Norwegian School of Economics
Bergen, Spring 2022

Machine Learning in Application-Based Case Management

A study on using machine learning to predict decision making in case management processes

Sindre Lien Oftebro & Adrian Rabben

Supervisor: Christian Langerfeld

Master thesis, Economics and Business Administration
Major: Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

Preface

This thesis is written as a part of our master's degree in Business and Administration with a Business Analytics major at The Norwegian School of Economics. We want to thank our supervisor Christian Langerfeld for his guidance throughout working on this thesis. At critical moments throughout this semester, he provided invaluable advice that proved vital for the thesis outcome.

Additionally, we extend our deepest gratitude to REC for providing their data, and especially Kjell Roald Langseth for helping us better understand the data through multiple meetings and interviews. We are also humbled by the opportunity to write this thesis as a case study for Machina AS, collaborating with them to identify key interest areas of machine learning in application-based case management processes. CEO at Machina, Helge Loy, has especially provided helpful advice regarding case management in general and potential applications of the findings.

Lastly, we would like to thank family and friends for being supportive throughout the writing of this thesis and though the entirety of our studies at The Norwegian School of Economics.

Norwegian School of Economics

Bergen, May 2022

Sindre Lien Oftebro

Adrian Rabben

Abstract

This thesis studies the possibility of using machine learning to predict the outcome of applications processed by the Regional Committees for Medical and Health Research Ethics (REK) in Norway. More specifically, the purpose is to predict rejections of medical research applications. Four supervised prediction methods are used to achieve this: Logistic regression, Naive Bayes, Random Forest, and XGBoost. Before training the models, a Latent Dirichlet Allocation topic model is implemented to extract structured features from the textual project description data, making it suitable for the supervised prediction models. The prediction models are evaluated and compared using metrics derived from the confusion matrix, namely Accuracy, ROC AUC, and Cohen's Kappa. The results show that the methods are suitable for predicting application outcomes, and XGBoost proves to have the best overall performance based on the selected metrics. Moreover, the topic variables from the LDA model prove to be influential to the predictions.

Based on the results, the thesis discusses some use cases of the XGBoost methodology, investigating the possibility of flagging applications predicted by the model to be rejected. Such an implementation aims to help case officers quickly identify applications that likely should be rejected, simplifying the work related to the initial assessment. The thesis finds this feasible but discusses some challenges of implementation. Subsequently, a discussion is made regarding the possibility of using the methodology to reject applications automatically. This is a more radical intervention in the case management system, and further clarification with REK is essential before real-world implementation.

Furthermore, the thesis looks at the weaknesses of the results. A discussion is made regarding the model's ineffectiveness in adapting to rapid changes in the environment, which is an inevitable issue when it comes to predicting the future based on historical data. In addition, the thesis examines which variables are ethically sound to include as predictors in predicting application rejections, and reflecting upon this issue before real-world implementation is advised.

Contents

FIGURES

TABLES

Equations

1 Introduction

This thesis explores the usage of machine learning in application-based case management processes. It is written in collaboration with Regional Committees for Medical and Health Research Ethics[1] (REK), an organization responsible for approving applications for all medical and health research projects in Norway, and Machina AS (Machina), an organization that specializes in digitalizing application-based case management processes. According to Machina, a case is "data or information that requires a form of processing" and case management is "the coordination of work related to evaluating, deciding, and following up a case" (H. Loy, CEO at Machina, personal communication, 25.04.2022). A *case* in this thesis refers to an application for permission to conduct a project, and the processing of such applications is referred to as *application-based case management*.

Case officers process applications, and to process effectively, they must possess extensive knowledge of the application subject area. Senior advisor at REK, K. Langseth, proclaims that case officers at REK often feel pressured on time when assessing applications due to having to meet strict deadlines (personal communication, 28.04.2022). This can potentially damage decision quality, but according to Chu & Spires, utilizing a sound decision support system "can induce decision makers to process more information and use more rigorous decision strategies, which can result in enhanced performance" (2001, p. 226).

In line with this, the Norwegian Ministry of Local Government and Regional Development has proposed a national strategy regarding the public use of artificial intelligence (AI), stating that AI can provide support for case officers by detecting anomalies, predicting outcomes, and improving the processing of natural language data (2020, p. 53). The strategy advocates for automation of decision processes and removing unnecessary discretionary assessment (Norwegian Ministry of Local Government and Regional Development 2020, p. 21), with the potential benefit of "more equal treatment [of applications] and more consistent implementation of regulations" (Norwegian Ministry of Local Government and Regional Development 2020, p.

[1] In Norwegian, Regionale komiteer for medisinsk og helsefaglig forskningsetikk - https://www.forskningsetikk.no/om-oss/komiteer-og-utvalg/rek/

21). However, according to Broomfield & Reutter, many public organizations lack the competency to implement AI successfully (2019, p. 3). The lack of competency leads them to conclude that "long-term, interdisciplinary holistic thinking about AI demands involvement from academia, private and public sector" (Broomfield & Reutter, 2019, p. 9). For this thesis, all three of these actors are involved, with academia providing the theoretical foundation and methodology, the public organization REK the data, and the private organization Machina their experience.

REK's objective is to ensure that medical research in Norway is conducted "in an acceptable manner" (Regional Committees for Medical and Health Research Ethics, 2022). All medical and health-related research projects completed on human beings, human biological material, or personal health data in Norway must be pre-approved by REK before the project can commence (Regional Committees for Medical and Health Research Ethics, 2022). Project leaders of such research projects must apply through REK's digital platform "Rekportalen" to get this approval, of which Machina is the provider.

To contend with the challenges that REK face and operationalize the means provided by the Norwegian Ministry's AI strategy, this thesis explores the value AI can provide in application-based case management processes. More specifically, machine learning, a subfield of AI, is used to extract insights from applications and predict outcomes. As the thesis uses data provided by REK, the results show the potential of implementing machine learning in their application processes specifically. However, implementation in other subject areas for application-based processing in the public sector, such as grants or licensing, is assumed to produce comparable results.

The thesis includes the implementation of an LDA topic model, used to engineer structured features from the application project description, which is crucial for the application outcome (K. Langseth, Senior advisor at REK, personal communication, 28.04.2022). Beyond this, the scope of the thesis is limited to supervised prediction methods, predicting whether applications are rejected or not. The predictions are solely based on data from the application forms and relevant metadata, all available to the case officer during their assessments. The thesis also discusses

some potential use cases of the predictions based on the proposals presented by the AI strategy (Norwegian Ministry of Local Government and Regional Development 2020).

Consequently, the research question of the thesis is: "*How can machine learning be used to predict whether applications sent to REK will be rejected?*".

The remainder of the thesis is structured as follows. In the next section, section two, the primary literature that the thesis is based upon is reviewed, looking at the theoretical foundations and related work. The third section examines the methodology used, presenting machine learning methodologies and metrics used for benchmarking. The fourth section discusses the data set, while the fifth section showcases the results and highlights the strengths and weaknesses of the models. The sixth section discusses potential use cases for the technology, weaknesses of the results, and ethical considerations. The seventh section summarizes the findings and concludes with closing remarks.

2 Literature Review

The literature review presents key literature and research used, laying the foundation for the methodology and analysis of the thesis.

2.1 Academic Field and Literature

For decades, machine learning has been used to automate and improve tasks, back to the 1950s when Frank Rosenblatt built a machine made to recognize letters (Fradkov, 2020, p. 1385). Since then, the popularity of the field has varied, hitting a golden age in the twenty-first century due to the access to big data, reduced cost of parallel computing, and development of deep machine learning methodologies such as deep neural networks (Fradkov, 2020, p. 1387).

Within machine learning, this thesis heavily relies on the ideas and methodology presented by James et al. (2013) in their book *An Introduction to Statistical Learning with Applications in R*, which presents an array of machine learning methods and their applications. Especially the theory and methodology related to classification (James et al., 2013, pp. 129-195) and tree-based methods (James et al., 2013, pp. 327-365) have been central to the thesis' approach. The work of Rhys (2020) in the book *Machine Learning with R, the tidyverse, and mlr* has also been of importance, acting as an inspiration and guideline, both in the aspect of model options and conveying the importance of model validation.

For implementing the methodology, the statistical programming language R, released in 1993 by Ihaka and Gentleman at the University of Auckland (Ihaka, 1998, p. 4), and R-studio, an integrated development environment (IDE) for R, released in 2011 (RStudio, PBC., 2022) has been significant. It has allowed powerful computations required for the methodology and is the primary tool used for the analysis. Concerning this, it is also worth mentioning the work of Wickham & Grolemund (2016) and their book *R for Data Science*, laying out the tidy method and the Tidyverse package, as its methods and practices are continuously used throughout the work done for this thesis.

Regarding topic modeling, the work of Silge & Robinson (2017) and their book *Text Mining with R: A Tidy Approach* has been influential. The preprocessing and topic model methodology used to analyze the textual data heavily resembles what is laid out in their book. Like this thesis, they also use The Latent Dirichlet Allocation method presented by Blei et al. (2003). In addition, the critiques offered by Vayansky & Kumar (2020) in their review of the LDA model are taken into consideration to handle some of the model's limitations.

