for data mining and machine learning, but it should not be the sole measure of ethics.

3.2 Discrimination

Discrimination is defined by the Oxford dictionary as 'The unjust or prejudicial treatment of different categories of people, especially on the grounds of race, age, or sex' (Dictionary, 2018). Discrimination is caused by many factors in society and as such this research project will not address the causes of societal discrimination, but rather how it is found in algorithms. A fair assumption to make is that discrimination is usually caused by human behavior and actions. The public often assume that algorithms remove issues like discrimination, but algorithms in fact do not always offer a solution that remove human bias. This section highlights how algorithms can also discriminate towards humans; it should not be misconstrued with discriminative models found in classification (Ng and Jordan, 2002).

3.2.1 Algorithm bias in the media

There have been many examples of algorithm bias that have appeared in the media, in recent years. This has brought the problems to the public's attention and fueled academic research. Recent news articles can assist in understanding the public's current perceptions. An investigation by Reuters discovered that Amazon's AI recruiting tool favoured men (Dastin, 2018). It was trained on data which was predominantly male resumes . The male dominated tech industry led to the hiring tool favouring men. The system learned to favour the applications of men, as words such as 'women's' led to a lower rating score meaning the system ranked the application as less desirable. An article by Aviva Rutkin brings the problem of sexism in text mining further into the public eye (Rutkin, 2016). It focuses on word-embedding and the formation of linguistic links. An example given is female names being more closely associated with home and male names with career. Language and works of text have biases which can cause algorithms to contribute towards stereotypes. Language is not gender neutral so gender bias should be taken into account when using text. An arguably more serious example of algorithm bias in the media was the investigative piece by ProPublica (Angwin et al., 2016). A computer algorithm was used to predict the chance of a criminal re-offending. The algorithm was found to be both unreliable and biased. Only 20% of those predicted to commit violent crimes went on to do so in the two years after the tool was used. More worryingly, black defendants were far more likely to be labelled as future high risk criminals, despite the re-offending rates not matching. The software investigated is made by a for-profit company and is used widely throughout the US. The main take from this is that a systems designed to combat racial inequality is in fact helping to extend it further. We can now look at how academic research has tried to fight algorithm bias.

3.2.2 Academic research into algorithm bias

Multiple studies have gone into the area of biased algorithms and discrimination. Calders and Žliobaitė discuss the issue of unbias computation processes in Chapter 3 of Discrimination and Privacy in the Information Society (Calders and Zliobaite, 2013). They argue that the assumptions that data mining methods make do not align with reality. Data is not unbiased and human populations are hard to represent. Classifiers can make their decision based on discriminatory generalisations found in the dataset. The chapter also discusses the fact that even after removing identifiable and sensitive attributes, discrimination can still occur due to algorithms treating all attributes from a neutral perspective. Algorithms do not have context to attributes, unless it is implicitly included in the model. They state that the main reason why discrimination in data mining occurs is because construction methods are usually based on assumptions which do not reflect reality. There needs to be greater focus on the context of assumptions which algorithms make and is then used to classify. A research article raises an important point about identifying discrimination (Veale and Binns, 2017). Organisations often do not hold the sensitive data required to identify and fix the issue of discrimination. The paper presents three potential ways to deal with the lack of sensitive data to compare results against. These include trusted third parties to store sensitive data and perform discrimination discovery, online platforms to promote fairnesss and unsupervised algorithms to create hypthosis fairness testing. Machine Learning algorithms are meant to discriminate; they used logic to classify. This does not make it socially acceptable when the data itself is not fair. Human data is never going to represent the whole population nor will it contain all factors that contribute towards social biases. Data is deliberately used in a subjective manner, we select the data which appears to be the most useful. Input variables are reconstructed to try to achieve the best outcome. Additionally, the article emphasises that the choice of modelling process is by humans. The system model can impact the level of algorithm bias that occurs. There needs to be a re-evaluation of how algorithm bias is avoided. The industry standard for protection against sensitive attributes is inadequate (Hardt, Price, and Srebro, 2016). Avoiding attributes will not remove the impact of the attributes, reality cannot be reduced to selected non-sensitive attributes. Removing sensitive attribute can however reduce the profiling nature of using personal features such as gender. The main focus of academic research is combating algorithm bias. This algorithm bias mostly relates to the data used but developer bias can exist as well. The assumption that a programmer is unbiased is naive. Research into this idea was found to be very limited. The only related academic research found was a 1995 article on cognitive bias in software engineering which causes humans to move away from optimal reasoning (Stacy and MacMillan, 1995). Some articles were found which relate to programmer bias such as how an overwhelming number of Machine Learning experts are male but discussion of this topic as a whole was minimal (Byrnes, 2016). This may be a good area of research to explore further. There is a conception that algorithms are unbias yet there is very little research into programmer bias.

Researchers have posed multiple ways of taking discrimination into account when using algorithms on data. Chouldechova discusses the different fairness criteria that have been used to prove a system's unbias (Chouldechova, 2016). The paper mentions that fairness is a social and ethical concept, rather than a statistical one. This is hard to disagree with and it makes measuring fairness a somewhat subjective task. The types of fairnesss assessment included are calibration, predictive parity, error rate balance and statisitcal parity. The context of the paper is focused around the ProPublica article mentioned previously (Angwin et al., 2016). Calibration measures if there is a disparity between the likehood of individual groups (races) reoffending, as defined by the model. Low disparity means that the system is well calibrated. Predictive parity is similar but instead of just looking at reoffending, it specifically has the condition of high risk offenders, in the example given. Error rate balance measures if the false positive and false negative rates are equal across all individual groups(races). Finally, statisticaly parity checks if the proportion of individual in each individual group are classified the same proportion as high-risk, using the ProPublica example. There are many ways of measuring fairness and as such no method can lead to conclusive evidence of fairness (Zliobaite, 2015). Other researchers have more proactively tried to modify the classifying models to take biases into account. This paper attempts to remove algorithmic discrimination while still minimising the individual error (Mahdi El Mhamdi et al., 2018). They mention the complication of reducing group discrimination and maintaining individual accuracy. Other researchers have focused on pre-processing the data, causal reasoning (learning phase) or removing sensitive attributes. This paper suggests making adjustments at post-processing, it poses making considerations after learning has been done, by applying a score function which adjusts results.

Algorithm bias is a problem that is both difficult to define and minimise. Discrimination must be avoided at all costs. Bias can occur in both the data and system. Data involving individuals should always be used with context in mind. Systems must be used with the biases of society in mind. A more well-defined standard for algorithm bias and human data usage would be useful. The status quo is that algorithms should be used until the problem is discovered. This project extends on the problems found by both finding the concerns of the public and starting the process of trying to find steps which can help improve ethics. There is a battle of what can be achieved with Machine Learning techniques and whether it should be used ,e.g. job hiring process, with the current lack of regulation which forces programmers to minimise algorithm bias. Further research, such as this project, is required because alarming aspects of Machine Learning have been discovered but there is still no consensus or legal guidelines which can turn Machine Learning into a more ethical sound field.

3.3 Algorithms Used for Experiment

Some background to the algorithms used is needed as they differ in how they work. The algorithms used were: K-Nearest Neighbour, Decision Tree, Random Forest, and Naive Bayes which were deemed to be widely used (Le, 2018).

1. K-Nearest Neighbour: This algorithm is one of the more simple ML algorithms. Classification is based on the position, of the instance you are trying to classify, in the feature space (Cover, Hart, et al., 1967). Each instance in the training data is positioned in the feature space based on its non-class attributes. The predicted class is classified based on the neighbour/neighbours the instance it is closest to. At its core it is distance-based classification. K refers to the number of neighbours, so for example if it is 3 it will classify based on the 3 closest neighbours (classes). If 2 of the 3 neighbours are the same class it will classify the instance as the same as the majority of the set number of closest neighbours.
2. Decision Tree: A decision tree is a form of classification in which the data is used to build a rule-based tree (Breiman et al., 2017). The root and intermediate nodes represent the non-class attributes. Each branch represents a decision (question) for that attribute such as yes or no (Quinlan, 1986). The leaf node represents the predicted class. When an instance is applied to the decision tree it is classified as one of the leaf nodes (classes) based on how its non-class attributes dictate the tree traversal. Observations are found in the data set during learning and are used to create branch rules which best represent the training data.

3. Random Forest: Random Forest is an extension of decision trees (Ho, n.d.). Rather than forming one tree many smaller trees are created. An instance is then classified using all the smaller trees and the mode class is chosen. For example if 10 decision trees were modelled and in 6 cases the tree classified the instance as class yes the the mode class would be yes. This is a type of ensemble learning which is when individually trained classifier, in this case decision trees, are combined when classifying instances (Maclin and Opitz, 2011). The decision trees are combined using bagging. Bagging trained the decision trees by providing it with a random subset of the training set. Random Forest helps to reduce overfitting by adding variance as subsets are used rather than the full dataset (Hastie, Tibshirani, and Friedman, 2016).
4. Naive Bayes: This algorithm is built on Bayes' law. It describes the probability of an event(class) based on prior knowledge (non-class attributes) (John and Langley, 1995). The naive aspect is that each feature is independent from each other. An example use of this is medical diagnostic where you give a variety of attributes as input and the diagnostic with the highest probability is given as output. The independence of each feature from each other allows high scalability. It is found to have good performance which is surprising as the assumption of independence between attributes is rarely true in real world applications (Zhang, 2004).

Chapter 4

Research Methodology

This chapter will discuss the research methods used to answer the research questions and achieve the aims and objectives. The methods were chosen to give a broad and varied approach to the research questions by both running technical experiments and gathering public input.

Relevant laws and licenses

- Data sets are released under CC0:Public Domain.
- SurveyJS (JavaScript library) provide a free Creative Commons license for non-commercial usage.
- Consent forms will be given for all questionnaires and user evaluations
- WEKA is released under GNU General Public License

4.1 Qualitative

Qualitative research is the methods used to gather non-numerical data. Quality is the what, where, when, how and why of a topic of discussion. It provides perspective and context to the research evaluation (Berg, 2011). There are generally three kinds of qualitative research which consists of: interviews, observations and documents. Interviews yield direct quote from participants, observations comprise of detailed descriptions and documents includes all data generated from written records, such as questionnaire (Patton, 2002). This project generated qualitative data from the questionnaire and observations found in data.

Qualitative data can be used by itself or in combination with quantitative research. In this research, the qualitative data was combined with quantitative data which was gathered along side it. The purpose of the qualitative research was to get the views of individuals and use them for exploratory research which may go beyond my research questions. The qualitative research was fairly minimal as only one question in the survey was qualitative. This contributed towards exploratory research as individuals have different concerns and perceptions of how their data is used. Attaching open-ended questions with choice questions can offer more in-depth observations. The qualitative data can help explain the choices made to closed questions (quantitative). The open-ended nature of qualitative data makes it harder to analysis. As a researcher you must be careful when making conclusions from qualitative data. You must be responsible and disciplined when using the data especially when there is an opinion shared by only one individual. It must not be manipulated or lead to untruthful conclusions. There is no scientific procedure you can use to correctly use qualitative data, in the same way you can with quantitative methods. Statistical modelling cannot be applied to textual responses. Qualitative methods was used to allow participants to offer their own insight and explore other lines of thought. Furthermore, using purely quantitative methods may answer my research questions but it won't explain why participants answers questions the way they did. Data collected from qualitative methods should be used to generate observations which are useful in answering my research questions.

4.2 Quantitative

Quantitative data is the methods used to collect numeric data. The results of this can be used to answer research questions and confirm hypothesis. The aim is to build accurate and reliable measurements that allow for statistical analysis (Goertzen, 2017). Statistical method can be applied to quantitative data to gain insight and generalise based on the data. Analysis of the data should highlight patterns which can help make conclusions. The questionnaire and data set experiments both involved quantitative data collection. Quantitative data was used to assess which of the research questions were able to be answered. Quantitative data is useful for confirmatory research, but it can also suggest future work. This project uses a mixed approach (qualitative and quantitative) but there is a greater focus on using quantitative data to address my research questions.

4.3 Questionnaire

Questionnaires are a series of questions which gathers the views of respondents. A questionnaire was released on social media to gather the thoughts of a sample of the public, in relation to the research questions outlined (Chapter 5.1 R-Q 1-3). The appendix includes screenshots of the survey(Appedix A.1) which was published. It used predominantly quantitative techniques (25 out of 26 questions) to discover the public's opinion on Machine Learning and data usage.

Participants were provided with a brief high level description of Machine Learning. The questionnaire mostly consisted of statements and a Likert scale which participants use to respond (Likert, 1932). Statements were presented in a way that aimed to avoid leading participants to answer in a way which confirms the author's own bias. The data produced should provide an honest overview rather than data which is desirable. In addition to the statements with Likert scale, there was an option for participants to share their own thoughts at the end. This was completely optional as the survey was designed to take an average time to complete of around

5-10 minutes. SurveyMonkey took a random sample of 100,000 surveys with 1 to 30 questions and found that abandonment rate went up when surveys take longer than 8 minutes; completion rates dropping anywhere from 5% to 20% (Chudoba, 2018). Furthermore, they discovered that time taken on each question decreases as the number of questions increase. Participants were able to complete the entire survey without having to offer any qualitative data. There were not sections that force you to write a response to continue. Open ended questions increase survey time and reduce the response rate (Boyer and Stron, 2012) . There was a plan to reduce the survey length if response rate was low, but sharing the survey achieved a sufficient number of participants.

The original plan was to conduct the survey in SurveyMonkey due to its widespread usage by both industry and academics. It has the most funding(investment) out of any data and analytics cloud-based company at $1.5 billion, highlighting its popularity. SurveyMonkey advertise that 20 million questions are answered daily (SurveyMonkey, 2018). Additionally, there are features that were found to be desirable. Their service makes creating surveys an easy task. SurveyMonkey provide all the question formats which were required already. A free service is available, but this is limited to only 10 questions, maximum 100 responses per survey, no survey analysis or question logic (e.g. skip if answered no). The free version didn't provide was the required features. The paid version had: unlimited questions, data exports, customised survey experience (including skip logic), advanced survey analysis, disqualification if consent is not given and £25 credit which can be used to send survey to their panel which includes millions of people worldwide. The survey was instead created using the free JavaScript library SurveyJS on the author's domain (dominiccalina.com). This had most of the benefits of the paid version of SurveyMonkey such as question logic but it was found to be superior in several ways. SurveyJS was free to use, survey participants recognised the url (unlike a string of characters in a survey monkey link), the survey was build using JSON allowing for easy modification and the data was not held by a third party but instead sent to a secure MySQL server. Furthermore it allowed more freedom in the visual aspects of the survey as it is fully customisable. The downside is that time had to be taken to both set the domain and website up correctly (programming) unlike a traditional questionnaire service which handles the publishing and server side. The code for the survey at be found at this Google Drive link

4.4 Run experiments on data sets

The experiment run on the data sets were set up to answer these research questions (Chapter 2.3 R-Q 4-6). All research was conducted using WEKA (Witten, Frank, and A. Hall, 2016). WEKA is a collection of Machine Learning algorithms which can be used for a wide range of data problems. WEKA in this case was mostly used for it's preprocessing and classification features. It is free software, released under the GNU General Public License, written in Java developed at the University of Waikato. It was planned that WEKA would be run from Python (Java virtual machine) using the python-weka-wrapper3 package (Reutemann, 2018). The reasoning for this is the ability to automate running the algorithms on multiple files and writing results to different unique text files. Using the WEKA GUI means you must manually insert a new file which is not ideal but it has more visualisation features.

4.4.1 Attribute Selection

Attribute selection is the process of selection specific features(columns) in a data set, thus producing a subset of the original data. This is generally done to remove irrelevant, redundant attributes or attributes which add noise(L. Bluma and Langley, 1997). This project involved removing sensitive data such as gender or age to produce data sets with less sensitive data. The main purpose of attribute selection is: improved accuracy, faster predictors and simplification of models (Guyon and Elisseeff, 2003).

4.4.2 Classification Accuracy

Experiments will be conducted to see how well algorithms can classify categoric attributes in the data sets. The goal of classification is to correctly predict the target class of each inputted instance. Classification is where a model is formed from training data which is used to predict the correct class (Kesavaraj and Sukumaran, 2013). This will observe both how well certain attributes are classified and how algorithms perform. When analysing the classifier accuracy not only will the correctly classified instances be discussed. The true positive, false positive and kappa statistic was also recognised as it provides further insight.

4.4.3 Classifier Testing Options

The test options used were leave-one-out cross-validation, 10-fold cross-validation, percentage split 66% (66% of the data set used for training) and percentage split 33% (33% if the data set used for training). A variety of test options were used for accuracy comparison. 10-fold cross-validation is generally considered a good choice as it is a good balance between reducing bias of the technique used, reducing variance, computational constraints and it includes all the data in classification and testing. Cross-fold allows you to use the data for both training and testing. Leave-one-out cross-validation was used when possible, due to varying data set size. Leave-one-out cross-validation is where the k in k-fold is the number of data points in the data set. This means that at any one time only one data instance is used for testing and the rest is used for training. This maximizes the amount of data used for training while still separating training and test data. This means that the computational complexity goes up significantly with each new data instance, so it is unrealistic for large data sets. Percentage split was used to have a completely separate training and testing set. Unlike cross-fold the data is not used for both testing and training. Additionally, percentage split 33% means that only 33% of the data set is used for training, which should lead to a lower accuracy. It was useful to also look at a scenario where there is less training than test data, which can exist when there is a lack of data. 10-fold cross-validation was the default choice when looking at the classifier accuracy but other options were used for more complete accuracy results.

4.4.4 Experiment Procedure

The main data used was the one containing student alcohol consumption, to allow one dataset to be discussed in detail. The remaining datasets followed the same method, but the results are summarised to support the main data set and deal with different research questions. A standard procedure was used when using data. The experiment procedure used throughout was as follows. The algorithms used were Decision Tree (J48), Naive Bayes, K-nearest neighbour and Random Forest which were discussed in the literature review. The default WEKA settings were used for the algorithm, unless specified. The data experiments did not deal with parameter tuning which may have improved accuracy. It was important to use a variety of widely used Machine Learning algorithms. This allowed comparison between algorithms and to measure the impact of attribute selection. This is a project about ML algorithms rather than one specific algorithm so it would be invalid to focus only on one.Furthermore, all data sets were randomised during pre-processing to minimise the impact of instance ordering, as this affects the training stage. Classifier accuracy is often impacted by the order of instances. The order in which data is stored can have a bias on results so this was avoided where possible. The default seed for randomising the data set is 42 in WEKA, this was used throughout the process of experimentation so that experiments could be repeated. To further explain the experiment procedure here is a generic process of how datasets were used.

1. Find publicly available data set with sensitive attributes
2. Observe the context of the data set e.g. how it was collected or where there may be ethical concern
3. Preprocess data set for WEKA by converting to CSV or ARFF(Attribute-Relation File Format). ARFF is a file format for WEKA in which there is a header containing the list of attributes and their types and the data below. Data is described as a fix number of attributes and its type e.g. nominal
4. The file is loaded into WEKA
5. Observations are made about the distribution of attributes and which attributes could be deemed sensitive
6. The four algorithms are used (separately) to build a classifier with 10-fold cross-validation as the default test option
7. The sensitive attributes are removed from the data set which is the attribute selection
8. The four algorithms are used to build a classifier with 10-fold cross-validation as the default test option on the new smaller subset
9. Results files are saved (for later use and as evidence)
10. The accuracy of the algorithm with and without sensitive attributes is recorded. This can be multiple different smaller data sets.
11. Accuracy change is analysed (including TP rates, FP rates and kappa statistic). The metric is not only accuracy but the other values in the output file as well.
12. Observations are made about whether the same/similar accuracy can be achieved with a smaller number of less sensitive attributes.

4.5 Evaluation Strategy

All the observations found from the research (experiments and survey) are discussed with the topics raised in the background research. The results of the questionnaire relate to the relevant sections of the background review such as anonymisation, consent and discrimination. Graphs and charts created from the quantitative survey responses help to find patterns which relate back to the project search questions. SurveyJS does not have built-in features for analysis like SurveyMonkey so analysis was completed using Microsoft Excel. There is a focus on visualising the results of the survey and finding the view of the majority, but the survey is not only interested in what the majority think. The result of evaluating the quantitative data helps answer research questions (R-Q 1-3). The qualitative data inspires new avenues of research and bring some context to the result of the research questions. The qualitative data will not be used to make conclusions as there is the difficulty of interpreting the meaning behind responses and selecting which textual responses to value more highly can be difficult.

The investigation into Machine Learning algorithms more directly addresses the remaining research questions (R-Q 4-6) than the background research does. Relating the results of the data set experiment with the background research helps to further emphasis the issues that have been raised. The public survey is used to find public perception towards the issues raised whereas the ML experiments give physical examples of ethical concern. The evaluation strategy focuses on confirming problems with data and how ML algorithms are impacted by it which can then be combined with public concern to discover where there are problems and how to try to solve them. The evaluation involves discovering both technical problems with ML and non-technical aspects which require regulation/reform.

Chapter 5

Survey

5.1 Survey

The survey was conducted between the 15th February and 3rd of April, with the vast majority of survey responses happening within the first week of when the survey was published. An excel spreadsheet containing the quantitative results of the survey as well as demographic filtering can be found at this Google Drive link. The survey was conducted on the author's domain at https://dominiccalina.com/ML/. The Appendix shows screenshots of all questions participants saw as well as the consent page (Appendix A.1). The survey finished with 145 respondents which provided enough data to find both trends and compare demographics. Furthermore, there was a good variety of respondents, as there was a satisfactory spread of gender, age and residency (location). The results showed some clear trends and shared opinions were discovered between participants. In most cases there was a weighting towards a certain response for a question, rather than an even spread of answers e.g. 20% each for strongly agree to strongly disagree. This allowed conclusions to be made about the common response. However even if answers were spread this can allow conclusions in the context of the survey as it means opinion varied. The concern of even a minority is important as ideally you'd want an overwhelming majority to be satisfied with the current state of the topics raised. The topics brought up in the survey had varying levels of concern. There was a link found between knowledge of the technology and the public's trust in it, which will be discussed later in this section. The survey produced results that help to support the data experiments conducted later in this document. Consistently throughout this chapter questions which asked the participant to strongly agree to strongly disagree were converted to numerical values so that the mean response could be found. We shall first discuss the demographics of the survey (Figure 5.1 and Figure 5.2).

The demographics which we were interested in recording were: gender, age, residency, education, internet usage, whether they had heard of ML (binary) and whether they currently work in the IT sector (binary). From these demographic factors it was decided that all of them except Internet usage would be useful for demographic analysis. Internet usage was excluded because 86% of respondents used the Internet frequently throughout the day and only 4 of 145 respondents were infrequent users. There was a good spread of demographics, but a more balanced spread of age and residency would be ideal. The survey was dominated by participants who reside in the United Kingdom, which was expected as this survey was conducted from a British University. Additionally, only 14 participants were within the age range of 26-35 years old; this age group would be improved by having more participants for a more balanced age distribution over the whole survey results. On the other hand, there was a good balance of gender, education and age between the youngest (18-25 years old) and oldest (50 and older) participants. An even balance of those who work in the IT sector and those who have heard of ML was not necessary. There just needed to be enough participants who answered yes to those two questions to allow for them to be used for demographic comparison. It just happened to be the case that there was close to a 50/50 split between those who had heard of ML and those who hadn't. Getting balanced demographics is a challenge unless you target specific groups, which are found to be underrepresented. This survey was shared via social media, so there is no control over the demographics of respondents. Allowing the general public to participate in the survey, rather than targeted groups was important. The survey results are not heavily dominated by one particular group, so it is valid to state that the survey results represent many different groups of demographics (Figure 5.1 and Figure 5.2) . Most notably the imbalance which affects demographic comparison can be found in age and residency.