Furthermore, the strategy by the Norwegian Ministry of Local Government and Regional Development (2020), presenting desired gains from implementing AI in the Norwegian public sector, is used. The strategy specifies potential use cases for case management processes and has provided this thesis with valuable guidelines and insights regarding the possibilities and value of implementing AI in REK's case management.

2.2 Similar Research

Besides the methodology literature, the thesis is also inspired by other research conducted on relevant topics. The first is a study done by Hubl & Merkert (2015), where they looked at 52 individual research papers which used machine learning to improve decision making in decision support systems. They concluded that the results of using machine learning in decision support systems "are better decision results in a faster way" (Hubl & Merkert, 2015, p. 13). This gives the thesis a promising foundation to build on as it also studies how machine learning can be applied to improve decisions. Hubl & Merkert also conclude that a combination of machine learning methods leads to higher effectiveness (2015, p. 13), supporting the combination of methods done in this thesis.

Two studies that use LDA topic modeling to support prediction methods have also been inspirational, the first being the study done by Geletta et al. (2019). They researched whether LDA topic modeling could improve the accuracy of a Random Forest prediction model, predicting whether a clinical study would be terminated (Geletta et al., 2019). They found that LDA topic modeling "significantly raises the utility of unstructured data in better predicting the completion vs. termination of studies." (Geletta et al., 2019, p. 10). This heavily relates to this thesis, as it also looks at data regarding clinical studies, trying to predict an outcome using

supervised prediction models. Their findings support the potential of including LDA modeling in the predictions (Geletta et al., 2019). This is also supported by the research conducted by Slof et al. (2021, p. 12), where they showed that prediction methods including LDA topic modeling outperformed prediction methods without topic modeling, predicting customer churn in the telecommunications business.

Finally, the work done by Etscheid (2019) in his conference paper "Artificial Intelligence in Public Administration" is considered. The study is used to discover the usefulness of artificial intelligence in public administration processes and how it should be approached. Etscheid points out that advances in AI in the last few years have opened the opportunity for more processes to be automated and highlights that not all processes can be automated due to the importance of people being the final decision-makers (2019). Building on this, he presents the importance of looking at public administration processes as a set of steps that each offer different challenges and opportunities (Etscheid, 2019). He encourages future research to delve into one of these specific steps for an "in-depth analysis of the individual phases and the development of concrete indicators for the degree of automation" (Etscheid, 2019, Conclusion, para. 3). This suggestion is highly in line with the scope and goals of this thesis, as it looks at the decision step of an application process to find the indicators that are important for the decision making and then predict the application outcome.

3 Data

Applications to REK are filled out using digital forms and sent in through Rekportalen. The data used for analysis consists of all the relevant application form fields. The forms allow for capturing data in a highly structured manner, resulting in data that require little preprocessing before being used to predict outcomes. Some metadata variables that are deemed relevant are also included in the analysis, such as the applicant organization, the processing organization, and the date applied.

The data set contains 23,719 applications which have all been processed by case officers in REK. In the preprocessing, quite a substantial proportion of applications are lost due to missing critical data (such as the project description or the decision variable), leaving 14,422 applications for the modeling. The data section is regarding these 14,422 applications, presenting all variable types used for modeling, with the most essential variables highlighted explicitly. Appropriate preprocessing steps are explained for each variable or variable type, with the goal of converting the data into formats that are well suited as input variables for the supervised machine learning methods. Data preprocessing is vital for the thesis, as high-quality data is a prerequisite for good analysis (Sesseions & Valtorta, 2006, Conclusion, para. 1). See Appendix 8 for an extensive list of the variables used in the analysis.

3.1 The Decision Variable

The decision variable represents the case officer's main decision regarding the application and is the response variable for the prediction methods. The variable is categorical and can take one of the following values: "Declined", "Rejected", "Approved", "Approved with conditions" and "Postponed".

Declined	**Rejected**	Approved	Approved with Conditions	Postponed
1,182	**3,796**	7,505	1,741	198

Table 1: The five outcomes of the application processing

The decisions are made based on how the research projects that are applied for satisfy the requirements of the "Act on Medical Research" (The Health Research Act). Applications are approved if the requirements are satisfied, declined if they are not satisfied, and postponed if further clarifications are required before deciding the outcome. A rejection is done if REK does not consider the research to fall under The Health Research Act (K. Langseth, personal communication, 24.05.2022). Table 1 shows that 3,796 out of the 14,442 applications are rejected, which represents a slight class imbalance in the data set when it comes to predicting rejections versus non-rejections. A total of 7,842 applications lacked a value for the decision variable and were filtered out as it is an essential variable for the analysis, accounting for 84,3% of the filtered applications.

3.2 Application Form Variables

The largest part of the data set is the values gathered from the applicants through the application form. The application form is extensive, containing 201 fields in total, and preprocessing of this data is therefore done based on the respective field types rather than for each variable. Table 2 shows all the distinct types of variables present in the application form. See Appendix 4 for an example of how the application form looks when applying for a research project through Rekportalen.

Field type	**Explanation**
Long text	Fields for either a short or a long answer
Short text	Fields for short answers, typically one sentence /1-3 words
Boolean	Fields with only yes/no options
Code list	Fields with a predefined list of options, where a maximum of one is selected
Code list multi	Fields with a predefined list of options, where 0-n are selected
Numeric	Fields where a numeric response is required
Date	Fields with date responses

Table 2: The types of variables gathered from the application form

3.2.1 Multiple Choice Variables

The data set contains 24 code list multi variables, where 13 have information regarding the actors in the project and the type of research to be conducted. The other 11 contain statistics regarding which countries the research project affects. Due to the high number of unique choices in these variables, all countries are aggregated up to their respective regions, as mapped by the United Nations Statistics Division (2022). Then, the code list variables are converted into dummy variables with one column for each choice to make them useful for prediction purposes.

3.2.2 Boolean, Numerical, and Date Variables

How the research project will be conducted and what preparations have been done is captured through a total of 76 boolean variables. These variables are already well suited for modeling purposes, so no preprocessing is performed.

There are two numerical values in the application form, the number of participants in the project in Norway and the number of participants in total. No preprocessing is necessary for these variables. There are also two date variables containing information concerning when the research is going to start and end. These variables are used to calculate the number of days of the project duration.

3.2.3 Text variables

The remaining 97 variables from the application data are text variables describing various aspects of the research. These variables are highly specific to certain types of projects, resulting in a considerable proportion of missing values, making it necessary to exclude them from the analysis. The exception is the project description which is present in almost all applications.

The project description is of particular interest as it contains explanations of the projects and how they will be conducted. Since the variable is in a textual format, it is not immediately well suited for decision outcome modeling. Still, by using topic modeling, structured features in the form of project topics are extracted. This is described in further detail in section 4.2 Topic Modeling.

3.3 Metadata Variables

The metadata variables are variables that are not present in the application form, but that still capture some of the context of the various applications, such as which organizations are involved and when the application was submitted.

Two main organizations are relevant to the applications, the applicant organization and the processing organization. The applicant organization is the organization responsible for the research project, typically universities. In total, there are 279 applicant organizations in the data set, and Figure 1 displays the 25 most frequent ones, with the rest being grouped in the category "Other". The figure shows that the applicant organization is unknown for about 3,000 applications and that the most active applicant organizations are Oslo University Hospital, the Norwegian University of Science and Technology, and Haukeland University Hospital.

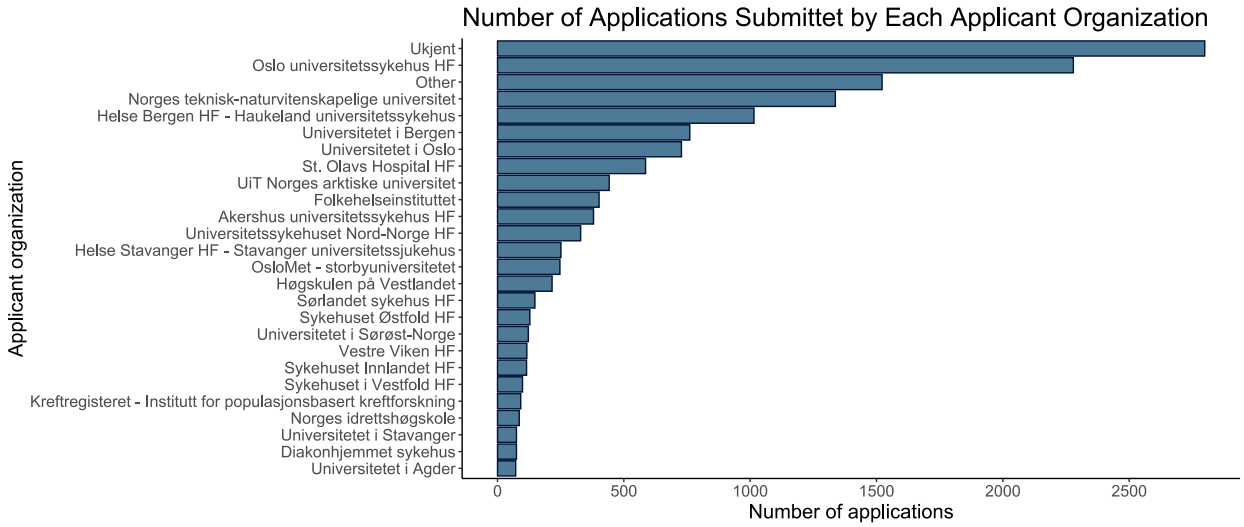

Figure 1: The total number of applications sent in by the most active organizations

The processing organization is the sub-section of REK that processed the application. There are four main processing organizations, REK Central, REK West, REK North, and REK South-East. The latter is split into four sub-organizations, giving a total of seven processing organizations. Each organization processes approximately the same number of applications, as shown in Table 3.

Processing Organization	Number of Applications
REK Central	1,941
REK North	2,126
REK South-East A	2,083
REK South-East B	2,062
REK South-East C	2,063
REK South-East D	2,149
REK West	1,998

Table 3: The processing organizations and number of applications processed by each

The last metadata variable is the applied year. REK started evaluating applications in 2009, the year that The Health Research Act came into force. The data set consists of applications from all years after 2009, but a larger proportion of the applications used in the modeling phase are from recent years compared to earlier years. Figure 2 shows that this is not because REK is processing more applications now than before. Instead, it stems from an improved data quality over the years, where more of the applications from early years have been filtered out due to missing data than applications from recent years.

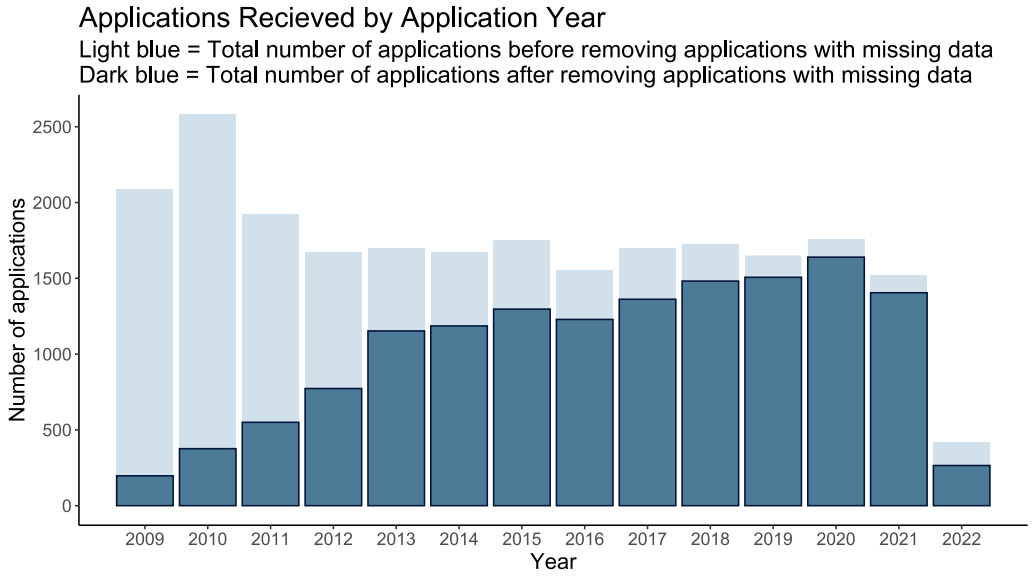

Figure 2: The number of applications received by REK each year

According to K. Langseth, there was a lack of competency regarding the legislation among researchers in the early years, which led them to apply unnecessarily (personal communication, 24.05.2022). This helps explain why there were more applications in 2009-2011 than in recent

years. K. Langseth points out that this is likely to also have affected the proportion of rejected applications in those early years (personal communication, 24.05.2022).

3.4 Application Language Variable

During the initial topic modeling, it was discovered that a considerable proportion of the applications were written in English rather than Norwegian. The difference in language proved to be a disturbance for the topic distribution as the topics were distinguished by language rather than semantics. Since this reduces the ability of the topic model to extract research subject areas from the applications, filtering out the observations with one of the languages before implementing the topic model is necessary. The data consists of far more applications in Norwegian than in English, and it is therefore decided that the English applications should be filtered out of the data set. The method used to filter the applications entails identifying the language of each application and then removing the applications labeled as being written in English. See Appendix 1 for more details on the language labeling methodology used. The language filtering is only done for the LDA topic model, and the English applications are re-introduced when predicting the application outcome. The reason for the re-introduction is that, according to information on Rekportalen, applications should only be written in English "if the project in its entirety is conducted abroad" (*REK-Portalen*, 2022). This indicates that the application language might be used as a rejection criterion in some instances, and the language variable might therefore be valuable for the prediction methods.

4 Methods

To explore how machine learning can be used to predict the outcome of new applications, four supervised prediction methods are implemented. In addition, an unsupervised classification method is used to extract useful features from the project descriptions in the application forms. In the following, the theory regarding the methods used in this thesis is presented as well as core decisions made regarding the implementation of each model.

4.1 Supervised and Unsupervised Learning

All machine learning methods can be categorized into one out of two categories, supervised or unsupervised (James et al., 2013, p. 26). The difference in how the methods work affects how they are used, and understanding these differences is important before presenting the methods themselves.

Supervised learning is the study of modeling the relationship between a set of predictor variables and a response variable (James et al., 2013, p. 26). For every value of the predictors x_i where $i = 1 \ldots, n,$ there is a response $y.$ The goal is to create a model that can predict the response of a new observation with a set of predictor variables or to better understand to what degree the predictor variables affect the response. A classic example of supervised learning is linear regression (James et al., 2013, p. 61).

With unsupervised learning "we lack a response variable that can supervise our analysis" (James et al., 2013, p. 26). When using an unsupervised learning method, the goal is instead to find the relationship between the observations. An illustrative example of an unsupervised learning method is clustering, where the goal is to label an observation into a distinct group based on its characteristics (James et al., 2013, p. 385). Another unsupervised learning method is topic modeling, which is presented in the following section.

4.2 Topic Modeling

Topic modeling is a clustering method especially suited for text data, used to determine which events or concepts documents concern by extracting latent variables from larger data sets (Vayansky & Kumar, 2020). In this thesis, topic modeling is used to categorize the different applications based on their project descriptions. The goal is to understand whether the project description topics are influential in predicting the application outcome.

4.2.1 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation is an unsupervised clustering method used to create a topic model of a given corpus (Newman et al., 2010). The method was first proposed by Blei et al. in 2003 and has since then been a frequently used topic modeling method (Vayansky & Kumar, 2020, p. 3). The assumptions of LDA are that every word in a document can be assigned to a topic, and that every document can belong to several topics (Blei et al., 2003). This is unlike many other document clustering methods which only allow for each document to belong to one topic (Blei et al., 2003, p. 997). The term *Latent* refers to the fact that the model contains variables "*which aim to capture abstract notions such as topics*" (Blei et al., 2003, p. 995). These variables are the unseen notional variables that link the topics and are never explicitly apparent in the model.

One of LDA's strengths compared to other topic models is its usage of the *Dirichlet* distribution, which is one of the reasons the method is chosen for this thesis. The Dirichlet distribution, $\mathrm{Dir}(\theta|\alpha)$, is a multinomial distribution used in Bayesian statistics (Liu, 2019). The distribution shows the probability of variables θ_i within *k* dimensions, over the simplex, where $\theta_i > 0, \sum_{i=1}^{\theta_i} = 1$ (Blei et al., 2003, p. 996). Figure 3 illustrates this distribution principle with three topics and observations.

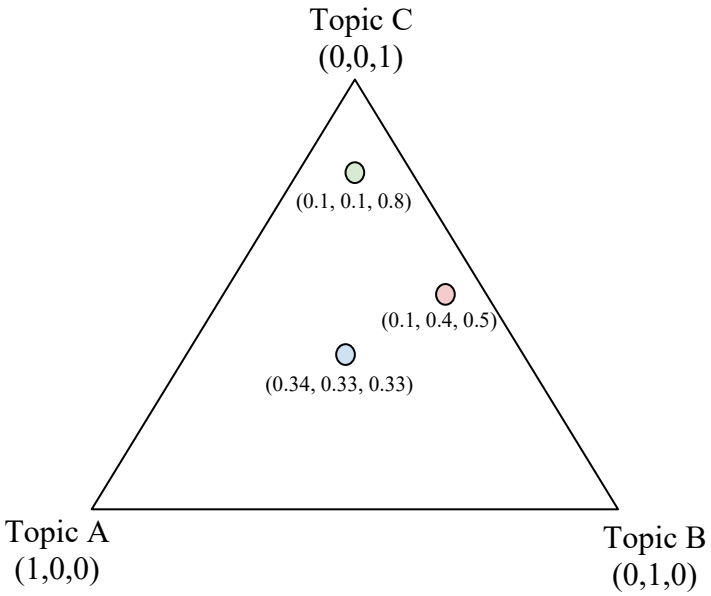

Figure 3: An example of three observations on the simplex

The LDA model uses Dirichlet distribution as it assumes that the sum of the *k topic* probabilities of each document in the corpus equals one (Blei et al., 2003, p. 996). In addition to the confined sum of θ, the Dirichlet distribution Dir(θ|α) also has the aspect of distribution density $\theta \sim Dir(\alpha)$ where $\alpha > 0$ (Blei et al., 2003, p. 996). The parameter α is a vector of *k* values and controls the distribution on the simplex (Liu, 2019). If $\alpha < 1$ the distribution spikes around the corners of the simplex, if $\alpha > 1$ the distribution will spike in the middle, especially for larger values (Liu, 2019).