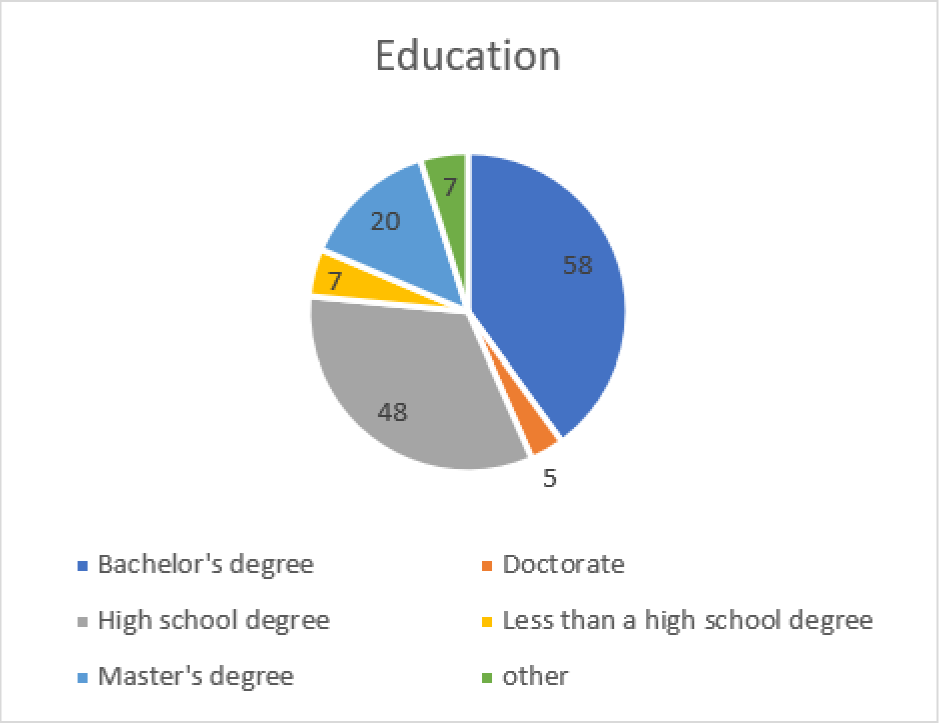


Figure 5.1: The Education Demographic

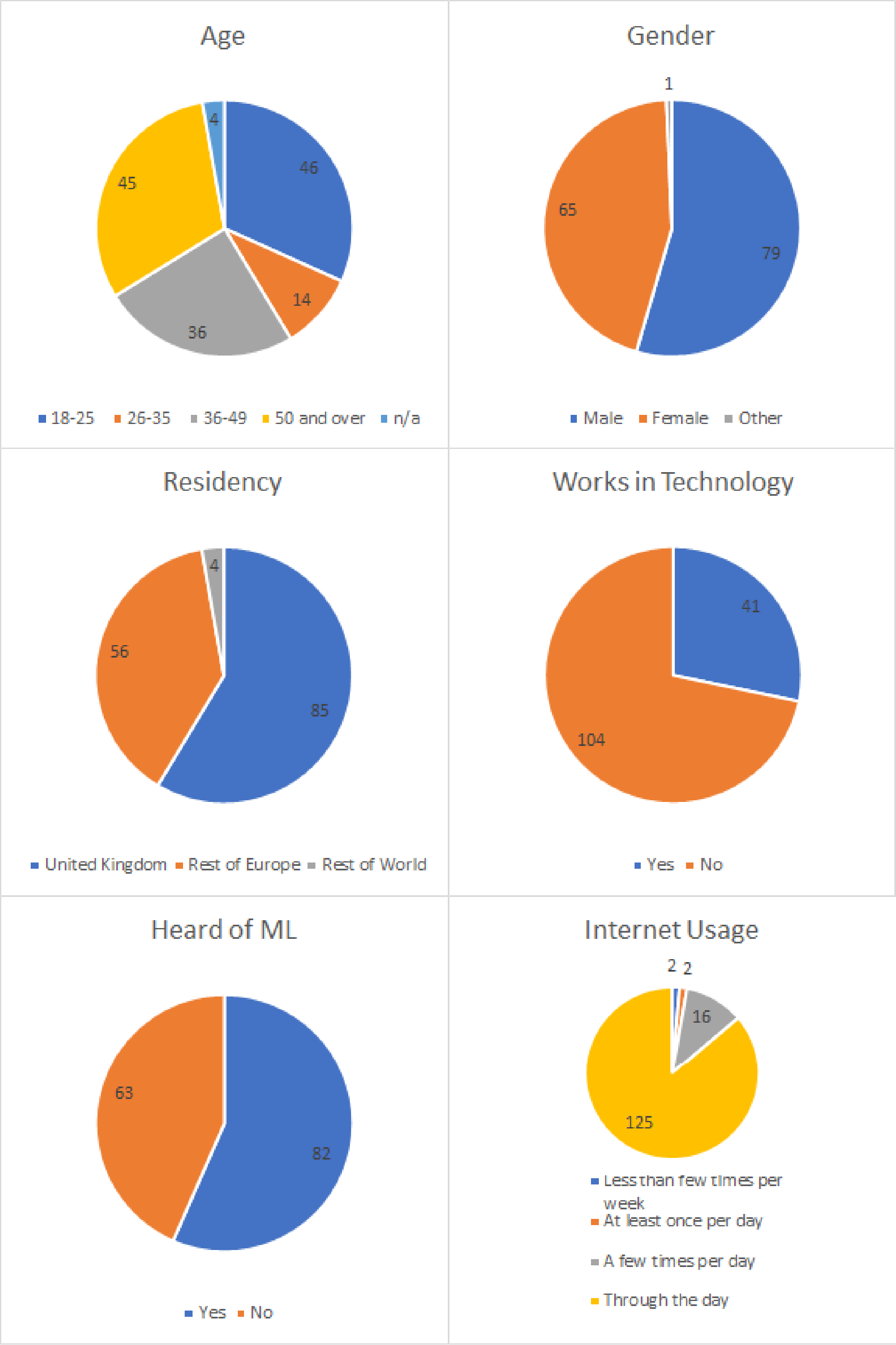
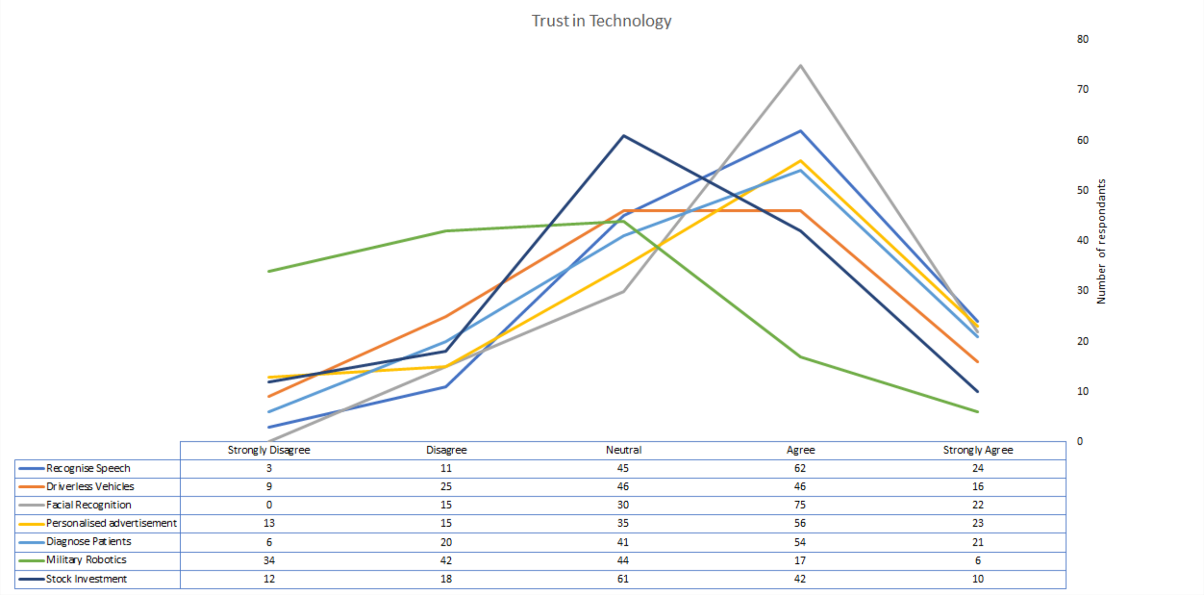


Figure 5.2: Demographics of survey participants

5.1.1 Analysis of survey results

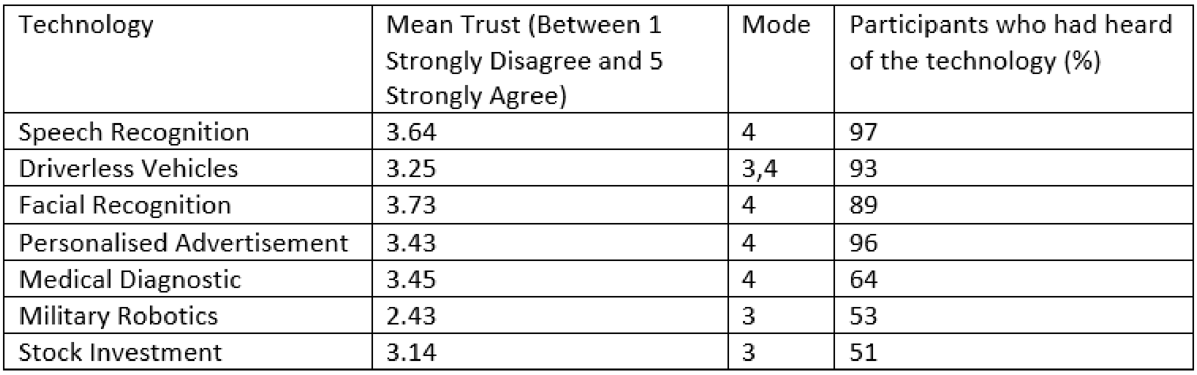
We shall now breakdown the trends found in the survey results. The survey discovered clear variation in the concerns of the public in relation to the different technologies and topics presented in the survey. It is understood from the results that public trust in the technologies discussed and the organisation which use them can be improved. There is room for organisations to improve the way the public perceive how they operate. There should be motivation for organisations to improve the public perception, as in turn it will improve their reputation and the willingness of the public to approve of the new technology. The results in the survey suggest that the public are more willing to trust technology they have a knowledge of. A part of getting the public to adopt new technology such as driverless cars or computer-based medical diagnosis is to help them understand what it is. Lack of trust in new technology can arise both from lack of exposure to the new technology and also how it is presented in the media. This can be seen as an issue of all new technology rather than ML specifically, people often fear what they do not understand. This does not mean that lack of trust in new technology is unfounded; it's a matter of whether it's based on evidence or lack of knowledge. The survey produced results which show areas in which further research can occur such as how to improve trust and/or how to better educate the public on new technology.



	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Recognise Speech	3	11	45	62	24
Driverless Vehicles	9	25	46	46	16
Facial Recognition	0	15	30	75	22
Personalised advertisement	13	15	35	56	23
Diagnose Patients	6	20	41	54	21
Military Robotics	34	42	44	17	6
Stock Investment	12	18	61	42	10

Figure 5.3: The trust participants had in each technology as a line graph

In most cases there was a leaning towards a certain view (agree or disagree) from the majority of participants. Above is a graph showing the trust participants had in each technology (Figure 5.3). It is clear that there was greater trust in facial recognition, speech recognition, personalised advertisement and patient diagnostic than driverless vehicles, military robotics and stock investment, as agree was the majority choice for those technologies. Facial recognition had only 15 people disagree, and no one strongly disagree, which means that 90% of the survey participants either strongly agreed, agreed or were neutral about whether facial recognition was trustworthy. Speech recognition was similar with only 11 people disagreeing and 4 people strongly disagreeing. By comparison, the other three technologies (driverless vehicles, military robotics and stock investment) did not have the majority agree or strongly agree. However, military robotics had the lowest trust by far. Over 50% of participants (76 out of 143) either disagreed or strongly disagreed that they trusted military robotics. This was the only technology of the 7 where the majority of participants did not trust the technology. The answers were converted to continuous numeric values: strongly agree (5), agree (4), neutral (3), disagree (2), strongly disagree (1). The mean and mode of each question was calculated, as shown by the table below (Figure 5.4). The technologies described as having a greater trust had the four highest means (between 3.43 and 3.73) and all had a mode of 4. The means of driverless cars and stock investment were not significantly lower, 3.25 and 3.14 respectively. Driverless cars had a mode of both 3 and 4 (46 instances of neutral and agree) and stock investment had a mode of 3. Military robotics had a mode of 3, but the mean was 2.43, which was significantly lower. This is due to only 16% of those who responded to the survey either strongly agreeing or agreeing.



Technology	Mean Trust (Between 1 Strongly Disagree and 5 Strongly Agree)	Mode	Participants who had heard of the technology (%)
Speech Recognition	3.64	4	97
Driverless Vehicles	3.25	3,4	93
Facial Recognition	3.73	4	89
Personalised Advertisement	3.43	4	96
Medical Diagnostic	3.45	4	64
Military Robotics	2.43	3	53
Stock Investment	3.14	3	51

Figure 5.4: The mean trust participants had in each technology and percentage who had heard it

Trust in driverless cars and stock investments was lower as more participants were neutral about their trust. For example, 61 out of 143 participants were neutral about their trust in computers which can make stock investment. This is 43% of participants, which suggests that participants had less of a weighted view (positive or negative) towards this technology. As the graph shows no two technologies had the exact same level of trust. This suggests that the public's trust is specific to each technology and as such the public's perception towards a technology should be treated independently. Furthermore, with the exception of driverless cars there appears to be a correlation between whether the participant has heard of the technology and how far they trust it. Arguably there has been many negative articles about driverless cars in recent years such as deaths (BBC News, 2018). It is a technology which will most likely have a more radical impact on day to day life, than for example speech recognition. The mean of the answers was measured against the percentage of people who had heard of the technology. A positive correlation of 0.68 was found which shows a strong correlation between the two values (Figure 5.5). Additionally, removing driverless cars from the values measured saw the correlation increase to 0.74. Without or without driverless cars the correlation is strong enough to suggest there is a link between the two. This line of reasoning somewhat goes against the concerns in this overall dissertation. Rather than focusing on changing technology there may be an improvement in public trust in technology with more education.

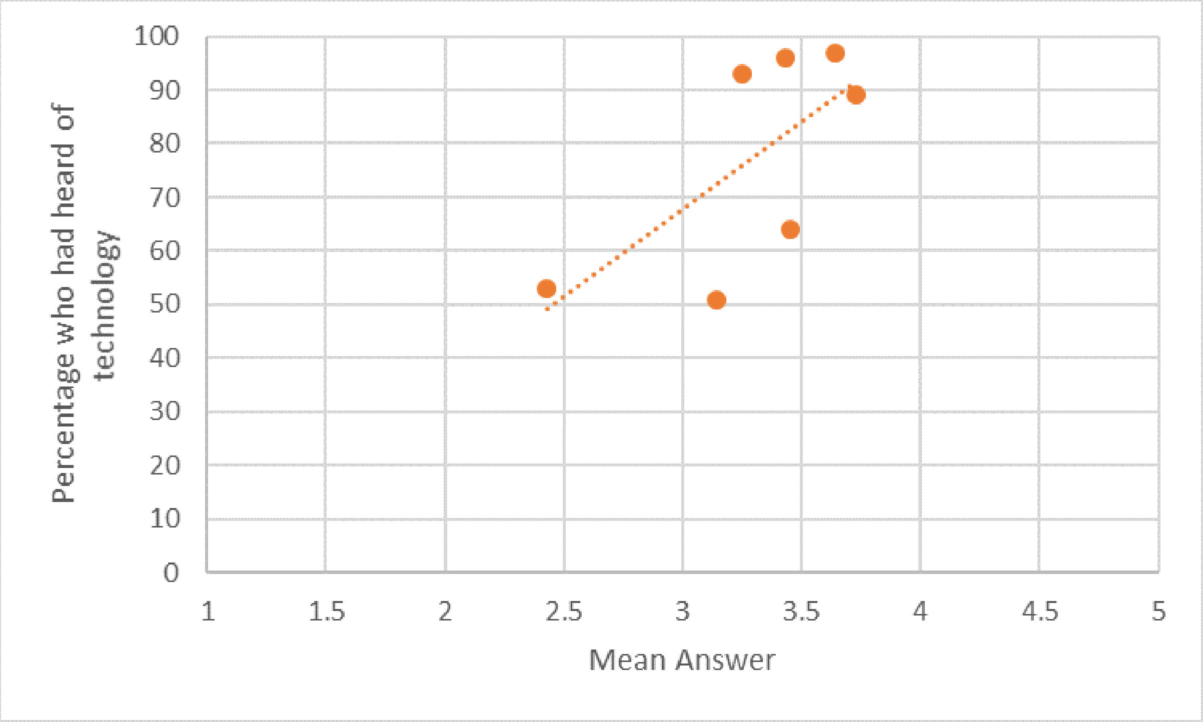
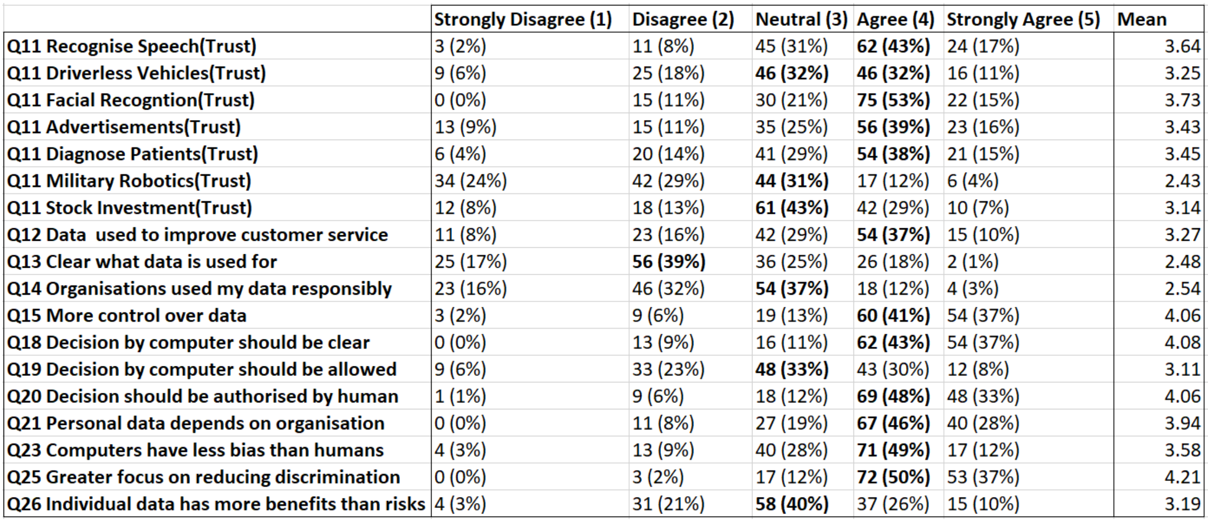


FIGURE 5.5: The mean trust in a technology plotted against the percentage of participants who had heard of it



	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)	Mean
Q11 Recognise Speech(Trust)	3 (2%)	11 (8%)	45 (31%)	**62 (43%)**	24 (17%)	3.64
Q11 Driverless Vehicles(Trust)	9 (6%)	25 (18%)	**46 (32%)**	**46 (32%)**	16 (11%)	3.25
Q11 Facial Recogntion(Trust)	0 (0%)	15 (11%)	30 (21%)	**75 (53%)**	22 (15%)	3.73
Q11 Advertisements(Trust)	13 (9%)	15 (11%)	35 (25%)	**56 (39%)**	23 (16%)	3.43
Q11 Diagnose Patients(Trust)	6 (4%)	20 (14%)	41 (29%)	**54 (38%)**	21 (15%)	3.45
Q11 Military Robotics(Trust)	34 (24%)	42 (29%)	**44 (31%)**	17 (12%)	6 (4%)	2.43
Q11 Stock Investment(Trust)	12 (8%)	18 (13%)	**61 (43%)**	42 (29%)	10 (7%)	3.14
Q12 Data used to improve customer service	11 (8%)	23 (16%)	42 (29%)	**54 (37%)**	15 (10%)	3.27
Q13 Clear what data is used for	25 (17%)	**56 (39%)**	36 (25%)	26 (18%)	2 (1%)	2.48
Q14 Organisations used my data responsibly	23 (16%)	46 (32%)	**54 (37%)**	18 (12%)	4 (3%)	2.54
Q15 More control over data	3 (2%)	9 (6%)	19 (13%)	**60 (41%)**	54 (37%)	4.06
Q18 Decision by computer should be clear	0 (0%)	13 (9%)	16 (11%)	**62 (43%)**	54 (37%)	4.08
Q19 Decision by computer should be allowed	9 (6%)	33 (23%)	**48 (33%)**	43 (30%)	12 (8%)	3.11
Q20 Decision should be authorised by human	1 (1%)	9 (6%)	18 (12%)	**69 (48%)**	48 (33%)	4.06
Q21 Personal data depends on organisation	0 (0%)	11 (8%)	27 (19%)	**67 (46%)**	40 (28%)	3.94
Q23 Computers have less bias than humans	4 (3%)	13 (9%)	40 (28%)	**71 (49%)**	17 (12%)	3.58
Q25 Greater focus on reducing discrimination	0 (0%)	3 (2%)	17 (12%)	**72 (50%)**	53 (37%)	4.21
Q26 Individual data has more benefits than risks	4 (3%)	31 (21%)	**58 (40%)**	37 (26%)	15 (10%)	3.19

FIGURE 5.6: The responses to the questions using Strongly Disagree to Strongly Agree as answer type

The next page of the survey had a series of questions about data usage. There was a concerningly negative attitude towards data usage. Below is a summary of the respondents to the survey questions on page 2 (Figure 5.6 Q12-Q15) (Appendix A.5). Almost none of the questions had a positive response towards the data usage by companies except for the first question. The first question had 47% of respondents either agree or strongly agree that their data is used to improve customer service. By comparison only 24% either disagree or strongly disagree with the statement. This is still a quarter of respondents, so a sizeable amount. Notably only 18% either strongly agreed or strongly disagreed which suggests most participants did not have a polarised view on this, especially considering 29% were neutral. This produced a mean of 3.27, which is closest to neutral further suggesting most participants did not feel strongly for or against the statement. There is still enough disagreement with the statement to suggest attitudes could improve. A company like Netflix do not want a quarter of the public questioning the use of the data they have collected. The next question (Q13) showed a more substantial trend towards one attitude. Over 50% (56%) of the survey respondents either disagreed or strongly disagreed that organisation made it clear what their data is used for. This is stark contrast to those who agreed or strongly agreed that it is clear what their data is used for, they made up only 19% of respondents; a quarter of respondents were neutral. This is a worryingly low percentage as it means many participants do not understand what they are consenting to when they give away their data. This produced a mean of 2.48, which is closest to disagree. The fact that the mean is disagree is problematic. Consenting to a service in return for your data should be an agreement with transparency. When clicking consent to a service it is assumed that you understand the agreement, but this may not be the reality (Berreby, 2017). The next question (Q14) was a continuation of this topic. The statement was “Organisations (such as Netflix, Facebook and Amazon) use my data responsibly” and a question asking, “how far do you agree?”. This also had a low level of agreement; it was even lower than the previous question with only 15% agreeing or strongly agreeing that their data is used responsibly. The mean was 2.54 due to the majority either being neutral (37%) or disagreeing (32%), which totalled 69% of participants. More people strongly disagreed (16%) than the combination of those who agreed (12%) or strongly agreed (3%). These two questions may be related. If the general public better understood what their data was used for, they may also more strongly believe their data is used responsibly. The lack of trust in how data is used may come from the unknown. Organisations don’t want a sizable minority lacking trust let alone most of the public. A survey of 145 participants does not represent the general population around the world, but it does not support trust in data usage. The overarching trend of this section is that the perception of data usage is subpar and organisations should do more to improve public opinion. The final data usage question (Q15) related to data control. The statement was “Organisations (such as Netflix, Facebook and Amazon) should give me more control over my data.”. The participants overwhelmingly agreed (41%) or strongly agreed (37%) with this statement. This amounted to 78% of participants which shows that a considerable amount of participants want more data control. Only 8% either disagreed (6%) or strongly disagreed (2%) which shows that only a small minority are against more control. This caused the mean answer to be 4.06, which is closest to agree. The mean further supports the conclusion that the public want more data control. What is meant by more data control is something that may differ from person to person. This could be easier options to request and delete your data or less power organisation have over your data. When looking at this collectively it appears that there is a problem with the current relationship the public have with the organisations they provide data to. All stakeholders involved would benefit from the public better understanding what they consent to when agreeing to a site’s terms and conditions for example. How this can be achieved is a big issue in of itself and much improvement is needed.

The rest of this page of the survey was about GDPR. It entailed two simple questions of whether the participant had heard of GDPR (Q16) and if so, how much they believed they knew about it (Q17). A positive amount (86%) of those surveyed had heard of GDPR. It’s important for the public to understand the laws which protect their data and which they must abide by. GDPR is one step towards greater data protection and more sound data usage universally. The public may not have a positive image of current data usage but understanding the laws which organisations must follow means that the public at least know the limits and constraints to what they give away. If only a low proportion of the public had heard of GDPR this may be one of the topics of education to help the public perception of how their data is used. When giving away data via an agreement you may not understand the specific agreement but understanding the laws in place to protect you means that you at least know a minimum standard all organisation must follow. What is done beyond this however will still be unknown, if you do not understand what you are consenting to.

Those who had heard of GDPR had varying levels of knowledge, as determined by the person answering. Those living within the EU would've most likely seen GDPR in their national news sources, but it doesn't mean they necessarily understood the implications. By contrast there are those who would've been forced to learn about GDPR due to their place of employment. 60% of the participants (including those who hadn't heard of GDPR) stated they knew at least a fair amount about GDPR. This is a relatively high amount of people, as most regulations do not have this amount of wide spread knowledge. This means that data usage as a topic is something that the public are already widely exposed to. Despite this very few people think their data is used responsibly. This may suggest that GDPR and similar regulations in other countries have not done enough to make the public believe that their data is used responsibly.