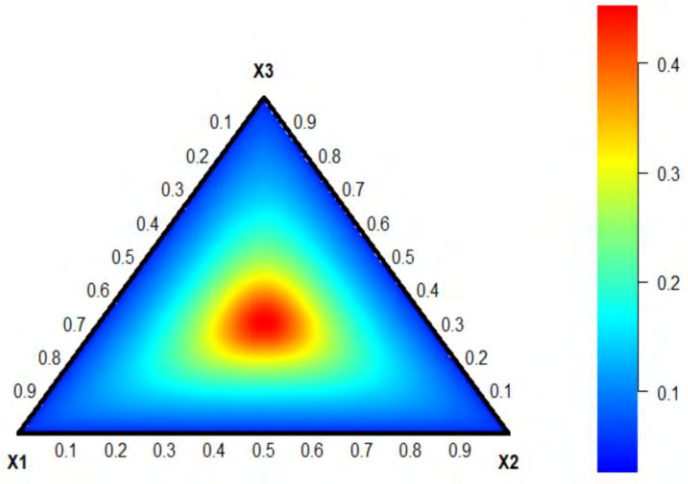

Figure 4: Distribution example with k = 3 and α = (5, 5, 5)

Figure 4 shows the probability distribution when $k = 3$, $\alpha = (5, 5, 5)$ and 500 random observations θ are drawn. This way of distributing topics is well suited for studying the project descriptions, as there is a chance that each of the project descriptions in the data set contains information from different areas of expertise. For a full description of the LDA topic model theory and methodology see the article "Latent Dirichlet Allocation" published by Blei et al. (2003).

4.2.1.1 Weaknesses of LDA

Although the LDA method can improve the ability to infer latent variables in unstructured data, it also has some challenges, as Vayansky & Kumar (2020) points out in their article "A review of topic modeling methods".

The first of these challenges is the method's dependency on hyperparameters, where using a different k (number of topics) and α (topic probability distribution) highly affect the outcome, even if the same data set is used (Vayansky & Kumar, 2020, p. 3). An ideal number of topics k for a corpus rarely exists (Vayansky & Kumar, 2020, p. 3). The thesis tries to offset this challenge by using the discovery methodology for number of topics presented by Cao et al. (2009). They illustrated that an indication of the optimal number of topics can be found using topic density, with the idea that the intra-cluster (within topic) similarity should be as high as possible, and inter-clusters (between topics) should be as low as possible (Cao et al., 2009, p. 1778). In combination with the density metric the method of using a Markov chain Monte Carlo algorithm to infer the number of topics presented by Griffiths & Steyvers (2004) is used. They found that such a method is effective for inferring the topics of scientific documents (Griffiths & Steyvers, 2004, p. 5235), as the project description of REK's applications are. When it comes to α, a symmetrical value can be beneficial because the real distribution of topics is unknown (Vayansky & Kumar, 2020, p. 3).

The second challenge pointed out by Vayansky & Kumar is the fact that the LDA method assumes that all topics and words are uncorrelated, disregarding context and semantics (2020, p. 4). Unfortunately, these assumptions are rooted in the nature of the LDA method and are difficult or even impossible to overcome. In this thesis, some correlation between the project descriptions

is likely, as some are written by the same person, some are follow-ups to previous applications, and some are project descriptions from previously declined applications that have been altered.

4.2.1.2 Preprocessing and Choosing Number of Topics

Before extracting topic information with LDA, the corpus is converted into unigrams and lemmatized, which is "a type of annotation that reduces inflectional variants of words to their respective lexemens (or lemmas) as they appear in dictionay entries" (McEnery et al., 2006, p. 35). In the context of topic modeling this is done "so that those representations are not undermined by a proliferation of words with similar meanings" (May et al., 2019, p. 1). To lemmatize the project description words, the R package "udpipe" published in 2021 by Jan Wijffels is used. Udpipe takes a list of words and returns the lemmatized version based on a Norwegian treebank (Wijffels, 2022). An example of a lemmatized word in this thesis is presented in Table 4, showing the original word and the lemmatized version.

Original Word	**Lemmatized Word**
opererer	operere
opereres	operere
operered	operere
operert	operere
opererte	operere
operertes	operere

Table 4: Example of lemmatization

Next, a common preprocessing step for LDA modeling, removal of stop words, is performed. Stop words are words that are common and contentless in a corpus (Schofield et al., 2017, p. 432). Stop words are identified using two methods, the first being the R package "stopwords" published by Benoit et al. in 2021. The package contains a collection of commonly used words for various languages, which often acts as noise in a topic model. Only the Norwegian words in the package were used. After these general-purpose stop words have been filtered out, corpus-specific stop words are identified using each words' inverse document frequency (IDF), representing how many documents (project descriptions) each word is used in. All words with IDF = 9.59, which are the words that are only present in one application, are added to the stop word list as they cannot be used to identify similarities between applications. It is also discovered through observation that some words were highly influential to many topics. To handle this, the

100 words with the highest IDF score are also added to the stop word list due to being over-represented in the corpus. An example of a word that was represented in nearly all applications was the word “project”, which does not provide any value for the topic model as all descriptions are about a research “project”. See Appendix 2 for more details on the IDF stop word removal methodology, and Appendix 3 for the full final list of the stop words used.

Finally, before the topic model is implemented, the number of topics *k* must be chosen. In Figure 5 the test metrics proposed by Cao et al. (2009) and Griffiths & Steyvers (2004) are plotted based on multiple runs of the LDA model with various values for *k*. The figure shows a diminishing improvement of both metrics for each additional topic, with almost no additional improvement for each additional topic above 50. This acts as a guide for choosing the final number of topics, but the final value is ultimately chosen based on qualitative inspections of the topics.

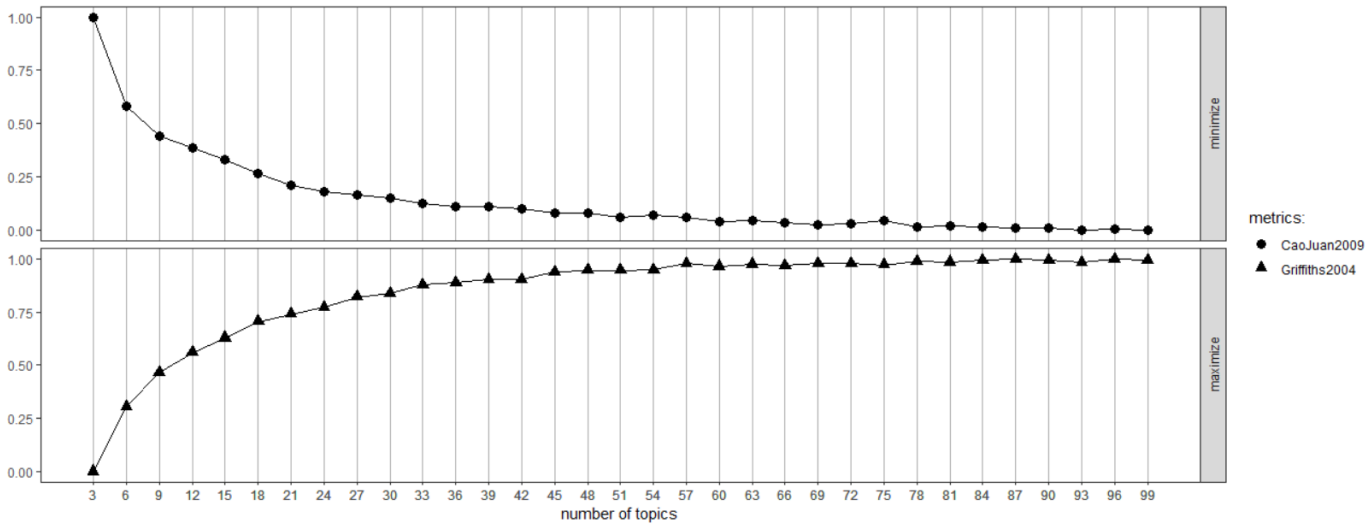

Figure 5: The topic optimization metrics for runs with different numbers of topics

Through the qualitative assessment, it is found that 40 topics is a reasonable number due to the topics proving to be meaningful when this number is used, and not being particularly computationally expensive to implement. After the LDA topic model is created, the topic probabilities for each of the applications are merged into the data set used for analysis. This results in 40 new numerical variables, ranging from zero to one (topic probability), with the combined sum of 1 per observation.

4.2.1.3 Presenting Example Topics

After creating the topics, it is of interest to infer what the typical contents of these topics are, both as a quality assurance but also to be able to better understand how they affect the predictions. The topics presented in Table 5 below show the ten most frequently used words in five of the topics. Each topic is clearly related to a specific subject within the field of medicine, with topic 6 relating to obesity, topic 11 to heart diseases, topic 15 to geriatrics, topic 17 to cancer treatment and topic 22 to mental health. The words in Table 5 are translated to English, see Appendix 6, Table 23, for an overview of the original Norwegian words.

Topic 6	**Topic 11**	**Topic 15**	**Topic 17**	**Topic 22**
obesity	heart	old	survival	mental
overweight	variety	relatives	relapse	disorder
vitamin	atrial fibrillation	dementia	cancer	depression
diet	heart failure	nursing homes	radiation therapy	anxiety
metabolic	vascular disease	care	side effect	symptom
nutrition	heart disease	palliative	combination	sleep
food	heart function	kidney	chemotherapy	psychological
weight	alcohol	home residents	chemotherapy	therapy
diabetes	genetically	life	tumor	psychiatric
eat	blood pressure	delirium	lung cancer	stress

Table 5: Top 10 words for five of the topics extracted with the LDA model

4.3 Supervised Machine Learning Methods

4.3.1 Classification methods

With the aim of predicting whether applications are rejected, four supervised classification methods are introduced: logistic regression, Naive Bayes, Random Forest and XGBoost. Using several methods enable model comparison through evaluating results, implementation process, and variable importance, helping to understand the strengths and weaknesses of the different methods. It also allows for showcasing whether outcome prediction is generalizable for a range of models with different underlying assumptions.

4.3.1.1 Logistic Regression

Logistic regression is a predictive method used to understand the relationship between the response variable and the predictor variables by estimating probabilities through a logistic regression equation (James et al., 2013, pp. 133-137). Logistic regression is well suited for binary classification, such as in this thesis, and is according to Hosmer et al. (2013, p.1) "the most frequently used regression model" for such classifications. Other than the assumption that the outcome is categorical or binary, logistic regression mostly follows the principles of linear regression (Hosmer et al., 2013, p.1).

4.3.1.1.1 *Logistic regression Method*

One type of predictive method that uses logistic regression is generalized linear models (GLM). All GLMs have in common that the mean of the outcome $E(Y)$ is modeled as a function of the predictors $X_1, \dots, X_p$ (James et al., 2013, p. 170). In the case of logistic regression, this function takes the form as seen in Equation 1.

$$E\left(Y\middle|X_1, \dots, X_p\right) = \Pr\left(Y = 1\middle|X_1, \dots, X_p\right) = \frac{\left(e^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}\right)}{\left(1+e^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}\right)}$$

Equation 1: Logistic regression function

In GLMs, the relationship between the predictors and the expected outcome is assumed to be linear, which can be seen from Equation 1 is not initially satisfied. This is solved through the *logit* transformation seen in Equation 2, which alters Equation 1 through a link function η, taking "the logarithm of the odds of the positive response" (Maalouf, 2011, p. 4).

$$\eta = logit\big(E(Y)\big) = \log\left(\frac{E(Y)}{1 - E(Y)}\right)$$

Equation 2: Logit transformation function

The transformation leads to Equation 3, where "the transformed mean is a linear function of the predictors" (James et al., 2013, p. 170).

$$\log\left(\frac{E(Y)}{1-E(Y)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Equation 3: Transformed logistic regression function

Equation 3 then satisfies the linearity assumption between the expected outcome and the predictors.

Various methods can be used to estimate the coefficients in logistic regression, but the *maximum likelihood* method is often preferred, yielding values "that maximize the probability of obtaining the observed set of data" (Hosmer, et al., 2013, p. 8). The method is applied through the likelihood function shown in Equation 4, Where the resulting estimators for the parameters are the values that maximize this function (Hosmer, et al., 2013, p. 8).

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} \left(1 - p(x_{i'})\right)$$

Equation 4: Likelihood function

Beyond the requirement of a categorical outcome and the linear relationship between predictors and the non-transformed outcome, all observations are assumed to be independent of each other, where the same subjects are not represented across several observations in the data. In addition, all predictors are assumed to be uncorrelated, and all errors to be independent of each other.

4.3.1.1.2 Logistic regression Model Implementation

When implementing logistic regression in this thesis, the format of the variables is first considered. To meet the linearity-assumption, all variables must be formatted in an ordinal fashion, meaning that categorical variables must be transformed into numerical ones. This is done through one-hot-encoding, transforming each categorical variable into multiple dummy variables (taking the value of 0 or 1). Prior to this, all uncommon factor levels for each of the categorical variables are grouped together into a new class "Other" to avoid previously unseen factor levels in the test set. To face the issue of the topic variables being highly correlated due to the nature of the Dirichlet distribution, an attempt to decorrelate the topic variables through

principal component analysis is made. The results show that this does not significantly improve the model results, and since principal components makes the feature interpretation less intuitive, this step is not included in the final model. However, another decorrelation and dimensionality reduction approach, backwards AIC variable selection, is performed. See Appendix 7 for more information regarding this technique and the final set of variables for the logistic regression model.

4.3.1.2 Naive Bayes Classifier

The Naive Bayes Classifier is a "simple and powerful machine learning algorithm" based on Bayes' theorem (Berrar, 2019, p. 1). The method is known for being "fast, easy to implement […], and effective" as well as being useful for high dimensional data (Taheri & Mammadov, 2013, p. 788).

4.3.1.2.1 *Naive Bayes Method*

Bayes' theorem regards using the initial probability (a priori) of a given class *P(A)* and the conditional probability of a given effect given that class *P(B|A)*, to estimate the probability A given B *P(A|B)*, exemplified with Equation 5 (Efron, 2013, p. 1178).

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Equation 5: Bayes' theorem

The Naive Bayes classifier simplifies the Bayes' theorem by assuming that features $B_1, B_2, \dots, B_n$ are conditionally independent of each other given the class A, leading to Equation 6 (Taheri & Mammadov, 2013, p. 788).

$$P(A|B) = \frac{\prod_{i=1}^{n} P(B_i|A)\,P(A)}{P(B)}$$

Equation 6: Naive Bayes Classifier

The independence assumption is often criticized as a poor assumption to make due to being unrealistic in real life (Taheri & Mammadov, 2013; Berrar, 2019; Rish, 2001). Nevertheless,

Naive Bayes is known to be a solid classification method that is theoretically well grounded (D'Agostini, 1995, p. 2) and that "often competes well with much more sophisticated techniques" (Rish, 2001, p. 41).

4.3.1.2.2 *Naive Bayes Model Implementation*

When preparing the data for the Naive Bayes model, attempts are made at decorrelating some of the variables. However, Naive Bayes is according to Taheri & Mammadov "remarkably robust" when violating the independence-assumption (2013, p.788), which is discovered in this thesis as well. Based on this, and to keep the implementation process simple, the final model does not utilize any decorrelation-techniques.

The methodology presented shows how Naive Bayes works for discrete variables, but the classifier is also designed to handle continuous variables (Berrar, 2019, p.6). However, Taheri & Mammadov show that the method tends to have a better performance when the continuous variables are discretized, "a process which transforms continuous numeric values into discrete ones" (Taheri & Mammadov, 2013, p.789). The continuous variables in this thesis are therefore discretized into bins based on quantiles before implementing the Naive Bayes model. In addition, uncommon factor levels in categorical variables are grouped together, as was done for the Logistic Regression model.

4.3.1.3 Tree-based Methods

The last methods used in this thesis are Random Forest and Extreme Gradient Boosting (XGBoost). These methods are decision tree ensembles, meaning that they are based on the combination of multiple decision trees. Each individual decision tree is called a "weak learner" since they only make mediocre predictions on their own (James et al., 2013, p 340), and by combining the predictions of many weak learners, the goal is "to obtain a single and potentially very powerful model" (James et al., 2013, p. 340).

Tree-based methods are widely recognized to be simple both in terms of implementation and interpretation and are considered as closely resembling the thought-process behind human decision-making (Hardman & Macchi, 2004, p. 191). The methods involve "segmenting the

predictor space into a number of simple regions" (James et al., 2013, p. 327), where each area in the predictor space is associated with a specific outcome based on the training observations in that area. This can be visualized as in Figure 6, where the predictor space is the area created by the variables along each axis. The segmentation of the predictor space is done based on specific values of the predictors, in this case *project_period_days* and *topic_27*, where the values are chosen based on what would lead to the largest improvement of the objective function. Each point in the predictor space of Figure 6 represents an actual training observation. Note that this is a simplified version of a predictor space as it is only made up of two variables.