The previous page dealt with data usage and this page dealt with data usage again and also computer decisions (Figure 5.6 Q18-21). This page had four statements and asks the participant how far they agreed with each. The first one (Q18) was related to when a decision is made by a human vs a computer and whether the distinction should be clearer. A large majority agreed or strongly agreed that “Organisations should make it clearer when a decision is made by a computer which directly affects an individual. For example, a bank loan application”, in fact they totalled 80% of all survey participants. This high agreement does not quantify how clear participants currently find decisions made by computers which affects an individual to be, but it does confirm a desire for change in how clear it is. Only 9% disagreed and no one strongly disagreed. This result would support a movement towards more clarity about decisions made by computers. Once again, the details of how this information is presented requires more thought and research. The next statement (Q19) was related to decisions by computers but it was instead about whether a computer should actually be used when the decision affects an individual. This had a very mixed response, the most mixed response of the entire survey. The mixed response is interesting as it showed a real divided opinion of the extent in which computers should be used. This is a longstanding battle of technology vs the status quo. This may be generational, and it will be discussed later in this section. More people either strongly agreed or agreed with the statement “Organisations should be allowed to use Machine Learning/computers to make decisions that impact individuals” than those who strongly disagreed or disagreed but with a difference of only 9% (13 participants) it is fair to not make conclusions on the default view on this topic. Moreover, the majority (33%) were neutral which further reinforces the fact that there was not a consistent response. The general public had varying opinions about the statement which is useful as well. Does greater involvement of computers in our lives come with the cost of disaffecting a proportion of the population? There will always be those who oppose new technology, but it may be a case of what quantity of people oppose the new technology. This statement is related to the question of what roles computers play in our society and may explain the mixed response. The next statement (Q20) almost bridges the gap between the first two statements. It’s a question of whether computers should be used without the constraint of humans. Unlike the previous statement this had a clear leaning towards agree or strongly agree. 81% either agreed (48%) or strongly agreed (33%) that “Decisions made by a computer should be authorised by a human when possible”. A large majority agree but they may agree with a different version of what is when possible. The when positive part of the statement is subjective as some may consider it not possible if additional humans had to be hired (monetary cost) for the authorisation of computer tasks or the time it takes meant that the operations of a company slowed down. Others may think a cost of any kind is worth it if it’s possible for a human to authorise the task. Rather than concluding the logistics of the statement we can conclude that complete automation, one without human authorisation, was not a popular. This fact that only 6% disagreed and under 1% strongly disagreed is evidence that the public are against the removal of human control and allowing decisions to be made by a computer without human authorisation.

The final question (Q21) related back to data usage discussed on the previous page of the survey. The way participants answered was as expected. Most people treat each organisation independently and as such it’s rational that the amount of data you provide varies. The majority, totalling 74%, strongly agreed (28%) or agreed (46%) that “The amount of personal data I provide depends on the organisation. For example, Amazon, Facebook and your health service provider”. GDPR, as discussed above, may have created a standard for organisations, but it does not stop most survey participants dictating the amount of data they provide based on the individual organisation. Only 8% disagreed with the statement and no one strongly disagreed, showing a clear majority who do vary how much data they provide. Terms and conditions are read by almost no one so the varying level of data provided must be based on the perceived usage of the data and reputation of the organisation rather than the known use of the data (Berreby, 2017).

The final page of survey questions more closely related to the research of the data experiments (Figure 5.6 Q23-Q26) (Appendix A.9). It covered reidentification, algorithm bias and data usage. The first question (Q22) on this page was about reidentification. It gave a brief description of how anonymised data can be reidentified. This was followed by the question "Does this impact your trust in an organisation's use of your data?". Most participants (61%) answered yes to this question. The remaining 39% either answered no (18%) or were unsure (21%). More participants answered unsure than no which suggests that a greater explanation of reidentification may have been required or they are did not want to answer a definitive yes or no. Furthermore, the fact that only 18% explicitly answered no compared to the 61% who answered yes shows that it is a factor which impacts a lot of participants trust in the use of their data. It is not clear whether this is a factor which participants already knew about or if this is new information that impacted their trust. Risk of reidentification, much like stolen data, is something which organisation will not want to publicise but it's important that there is transparency. There is often a presumption that data is highly secure when it is provided to an organisation, especially a large one. Most companies have extensive procedures to prevent this, but accidents do happen. There isn't much someone can do if there is a data breach/mishandling of their data, but if the public had a better understanding of the risks involved when giving their data away it may make the agreement between the individual and organisation more honest. Furthermore, it may make individuals more conservative about the data given away. Data breaches, data reidentification and data mishandling is hard to eradicate completely and as such should be known.

The next question (Q23) was related closely to the ethics of Machine Learning. It was a statement of "Computer systems generally have less bias than humans" and asked how far they agree. Most participants either agreed or strongly agreed with this statement. This was expected when the survey was written, there was an assumption that most of the public believe computers to be less bias. This becomes a somewhat disputed topic of what a computer system is and whether it is separate from or linked to those who built it. Systems are built by humans and as such their biases come with it. Likewise, the data set used is often based on human society and the biases that come with it, such as crime or employment data. When a computer system is known to have a bias such as discrimination towards a gender is this a bias of humans, the computer system or both? This is a contentious issue and impacts the role of computer systems but is its own research topic. 61% of participants strongly agreed (12%) or agreed (49%) with the statement compared to only 12% who either disagreed (9%) or strongly disagreed (3%). Notably the statement has generally in it which means that participants did not agree or disagree as to whether computers are always less bias than humans. Regardless, it showed that the majority shared the attitude that computers are less bias than humans. This leads onto the next two questions about algorithm bias and discrimination. Machine Learning algorithm will often find the biases/patterns in a data set as it is meant to and as such this can lead to unwanted side effects. The next question (Q24) had a statement of "Computer systems have been known to amplify discrimination in society" and gave two examples (an AI recruiting tool and an algorithm used to predict if a defendant was likely to reoffend). Only 24% of survey respondents had heard of the discrimination described. 61% of participants strongly agreed or agreed that "Computer systems generally have less bias than humans" but only 24% had heard of the algorithm bias mentioned. We shall not conclude that less participants would strongly agree or agree with more knowledge of algorithm bias, but it is a topic of further research as it may contradict their view of computer systems. The public overwhelmingly agreed with the next statement (Q25), however. 87% of participants either strongly agreed or agreed "A greater focus should be placed on ensuring computer system do not reflect the biases and discrimination in society" despite many (76%) not being aware of the discrimination mentioned in the previous question. Only 2% of participants disagreed and no one strongly disagreed, with the rest being neutral (11%). The high percentage of agreement indicates that the bias and discrimination found in computer systems is a matter which the public care about and would like to see more be done.

The final quantitative question (Q26) related back to data usage. It asked participants how far they agree with "Using the data of individuals generally has more benefits than risks to society". This had a very mixed response and as such shows the attitude towards the data of individuals is very mixed. The majority were neutral (40%) and a similar number of participants both agreed (26%) and disagreed (21%). The mean answer was 3.19, which is closest to neutral. The mean is slightly closer to agree than disagree which shows a small leaning towards using the data of individuals having more benefit than risk. This is minimal and shows opinion is both divide and not polarised towards a particular view; few people strongly agreed (10%) or strongly disagreed (3%). The fact that there is not a large leaning towards agree shows that more can be done to reduce the risk of using the data of individuals and improve the public trust towards the use of this data.

The last question of the whole survey allowed participants to optionally give any thoughts they wanted to share on Machine Learning or the use of their data. Most participants chose to leave this section blank. Of those who chose to leave a response some interesting points were raised which is summarised below. Most opinions below were that of a single individual. Multiple participants showed concern towards data usage, automated computer systems and consent.

1. There should be a limitation on how much data can be obtained about individuals for advertisement.
2. More steps should be put in place to explain how the data is collected.
3. Humans should stay "above" computers. Computers shouldn't control human lives.
4. Focusing on the safety of a computer system is important.
5. Allowing users to opt into data related features rather than out should be the default privacy setting. You should be able to access more services with having to share less data. Default settings should be least invasive e.g. recommended social media posts.
6. There is no guarantee that a computer system is unbiased, so a final decision should not be made by a computer.
7. Watching one video on a topic doesn't mean you want to be recommended topics in that video afterwards. Certain personalised advertisement generated by an algorithm can be unwanted, so a blacklist for personalised ads would be useful.
8. A human may not give an accurate account of their health and habits which a doctor may recognise when a computer would not.
9. Computers only do what a human decides. A computer cannot have bias, but the programmer can.
10. It should be people with technology rather than people or technology.
11. Companies make it hard to understand what you are consenting to, so you click OK to get quick access to the services essentially clicking OK to the unknown.
12. GDPR is a good step in the right direction and there needs to be accountability for what happens with data.
13. The discrimination found is caused by 'bad' data rather than the algorithm as surely an ML algorithm's goal is to find bias.
14. Most big companies do a good job of anonymising data.

15. The trust in the technology is based on the trust in the corporation.
16. Machine Learning will get better with time and it is the way towards improving quality of life and efficiency in industry.
17. Every process needs auditing and machines will requires the ongoing input from humans.
18. It is an individual's choice to provide data in the knowledge that personal preferences are used in a multitude of ways.
19. Netflix, Amazon and Facebook are separate in the way in which they influence the choice and spending of their customers and as such should not be grouped together.

5.1.2 Demographics

After analysing the survey as a whole the survey results were split into different data sets which include: gender, age, location, whether they work in the IT sector, whether they had heard of ML and their educational background. There had to be care taken when discussing the disparity between demographics as there were 145 participants, which is a sizeable amount but the difference in response between demographics may be reduced with more participants. The mean response to a question is impacted less by outliers when there is more participants. Below is a summary of the observations found.

5.1.2.1 Gender

The first demographic assesed was gender. The survey had 79 male, 65 female and 1 other respondent. The respondent who selected other identified as female, so they were included in the female demographic as using other as a demographic was not possible due to the low number, but their opinion should be considered as well. The first difference between demographics noticed was that a far higher percentage of the men (41.8%) were 18-25 in comparison to the women (19.7%). As well as this a far lower percentage of women (13.6%) worked in the technology sector in comparison to men (40.5%). This means that the difference in response from men and women is influenced by other factors as well.

Comparing gender appears to further support a correlation of knowledge in the technology with trust. In all cases (7 technologies) a higher percentage of men had heard of the technology and in 6 out of 7 cases they also had higher trust. The only exception in trust was personalised advertisements. Women had a mean response of 3.63 whereas men had a mean response of 3.26, a difference of 0.37. The largest difference in trust was medical diagnostic with a difference of 0.55, with a higher trust from men. Additionally, a higher percentage of men had heard of ML which helps explain the higher knowledge of the technologies mentioned. The means were plotted on the graph like in the previous section to see if there was a correlation between the mean trust and if they had heard of the technology (Figure 5.7). There was a correlation of 0.71 when using the 7 mean trust responses for each gender (14 total data points).

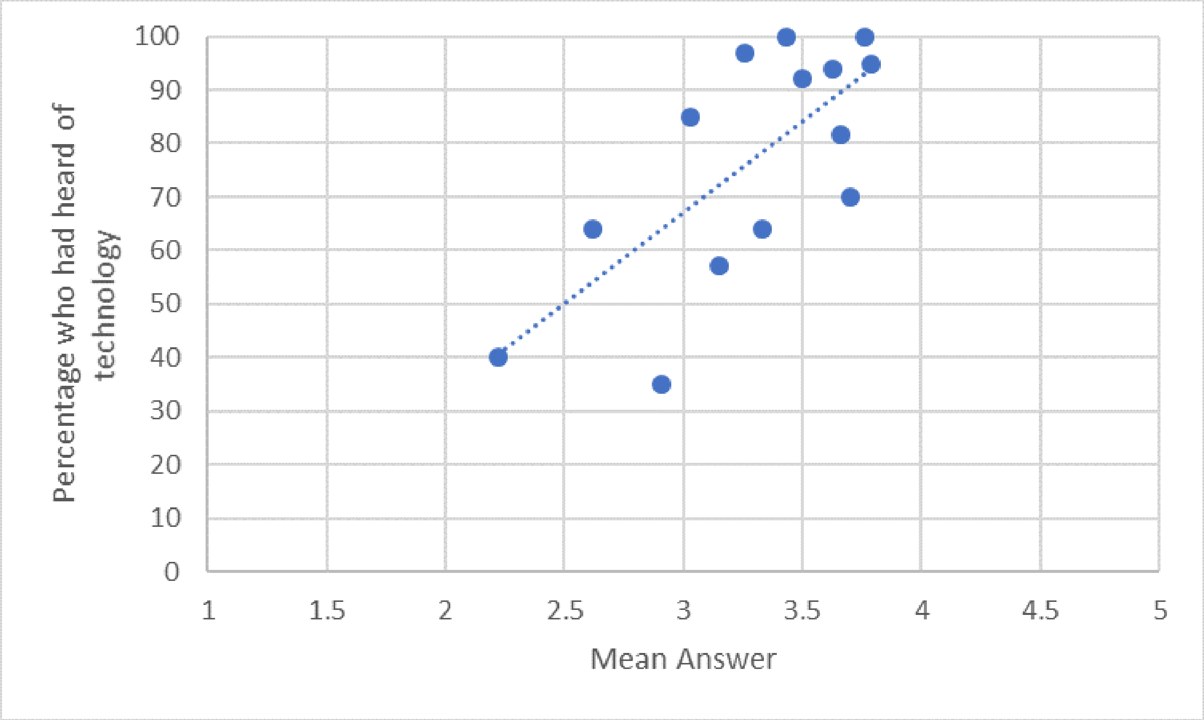


FIGURE 5.7: The mean trust in a technology plotted against the percentage of participants who had heard of it using both Male and Female survey results

There were only three cases in the rest of the survey where the difference in mean response was more than 0.3. This is still a fairly low difference as both means are most likely closest to the same answer e.g. 3/neutral. Men more strongly disagreed than women with the two statements about organisations making data usage clear (Q13) and decisions authorised by a human (Q20). These had a difference in mean of 0.3 and 0.36 respectively. The other case had a far larger difference in response of 0.76 which was the statement "Organisations should be allowed to use Machine Learning/computers to make decision that impact individuals". Women were more against this as only 22% of women either agreed or strongly agreed in comparison to 51% of men. Men and women had similar views when looking at the responses of most questions except this one. The fact that only this question had a large difference is compelling. The role of computers may be the most controversial topic.

5.1.2.2 Age

Age appears to have the largest disparity in responses out of any demographic factor. 26-35 year olds were unrepresented in this survey but this did not cause the large disparity in responses, which could have been caused by outliers in this demographic. The mean response from each age demographic varied for every single question in the survey with 0.09 being the smallest difference in mean responses, 0.23 was the next smallest. The difference however was not a linear trend from young to old. It was not a case that the mean went up or down from 18-25 to 50 and older. There were cases where 18-25 year olds had a similar response to 50 and over and it was the middle two age groups which caused the difference. For example, there is a case where each of the four age groups had the highest trust. It was not a case that older or younger people consistently had higher trust. 18-25 year olds had the lowest trust in driverless cars which could be deemed surprising as it is an emerging technology. They did however have the highest trust in medical diagnostic and stock investment.

The lack of trend from young to old was true for most questions in the rest of the survey as well. There were three examples where younger people most strongly agreed that: "Organisations use my data to improve customer service", "Organisations should give me more control over my data" and "Organisations should be allowed to use Machine Learning/computers to make decisions that impact individuals". In these cases, the mean response went down with each age group e.g. 36-49 year olds and 50 and over. The first statement had a difference of 0.7 from 18-25 year olds (3.72) to 50 and over (3.02). The 50 and over response was much closer to neutral. The third statement went from agree/neutral to disagree/neutral. The 18-25 year olds had a mean response of 3.5 whereas 50 and over had a mean of 2.84. This was the only shift from agree to disagree found between age groups, but it was minor as both means are still close to neutral. The opposite was true in one example with agreement going from older to younger. 18-25 year olds most strongly disagreed with "Organisations (such as Netflix, Facebook and Amazon) use my data responsibly" (Q14). Even though the trend was not generally from young to old there was an evident difference in mean response for all questions. The difference varied from 0.09 to 0.7 so there was always a difference in response, but the disparity varied. Overall these results suggest that trust and attitude towards the topics discussed were not strictly generational, a greater pattern may be found with more respondents. Additionally difference in mean response should be expected with only 145 participants and 4 age groups. The mean difference above a certain amount e.g. 0.5 is enough to suggest some difference in views.

5.1.2.3 Location

This section is concerned with the difference in opinion of those living in and outside of the UK. No significant trend was found between those living inside and outside of the United Kingdom. There was a difference in the mean response for each question, but this is expected. The only question where the difference in mean response was remarkable was the question with the statement “Organisations should be allowed to use Machine Learning/computers to make decisions that impact individuals”, which has consistently had a difference in opinion between demographic groups. The mean response for those living within the United Kingdom was 3.34 whereas those living outside of the United Kingdom had a mean response of 2.78, so there was a small change from agree (4) to disagree (2), but both were closest to neutral (3). The attitude towards the questions did not greatly vary based on whether the participant lived inside or outside of the UK.

5.1.2.4 Works in IT Sector

It was of interest to see if those working in the IT sector had a different view to those who didn’t. This demographic comparison mostly promoted the observation previously found that those who had heard of the technology also had higher trust in the technology. For all technology, except for facial recognition, a higher percentage of those who worked in the IT sector had heard of it. In turn those working in the IT sector had higher trust for 5 out of the 7 technologies; the other two technologies had very similar trust between the two demographic groups. Beyond this section most responses were similar with the biggest exception being those who don’t work in the IT sector more strongly agreeing that decisions made by a computer should be authorised by a human when possible. Working in the IT sector did not appear to significantly impact the attitude of participants.

5.1.2.5 Heard of ML

The difference between those who had heard of ML and those who hadn’t was similar to those who do and don’t work in the IT sector. Those who had heard of ML also had a greater knowledge of the technologies. Much like the IT sector, those who had heard of ML also had a higher trust in 5 out of the 7 technologies; it was not the same 5 technologies. Additionally, the biggest difference in response was for the question about decisions made by a computer which impact individuals and whether it should be authorised by a human when possible, which was also the case for whether the participant worked in the IT sector. Whether or not the participant had heard of ML did not appear to greatly impact their response beyond the expected different in mean response.

5.1.2.6 Education

The final demographic factor looked at was education. The participants were split into three distinct categories: high school (or equivalent) or less, undergrad or postgrad education. A major pattern was not found between response and education, but some observations were found. Participants with a postgrad education had a notably low trust in driverless cars, with a mean trust of 2.85. The other two largest difference in mean response was how far they agreed with "Organisations should make it clearer when a decision is made by a computer which directly affects an individual" and "Computer systems generally have less bias than humans". The education groups had different questions have disparity, in comparison to the other demographic factors. Educational background did not have a major trend but the differences it did have were different to the other demographic factors.

5.1.3 Conclusion

The survey was successful in finding the public standpoint of the issues raised. 145 participants was enough participants for trends to form. The current state of public opinion of the topics raised is not adequate. There is a clear area for improvements to be made both by continuing the gains made by GDPR and educating the public. Education means companies making the usage of the data clearer and also knowledge of new technology being widespread amongst the public. A strong correlation (0.68) was found between whether participants had heard of the technology and trust.This suggests that part of trust in technology improving is not only the technology but how it is shared to the public. Technology is widely adopted when the general public have a understanding of it rather than just enthusiasm from early adopters.The large proportion of survey respondents having heard of GDPR means that progress is being made. Educating the public does not take away the areas in which organisations and the tech industry as a whole need to change. Companies should be actively trying to boost the relationship customer's have with the data and technologies they use. It shouldn't be a scenario of giving away your data to use a service when you don't know what you are giving away. Consistently throughout the survey the public exhibited that there was a general consensus that more can be done to improve clarity of both data usage and when decisions are made by a computer. Extending from this the public want more control over their data. Furthermore the role of computers is divided, especially based on the limitations of current technology based on the qualitative feedback. Automating decisions made by a computer which affect an individual was the most split question. It stood out out as the one question which did not have a majority share a view. Relating back more closely to the topic of Machine Learning algorithms the survey helped motivate the purpose of this dissertation. 87% of participants agreed or strongly agreed that ”A greater focus should be placed on ensuring computer system do not reflect the biases and discrimination in society” meaning that the aim of this dissertation is in the interest of many. Demographics did not greatly impact the survey responses as it was more a case of a slight shift in mean response e.g. mean of 3.5 to 3 rather than a different average survey response. The vast majority of individuals had similar views. The survey exposed that there are issues which the participants pretty universally agreed on.

Chapter 6

Data Set Experiments

6.1 Data Set Experiments

During the process of experimenting with data sets, 18 data sets were found and used with varying levels of success and as such only the most useful and distinct are described. The experiments were conducted in WEKA, a Machine Learning software written in Java. It contains a large range of Machine Learning algorithms and good pre-processing options. It was ideal for this project because it allows the user to quickly change datasets, preprocess and select specific attributes. The plan originally was to use the WEKA library within python, but automating the process was difficult when every data set was different so a standard couldn't be programmed; WEKA GUI became the better option. Output files of the data experiments can be found at Experiment Results. The rest of this chapter breaks down the observations found in each data set used.

6.1.1 Main Data Set: Student Alcohol Consumption (UCI Machine Learning, 2018)

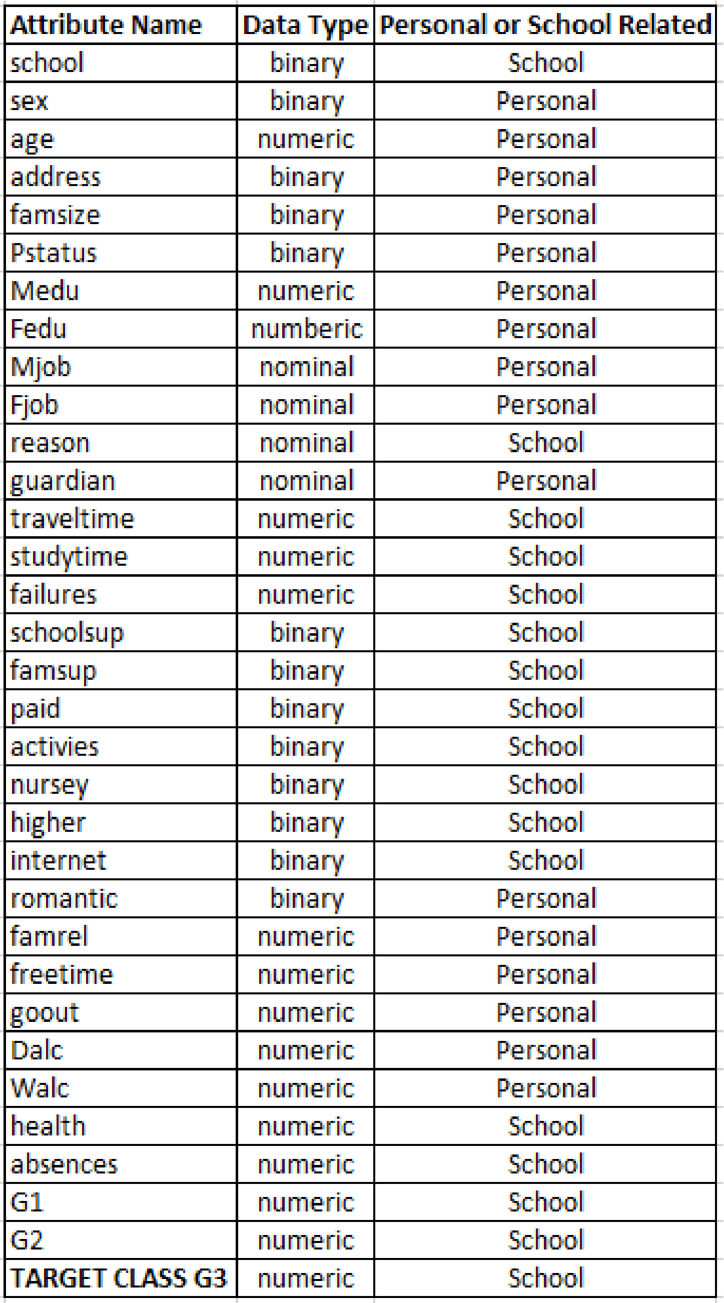
6.1.1.1 Data Description

This data set was created by combining a survey of Math and Portuguese language students with their final grades in each subject. This data set was released under CC0: Public Domain (UCI Machine Learning, 2018). The data is presented as two separate CSV files (one for Math grades and one for Portuguese). Despite the data set name, its focus is not only alcohol consumption. Alcohol consumption is one of many attributes within the data set. The attributes include many social factors, demographics and study information. The data can be used to predict a student's final grade or find correlation between attributes such as alcohol consumption and final grade. The Math data set consists of 395 instances and the Portuguese data consists of 649 instances. The two different subjects allows for good comparison. The main limitation of the data set is the number of instances, as a larger number of instances would be desirable. It is worth noting that the first and second period grade in the data set appear to highly correlate with the final grade and as such will be addressed.

6.1.1.2 Attribute Description

The data set contains data which can be categorized as personal or school related and as such will be differentiated between (Figure 6.1). Details of each attribute can be found in the Appendix.