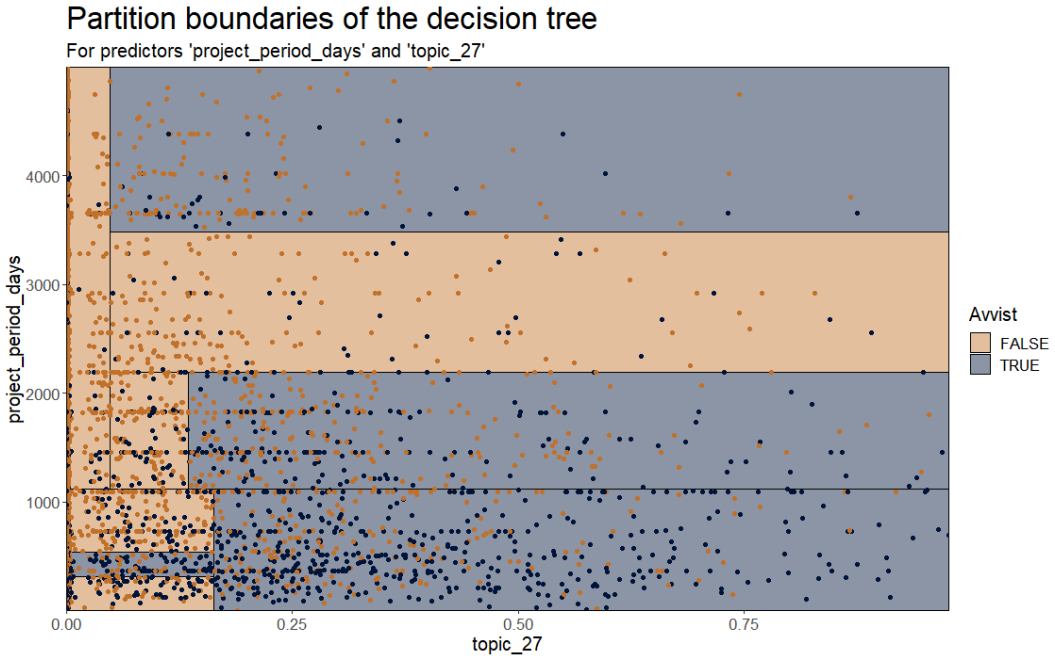

Figure 6: Two-dimensional predictor space segmented into multiple prediction areas

To better understand how decision trees work, see Figure 7 which displays the decision-tree corresponding to the predictor space of Figure 6. The tree consists of several partitioning rules represented by the nodes with outgoing sequence flows. The tree starts with a single node representing the root of the tree and given a criterion relating to the predictors in the model, the tree splits into two new nodes. In the predictor space of Figure 6, this initial split is represented by the horizontal line where *project_period_days* is equal to 1119, and the nodes are represented by the areas of the predictor space that result from the split. The splitting process of the tree is repeated several times, and each of the nodes where a new split is performed is referred to as an

"Internal node" (James et al., 2013, p. 329). In the example, there are nine internal nodes in the tree which creates 11 segments.

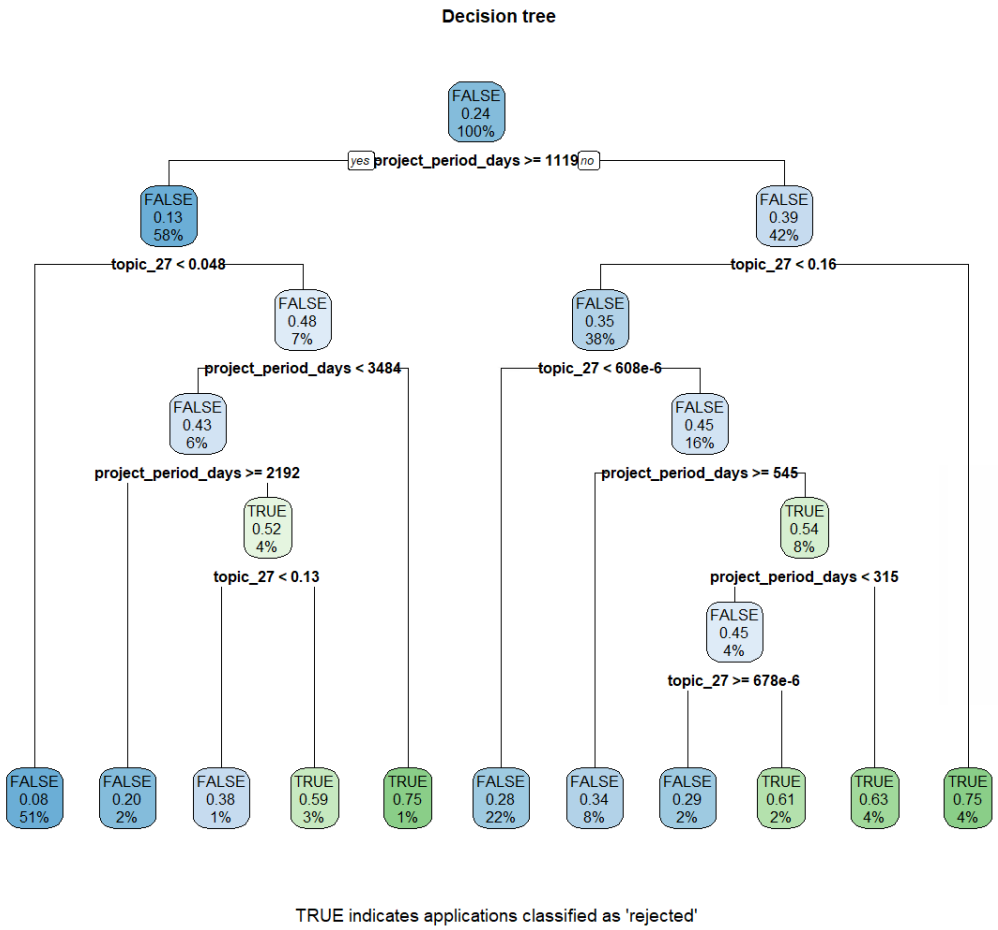

Figure 7: A decision tree used to divide the predictor space

In the decision tree displayed in Figure 7, the label TRUE and FALSE represent which outcomes the nodes are associated with, with TRUE meaning that a node is associated with rejections. The decimal number represents the proportion of training observations within that node that are rejections, and the percentages in each node represent the proportion of the total observations that belong to that node.

The decision rules for the splits in the internal nodes are identified by minimizing an objective function, most often the Gini index, which is derived by Equation 7 (James et al., 2013, p. 336).

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

Equation 7: Gini index function

Where $\hat{p}_{mk}$ is the "proportion of training observations in the *m*-th region that are from the *k*-th class*"* (James et al., 2013, p. 336). As can be seen from the Gini index, low or high values for $\hat{p}_{mk}$ leads to a low Gini index, meaning that partitions that lead to nodes with an overwhelming number of observations from a specific class are favored. This is also referred to as the *purity* of the nodes (James et al., 2013, p. 336).

In Figure 7, the top node represents the decision that led to the minimization of the Gini index at that specific node, which would implicate that the decision rule *project_period_days* >= 1119 (about three years) is the decision rule that leads to the split that minimizes the Gini index at that node. The two resulting nodes split up even more if it leads to an improvement in the Gini index, based on another decision rule. If a split would not lead to an improved Gini index, the node is called a "leaf" (James et al., 2013, p. 329), and all new observations in these nodes are predicted as belonging to "the most commonly occurring class of training observations that belong to the same node" (James et al., 2013, p. 335).

While decision trees are easy to understand, great for interpreting variable importance, and can handle both qualitative and quantitative data, their predictive accuracy are not particularly good due to having a high variance (James et al., 2013, p. 340). However, through ensemble methods such as bagging and boosting, Random Forest and XGBoost aims to solve this variance issue and improve the accuracy of the models (James et al., 2013, p. 340).

4.3.1.4 Random Forest

Radom forests are ensemble methods known for handling data sets with many predictors exceptionally well (Biau & Scornet, 2016, p. 1). The methods combine several randomized decision trees and then averages their predictions, which has the benefit of reducing variance compared to single decision trees (Biau & Scornet, 2016, p. 10). There are many variations of Random Forests, but this thesis implements the original method introduced by Breiman in 2001.

4.3.1.4.1 *Random Forest Method*

The core idea of the Random Forest method is the concept of bagging. Bagging is "a general-purpose procedure for reducing the variance of a statistical learning method" (James et al., 2013, p. 340). The method entails generating B different bootstrap training data sets, training a model on each of the bootstraps, and then averaging the model results to obtain predictions which are the average of all the trained models (James et al., 2013, p. 341). The objective function used to decide the splits at each node for each of the decision trees is the Gini index, as presented in Equation 7.

Although bagging is useful for reducing variance, boosting methods alone have shown to not be sufficient for decision tree ensembles when there is one very strong and several moderately strong predictors in the data set (James et al., 2013, p.344). This is because the strongest predictor is likely to be used at the root node for almost all individual decision trees, which in turn means that the trees inevitably will be highly correlated. To handle this, Random Forest utilizes random sampling of predictors at each split in the decision trees (James et al., 2013, p. 343). At each split in a tree, only a random sample *m* of the total number of predictors *p* is considered, which has the effect of reducing the number of trees that are highly influenced by the strongest predictors. This decorrelation technique leads to a lower variance when averaging the predictions of all individual trees, which makes the results more reliable (James et al., 2013, p. 344). When training a Random Forest model, the choice of the number of predictors to consider at each split is therefore vital to the model's performance, and according to James et al., (2013 p. 345), *"*Using a small value of *m* [number of predictors considered] in building a Random Forest will typically be helpful when we have a large number of correlated predictors".

4.3.1.4.2 *Random Forest Model Implementation*

A key feature of the Random Forest method is that it does not make any formal assumptions about the data, and because of this, extensive data preprocessing is not necessary. However, as with the previous models, uncommon factor levels in categorical variables are grouped together before the implementation of the Random Forest model to reduce dimensionality.