Personal in this context is any attribute that is not directly related to the school rather than just personal data such as age. Differentiation between personal and school attributes will be used when experimenting with the data.



Attribute Name	Data Type	Personal or School Related
school	binary	School
sex	binary	Personal
age	numeric	Personal
address	binary	Personal
famsize	binary	Personal
Pstatus	binary	Personal
Medu	numeric	Personal
Fedu	numberic	Personal
Mjob	nominal	Personal
Fjob	nominal	Personal
reason	nominal	School
guardian	nominal	Personal
traveltime	numeric	School
studytime	numeric	School
failures	numeric	School
schoolsup	binary	School
famsup	binary	School
paid	binary	School
activies	binary	School
nursey	binary	School
higher	binary	School
internet	binary	School
romantic	binary	Personal
famrel	numeric	Personal
freetime	numeric	Personal
goout	numeric	Personal
Dalc	numeric	Personal
Walc	numeric	Personal
health	numeric	School
absences	numeric	School
G1	numeric	School
G2	numeric	School
TARGET CLASS G3	numeric	School

Figure 6.1: The attributes in the Student Alcohol Consumption Dataset

6.1.1.3 Observations found in data set before experiments

This data set was chosen because it had a large variety of personal attributes with varying levels of direct correlation with a student's final grade. The data set has many personal attributes which are being associated with a final grade which are not directly related to their school life. This is not a proposal that social factors do not impact a student's academic performance as there has been much discussion on this topic. The scope of this experiment is not a discussion of the social factors that affect students but rather whether they should be used to predict their academic performance using an algorithm. The data set is a closed world and as such there is a danger in classifying a student's final grade based solely on the provided attributes. Using demographics may help to further project stereotypes found in society. If this data was used for example to highlight students early on who may need additional support to get a higher final grade, it begs the question of whether or not an algorithm should be basing this on attributes which require greater context. This is the kind of context a support worker may understand which a system would not. Furthermore, it becomes an issue of to what extent a computer algorithm should be making decisions which are based on societal factors which are not a directly correlating numeric value but rather a factor which cannot be treated as an independent attribute.

This data set is interesting because it has two distinct problems of whether these personal attributes are needed and if they are, then whether the classifier should be used in the first place. The second problem is more contentious and still an area of great dispute. It's this assumption by some that a computer is not bias and as such it can use personal data without causing harm to society. The counter argument is that data is bias and as such a computer is as well. It becomes an issue for example if a system consistently predicted a certain gender would get a higher final grade. If the system consistently correctly predicted a certain gender would get a higher final grade without gender being a factor in training, you can argue that it is the general pattern of a certain gender outperforming the other rather than the system itself. One might argue that if the goal is to help as many students as possible, it is more important than avoiding stereotypes or disparity between demographics. This is a more general observation of the overall data set.

During preprocessing the distribution within each attribute was noted. Most notable was that G1, G2 and G3 (final grade) all share the classic bell curve distribution. This logically exists due to the attributes being school grades. School grading is usually designed so that the majority of students achieve around the average grade, for example a B ,with the minority being either being high or low achievers. The bell curve distribution in all three attributes means that G1 and G2 most likely highly correlate with G3.

Finally, it appeared having looked at the attributes that it would be possible to also predict attributes other than the final grade such as how much the student drank per week or how often they go out. It is not clear what sort of accuracy would be achieved however. Most of the attributes are numeric but they can easily be turned into nominal values using binning. These observations lead to the four specific research questions.

6.1.1.4 Data set research questions

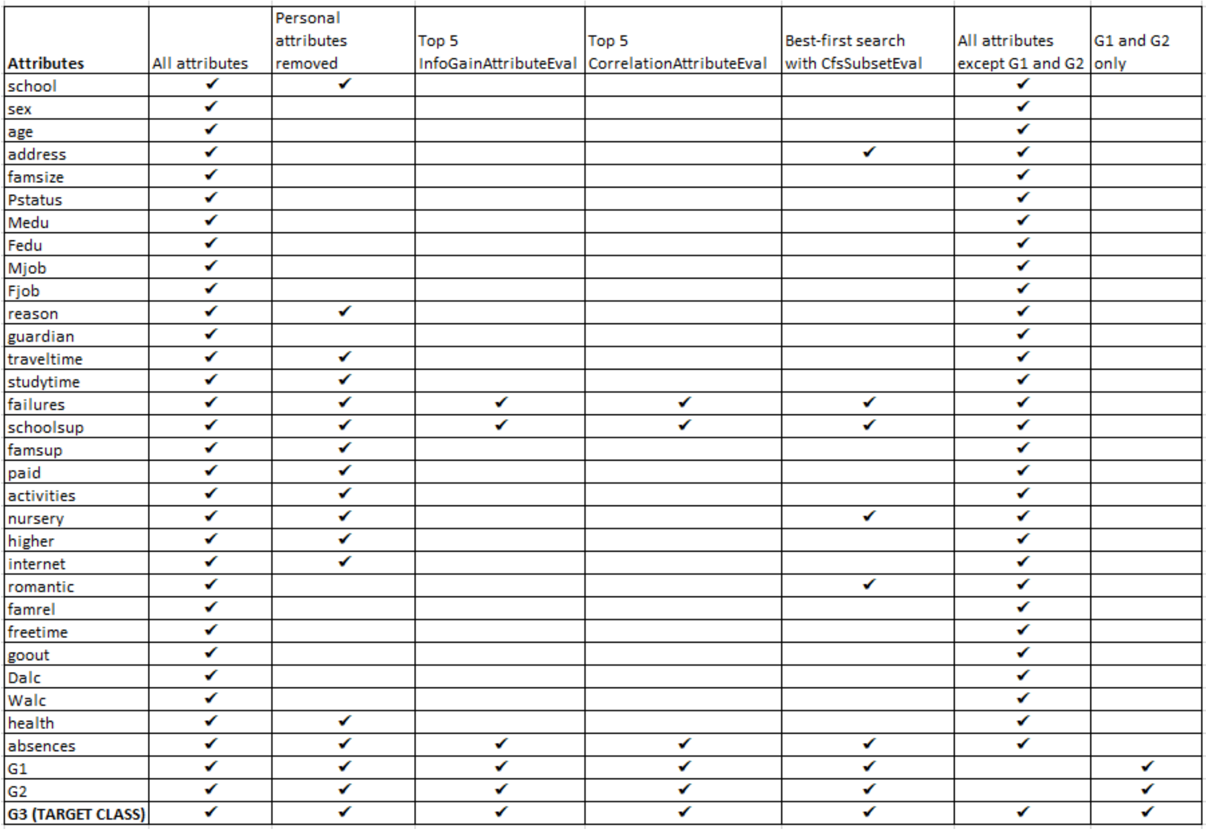
1. Does removing personal attributes (as defined in the attribute description) impact the accuracy of classifiers?
2. Can a high accuracy be achieved using only the G1 and G2 attributes?
3. Can a high accuracy be achieved without using the G1 and G2 attributes?
4. Can the data set be used to accurately predict personal attributes, rather than predicting the intended target class?

6.1.1.5 Result of experiments (Machine Learning)

All the results of the experiments can be found in the provided text files. This ended up totalling 240 text files and the most relevant observations will be summarised.

Selecting attributes

Five different data sets were created by selecting attributes based on a variety of methods. These data sets consisted of a dataset with all attributes, personal data removed, the top 5 attributes after personal data was removed using both InfoGainAttributeEval and CorrelationAttributeEval in WEKA with the ranker search method and best-first with CfsSubsetEval in WEKA. The ranker search method ranks attributes by their individual contribution. Both methods were used to reduce reliance on one method. InfoGainAttributeEval measures the information gained on the class based on the individual attribute. By comparison, CorrelationAttributeEval measures the Pearson correlation coefficient between the attribute and the class. The best-first search method was used with CfsSubsetEval attribute evaluator setting in WEKA. This evaluates a subset of attributes (from the complete data set) by weighing both the individual predictive ability of an attribute and also the correlation between attributes (removing redundancy if two attributes have the same predictive ability). This is different to the ranker method as it both evaluates the usefulness of individual attributes and the relationship between them. Finally, data sets were created for experimenting with G1 and G2. A dataset was created with only G1, G2 and the target class G3, whereas the other is the opposite; it has all attributes except G1 and G2. Figure 6.2 and Appendix A.11 is a summary of the data sets produced from the different attribute selection methods. All three attribute selection methods lead to the selection of schoolsup, absences, G1 and G2 in the Math grade data set. This confirms that G1 and G2 highly correlate with the final grade as suggested by the bell curve distribution in the Math grade data set. All three attribute selection methods lead to the selection of failures, higher and G2 in the Portuguese grade data set. G1 was not chosen by the best-first search method. It is most likely not chosen because it correlates so highly with G2, rather than because it does not correlate with G3.

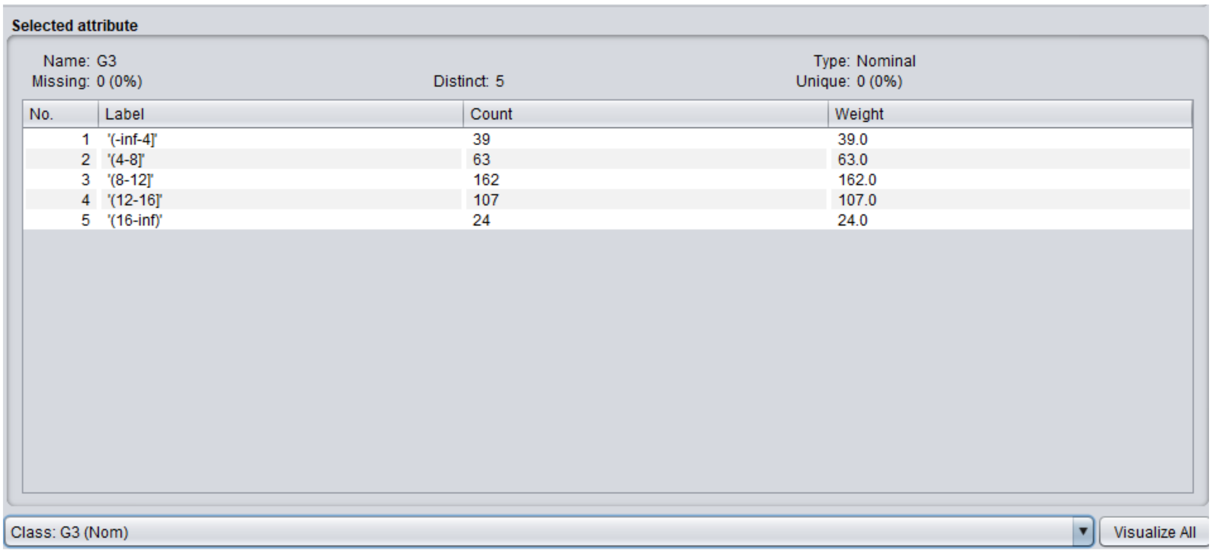


Attributes	All attributes	Personal attributes removed	Top 5 InfoGainAttributeEval	Top 5 CorrelationAttributeEval	Best-first search with CfsSubsetEval	All attributes except G1 and G2	G1 and G2 only
school	✓	✓				✓	
sex	✓					✓	
age	✓					✓	
address	✓				✓	✓	
famsize	✓					✓	
Pstatus	✓					✓	
Medu	✓					✓	
Fedu	✓					✓	
Mjob	✓					✓	
Fjob	✓					✓	
reason	✓	✓				✓	
guardian	✓					✓	
traveltime	✓	✓				✓	
studytime	✓	✓				✓	
failures	✓	✓	✓	✓	✓	✓	
schoolsup	✓	✓	✓	✓	✓	✓	
famsup	✓	✓				✓	
paid	✓	✓				✓	
activities	✓	✓				✓	
nursery	✓	✓			✓	✓	
higher	✓	✓				✓	
internet	✓	✓				✓	
romantic	✓				✓	✓	
famrel	✓					✓	
freetime	✓					✓	
goout	✓					✓	
Dalc	✓					✓	
Walc	✓					✓	
health	✓	✓				✓	
absences	✓	✓	✓	✓	✓	✓	
G1	✓	✓	✓	✓	✓		✓
G2	✓	✓	✓	✓	✓		✓
G3 (TARGET CLASS)	✓	✓	✓	✓	✓	✓	✓

FIGURE 6.2: The selected attributes for the Math grade data set

Classifying final grade (G3)

To be able to calculate the classifier accuracy of the final grade (G3) it was changed from a numeric to nominal value using equal-width binning. It was decided that five bins was an appropriate amount as it kept the bell curve and represented the grading system well as most students are within the middle three bins (Figure 6.3). The Math grades had more of a bell curve than the Portuguese grades. The majority of the Portuguese grades were in the third or fourth bin, so it was harder to achieve a nice bell curve (Appendix A.15). Additionally, only five bins were used, as more than five would start to see a drop in accuracy, as more classes is harder to predict. The original grading system in the data set was between 0 and 20, which means that the bins represented Class 1 0-4(low), Class 2 4-8(low/average), Class 3 8-12(average), Class 4 12-16(average/high) and Class 5 16-20(high). The data set is relatively small (395 Math grades and 649 Portuguese grades), which means that each correctly classified instance has a big impact on accuracy which has been taken into account when discussing accuracy. We shall now discuss the results of the classification.



Selected attribute

Name: G3 Type: Nominal
Missing: 0 (0%) Distinct: 5 Unique: 0 (0%)

No.	Label	Count	Weight
1	'(-inf-4]'	39	39.0
2	'(4-8]'	63	63.0
3	'(8-12]'	162	162.0
4	'(12-16]'	107	107.0
5	'(16-inf)'	24	24.0

Class: G3 (Nom)

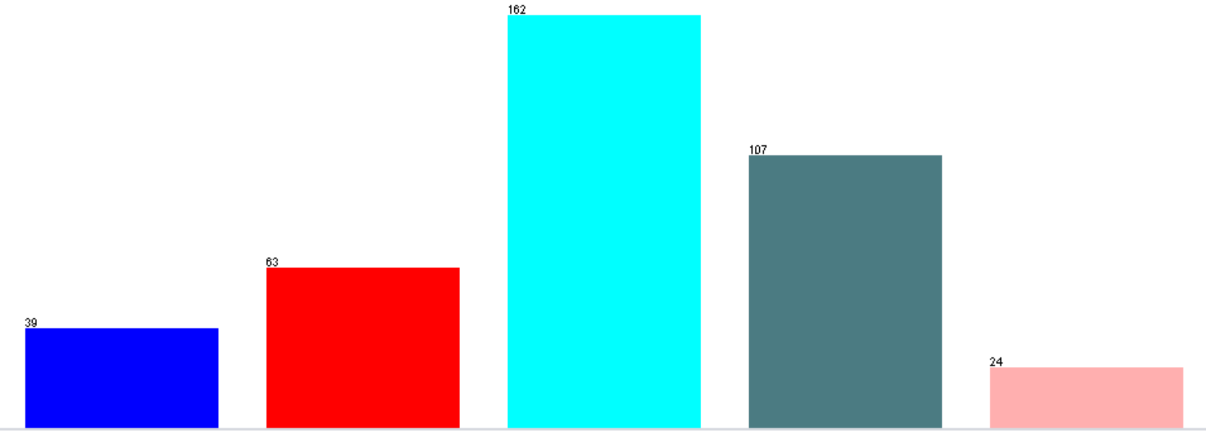
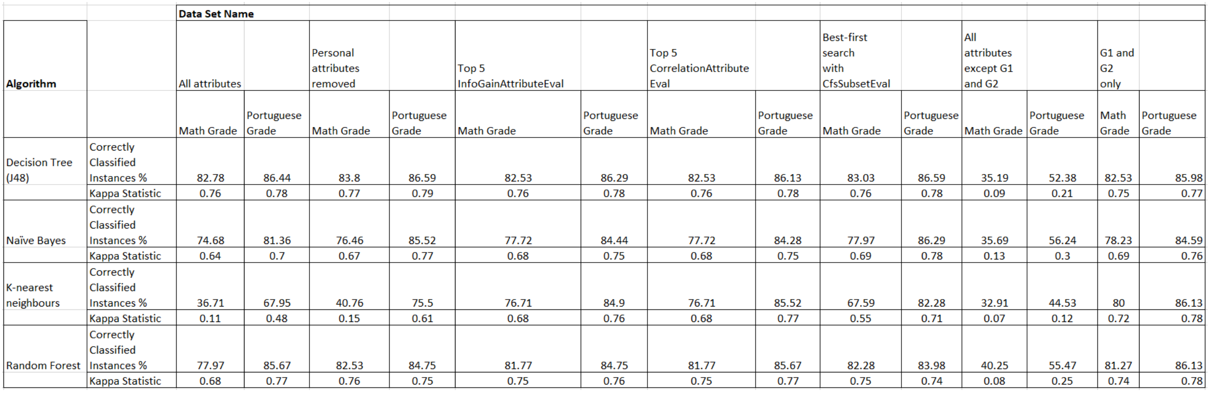


FIGURE 6.3: Target class distributed to five bins using equal-width binning in the Math student data



		Data Set Name													
Algorithm		All attributes		Personal attributes removed		Top 5 InfoGainAttributeEval		Top 5 CorrelationAttributeEval		Best-first search with CfsSubsetEval		All attributes except G1 and G2		G1 and G2 only	
		Math Grade	Portuguese Grade	Math Grade	Portuguese Grade	Math Grade	Portuguese Grade	Math Grade	Portuguese Grade	Math Grade	Portuguese Grade	Math Grade	Portuguese Grade	Math Grade	Portuguese Grade
Decision Tree (J48)	Correctly Classified Instances %	82.78	86.44	83.8	86.59	82.53	86.29	82.53	86.13	83.03	86.59	35.19	52.38	82.53	85.98
	Kappa Statistic	0.76	0.78	0.77	0.79	0.76	0.78	0.76	0.78	0.76	0.78	0.09	0.21	0.75	0.77
Naïve Bayes	Correctly Classified Instances %	74.68	81.36	76.46	85.52	77.72	84.44	77.72	84.28	77.97	86.29	35.69	56.24	78.23	84.59
	Kappa Statistic	0.64	0.7	0.67	0.77	0.68	0.75	0.68	0.75	0.69	0.78	0.13	0.3	0.69	0.76
K-nearest neighbours	Correctly Classified Instances %	36.71	67.95	40.76	75.5	76.71	84.9	76.71	85.52	67.59	82.28	32.91	44.53	80	86.13
	Kappa Statistic	0.11	0.48	0.15	0.61	0.68	0.76	0.68	0.77	0.55	0.71	0.07	0.12	0.72	0.78
Random Forest	Correctly Classified Instances %	77.97	85.67	82.53	84.75	81.77	84.75	81.77	85.67	82.28	83.98	40.25	55.47	81.27	86.13
	Kappa Statistic	0.68	0.77	0.76	0.75	0.75	0.76	0.75	0.77	0.75	0.74	0.08	0.25	0.74	0.78

FIGURE 6.4: 10-Fold Cross Validation Student Results

The main goal of this experiment was to see if removing personal attributes would impact accuracy. The discussion in this paragraph refers to the results of the test option 10-fold cross validation in Figure 6.4. It was found that it improved accuracy in 7 out of 8 cases (the 4 algorithms used on both datasets). The Random Forest algorithm applied to the Portuguese data with personal attributes removed was the exception. It saw a decrease in accuracy of 85.67% to 84.75%, which was only 6 less attributes than the original data set, due to the small size of the data set. The amount that accuracy improved by varied quite a bit. For example, the K-nearest neighbour algorithm applied to the Portuguese data set saw a significant increase in accuracy from 67.95% to 75.5%. This is not surprising however, due to the nature of K-nearest neighbour. Having less attributes makes it easier to create centroids and find a relationship between the target class and attributes. The K-nearest neighbour algorithm uses the Euclidean distance for it's nearest neighbour algorithm. This is the linear (straight line) distance between attributes in the Euclidean space and as such is not well equipped for many attributes. The other algorithms have improved accuracy due to a reduction in noise, but K-nearest neighbour is improved by both this and less attributes. This may explain why its highest accuracy (86.13%) is achieved when using only G1 and G2. By comparison, applying the decision tree algorithm to the Portuguese grade data set with personal attributes removed saw an increase of just 86.44% to 86.59%. This is small enough to suggest no real increase in accuracy, as this could be the reverse with a few small changes to the order of the data set or parameter tuning.

It is important not to just look at the classifier accuracy when analysing the results of an ML algorithm. Sometimes a classifier's accuracy can go up due to the true positive rate of just one of the classes improving. This was generally found to not be the case, but it was also not the case that the true positive rates went up for all classes when the accuracy went up. There was actually only one instance (out of eight) in which the true positive rate went up for all classes, which was Naive Bayes applied to the Portuguese dataset. By comparison Naive Bayes applied to the Math dataset saw the true positive rate of classes go up and down: Class 1 (0.667 → 0.564), Class 2 (0.556 → 0.651), Class 3 (0.833 → 0.802), Class 4 (0.766 → 0.832) and Class 5 (0.708 → 0.833). This meant the average true positive rate went up, but not the true positive rate of all classes. Additionally, in some cases the true positive rate did not change for some of the classes. There was only one example when at least 4 of the classes did not have a change in true positive rate. This was the decision tree applied to the Portuguese data set. Class 4 and Class 5 remained at 0.861 and 0.707 respectively. Ideally when when trying to improve the accuracy of the classier you don't want to be lowering the true positive rate of some classes in the process. Looking at the true positive rates showed that the change in accuracy was not due to a single class, but all the classes. The main take away from using the full data set vs personal attributes removed data set is that in general the accuracy did not go down, which would suggest the personal attributes are not a key component of classifying the final grade. From an ethical standpoint it puts the case forward that personal attributes should only be used if necessary.

As well as comparing the full data set to a data set with personal attributes removed, there was further attribute selection by using a data set with the top 5 attributes of the data set with the personal attributes removed. This produced mixed results. The table shows that there was not a significant enough change in accuracy to warrant any conclusions for the Decision Tree, Naive Bayes or Random Forest algorithm (Figure 6.4). The k-nearest neighbour algorithm saw a large improvement in accuracy which is likely due to its linear nature previously discussed. K-nearest neighbour generally performs well when there is fewer attributes. It saw a very large increase of 35.96% in the Math grade data set with the kappa statistic going up from 0.15 to 0.68. This was the only data set where further attribute selection saw a notable change in accuracy.

Selecting attributes with best-first search was a different kind of attribute selection as this started with the full data set. In both the case of the Math and Portuguese data sets it chose a mixture of personal and school related attributes. This would suggest that there are some personal attributes have help in classifying the final grade G3. The accuracy's produced were similar to the data set with personal attributes removed. There was not enough of an improvement in accuracy to conclude that personal attributes were required to produce the best accuracy's.

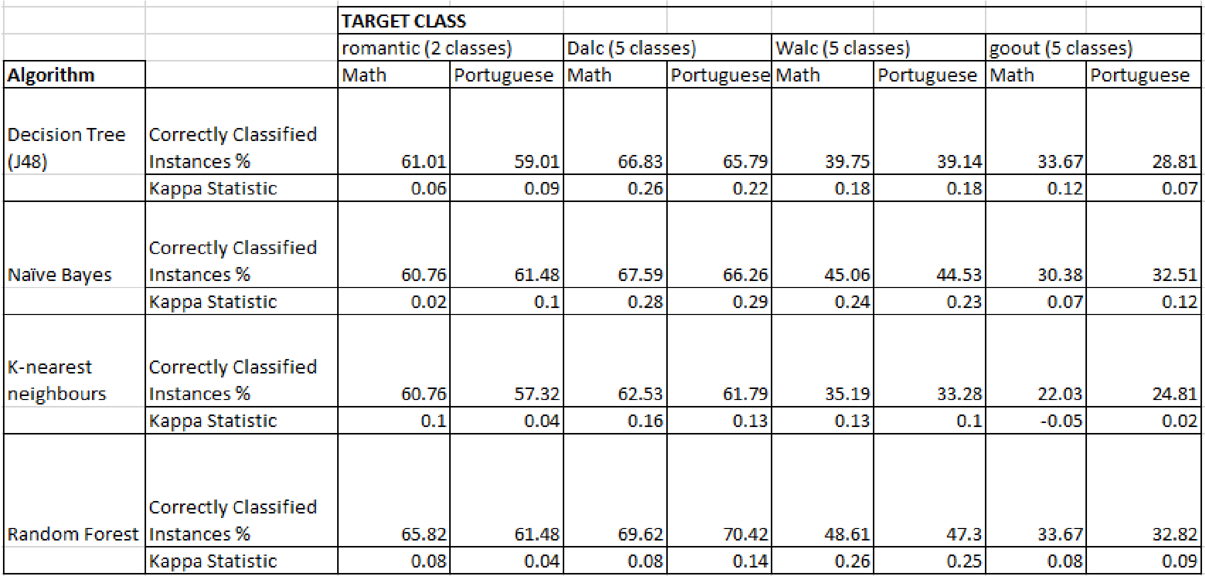
The assumption when this data was chosen was that final grade would correlate most with G1 and G2 and as such these were necessary if you wanted to achieve a high accuracy. This was found to be true. The data set with G1 and G2 removed produced very poor results. The kappa statistic was between 0.07 and 0.25 for the 4 algorithms applied to the 2 data sets. G1 and G2 are essential to classifying the final grade and as such without them the accuracy is very low.