A difference between Random Forest compared to Logistic regression and Naive Bayes is the need to decide values for the *hyperparameters* for the model, namely the number of variables to consider at each split (*mtry*), and the minimum number of data points required to perform a split (*min_n*). These parameters are decided through hyperparameter tuning with 5-fold cross-validation and can be seen in Table 6. The parameter *trees*, which refers to the number of decision trees to use in the model, is chosen manually to be 1,000. This is done based on the argument that Random Forests are not susceptible to overfitting if the number of trees is sufficiently large (James et al., 2013, p.341).

trees	**mtry**	**min_n**
1,000	75	15

Table 6: The hyperparameter values for the Random Forest model

4.3.1.5 XGBoost

XGBoost is another method based on decision tree ensembles, known to produce "state-of-the-art results on a wide range of problems" (Chen & Guestrin, 2016, p. 1). The method has dominated the winning solutions of the machine learning competitions hosted by Kaggle, with seventeen out of twenty-nine winning solutions using XGBoost in 2015 (Chen & Guestrin, 2016, p. 1).

4.3.1.5.1 *XGBoost Method*

XGBoost is implemented under the Gradient boosting framework (*XGBoost 1.6.0 Documentation*, 2021), which entails training trees in an additive manner with each tree learning from previous trees (Chen & Guestrin, 2016, p. 786). While the specific objective function in the XGBoost method can vary based on the available data (*XGBoost 1.6.0 Documentation*, 2021), the principle of the gradient tree boosting objective is represented as seen in Equation 8.

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)$$

Equation 8: XGBoost objective function, consisting of a loss function and a regularization term

Where $\hat{y}_i^t$ represents the prediction made for the *i*-th tree in the *t*-th training iteration (Chen & Guestrin, 2016, p. 786). *l* is a loss function "that measures the difference between the prediction $\hat{y}_t$ and the target y_i" (Chen & Guestrin, 2016, p. 786), and in this thesis, log loss is chosen as the loss function as it is dealing with a binary classification issue. $\Omega(f_t)$ is a regularization term that "penalizes the complexity of the model [the decision trees]", which helps prevent overfitting (Chen & Guestrin, 2016, p. 786). This regularization term is seen in Equation 9.

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \mathrm{w_j^2}$$

Equation 9: XGBoost regularization term

Where γ is a pruning factor, *T* is the number of leaves in each decision tree, j = 1, 2, ..., T is the individual leaves, *w* is the weight of each leaf and λ is the regularization term for the weights.

The gradients g_i (first order) and h_i (second order) are used to optimize the objective, giving the following simplified objective function at step *t* (after removing constants) (Chen & Guestrin, 2016, p. 786).

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} [\mathrm{g_i f_t(x_i)} + \frac{1}{2}\mathrm{h_i f_t^2(x_i)}]\, \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \mathrm{w_j^2}$$

Equation 10: XGBoost simplified objective function at step t

In practice, this means that the tree that minimizes Equation 10 at iteration *t* is the tree that is added to the ensemble (Chen & Guestrin, 2016, p. 787). To identify the optimal tree for each iteration, it is necessary to compute the optimal leaf weights w_i that lead to the optimal tree. For a fixed tree structure *q*, this is given by Equation 11 (Chen & Guestrin, 2016, p. 787).

$$\mathrm{w_j^*} = -\frac{\sum_{i \in I_j} \mathrm{g_i}}{\sum_{i \in I_j} \mathrm{h_i} + \lambda}$$

Equation 11: XGBoost optimal leaf weights for a fixed tree structure

And based on the optimal weights, the optimal tree is the tree structure q that minimizes Equation 12 (Chen & Guestrin, 2016, p. 787).

$$\tilde{\mathcal{L}}^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} \mathrm{g_i}\right)^2}{\sum_{i\in I_j} \mathrm{h_i} + \lambda} + \gamma T$$

Equation 12: XGBoost finding optimal tree structure

The problem with Equation 12 is that it usually "is impossible to enumerate all the possible tree structures q" (Chen & Guestrin, 2016, p. 787). Which is why XGBoost instead utilizes a greedy algorithm that starts from a single leaf and iteratively adds branches to the tree. In this algorithm, I_L and I_R are instance sets of the left and right nodes for each split, and $I = I_L \cup I_R$. The split candidates are evaluated as shown in Equation 13.

$$\tilde{\mathcal{L}}_{split} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} \mathrm{g_i}\right)^2}{\sum_{i\in I_L} \mathrm{h_i} + \lambda} + \frac{\left(\sum_{i\in I_R} \mathrm{g_i}\right)^2}{\sum_{i\in I_R} \mathrm{h_i} + \lambda} - \frac{\left(\sum_{i\in I} \mathrm{g_i}\right)^2}{\sum_{i\in I} \mathrm{h_i} + \lambda}\right] - \gamma$$

Equation 13: XGBoost split-finding algorithm

The first fraction inside the parenthesis is the gain from the left node, and the second fraction is the gain from the right node. The third fraction represents the original leaf before the split and is subtracted from the total gain of the left and right nodes. Thus, if the gain from performing the split is lower than the pruning factor γ, the split will not be performed.

In practice there are many ways to identify the best splits in gradient boosting techniques, and one should refer to Chen & Guestrin, (2016) for further reading on some of these split-finding algorithms.

4.3.1.5.2 *XGBoost Model Implementation*

XGBoost is similar to Random Forest as it does not make any major assumptions about the formats of the data. There is however one exception to this, as XGBoost does require numerical values as input (*XGBoost 1.6.0 Documentation*, 2021), meaning that all categorical variables must be converted to dummy variables if they are not ordinal. In the case of the data used in this

thesis, none of the categorical variables are ordinal, so all categorical variables are converted into dummy variables through one-hot encoding.

XGBoost also requires that values for its hyperparameters are chosen before the model is trained. The hyperparameters are the same as in Random Forest, with four additions. The first of these is *tree depth*, regarding the maximum number of splits that can be made for each tree. The second is the *learning rate* which considers how much previously grown trees should contribute to each additional tree. The third is the *loss reduction*, which is the minimum reduction in the objective function required to make a new split. The last of the hyperparameters is the *sample size*, the proportion of the training set that the model should sample prior to growing the trees (*XGBoost 1.6.0 Documentation*, 2021). Although each parameter impacts the model in its own ways, they are not discussed in detail because suitable parameter values can be found through hyperparameter tuning, meaning they are chosen automatically through optimization. The exception is the parameter *trees*, which is set manually to 1000 just as in the Random Forest model. The hyperparameter values resulting from the tuning process are shown in Table 7.

trees	**mtry**	**min_n**	**tree depth**	**learn rate**	**loss reduction**	**sample size**
1,000	86	18	11	0.014	2.64e-05	0.96

Table 7: The hyperparameter values for the XGBoost model

4.3.2 Prediction and Thresholds

The prediction methods used in this thesis have the goal of predicting whether an application is rejected or not. The methods produce a probability based on the predictors to indicate the likelihood of each application being rejected. The probabilities are weighed up against a threshold, where predicted probabilities above the threshold are classified as rejections. Most classification methods use a threshold of 0.5 by default, but in instances where the data set is unbalanced, a different threshold can be optimal (Esposito et al., 2021, p. 3623).

The prediction outcomes are classified as true positives and true negatives for correctly predicted outcomes, and false positives and false negatives for wrongly predicted outcomes. Models with a low threshold will reject more applications and will therefore have a higher number of false

positives. With higher thresholds, the criteria for being classified as a rejection is stricter, leading to fewer true positives, and a lower proportion of false positives.

Table 8 shows a confusion matrix displaying the relationships between true/false positives and negatives.

		True Class	
		Positive	Negative
Predicted Class	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Table 8: A confusion matrix showing the relationship between the predicted and true class

Since different thresholds lead to different results, it is natural to set thresholds according to the use cases to achieve the desired model behavior.

4.4 Benchmarking Metrics

Method benchmarking regards the comparison of performances for different methods using the same data set (Weber et al., 2019, p. 1). As this study uses binomial supervised prediction methods, the three metrics, accuracy, ROC AUC, and Cohen's Kappa are used to evaluate and compare the performance of each of the models.

4.4.1 Accuracy

Accuracy refers to the ratio of correct predictions a method makes, which can be formulated as in Equation 14.

$$\textbf{Accuracy} = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

Equation 14: Accuracy

The accuracy gives an impression of the model's prediction ability, giving an easily comprehensible value in percentage of correctly predicted cases. The simplicity of the metric is also one of its disadvantages, as a high accuracy score does not necessarily signify a good model,

leading to the potential of falsely verifying poor models. A high value only indicates that the model often predicts correctly but does not inform on how educated the predictions are.

4.4.2 ROC Curve

The receiver operating characteristic (ROC) curve is a measure of the quality of binomial classification models (Mandrekar, 2010), derived by looking at the results of different classification thresholds (Karimollah, 2013).