This led to creating a data set with only G1 and G2 to see if the opposite was true and it was found to generally be. The accuracy was found to be fairly high using only G1 and G2. By comparison the kappa statistic was between 0.69 and 0.78. In some cases it even produced the best accuracy out of all versions of the data set. This was true for Random Forest applied to the Portuguese data set, K-nearest neighbour applied to both data sets and Naive Bayes applied to the Math data set. High accuracy achieved from a minimal amount of attributes is good not only for ethical reasons; it helps to decrease the computation time of an algorithm. This is important when you have large data sets. The personal attributes are likely to be adding noise to the data set, bringing down accuracy in some cases. Some algorithms are more sensitive to noise than others such as K-nearest neighbour and decision trees whereas random forest can usually better handle it because it forms random subsets rather than one large tree.

The observations found using 10-fold cross validation were similar with the other test options. Removing the personal attributes did cause a notable drop in accuracy. Much like 10-fold the other three test options generally had versions other than the full version of the data set which produced a better accuracy. This was true for 22 out of the 24 results in Figure A.12-A.14 . The 24 results references the three different test options, with four different algorithms and the two data sets (Portuguese and Math grade) it is clear that removing the personal attributes did not have a clear negative impact on the accuracy.

Classifying personal attributes

Data sets are often published/made public without any guidelines of how the data set should and shouldn't be used. It was recommended on Kaggle.com that this data set be used to predict the student's final grade, but the data user was not limited in how they could use the data. This led to a path of investigation to see if it was possible to classify the romantic, Walc, Dalc and goout attributes (separately) to a certain level of accuracy. Equal width binning was used where required to turn numeric into nominal. It was decided that it was more unethical to try and classify sex or age, so these personal attributes were not a target class. Below is a summary of the accuracy's produced using 10-fold cross validation as the testing option (Figure 6.5).



		TARGET CLASS							
		romantic (2 classes)		Dalc (5 classes)		Walc (5 classes)		goout (5 classes)	
Algorithm		Math	Portuguese	Math	Portuguese	Math	Portuguese	Math	Portuguese
Decision Tree (J48)	Correctly Classified Instances %	61.01	59.01	66.83	65.79	39.75	39.14	33.67	28.81
	Kappa Statistic	0.06	0.09	0.26	0.22	0.18	0.18	0.12	0.07
Naïve Bayes	Correctly Classified Instances %	60.76	61.48	67.59	66.26	45.06	44.53	30.38	32.51
	Kappa Statistic	0.02	0.1	0.28	0.29	0.24	0.23	0.07	0.12
K-nearest neighbours	Correctly Classified Instances %	60.76	57.32	62.53	61.79	35.19	33.28	22.03	24.81
	Kappa Statistic	0.1	0.04	0.16	0.13	0.13	0.1	-0.05	0.02
Random Forest	Correctly Classified Instances %	65.82	61.48	69.62	70.42	48.61	47.3	33.67	32.82
	Kappa Statistic	0.08	0.04	0.08	0.14	0.26	0.25	0.08	0.09

FIGURE 6.5: 10-Fold Cross Validation Student Results for Personal Attributes as Target Class

Using the personal attributes as a target class had mixed results, which is expected. Predicting personal attributes is harder than predicting the final grade, which has some known correlating attributes. It is less obvious how to classify an attribute like alcohol consumption. The romantic target class had an accuracy of between 57.32% and 65.82% in the 8 instances (over the 4 algorithms in both data sets), which may initially appear to be an adequate accuracy, but it is not. This is because there are only 2 classes which is why the kappa statistic varies between 0.02 and 0.1, which is very low. In all 8 instances the class no has a very high true positive rate and class yes has a low true positive rate. The most extreme case of this is random forest applied to the Math data set. The class no has a true positive rate of 0.909 and class yes has a true positive of 0.159, but an accuracy of 65.82% is still achieved. In all 8 instances it has a bias towards predicting no which still leads to an accuracy of above 50% in all examples. When predicting only two classes the accuracy should be closer to 100% than 50%.

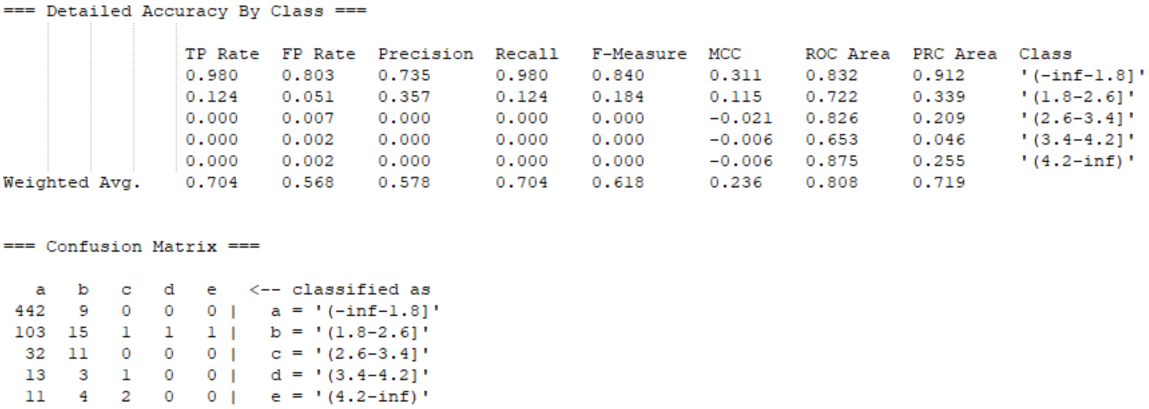
Predicting both Walc and goout was found to be difficult, as both produced a low accuracy with neither achieving above 50%. On the other hand, it appeared that some success was found with classifying the Dalc attribute. This was however found to be similar to classifying the romantic data set. There was five classes to be classified, but the classes were far from balanced. In the Portuguese grade data set the large majority (461 out of 649) of data instances had the value 1 in the Dalc attribute. The Walc attribute had a smaller majority class, which made it harder to predict and explains the large difference in accuracy. This meant that high accuracy could be achieved by over-predicting the majority class, which was the case. This is shown by the confusion matrix (Figure 6.6) of Random Forest applied to the Portuguese data set, with Dalc as the target class. This had an accuracy of 70.42%, but very poorly predicted any other class, as class a (as defined by the confusion matrix) had a true positive rate of 0.98 but the other true positive rates were 0.124, 0.0, 0.0 and 0.0. None of the algorithms actually produced an accuracy better than simply predicting the majority class, which would be 71.03% in the Portuguese grade data set and 69.87% in the Math grade data set.

Further experimentation would be required to confirm this but this data set is an example of one in which the specified target class is the one in which high accuracy will be achieved and predicting other attributes is more of a challenge. There is fairly low risk of predicting personal attributes with a high accuracy, based on the attributes tested. Other personal attributes may produce higher accuracy which could be further research.

6.1.1.6 Conclusions from experiment

The experiment produced some positive results and we are able to discuss the research questions (R-Q 4 and R-Q 6). This experiment does not suggest that removing personal attributes will improve accuracy in all cases, but it does beg the question of whether you should use personal attributes if the improvements are minimal, or not at all. Part of the problem is that when collecting data it is unknown what attributes will be needed when building a classifier. As a data user it may be appropriate to actively try to make a data set more generic by removing personal attributes, if it is not required for the classification to function accurately. When data is used with a Machine Learning algorithm it should first be accessed if using the full data set is appropriate as data is not always collected with the intended use of Machine Learning algorithms. R-Q 4 in relation to this experiment suggests that the context should be considered. Associating non-school related information with final grade should be done with care. An algorithm does not understand how the student's home live can impact their grade in the way that a support worker would. The data user may even find, like with this data set, that higher accuracy can be achieved by removing personal attributes. It was found that only a smaller subset of the data was needed. The removal of personal attributes from an ethical standpoint may help to increase trust in the use of ML techniques. Additionally, it was found that attributes G1 and G2 were the key attributes needed for classification as without them accuracy was low and with them alone you could achieve a relatively high accuracy. This highly supports the process of removing senstiive attributes and maintaining accuracy (R-Q 6). Beyond working with different versions of the full data set (all attributes) we also looked at classifying personal attributes. After analysing the results it was decided that the accuracy's produced were poor and as such the data set was best utilised by predicting the target class. The limitations of what a data set can be used to predict should be viewed as a positive when the data set has personal attributes. If it was found that the chosen personal attributes produced a high accuracy as a target class it may be proposed that the data set is remade or the attributes are excluded. Further research into the intended purpose of a data set and changing the intended target class

.



```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.980    0.803    0.735      0.980    0.840      0.311    0.832     0.912     '(-inf-1.8]'
                 0.124    0.051    0.357      0.124    0.184      0.115    0.722     0.339     '(1.8-2.6]'
                 0.000    0.007    0.000      0.000    0.000      -0.021   0.826     0.209     '(2.6-3.4]'
                 0.000    0.002    0.000      0.000    0.000      -0.006   0.653     0.046     '(3.4-4.2]'
                 0.000    0.002    0.000      0.000    0.000      -0.006   0.875     0.255     '(4.2-inf)'
Weighted Avg.    0.704    0.568    0.578      0.704    0.618      0.236    0.808     0.719

=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
 442   9   0   0   0 |   a = '(-inf-1.8]'
 103  15   1   1   1 |   b = '(1.8-2.6]'
  32  11   0   0   0 |   c = '(2.6-3.4]'
  13   3   1   0   0 |   d = '(3.4-4.2]'
  11   4   2   0   0 |   e = '(4.2-inf)'
```

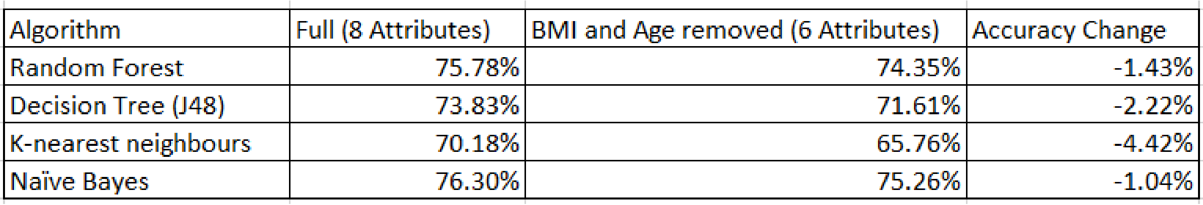
FIGURE 6.6: The Accuracy by Class and Confusion Matrix of Random Forest applied to Portuguese Data Set with Dalc as Target Class

6.1.2 Other data sets

The datasets below were a brief explored and are an expansion of the student alcohol consumption dataset. Random Forest, Decision Tree, K-nearest neighbours and Naïve Bayes were used in each example with 10 cross-fold validation for testing.

6.1.2.1 Diabetes (UCI Machine Learning, 2019b)

Medical data is often data which contains sensitive/personal data such as sex, age, race, age or mass/BMI. Unlike for example car insurance or defaulting on a loan there is often scientific evidence to support using demographics to diagnose or predict health problems. This dataset was created by the National Institute of Diabetes and Digestive and Kidney Diseases. It can be used to predict whether a patient has diabetes, using the 8 attributes for each instance. The dataset included both age and mass (BMI). Both age and BMI are known diabetes risk factors, so it makes sense to use both to help predict whether a patient has diabetes (Association, 2000). Medical diagnosis is concerned with finding as many patients with medical conditions as it can, misdiagnosis is better than missing a diagnosis. Using age, BMI or other sensitive details to build a classifier is different for medical purposes. It is not stereotyping when there is medical backing. Removing age and BMI reduced the accuracy of all algorithms (standard four) as expected, but with varying level of decrease from between 1.04% and 4.42% (Figure 6.7). The decrease is not important as there are only 768 instances, the key point is that removing age and BMI saw less patients be diagnosed. Machine Learning used for medical purposes is not an area which is of great concern. It could be viewed as one of the more objectively beneficial applications. The aim of using medical datasets is to provide a counter argument to removing sensitive attributes.

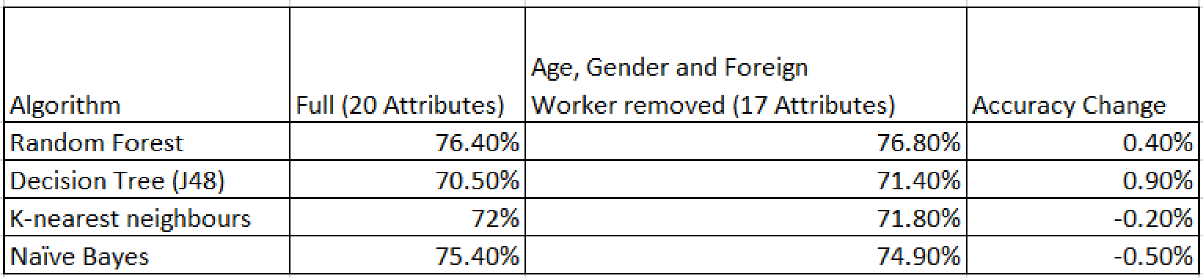


Algorithm	Full (8 Attributes)	BMI and Age removed (6 Attributes)	Accuracy Change
Random Forest	75.78%	74.35%	-1.43%
Decision Tree (J48)	73.83%	71.61%	-2.22%
K-nearest neighbours	70.18%	65.76%	-4.42%
Naïve Bayes	76.30%	75.26%	-1.04%

Figure 6.7: The Accuracy of the Classifiers with different attribute selection

6.1.2.2 Credit Risk (Hofmann, 1994)

This data set contained 1000 customers and can be used to classify whether the customer is good or bad credit risks. The dataset contains 20 attributes which can be used for training and includes obvious factors such as credit history, credit amount and other financial information. It also however includes gender, age and whether the customer is a foreign worker. These do not dictate a customer's ability to pay a loan etc. EU ruling since 2012 for example no longer allows car insurance companies in the EU to use sex as a factor in the assessment of insurance risks (BBC News, 2012). Furthermore, minimising financial risk to the bank may not need to be done by using the uncontrollable aspects of the customer such as age.



Algorithm	Full (20 Attributes)	Age, Gender and Foreign Worker removed (17 Attributes)	Accuracy Change
Random Forest	76.40%	76.80%	0.40%
Decision Tree (J48)	70.50%	71.40%	0.90%
K-nearest neighbours	72%	71.80%	-0.20%
Naïve Bayes	75.40%	74.90%	-0.50%

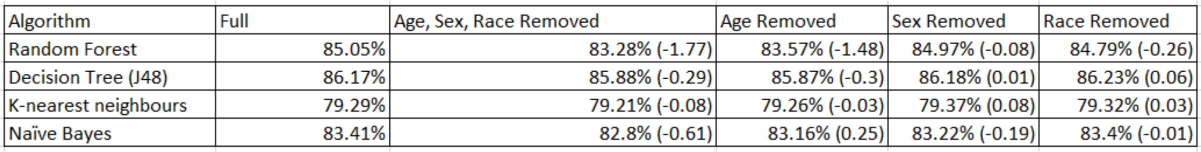
Figure 6.8: The Accuracy of the Classifiers with different attribute selection

This led to finding the difference in algorithm accuracy with and without using the attributes age, gender and whether they were a foreign worker (binary). Removing those values saw only a minor increase or decrease in accuracy. All four algorithms saw an increase or decrease of less 1% which means that the same accuracy could most likely be achieved with some parameter tuning (Figure 6.8). A decrease in accuracy may be a sacrifice for more equality between demographics but this sacrifice is not needed in this case. The minimal number of sensitive attributes should be desired.

6.1.2.3 Adult Income (UCI Machine Learning, 2019a)

The 1994 US Census bureau database was used to create this dataset which can be used to predict whether the person earns more than $50,000. The dataset contained 32561 individuals of which 29710 (91.24%) were born in the United States. The classes under and over 50k were not even, as 24720 (75.92%) earn under 50k. This means that around 75.92% classifier accuracy could be achieved, depending on the test option, simply by predicting the majority class (under 50k). The aim of the classifier in this case is to differentiate what makes the minority of individuals earn over 50k.

Adult income is impacted by societal factors beyond the simple closed world of the dataset. Both gender and race unfortunately correlate with adult income in the United States and as such it is an attribute that can be used to predict in this case whether the person earns more than $50,000 (Altonji and Blank, 1999). This does not mean that one gender earns over 50k and the other earns under 50k but combined with the other attributes it may help to improve accuracy. Demographics exist in this case to consider the biases rather than to cause them.



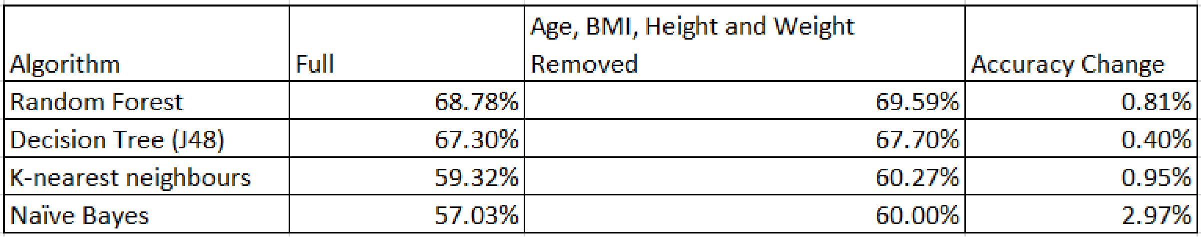
Algorithm	Full	Age, Sex, Race Removed	Age Removed	Sex Removed	Race Removed
Random Forest	85.05%	83.28% (-1.77)	83.57% (-1.48)	84.97% (-0.08)	84.79% (-0.26)
Decision Tree (J48)	86.17%	85.88% (-0.29)	85.87% (-0.3)	86.18% (0.01)	86.23% (0.06)
K-nearest neighbours	79.29%	79.21% (-0.08)	79.26% (-0.03)	79.37% (0.08)	79.32% (0.03)
Naïve Bayes	83.41%	82.8% (-0.61)	83.16% (0.25)	83.22% (-0.19)	83.4% (-0.01)

Figure 6.9: The Accuracy of the Classifiers with different attribute selection. The values in brackets are the difference in accuracy from the full data set.

Removing age, sex and race from the data set impacted the accuracy which was expected, due to already existing research on the gender and race wage gap. Removing age, sex and race saw a decrease in all four algorithms (Figure 6.9). When removing the attributes separately it was a mixed increase or decrease in accuracy. None of these attributes saw an increase in accuracy for all four algorithms when they were removed. The decrease was mostly minimal, but it was a decrease nonetheless. It is hard to predict a person's wage without considering the societal weightings which may benefit or disadvantage them such as the gender wage gap. Part of having unbiased systems is to acknowledge the biases found in society.

6.1.2.4 Absenteeism (Martiniano et al., 2012)

This dataset can be used to predict the absenteeism in hours of employees. The dataset was created using the absenteeism records of a courier company in Brazil. As well as factors such as distance from work, day of the week and reason for absence there was also attributes such as age and BMI. Predicting an employee's hours of absence based on age or weight would most likely not be allowed in the United Kingdom. Predicting an employee's absence using an algorithm is arguably both intrusive and unnecessary. There is an ethical issue with the premise of using this to predict absenteeism. Just because you can predict absenteeism it doesn't mean an employer should.



Algorithm	Full	Age, BMI, Height and Weight Removed	Accuracy Change
Random Forest	68.78%	69.59%	0.81%
Decision Tree (J48)	67.30%	67.70%	0.40%
K-nearest neighbours	59.32%	60.27%	0.95%
Naïve Bayes	57.03%	60.00%	2.97%

FIGURE 6.10: The Accuracy of the Classifiers with different attribute selection.

This led to an attempt to at least reduce some of the profiling nature of the classifier by removing age, BMI, height and weight. Notably the dataset also had attributes such as whether the employee was a social drinker, social smoker, owned a pet or their number of children but these were kept. Binning was used to split the numeric class into three bins (nominal) using equal frequency binning representing 0-2.5 hours, 2.5-7.5 hours and more than 7.5 hours. Not only were the attributes in the dataset questionable they did not appear to greatly improve accuracy. Accuracy using this dataset with the three bins was found to be relatively low in all cases (with or without attribute removal) (Figure 6.10). It was discovered that removing age, BMI, height and weight did not see a great decline in accuracy. Removing these attributes saw an increase in accuracy using all four algorithms. This outcome does not justify the use of these sensitive and personal attributes. A human predicting an employee would have a longer time off work due to their BMI or age is not appropriate, so it shouldn't be appropriate for a computer either.

6.1.2.5 Police Stop and Search Data (Home Office, 2019a)

The last data set to discuss was created using publicly available data from data.police.uk. It was stop and search data for all forces (regions). The exact data used can be generated using https://data.police.uk/data/fetch/5da54aaa-ed6b-43bb-9607-62c7d5bcbf6c/. The reason why this dataset was chosen is due to a known bias in stop and search. “Black people were 9 and a half times as likely to be stopped and searched as White people in 2017/18” according to government data (Home Office, 2019b). This leads onto the argument of whether computer systems are bias. If this data was used for a Machine Learning application of any kind there must be an understanding of the context. This dataset for example can be used to classify the outcome of the stop and search e.g. nothing found or suspect arrested. Similarly to the ProPublica article this data represents a systematic bias between races. Using data which can have bias between demographics contradicts the idea that a computer algorithm will not behave negatively in the way humans do. It is not the job of a programmer to fix the biases in society, but they should be aware of its impact on the data they may be using.

Chapter 7

Evaluation

7.1 Research Questions

The objectives were met to a satisfactory level. The public survey closed with 145 participants which was enough to discuss the research questions and find valid patterns. 145 was a large enough number to negate anomalies and prove a clear majority view in some cases. The other main objective was answering the remaining research questions by running Machine Learning algorithms on a variety of data sets. Simply running algorithms was not important, it was the ability to discuss the ethical aspects of Machine Learning which was the goal. This was achieved, and the results were encouraging in relation to R-Q 6. The dissertation was more interested in whether the research questions were answered, than the objectives themselves, which they all were to some extent. We shall discuss each research question below.

7.1.1 Do the public trust how their data is used?

The public do not trust how their data is used to a level that most organisations would most likely be satisfied with. Only 19% of participants either agreed or strongly agreed that organisations make it clear what their data is used for. As well as this only 15% either agreed or strongly agreed that their data is used responsibly and only 36% either agreed or strongly agreed that using the data of individuals has more benefits than risks. This hardly suggests that the public trust how their data is used. The public survey found that a sizeable number of participants felt that the current state of data usage is not satisfactory. The relationship the public have with the organisations they give their data to needs to be improved. Transparency seemed to be the clear issue as the public could not trust what they did not have knowledge of. Organisations need to make it clearer how data is used if they want public trust to improve.

7.1.2 How much do the public know about Machine Learning?

A relatively high percentage (57%) of participants had heard of Machine Learning and 54% self-described that they knew at least a little about it. The high percentage of participants who had heard of driverless vehicles, speech recognition, personalised advertisement and facial recognition shows that most of the public were at least aware of technologies which are known to incorporate Machine Learning. For a fairly specialised technology and not widely taught technology (outside of the IT sector) knowledge was noticeable high. Much of the public may not know the logistics of Machine Learning, but they are at least aware of its applications.

7.1.3 Is the public concerned about Machine Learning?

The survey did not discover that there was concern about Machine Learning as an overall concept but rather varying trust between technologies was found. Public concern for certain applications of Machine Learning was found rather than its broader use. The public for example showed far greater concern towards military robotics than speech recognition. A strong correlation (0.68) was found between whether the participant had heard of the technology and their trust of it. The technologies that a higher percentage of participants had heard of also had higher trust. There was a polarised view about whether "Organisations should be allowed to use Machine Learning/computers to make decisions that impact individuals". This is arguably public concern about the role of computers in decision making rather than Machine Learning specifically. Machine Learning can certainly be used in systems which make automated decisions impacting individuals, so it falls within this. The public overwhelming agreed that "A greater focus should be placed on ensuring computer system do not reflect the biases and discrimination in society". This suggests that the bias in computer systems, which can include Machine Learning, is of concern if the public want a greater focus placed on it. Furthermore it validates the motivation behind this project if the public want change as well. Overall it is not fair to say the public's concern discovered was strictly related to Machine Learning but rather Machine Learning was related to the topics of concern.

7.1.4 Should the surrounding context of a dataset be considered when using a dataset?

Generally, this was found to be true when using the selected data sets. It was difficult to use the datasets without understanding why the attributes were included in the dataset. Medical data was a simple example of this as "sensitive" attributes such as age, sex and BMI were attributes used for building a classifier. This was logical as any attribute which helps to diagnose a patient is desirable when the goal is helping patients. By comparison the adult census data was less clear in the need for "sensitive" attributes. On the surface it is less obvious as to why gender or race was needed when predicting whether a person earned more than 50k. It was proposed that the gender and race wage gap in the US may explain why it was a correlating factor. Using gender or race is less straightforward than using for example the person's occupation. There should be some level of explanation of the inclusion or removal of attributes. A computer algorithm treats data as a closed world when it often is not. As such it is up to the data collector and programmer to consider how closed world data should consider the greater context and be used in real world applications.

7.1.5 Are Machine Learning algorithms impacted by the same biases that exist in society?

This was probably the hardest question to answer and as such has the least clear answer. In the case of predicting income using adult census data it was hard to ignore the biases which impact the average wage of demographics. Accuracy went down when these attributes (race, age and sex) were removed. Machine Learning algorithms do not automatically fix the issues that exist in current decision making which can have discrimination. Data can contain the biases found in society. Machine Learning algorithms are built around data and are not immune to the biases.

7.1.6 Can you reduce the number of sensitive attributes such as gender or age in data sets and still maintain accuracy when using Machine Learning algorithms?

There were multiple datasets found where removing sensitive attributes did not see a drop-in accuracy, there were actually many cases where accuracy improved. It was found that it was possible to make datasets with sensitive attributes removed that functioning in the same way that the dataset would otherwise. The main exception to this was medical datasets but this is understandable as personal attributes often correlate with health. Furthermore, medical data has a more universal benefit than for example data used for personalised advertisement and as such there is less need to make the data a smaller subset. Removing sensitive attribute should not be a standard, as it not possible to generalise that accuracy can be maintained in all cases. There is a discussion however of whether it should be more common practise to try and remove sensitive attributes where possible.

7.2 Conclusion

This dissertation aimed to address the current state of ethics in Machine Learning. Concerning articles in the news such as discriminatory computer systems and a lack of standard practise motivated this. It was found that there is a problem with how Machine Learning algorithms are used with data, if there is an assumption that the algorithm is unbiased. The algorithm will most often build itself by discovering biases/differences in the data. Machine Learning algorithms cannot be used with bias data without dealing with the bias. You will not fix bias in a job hiring process for example by building a computer system without considering how to counteract the bias that exists. There is certainly room for greater consideration to be taken when creating systems which classify an individual. As well as this it was found that there were datasets which were not impacted by removing personal/sensitive attributes such as gender or age. There is an argument that programmers should actively try to reduce sensitive attributes where possible. It was found that it can even improve accuracy and if the opposite is true there may have to be a trade-off. Accuracy could potentially not always be the most important factor and removing sensitive attributes where possible may be more important. The public survey highlighted their concerns as well. The general consensus is that it is not clear what their data is used for nor was there a common held trust in data usage. Trust in the technology examples was generally higher than that of data usage as a whole. Public perception and trust in new technology and data usage was not intended to be a key component of this dissertation but as many discoveries were made about this as there was about the public's attitude towards Machine Learning. The defined aim of this project was to contribute towards safer data usage. Ultimately this dissertation did more to confirm topics which need more research than it did to find solutions to these problems, simply creating data sets with sensitive attributes removed will not prevent bias. It was abundantly clear that each Machine Learning application (different data sets) was unique and the handling the context of the data is a vital step in ethical Machine Learning. The main achievement was finding the areas where improvement can be made, which will hopefully lead to further research which will be laid out in future work.

7.3 Limitations

Several limitations were found, some can be solved by future work whereas others are more challenging. The dissertation dealt with data sets where problems could be found in preprocessing. In reality, it's difficult to prove a system will discriminate until it does. This is difficult to prevent without extensive testing. When a programmer builds a system which as far as they are aware is fair it doesn't always mean it is. Finding data sets which have been known to negatively discriminate between demographics may be a way to conduct research on how to identify and prevent discrimination. Furthermore, finding larger datasets would be ideal as ML algorithms are often used on data sets with 100,000s of instances rather than closer to 1000. The small data sets were useful in the time taken to complete. The more simple limitation to change is the survey. The survey was limited to the number of questions that it was due to completion rate. When writing the survey it was known that some topics/level of detail would have to excluded. If the survey could be an unlimited length the questions could be split up by organisation (Netflix, Facebook and Amazon) rather than grouping them together to see if opinion was varied for example. It is also worth noting that both the data set experiments and the survey required extensive analysis, so both could have been their own project. It may have been better to focus only on one of the two. This leads on to further work.

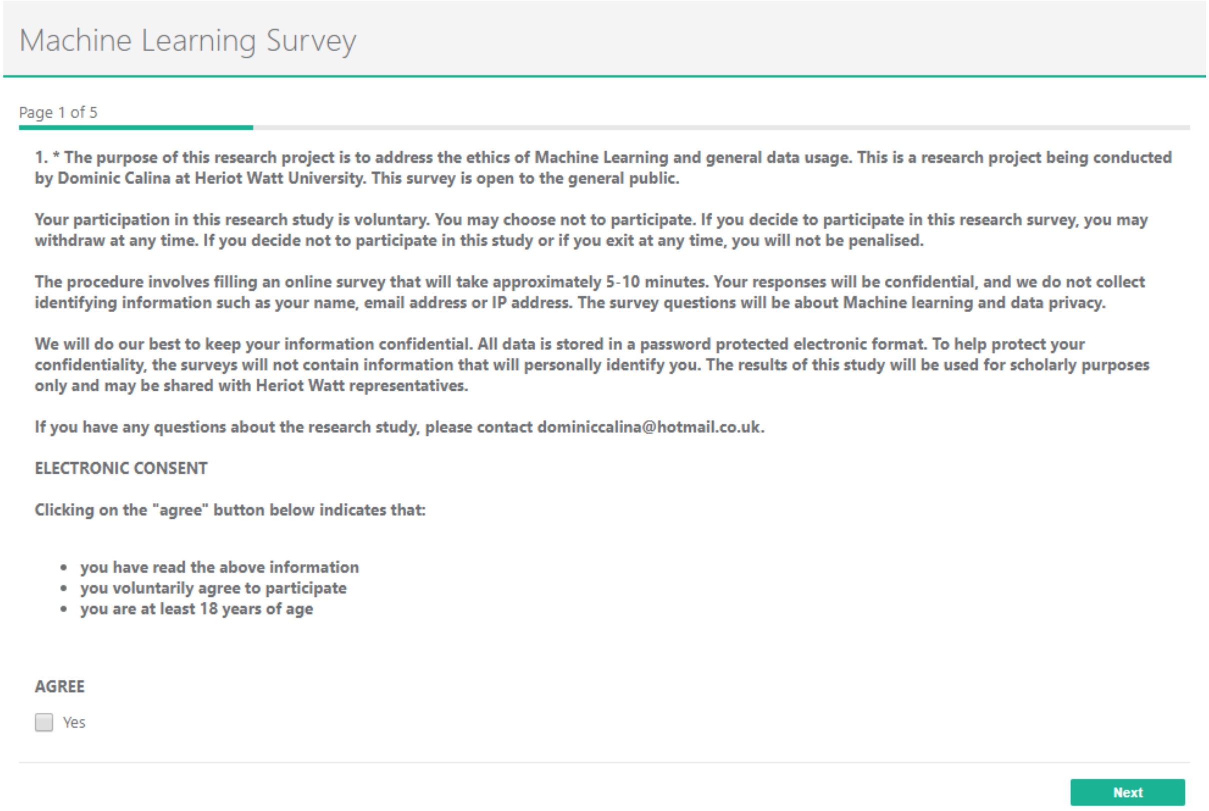
7.4 Future work

The future work can be split into three sections: the public, organisations and ML algorithms/programmers. As this was more exploratory research all the research questions could do with further research. The public do not understand how their data is used so research could be conducted into how the public can be educated to better understand how their data is used. Some of this responsibility lies with the organisations who collect this data. There could be investigation into how online agreements can be more transparent in how the public give their data away for services. Leading on from this the correlation found between knowledge and trust could be its own topic. Fear of new technology could be compared with public knowledge. The overarching research topic found from the survey is improving public trust and understanding. Relating more closely back to Machine Learning the concerns raised need a solution. Work can be conducted to create a Machine Learning guidance which would hopefully help improve the standard ethics. This project aimed to create guidance but the data set experiments need to cover more areas. Machine Learning algorithms need ethical standards because the programmer, data and application of the algorithm can be bias. Research into how computer systems are bias also requires additional work although there is currently ongoing research by others. Data sets are a closed world so further work can be conducted into how context (e.g. social bias) can be incorporated into systems to help reduce unwanted bias. Finally, the background research found that there was little research into developer/programmer bias so this a topic that would cover new grounds. There is currently only limited knowledge of how a programmer can willingly or unwillingly program bias into a system.

Appendix A

Appendix

A.1 Survey Screenshots



Machine Learning Survey

Page 1 of 5

1. * The purpose of this research project is to address the ethics of Machine Learning and general data usage. This is a research project being conducted by Dominic Calina at Heriot Watt University. This survey is open to the general public.

Your participation in this research study is voluntary. You may choose not to participate. If you decide to participate in this research survey, you may withdraw at any time. If you decide not to participate in this study or if you exit at any time, you will not be penalised.

The procedure involves filling an online survey that will take approximately 5-10 minutes. Your responses will be confidential, and we do not collect identifying information such as your name, email address or IP address. The survey questions will be about Machine learning and data privacy.

We will do our best to keep your information confidential. All data is stored in a password protected electronic format. To help protect your confidentiality, the surveys will not contain information that will personally identify you. The results of this study will be used for scholarly purposes only and may be shared with Heriot Watt representatives.

If you have any questions about the research study, please contact dominiccalina@hotmail.co.uk.

ELECTRONIC CONSENT

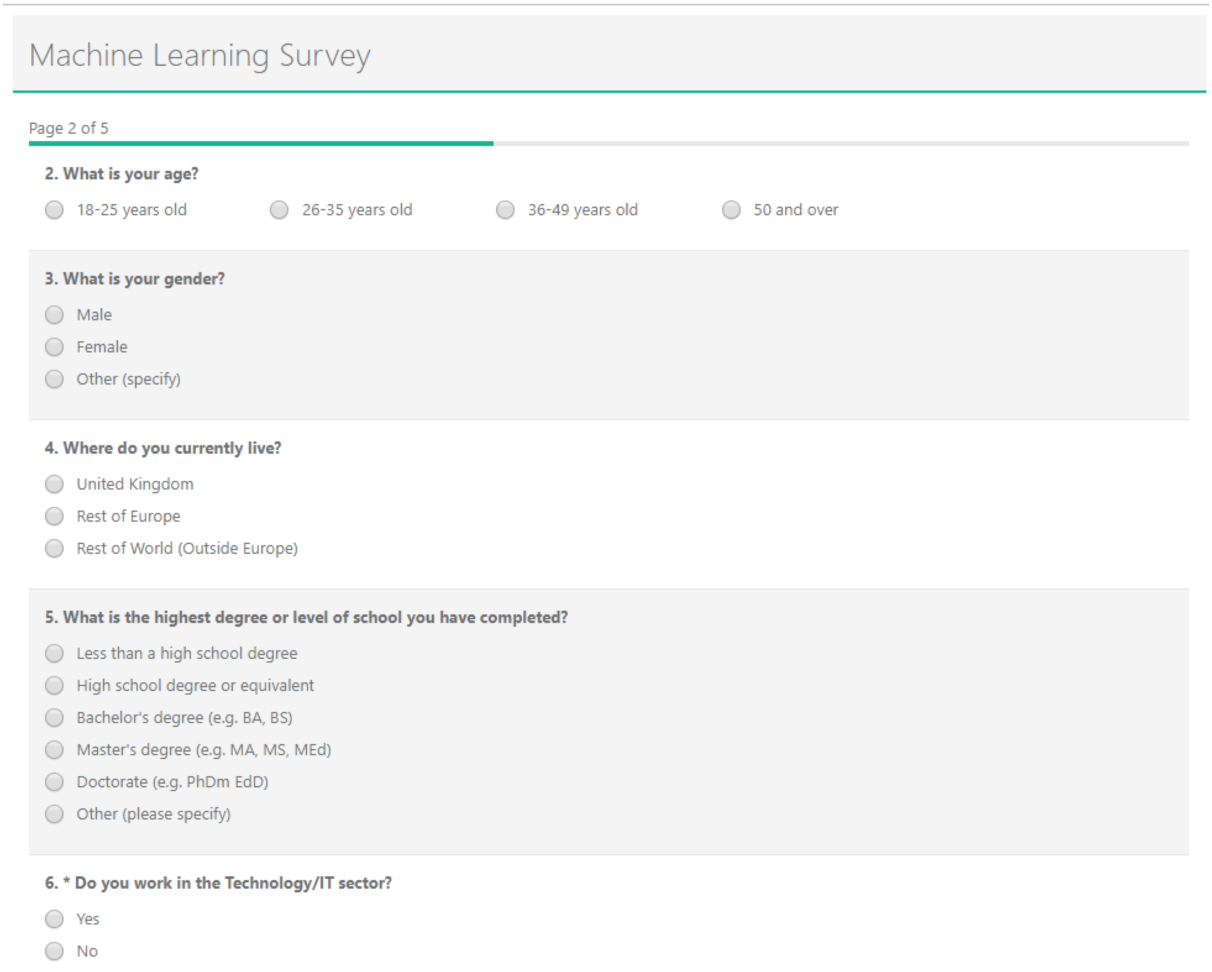
Clicking on the "agree" button below indicates that:

- you have read the above information
- you voluntarily agree to participate
- you are at least 18 years of age

AGREE

□ Yes

Figure A.1: The Survey consent form



Machine Learning Survey

Page 2 of 5

2. What is your age?

○ 18-25 years old ○ 26-35 years old ○ 36-49 years old ○ 50 and over

3. What is your gender?

○ Male
○ Female
○ Other (specify)

4. Where do you currently live?

○ United Kingdom
○ Rest of Europe
○ Rest of World (Outside Europe)

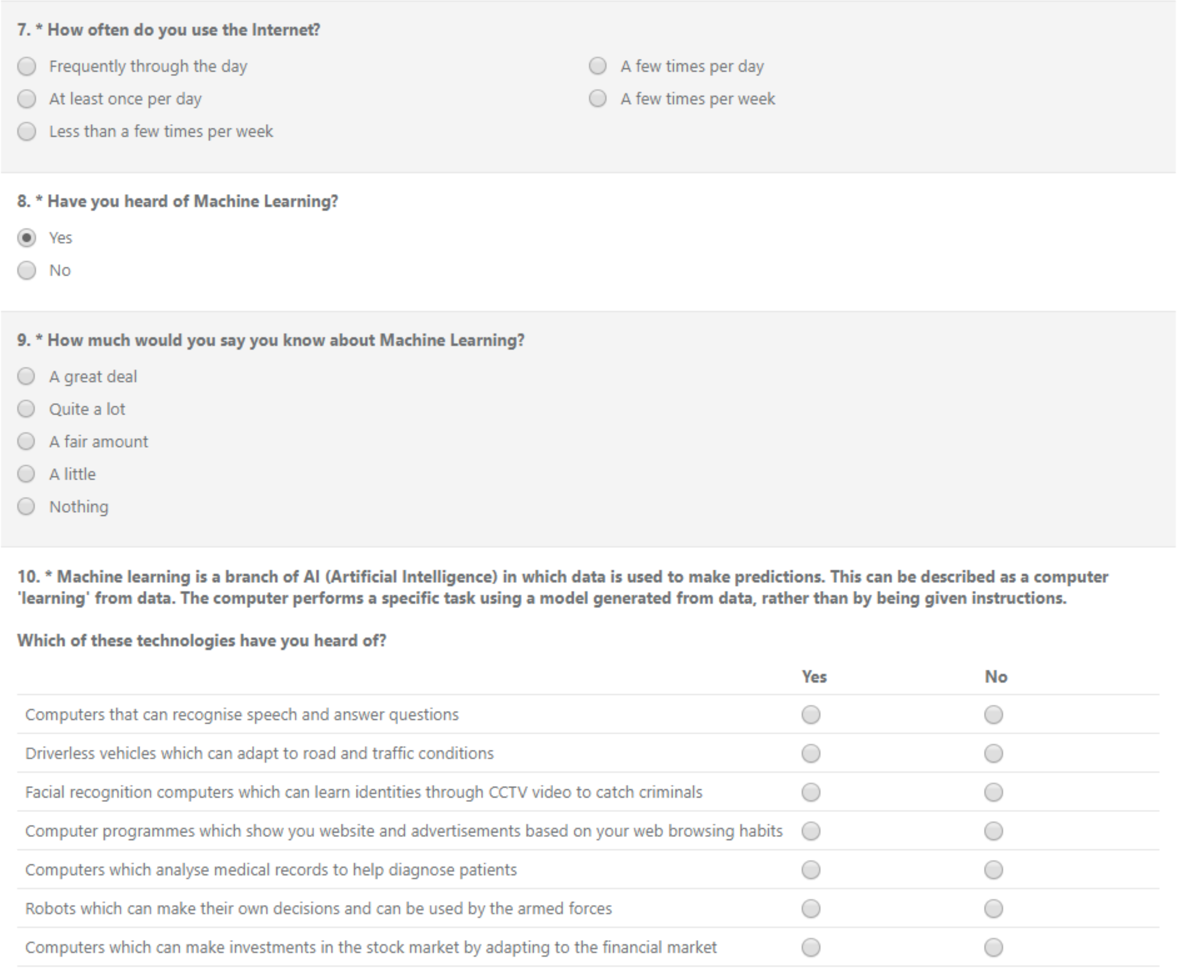
5. What is the highest degree or level of school you have completed?

○ Less than a high school degree
○ High school degree or equivalent
○ Bachelor's degree (e.g. BA, BS)
○ Master's degree (e.g. MA, MS, MEd)
○ Doctorate (e.g. PhDm EdD)
○ Other (please specify)

6. * Do you work in the Technology/IT sector?

○ Yes
○ No

Figure A.2: The Survey Part 1



7. * How often do you use the Internet?

○ Frequently through the day

○ At least once per day

○ Less than a few times per week

○ A few times per day

○ A few times per week

8. * Have you heard of Machine Learning?

● Yes

○ No

9. * How much would you say you know about Machine Learning?

○ A great deal

○ Quite a lot

○ A fair amount

○ A little

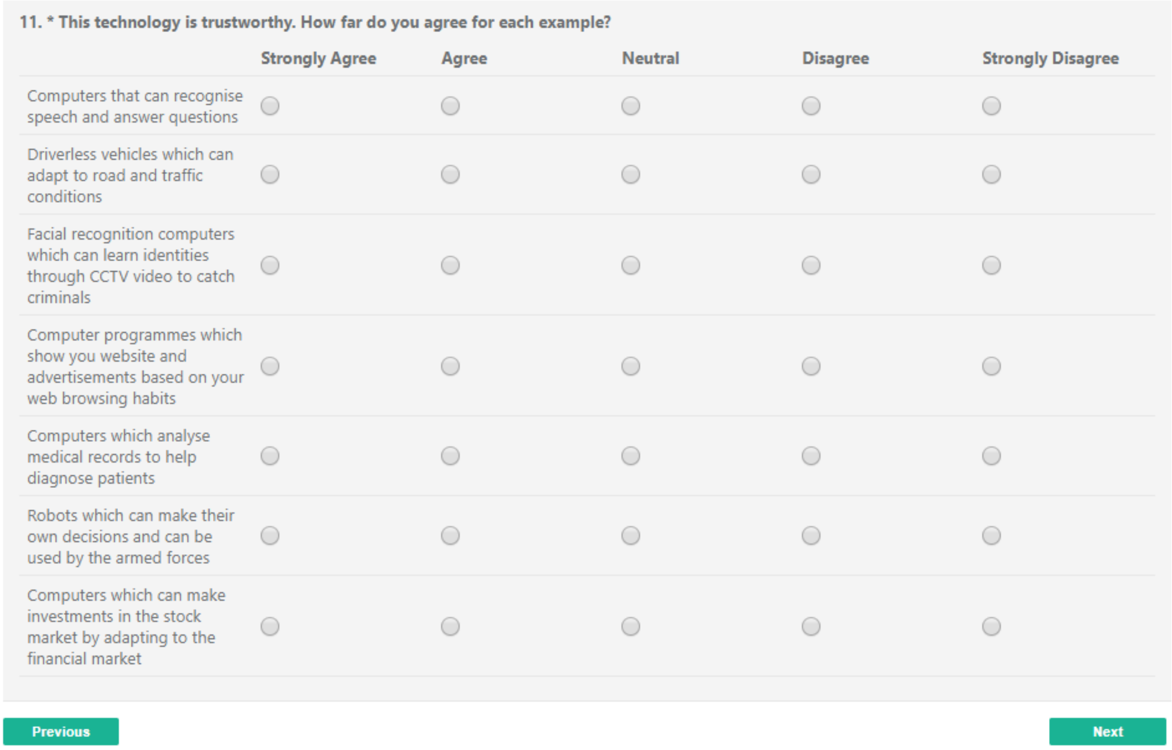
○ Nothing

10. * Machine learning is a branch of AI (Artificial Intelligence) in which data is used to make predictions. This can be described as a computer 'learning' from data. The computer performs a specific task using a model generated from data, rather than by being given instructions.

Which of these technologies have you heard of?

	Yes	No
Computers that can recognise speech and answer questions	○	○
Driverless vehicles which can adapt to road and traffic conditions	○	○
Facial recognition computers which can learn identities through CCTV video to catch criminals	○	○
Computer programmes which show you website and advertisements based on your web browsing habits	○	○
Computers which analyse medical records to help diagnose patients	○	○
Robots which can make their own decisions and can be used by the armed forces	○	○
Computers which can make investments in the stock market by adapting to the financial market	○	○

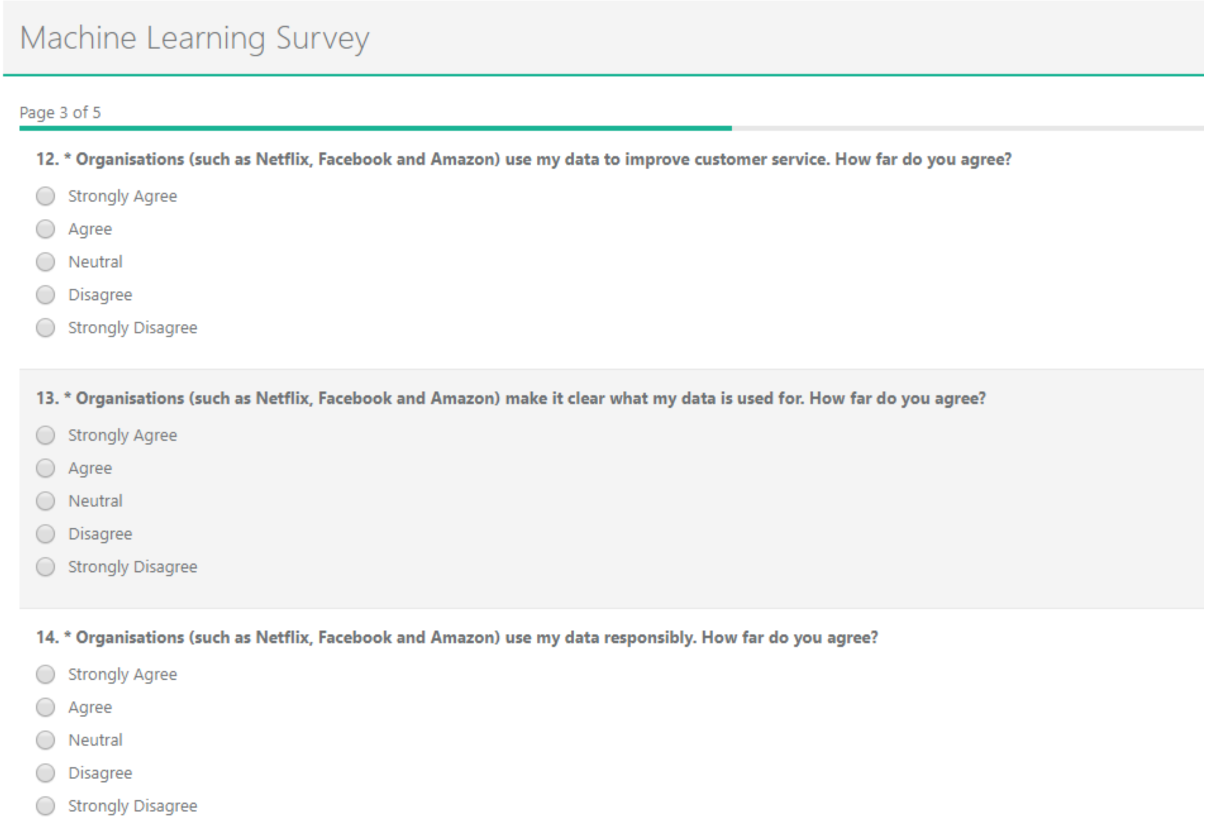
Figure A.3: The Survey Part 2



11. * This technology is trustworthy. How far do you agree for each example?

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
Computers that can recognise speech and answer questions	○	○	○	○	○
Driverless vehicles which can adapt to road and traffic conditions	○	○	○	○	○
Facial recognition computers which can learn identities through CCTV video to catch criminals	○	○	○	○	○
Computer programmes which show you website and advertisements based on your web browsing habits	○	○	○	○	○
Computers which analyse medical records to help diagnose patients	○	○	○	○	○
Robots which can make their own decisions and can be used by the armed forces	○	○	○	○	○
Computers which can make investments in the stock market by adapting to the financial market	○	○	○	○	○

Figure A.4: The Survey Part 3



Machine Learning Survey

Page 3 of 5

12. * Organisations (such as Netflix, Facebook and Amazon) use my data to improve customer service. How far do you agree?

○ Strongly Agree
○ Agree
○ Neutral
○ Disagree
○ Strongly Disagree

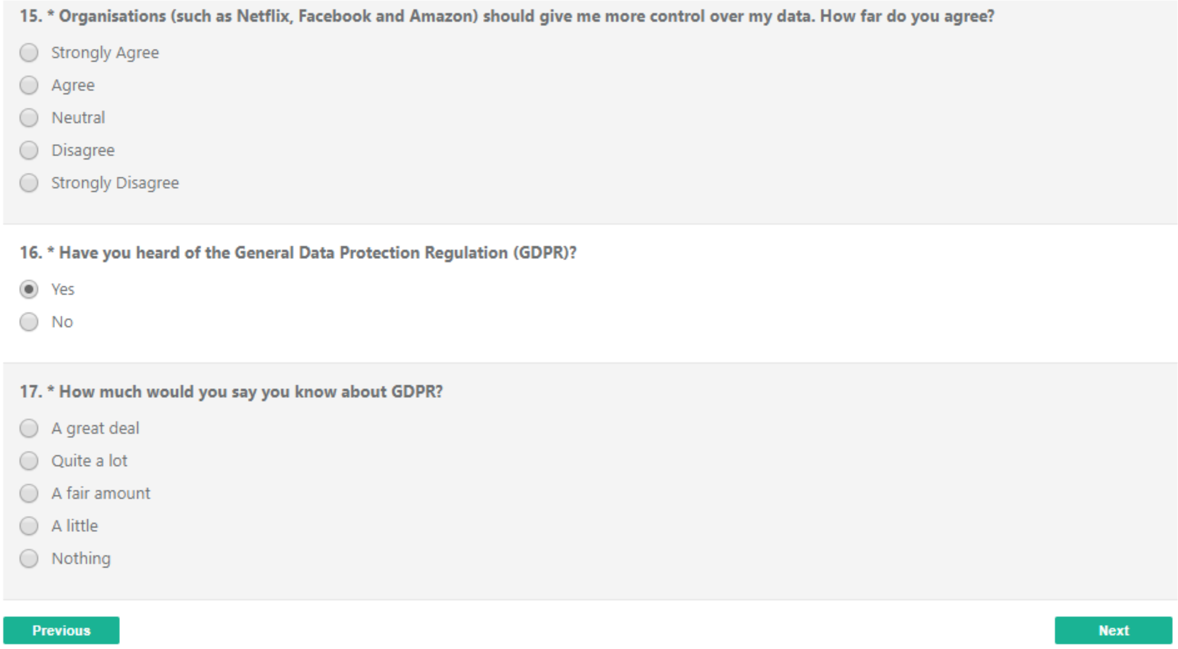
13. * Organisations (such as Netflix, Facebook and Amazon) make it clear what my data is used for. How far do you agree?

○ Strongly Agree
○ Agree
○ Neutral
○ Disagree
○ Strongly Disagree

14. * Organisations (such as Netflix, Facebook and Amazon) use my data responsibly. How far do you agree?

○ Strongly Agree
○ Agree
○ Neutral
○ Disagree
○ Strongly Disagree

Figure A.5: The Survey Part 4



15. * Organisations (such as Netflix, Facebook and Amazon) should give me more control over my data. How far do you agree?

○ Strongly Agree
○ Agree
○ Neutral
○ Disagree
○ Strongly Disagree

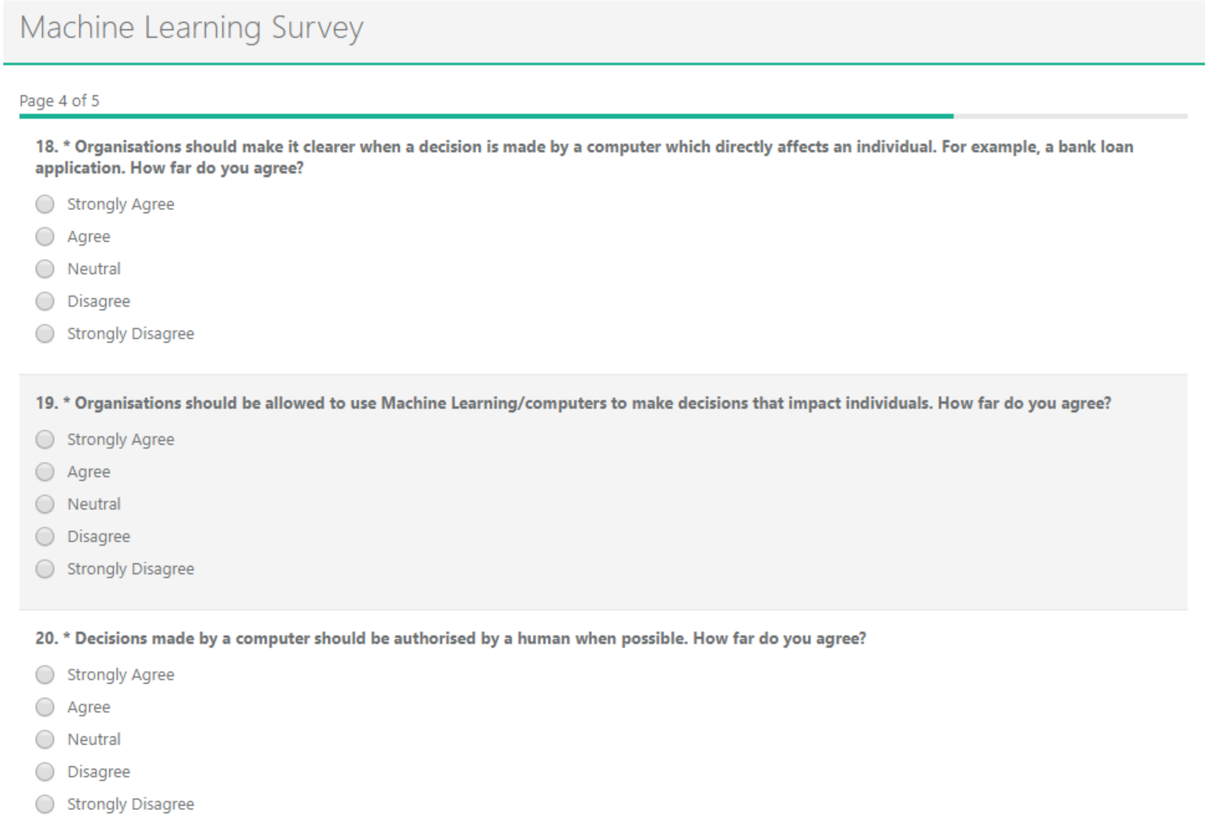
16. * Have you heard of the General Data Protection Regulation (GDPR)?

● Yes
○ No

17. * How much would you say you know about GDPR?

○ A great deal
○ Quite a lot
○ A fair amount
○ A little
○ Nothing

Figure A.6: The Survey Part 5



Machine Learning Survey

Page 4 of 5

18. * Organisations should make it clearer when a decision is made by a computer which directly affects an individual. For example, a bank loan application. How far do you agree?

○ Strongly Agree
○ Agree
○ Neutral
○ Disagree
○ Strongly Disagree

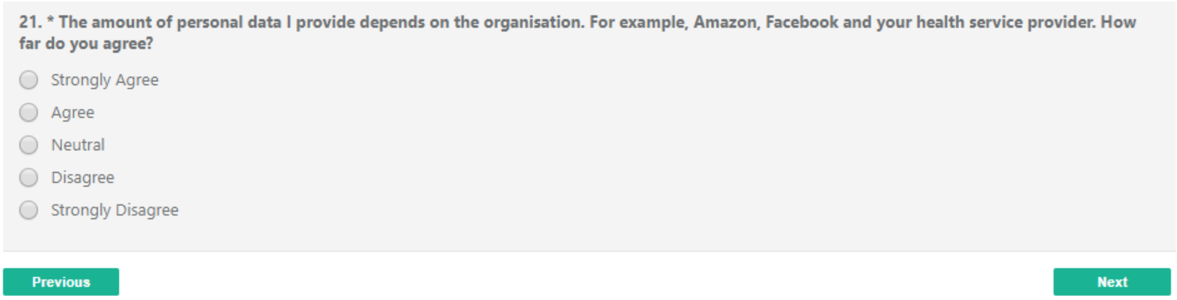
19. * Organisations should be allowed to use Machine Learning/computers to make decisions that impact individuals. How far do you agree?

○ Strongly Agree
○ Agree
○ Neutral
○ Disagree
○ Strongly Disagree

20. * Decisions made by a computer should be authorised by a human when possible. How far do you agree?

○ Strongly Agree
○ Agree
○ Neutral
○ Disagree
○ Strongly Disagree

Figure A.7: The Survey Part 6



21. * The amount of personal data I provide depends on the organisation. For example, Amazon, Facebook and your health service provider. How far do you agree?

○ Strongly Agree

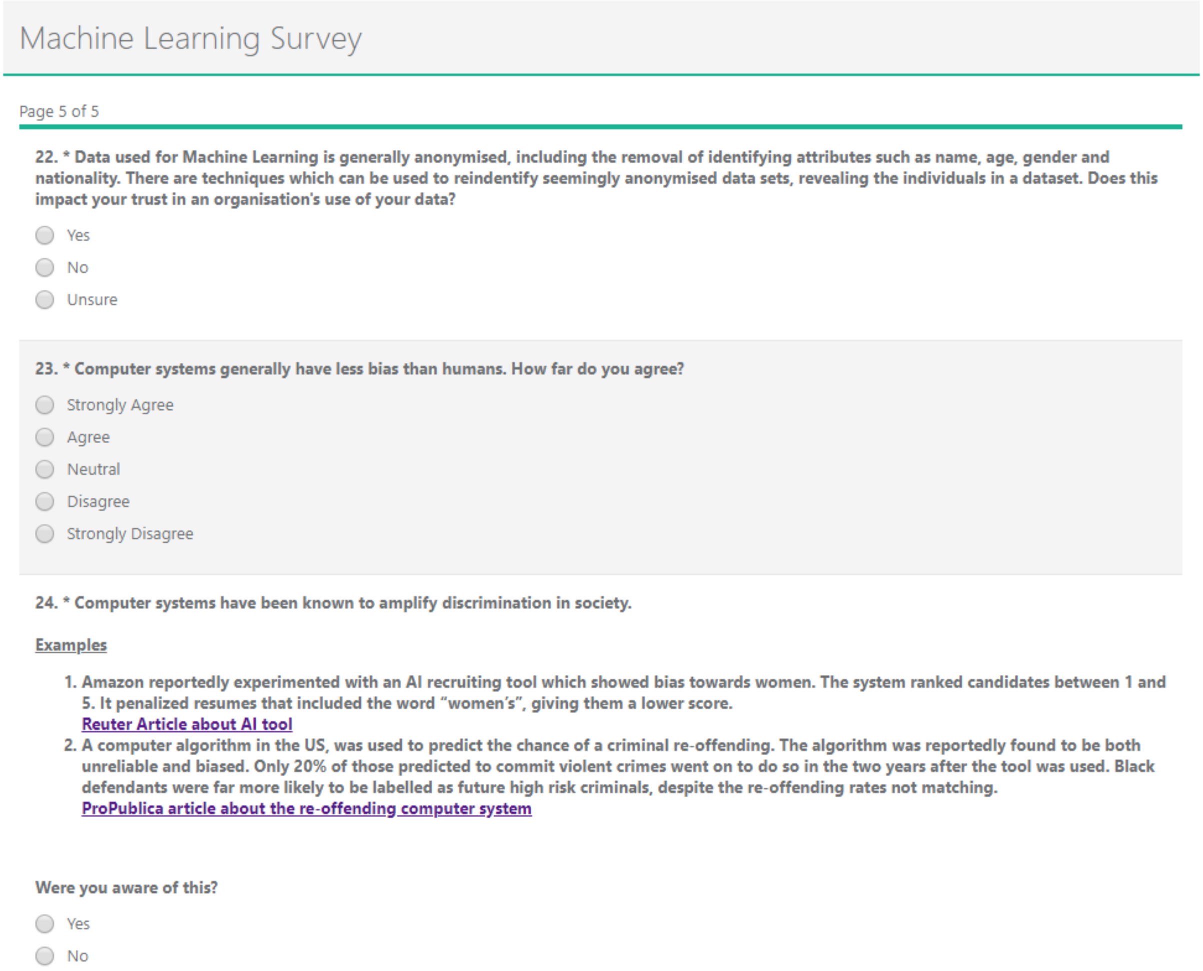
○ Agree

○ Neutral

○ Disagree

○ Strongly Disagree

Figure A.8: The Survey Part 7



Machine Learning Survey

Page 5 of 5

22. * Data used for Machine Learning is generally anonymised, including the removal of identifying attributes such as name, age, gender and nationality. There are techniques which can be used to reindentify seemingly anonymised data sets, revealing the individuals in a dataset. Does this impact your trust in an organisation's use of your data?

○ Yes
○ No
○ Unsure

23. * Computer systems generally have less bias than humans. How far do you agree?

○ Strongly Agree
○ Agree
○ Neutral
○ Disagree
○ Strongly Disagree

24. * Computer systems have been known to amplify discrimination in society.

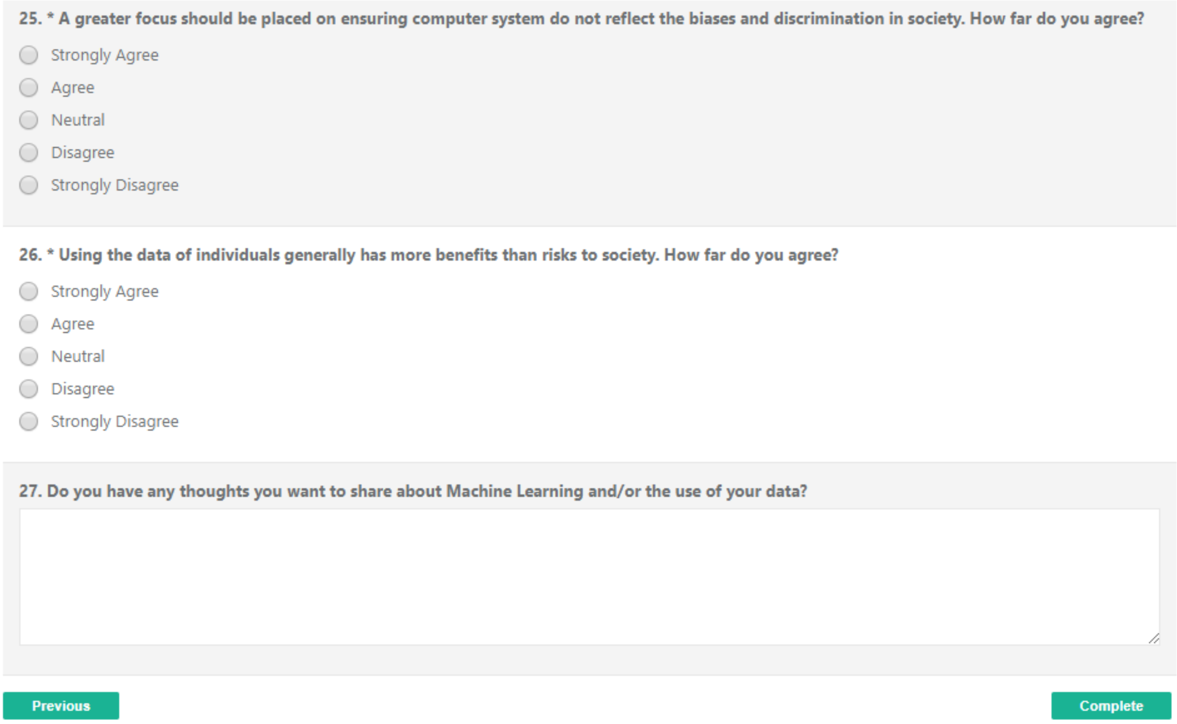
Examples

1. Amazon reportedly experimented with an AI recruiting tool which showed bias towards women. The system ranked candidates between 1 and 5. It penalized resumes that included the word "women's", giving them a lower score. Reuter Article about AI tool
2. A computer algorithm in the US, was used to predict the chance of a criminal re-offending. The algorithm was reportedly found to be both unreliable and biased. Only 20% of those predicted to commit violent crimes went on to do so in the two years after the tool was used. Black defendants were far more likely to be labelled as future high risk criminals, despite the re-offending rates not matching. ProPublica article about the re-offending computer system

Were you aware of this?

○ Yes
○ No

FIGURE A.9: The Survey Part 8



25. * A greater focus should be placed on ensuring computer system do not reflect the biases and discrimination in society. How far do you agree?

○ Strongly Agree
○ Agree
○ Neutral
○ Disagree
○ Strongly Disagree

26. * Using the data of individuals generally has more benefits than risks to society. How far do you agree?

○ Strongly Agree
○ Agree
○ Neutral
○ Disagree
○ Strongly Disagree

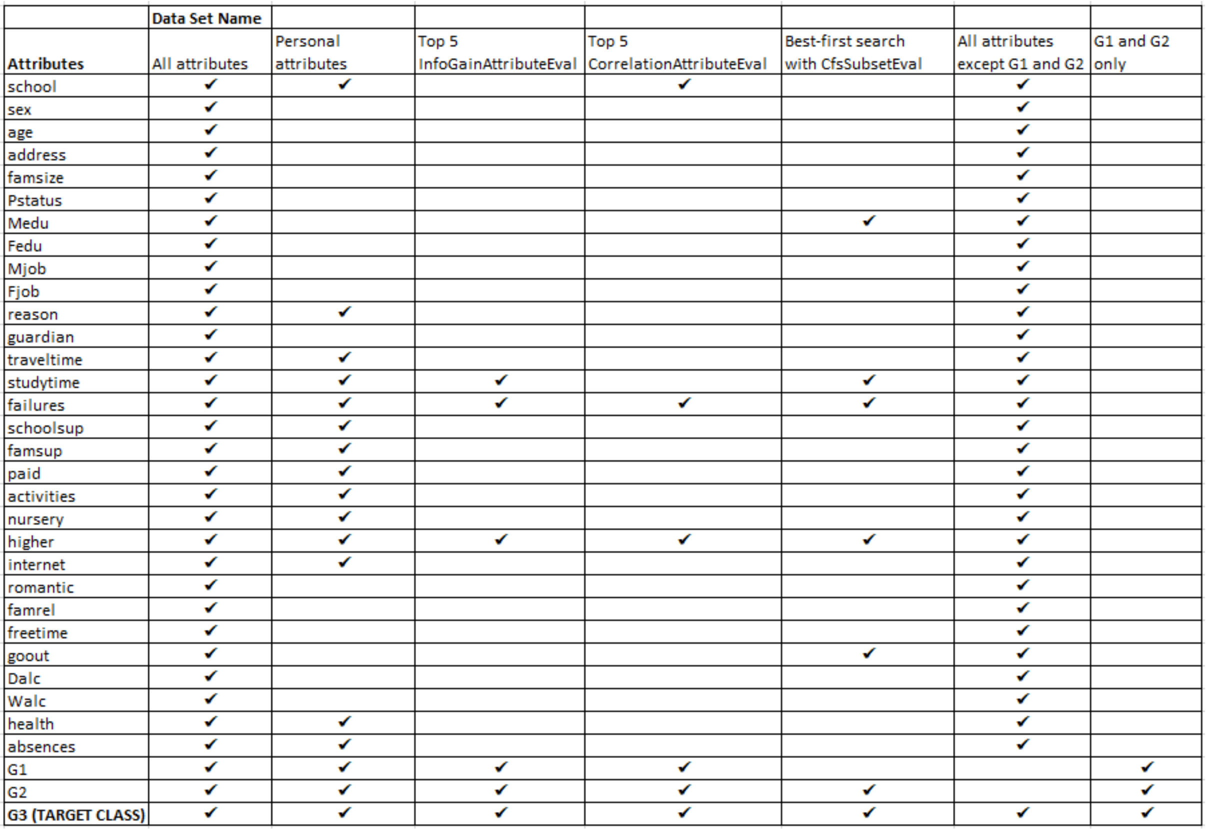
27. Do you have any thoughts you want to share about Machine Learning and/or the use of your data?

Figure A.10: The Survey Part 9

A.2 Dataset Experiments

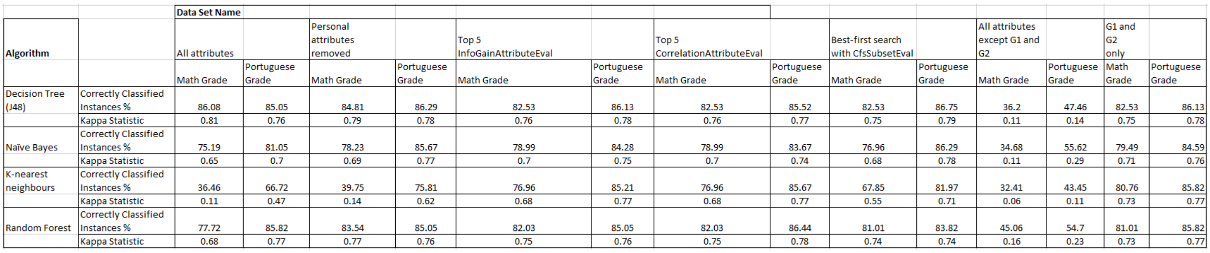
1. school (binary): Student's school (School related)
2. sex (binary): Student's sex (Personal)
3. age (numeric): Student's age (Personal)
4. address (binary): Student home address type which can be urban or rural (Personal)
5. famsize (binary): The student's family size (Personal)
6. Pstatus (binary): Their parent's cohabitation status which means living together or apart (Personal)
7. Medu (numeric): Mother's highest level of education (Personal)
8. Fedu (numeric): Father's highest level of education (Personal)
9. Mjob (nominal): Mother's job which can be teacher, civil services, health, at home or other, notably fairly limited (Personal)
10. Fjob (nominal): Father's job which can be teacher, civil services, health, at home or other, notably fairly limited (Personal)
11. reason (nominal): The reason for the student to choose their school which was either course, home, reputation or other. It isn't clear what is meant by home or course (School related)
12. guardian (nominal): The student's guardian which is mother, father or other. This attribute has limited information about the type of non-parental guardians (Personal)
13. traveltime (numeric): The student's home to school travel time. It is numeric categories rather than continuous e.g. 1 is ¡15 minutes (School related)
14. studytime (numeric): The student's weekly study time. It is numeric categories rather than continuous e.g. 1 is ¡ 2 hours (School related)
15. failures (numeric): The number of failures between 0 and 3 (School related)
16. schoolsup (binary): Whether or not the student received extra educational support (School related)
17. famsup (binary): Whether or not the student received family educational support, it is unclear what this entails (School related)
18. paid (binary): Whether or not the student received extra paid classes within the course subject, either Math or Portuguese (School related)
19. activities (binary): Whether or not the student participated in extra curricular activities (School related)
20. nursery (binary): Whether or not the student attended nursery school (School related)
21. higher (binary): Whether or not the student wants to continue into higher education (School related)

22. internet (binary): Whether or not the student has access to the Internet at home (School related)
23. romantic (binary): Whether the student have a romantic relationship (Personal)
24. famrel (numeric): The quality of the family relationship from very bad to excellent. The attribute scales from 1 to 5 which is very bad to excellent (Personal)
25. freetime (numeric): The amount of free time the student has after school which is from very low to very high. The attribute scales from 1 to 5 which is very low to very high (Personal)
26. goout (numeric): How often the student goes out with friends from very low to very high. The attribute scales from 1 to 5 which is very low to very high (Personal)
27. Dalc (numeric): The student's workday alcohol consumption which is very low to very high. The attribute scales from 1 to 5 which is very low to very high (Personal)
28. Walc (numeric): The student's weekend alcohol consumption which is very low to very high. The attribute scales from 1 to 5 which is very low to very high (Personal)
29. health (numeric): The student's current health status from very bad to very good. This attribute scales from 1 to 5 whcih is very bad to very good. It was decided that this be labelled as school realted as it directly impacts a student's ability to attend school (School related)
30. absences (numeric): The number of school absences which is from 0 to 93 (School related)
31. G1 (numeric): The student's first period grade from 0 to 20. It is unclear if period is year or term (School related)
32. G2 (numeric): The student's second period grade from 0 to 20. It is unclear if period is year or term (School related)
33. TARGET CLASS G3 (numeric): The student's final grade in either Math or Portuguese from 0 to 20. This was converted to 5 nominal classes using equal-width binning (School related)



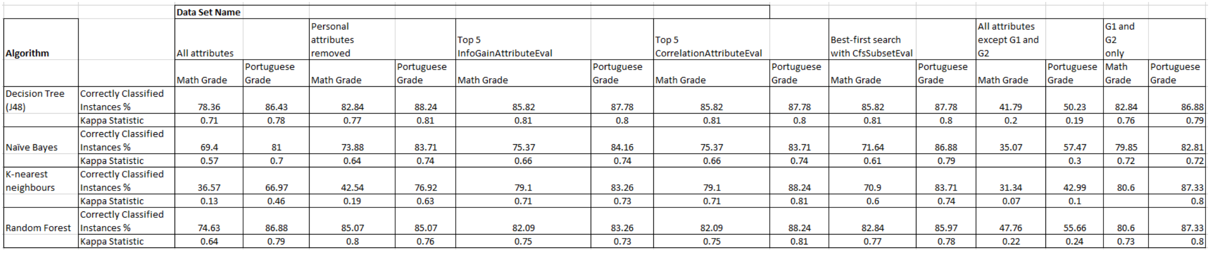
	Data Set Name						
Attributes	All attributes	Personal attributes	Top 5 InfoGainAttributeEval	Top 5 CorrelationAttributeEval	Best-first search with CfsSubsetEval	All attributes except G1 and G2	G1 and G2 only
school	✓	✓		✓		✓	
sex	✓					✓	
age	✓					✓	
address	✓					✓	
famsize	✓					✓	
Pstatus	✓					✓	
Medu	✓				✓	✓	
Fedu	✓					✓	
Mjob	✓					✓	
Fjob	✓					✓	
reason	✓	✓				✓	
guardian	✓					✓	
traveltime	✓	✓				✓	
studytime	✓	✓	✓		✓	✓	
failures	✓	✓	✓	✓	✓	✓	
schoolsup	✓	✓				✓	
famsup	✓	✓				✓	
paid	✓	✓				✓	
activities	✓	✓				✓	
nursery	✓	✓				✓	
higher	✓	✓	✓	✓	✓	✓	
internet	✓	✓				✓	
romantic	✓					✓	
famrel	✓					✓	
freetime	✓					✓	
goout	✓				✓	✓	
Dalc	✓					✓	
Walc	✓					✓	
health	✓	✓				✓	
absences	✓	✓				✓	
G1	✓	✓	✓	✓			✓
G2	✓	✓	✓	✓	✓		✓
G3 (TARGET CLASS)	✓	✓	✓	✓	✓	✓	✓

Figure A.11: The selected attributes for the Portuguese data set



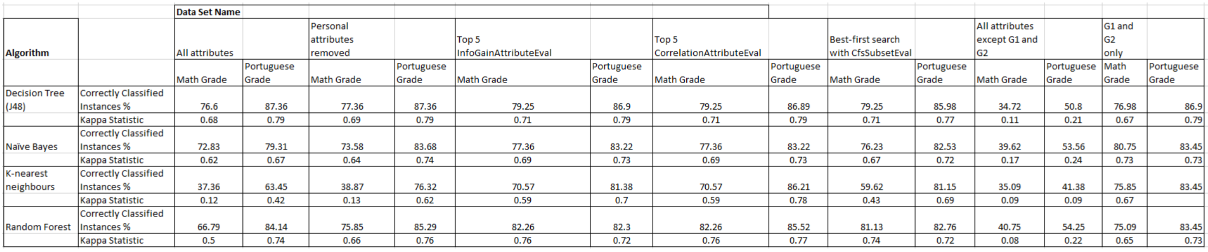
		Data Set Name													
Algorithm		All attributes		Personal attributes removed		Top 5 InfoGainAttributeEval		Top 5 CorrelationAttributeEval		Best-first search with CfsSubsetEval		All attributes except G1 and G2		G1 and G2 only	
		Math Grade	Portuguese Grade	Math Grade	Portuguese Grade	Math Grade	Portuguese Grade	Math Grade	Portuguese Grade	Math Grade	Portuguese Grade	Math Grade	Portuguese Grade	Math Grade	Portuguese Grade
Decision Tree (J48)	Correctly Classified Instances %	86.08	85.05	84.81	86.29	82.53	86.13	82.53	85.52	82.53	86.75	36.2	47.46	82.53	86.13
	Kappa Statistic	0.81	0.76	0.79	0.78	0.76	0.78	0.76	0.77	0.75	0.79	0.11	0.14	0.75	0.78
Naïve Bayes	Correctly Classified Instances %	75.19	81.05	78.23	85.67	78.99	84.28	78.99	83.67	76.96	86.29	34.68	55.62	79.49	84.59
	Kappa Statistic	0.65	0.7	0.69	0.77	0.7	0.75	0.7	0.74	0.68	0.78	0.11	0.29	0.71	0.76
K-nearest neighbours	Correctly Classified Instances %	36.46	66.72	39.75	75.81	76.96	85.21	76.96	85.67	67.85	81.97	32.41	43.45	80.76	85.82
	Kappa Statistic	0.11	0.47	0.14	0.62	0.68	0.77	0.68	0.77	0.55	0.71	0.06	0.11	0.73	0.77
Random Forest	Correctly Classified Instances %	77.72	85.82	83.54	85.05	82.03	85.05	82.03	86.44	81.01	83.82	45.06	54.7	81.01	85.82
	Kappa Statistic	0.68	0.77	0.77	0.76	0.75	0.76	0.75	0.78	0.74	0.74	0.16	0.23	0.73	0.77

Figure A.12: Leave-one-out Cross Validation Student Results



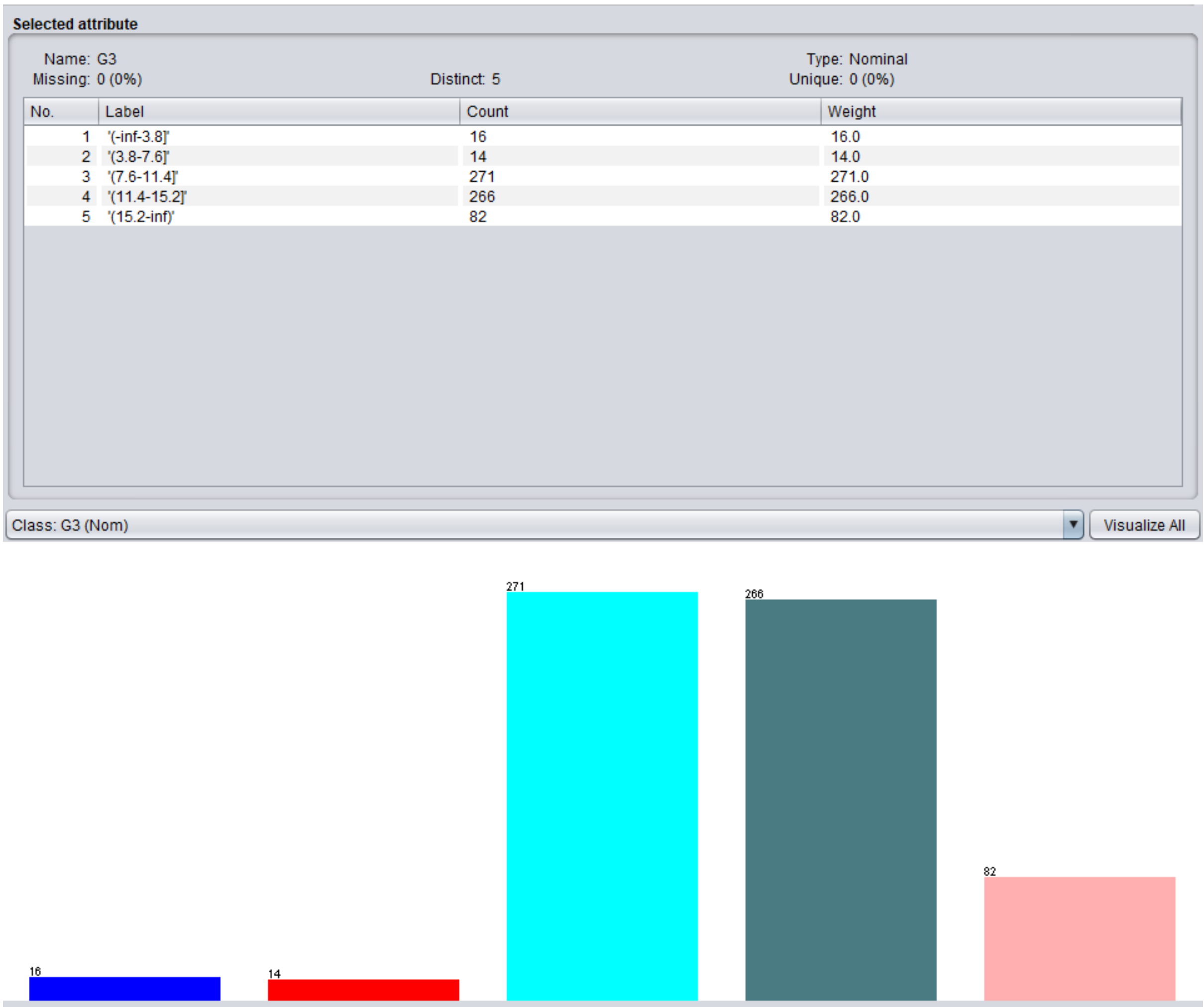
		Data Set Name													
Algorithm		All attributes		Personal attributes removed		Top 5 InfoGainAttributeEval		Top 5 CorrelationAttributeEval		Best-first search with CfsSubsetEval		All attributes except G1 and G2		G1 and G2 only	
		Math Grade	Portuguese Grade	Math Grade	Portuguese Grade	Math Grade	Portuguese Grade	Math Grade	Portuguese Grade	Math Grade	Portuguese Grade	Math Grade	Portuguese Grade	Math Grade	Portuguese Grade
Decision Tree (J48)	Correctly Classified Instances %	78.36	86.43	82.84	88.24	85.82	87.78	85.82	87.78	85.82	87.78	41.79	50.23	82.84	86.88
	Kappa Statistic	0.71	0.78	0.77	0.81	0.81	0.8	0.81	0.8	0.81	0.8	0.2	0.19	0.76	0.79
Naïve Bayes	Correctly Classified Instances %	69.4	81	73.88	83.71	75.37	84.16	75.37	83.71	71.64	86.88	35.07	57.47	79.85	82.81
	Kappa Statistic	0.57	0.7	0.64	0.74	0.66	0.74	0.66	0.74	0.61	0.79		0.3	0.72	0.72
K-nearest neighbours	Correctly Classified Instances %	36.57	66.97	42.54	76.92	79.1	83.26	79.1	88.24	70.9	83.71	31.34	42.99	80.6	87.33
	Kappa Statistic	0.13	0.46	0.19	0.63	0.71	0.73	0.71	0.81	0.6	0.74	0.07	0.1		0.8
Random Forest	Correctly Classified Instances %	74.63	86.88	85.07	85.07	82.09	83.26	82.09	88.24	82.84	85.97	47.76	55.66	80.6	87.33
	Kappa Statistic	0.64	0.79	0.8	0.76	0.75	0.73	0.75	0.81	0.77	0.78	0.22	0.24	0.73	0.8

Figure A.13: Test Split 66% Student Results



		Data Set Name													
Algorithm		All attributes		Personal attributes removed		Top 5 InfoGainAttributeEval		Top 5 CorrelationAttributeEval		Best-first search with CfsSubsetEval		All attributes except G1 and G2		G1 and G2 only	
		Math Grade	Portuguese Grade	Math Grade	Portuguese Grade	Math Grade	Portuguese Grade	Math Grade	Portuguese Grade	Math Grade	Portuguese Grade	Math Grade	Portuguese Grade	Math Grade	Portuguese Grade
Decision Tree (J48)	Correctly Classified Instances %	76.6	87.36	77.36	87.36	79.25	86.9	79.25	86.89	79.25	85.98	34.72	50.8	76.98	86.9
	Kappa Statistic	0.68	0.79	0.69	0.79	0.71	0.79	0.71	0.79	0.71	0.77	0.11	0.21	0.67	0.79
Naïve Bayes	Correctly Classified Instances %	72.83	79.31	73.58	83.68	77.36	83.22	77.36	83.22	76.23	82.53	39.62	53.56	80.75	83.45
	Kappa Statistic	0.62	0.67	0.64	0.74	0.69	0.73	0.69	0.73	0.67	0.72	0.17	0.24	0.73	0.73
K-nearest neighbours	Correctly Classified Instances %	37.36	63.45	38.87	76.32	70.57	81.38	70.57	86.21	59.62	81.15	35.09	41.38	75.85	83.45
	Kappa Statistic	0.12	0.42	0.13	0.62	0.59	0.7	0.59	0.78	0.43	0.69	0.09	0.09	0.67	
Random Forest	Correctly Classified Instances %	66.79	84.14	75.85	85.29	82.26	82.3	82.26	85.52	81.13	82.76	40.75	54.25	75.09	83.45
	Kappa Statistic	0.5	0.74	0.66	0.76	0.76	0.72	0.76	0.77	0.74	0.72	0.08	0.22	0.65	0.73

Figure A.14: Test Split 33% Student Results



Selected attribute

Name: G3 Type: Nominal
Missing: 0 (0%) Distinct: 5 Unique: 0 (0%)

No.	Label	Count	Weight
1	'(-inf-3.8]'	16	16.0
2	'(3.8-7.6]'	14	14.0
3	'(7.6-11.4]'	271	271.0
4	'(11.4-15.2]'	266	266.0
5	'(15.2-inf)'	82	82.0

Class: G3 (Nom)

Figure A.15: Target class distributed to five bins using equal-width binning in the Portuguese student data

Bibliography

Altman, Russ B. et al. (2013). "Data Re-Identification: Societal Safeguards". In: *Science* 339.6123, pp. 1032–1033. ISSN: 0036-8075. DOI: 10.1126/science.339.6123.1032-c. eprint: https://science.sciencemag.org/content/339/6123/1032.3.full.pdf. URL: https://science.sciencemag.org/content/339/6123/1032.3.

Altonji, Joseph G. and Rebecca M. Blank (1999). "Chapter 48 Race and gender in the labor market". In: vol. 3. Handbook of Labor Economics. Elsevier, pp. 3143–3259. DOI: https://doi.org/10.1016/S1573-4463(99)30039-0. URL: http://www.sciencedirect.com/science/article/pii/S1573446399300390.

Angwin, Julia et al. (2016). *Machine Bias.* URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (visited on 10/01/2018).

Association, American Diabetes (2000). "Type 2 Diabetes in Children and Adolescents". In: *Pediatrics* 105.3, pp. 671–680. ISSN: 0031-4005. DOI: 10.1542/peds.105.3.671. eprint: https://pediatrics.aappublications.org/content/105/3/671.full.pdf. URL: https://pediatrics.aappublications.org/content/105/3/671.

BBC News (2012). *Insurance gender ruling and you.* URL: Insurance%20gender%20ruling%20and%20you.

– (2018). *Uber halts self-driving car tests after death.* URL: https://www.bbc.co.uk/news/business-43459156.

Berg, Bruce (2011). *Qualitative Research Methods for the Social Sciences*, pp. 3–5. ISBN: 0205809383.

Berreby, David (2017). *Click to agree with what? No one reads terms of service, studies confirm.* https://www.theguardian.com/technology/2017/mar/03/terms-of-service-online-contracts-fine-print. (Accessed on 04/20/2019).

Bohannon, John (2015). "Privacy. Credit card study blows holes in anonymity." In: *Science (New York, N.Y.)* 347.6221, p. 468. ISSN: 1095-9203. DOI: 10.1126/science.347.6221.468. URL: http://www.ncbi.nlm.nih.gov/pubmed/25635068.

Boyer, Susan and Mark Stron (2012). *Best Practices for Improving, Survey Participation.*

Breiman, Leo et al. (2017). *Classification And Regression Trees.* Routledge. DOI: 10.1201/9781315139470. URL: https://doi.org/10.1201/9781315139470.

Brown, Ian, Lindsey Brown, and Douwe Korff (2010). "Using NHS Patient Data for Research Without Consent". In: *Law, Innovation and Technology* 2.2, pp. 219–258. ISSN: 0959-8138.

DOI: 10.1136/bmj.d973. URL: https://www.tandfonline.com/doi/pdf/10.5235/175799610794046186.

Brown, Ian, Lindsey Brown, and Douwe Korff (2011). “Limits of anonymisation in NHS data systems”. In: *BMJ* 342. ISSN: 0959-8138. DOI: 10.1136/bmj.d973.

Byrnes, Nanette (2016). *Why We Should Expect Algorithms to Be Biased - MIT Technology Review*. (Accessed on 04/22/2019).

Calders, Toon and Indre Zliobaite (2013). *Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures*. Springer, pp. 43–57. ISBN: 978-3-642-30486-6. DOI: 10.1007/978-3-642-30487-3.

Chouldechova, A. (2016). “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments”. In: *ArXiv e-prints*. arXiv: 1610.07524 [stat.AP].

Chudoba, Brent (2018). *How much time are respondents willing to spend on your survey?* URL: https://www.surveymonkey.com/curiosity/survey_completion_times/ (visited on 11/19/2018).

Cover, Thomas M, Peter E Hart, et al. (1967). “Nearest neighbor pattern classification”. In: *IEEE transactions on information theory* 13.1, pp. 21–27.

Dastin, Jeffrey (2018). *Amazon scraps secret AI recruiting tool that showed bias against women*. URL: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (visited on 11/21/2018).

Dictionary, Oxford (2018). *discrimination*. URL: https://en.oxforddictionaries.com/definition/discrimination (visited on 11/15/2018).

DLA Piper (2018). *DATA PROTECTION LAWS OF THE WORLD*, p. 624.

DOMO (2018). *Data Never Sleeps 6.0*. URL: https://www.domo.com/learn/data-never-sleeps-6 (visited on 11/18/2018).

El Emam, Khaled et al. (2011). “A Systematic Review of Re-Identification Attacks on Health Data”. In: *PLOS ONE* 6.12, pp. 1–12. DOI: 10.1371/journal.pone.0028071. URL: https://doi.org/10.1371/journal.pone.0028071.

European Parliament (2016). “REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL”. In: *Official Journal of the European Union* L116.12, pp. 1–88. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.

Goertzen, Melissa J (2017). "Introduction to Quantitative Research and Data". In: *Library Technology Reports* 53.4.

Guyon, Isabelle and Andre Elisseeff (2003). *An Introduction to Variable and Feature Selection.*

Hansson, Mats G. et al. (2016). "The risk of re-identification versus the need to identify individuals in rare disease research". In: *European Journal Of Human Genetics* 24. Article, p. 1553. URL: https://doi.org/10.1038/ejhg.2016.52.

Hardt, Moritz, Eric Price, and Nathan Srebro (2016). "Equality of Opportunity in Supervised Learning". In: *CoRR* abs/1610.02413. arXiv: 1610.02413. URL: http://arxiv.org/abs/1610.02413.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2016). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)*. Springer, p. 596. ISBN: 0387848576. URL: https://www.amazon.com/Elements-Statistical-Learning-Prediction-Statistics/dp/0387848576?SubscriptionId=AKIAIOBINVZYXZQZ2U3A&tag=chimbori05-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=0387848576.

Ho, Tin Kam. "Random decision forests". In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. IEEE Comput. Soc. Press. DOI: 10.1109/icdar.1995.598994. URL: https://doi.org/10.1109/icdar.1995.598994.

Hofmann, Hans (1994). *Statlog (German Credit Data) Data Set*. URL: https://www.openml.org/d/31.

Home Office (2019a). *DATA.POLICE.UK*. URL: https://data.police.uk/.

– (2019b). *Stop and search*. URL: https://www.ethnicity-facts-figures.service.gov.uk/crime-justice-and-the-law/policing/stop-and-search/latest.

John, George H. and Pat Langley (1995). "Estimating Continuous Distributions in Bayesian Classifiers". In: *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. UAI'95. Montr&#233;al, Qu&#233;, Canada: Morgan Kaufmann Publishers Inc., pp. 338–345. ISBN: 1-55860-385-9. URL: http://dl.acm.org/citation.cfm?id=2074158.2074196.

Kesavaraj, G. and S. Sukumaran (2013). “A study on classification techniques in data mining”. In: *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, pp. 1–7. DOI: 10.1109/ICCCNT.2013.6726842.

L. Bluma, Avrim and Pat Langley (1997). *Selection of relevant features and examples in machine learning.*

Le, James (2018). *A Tour of The Top 10 Algorithms for Machine Learning Newbies.* (Accessed on 04/22/2019).

Likert, R (1932). “A Technique for Measurement of Attitudes”. In: *Archives of Psychology.*

Maclin, Richard and David W. Opitz (2011). “Popular Ensemble Methods: An Empirical Study”. In: *CoRR* abs/1106.0257. arXiv: 1106.0257. URL: http://arxiv.org/abs/1106.0257.

Mahdi El Mhamdi, E. et al. (2018). “Removing Algorithmic Discrimination (With Minimal Individual Error)”. In: *ArXiv e-prints.* arXiv: 1806.02510 [cs.AI].

Martiniano, A. et al. (2012). “Application of a neuro fuzzy network in prediction of absenteeism at work”. In: *7th Iberian Conference on Information Systems and Technologies (CISTI 2012)*, pp. 1–4. URL: https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work.

Montjoye, Yves-Alexandre de et al. (2015). *Identity and privacy. Unique in the shopping mall: on the reidentifiability of credit card metadata.* 6221, pp. 536–9. DOI: 10.1126/science.1256297. URL: http://www.ncbi.nlm.nih.gov/pubmed/25635097.

Ng, Andrew Y. and Michael I. Jordan (2002). “On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes”. In: *Advances in Neural Information Processing Systems 14*. Ed. by T. G. Dietterich, S. Becker, and Z. Ghahramani. MIT Press, pp. 841–848. URL: http://papers.nips.cc/paper/2020-on-discriminative-vs-generative-classifiers-a-comparison-of-logistic-regression-and-naive-bayes.pdf.

Patton, Michael Quinn (2002). *Qualitative Research & Evaluation Strategy.* 3rd ed., pp. 4–5.

Personal Data Protection Commission of Singapore (2018). *GUIDE TO BASIC DATA ANONYMISATION TECHNIQUES.* URL: https://iapp.org/resources/article/guide-to-basic-data-anonymization-techniques/.

Porter, C Christine (2008). *De-Identified Data and Third Party Data Mining: The Risk of Re-Identification of Personal Information ¿¿ Shidler Journal of Law, Commerce & Technology DE-IDENTIFIED DATA AND THIRD PARTY DATA MINING: THE RISK OF RE-IDENTIFICATION OF PERSONAL INFORMATION*, p. 8. URL: http://www.lctjournal.washington.edu/Vol5/a03Porter.html..

Quinlan, J. R. (1986). "Induction of decision trees". In: *Machine Learning* 1.1, pp. 81–106. ISSN: 1573-0565. DOI: 10.1007/BF00116251. URL: https://doi.org/10.1007/BF00116251.

Reutemann, Peter (2018). *python-weka-wrapper3 0.1.6*. URL: https://pypi.org/project/python-weka-wrapper3/ (visited on 11/05/2018).

Rutkin, Aviva (2016). *Lazy coders are training artificial intelligences to be sexist*. URL: https://www.newscientist.com/article/2115175-lazy-coders-are-training-artificial-intelligences-to-be-sexist/ (visited on 11/20/2018).

Smart, J.C. (2016). *Ethical Reasoning in Big Data*. Ed. by Sorin Adam Matei Jeff Collmann. Springer, p. 95. ISBN: 978-3-319-28422-4. DOI: 10.1007/978-3-319-28422-4.

Smyth, Rosalind L (2011). "Regulation and governance of clinical research in the UK". In: *BMJ* 342. ISSN: 0959-8138. DOI: 10.1136/bmj.d238. eprint: https://www.bmj.com/content. URL: https://www.bmj.com/content/342/bmj.d238.

Stacy, Webb and Jean MacMillan (1995). "Cognitive Bias in Software Engineering". In: *Commun. ACM* 38.6, pp. 57–63. ISSN: 0001-0782. DOI: 10.1145/203241.203256. URL: http://doi.acm.org/10.1145/203241.203256.

SurveyMonkey (2018). URL: https://www.surveymonkey.com/ (visited on 11/20/2018).

Tockar, Anthony (2014). *Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset*. URL: https://research.neustar.biz/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/.

UCI Machine Learning (2018). *Student Alcohol Consumption*. URL: https://www.kaggle.com/uciml/student-alcohol-consumption/home.

– (2019a). *Adult Census Income*. URL: https://www.kaggle.com/uciml/adult-census-income.

– (2019b). *Pima Indians Diabetes Database*. URL: https://www.kaggle.com/uciml/pima-indians-diabetes-database.

Veale, Michael and Reuben Binns (2017). “Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data”. In: *Big Data & Society* 4.2, p. 205395171774353. DOI: 10.1177/2053951717743530. URL: https://doi.org/10.1177/2053951717743530.

Witten, Ian H., Eibe Frank, and Mark A. Hall (2011). *Data mining : practical machine learning tools and techniques*, p. 621. ISBN: 9780128043578.

– (2016). *The WEKA workbench*. 4th ed. Morgan Kaufmann. 128 pp.

Zhang, Harry (2004). “The Optimality of Naive Bayes”. In: vol. 2.

Zliobaite, Indre (2015). “A survey on measuring indirect discrimination in machine learning”. In: *CoRR* abs/1511.00148. arXiv: 1511.00148. URL: http://arxiv.org/abs/1511.00148.