In ROC curves there are two values of interest: the sensitivity, which is the true-positive fraction, and the specificity, the true-negative fraction (Karimollah, 2013). To derive the curve, the ratios are calculated as in Equation 15 and Equation 16 respectively, at every threshold of the prediction model (Karimollah, 2013).

$$\textbf{Sensitivity} = True\ Positive\ Fraction\ = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Equation 15: Sensitivity

$$\textbf{Specificity} = True\ Negative\ Fraction\ = \frac{True\ Negative}{True\ Negative + False\ Positive}$$

Equation 16: Specificity

When graphing the ROC curve, one plots the sensitivity on the y-axis and the false positive fraction expressed as “1 - specificity” on the x-axis, shown in Equation 17 (Karimollah, 2013).

$$\textbf{False Positive Fraction} = 1 - \text{Specificity} = \frac{False\ Positive}{True\ Negative + False\ Positive}$$

Equation 17: False positive fraction

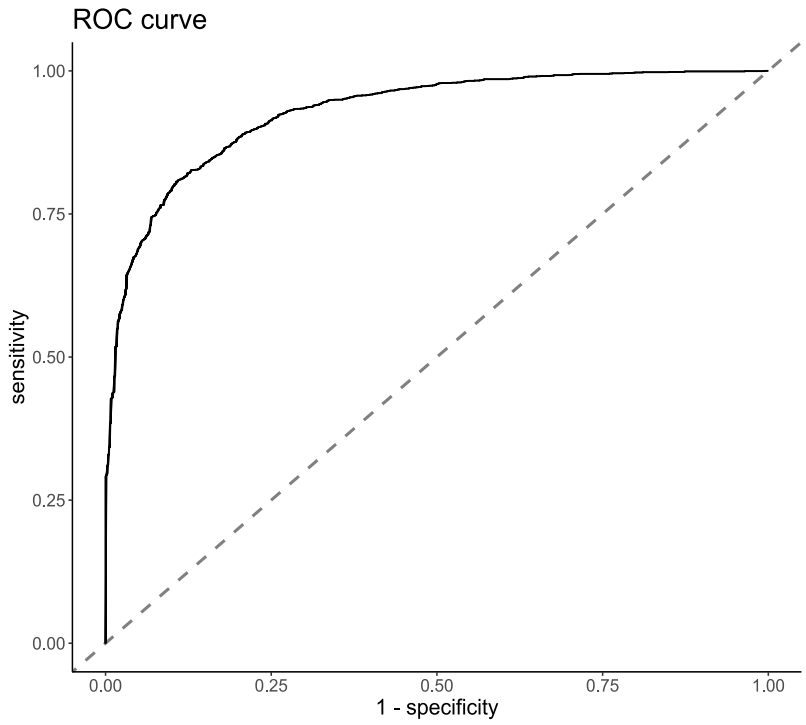

Figure 8: Example of a ROC curve

The result is a graph like in Figure 8 showing the trade-off between correctly classifying positives versus incorrectly classifying negatives for various thresholds. At higher thresholds, a higher fraction of observations is correctly classified as positive, but a lower fraction is correctly classified as negative. The opposite is true for low thresholds, with a lower fraction of correct positive classifications and a higher fraction of correct negatives. The dotted line from the origin illustrates the expected correct classification ratio achieved by random chance. The solid line arching towards the upper left corner is the observed ROC curve (Karimollah, 2013). The further the graph arcs towards the upper left corner the better, as it indicates a better trade-off ratio and a greater discriminant capacity of the model.

When using ROC to compare models, like in this thesis, the numeric metric for the area under the curve (AUC) is often used, where a greater area indicates a model that is better at discriminating the observations at different thresholds (Karimollah, 2013). In this thesis ROC AUC is used to measure the discriminatory capacity of the supervised classification models.

4.4.3 Cohen's Kappa

Cohen's kappa, named after Jacob Cohen, is a metric used to study the reliability of the decision making of two actors (McHugh, 2012). Initially it was used to figure out how much of the coherence between two deciding parties were due to chance, but later it has also been used in model verification, looking at predicted versus actual values (Widmann, 2020). This thesis uses the latter as a part of the model verification. The metric is expressed as a value *k* which is the "proportion of agreement after chance" (Cohen, 1960, p. 40). It is calculated by solving for *k* in Equation 18.

$$k = \frac{p_0 - p_c}{1 - p_c}$$

Equation 18: Cohen's Kappa

Where:

p_0 = The proportion of agreed decisions.

p_c = The proportion of decisions to be expected by chance.

The limits of the calculations make sure $-1 < k < 1$ (even though values $k < 0$ is unlikely in practice) and Cohen proposed the following assessment of the value shown in Table 9 (McHugh, 2012).

k	**Indication of agreement**
≤ 0	None
0.01–0.20	None to slight
0.21–0.40	Fair
0.41– 0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

Table 9: Interpretation of the Cohen's Kappa coefficient

This thesis utilizes Cohen's kappa as a benchmark metric to compare which method is the best at predicting the application outcome after chance. The metric can therefore indicate how good the models are at predicting application outcomes beyond predicting the most common outcome (i.e., always predicting "not rejected"). This enables us to understand how well each method predicts uncommon cases, which is especially important for the analysis as there are relatively few applications in the data set that are rejected. In such cases, a method that never predicts rejection will have a high accuracy and ROC AUC, as the predictions will be correct most of the time. However, such a model would be assigned a Cohen's Kappa score of zero.

4.4.4 Feature Importance

Feature importance regards identifying influential variables in a data set, giving an understanding of what basis the models make their predictions on.

Due to the models used in this thesis having different underlying assumptions, their feature importance metrics cannot be compared using their nominal values. Feature importance for logistic regression is measured in log loss, Random Forest in impurity and XGBoost in gain. Naive Bayes does not have an intrinsic way to measure feature importance, but it can be derived by studying changes in a model metrics through bootstrap resampling. Although the feature importance metrics vary for all the models in this thesis, it is possible to compare which variables are assessed as important across the models by studying the impact the variables have for each of the models.

5 Results

This section looks at the results of the machine learning models that have been implemented. When comparing the models, the metrics presented earlier are used in combination with a qualitative assessment of feature importance, model implementation process and model interpretability. The goal is to identify the superior model for predicting rejections for REK specifically, as well as for application-based case processing in general. The superior model is then used as the basis for discussing what benefits outcome prediction can provide in application-based case management processes.

5.1 Model Performance

All models are trained on a training set containing approximately 80% of the observations, totaling 11,536 applications, and the performance metrics are calculated based on predictions made on a test set containing approximately 20% of the total observations, totaling 2,886 applications. When the data split is performed, the decision variable is used as a stratification variable, meaning that the proportion of rejected applications should be the same in both the training and test set. The test data has been held out of the training phase so that the models could be evaluated on unseen data that has not affected the model training. This is standard practice in model evaluation, where one should never test a model on the same data as it is trained on (Rhys, 2020).

5.1.1 Confusion Matrix

Before looking at the metrics it can be beneficial to look at the confusion matrix of the predictions to get an overview of the nominal model performances. In Table 10 it can be read that out of the 760 rejected applications in the test set, 528 of them are correctly identified as such by XGBoost, while only 151 applications are wrongfully predicted as being rejected. This is indeed a substantial performance, indicating that the model is making informed decisions regarding rejected applications. The other models also show a decent performance in predicting rejections, but Naive Bayes has a relatively high number of false positives.

	Predictions							
	Logistic regression		Naive bayes		Random Forest		XGBoost	
Truth	**False**	**True**	**False**	**True**	**False**	**True**	**False**	**True**
False	2,000	122	1,871	255	2,003	123	1,975	151
True	275	479	284	476	266	494	232	528

Table 10: Confusion matrix for the four models

The confusion matrices provide a decent overview of the model performances. However, to more easily be able to compare the models, the following sections will discuss the models' performance as measured in the metrics introduced in section 4.4, namely the accuracy, ROC AUC and Cohen's Kappa.

5.1.2 Accuracy

Table 11 shows that the highest accuracy is achieved by XGBoost, with 86.7% correctly predicted outcomes. Logistic regression and Random Forest have slightly lower accuracies, while Naive Bayes performs significantly worse, predicting the correct outcome 81.3% of the time.

Logistic regression	**Naive Bayes**	**Random Forest**	**XGBoost**
0.862	0.813	0.865	0.867

Table 11: The accuracy metric for each of the models

These are promising results, but the accuracy metric should not be used blindly. A model that predicts "Not rejected" in every case would achieve an accuracy of 0.737, as this is the proportion of applications that are not rejected. Therefore, the accuracy cannot single-handedly be used to determine how good a model is to predict application rejections.

5.1.3 Cohen's Kappa

The ability to predict uncommon outcomes, however, can be measured with Cohen's Kappa. As seen from Table 12, Naive Bayes, has a "moderate" performance according to the Cohen's Kappa metric, signaling a moderate indication of agreement between the model's predictions and

the case officer assessments. Meanwhile, Logistic regression, Random Forest, and XGBoosts all perform "substantially" well with scores of 0.618, 0.630, and 0.646, respectively. These performances indicate that the models are decent at predicting rejections, which can also be seen from the confusion matrices looking back at Table 10.

Logistic regression	**Naive Bayes**	**Random Forest**	**XGBoost**
0.618 (substantial)	0.513 (moderate)	0.630 (substantial)	0.646 (substantial)

Table 12: The Cohen's Kappa value for each of the models

5.1.4 ROC and ROC AUC

The ROC differs from the Accuracy and Cohen's Kappa metrics, as it measures the classification abilities of the models at various thresholds. Figure 9 displays the ROC curves for each model.

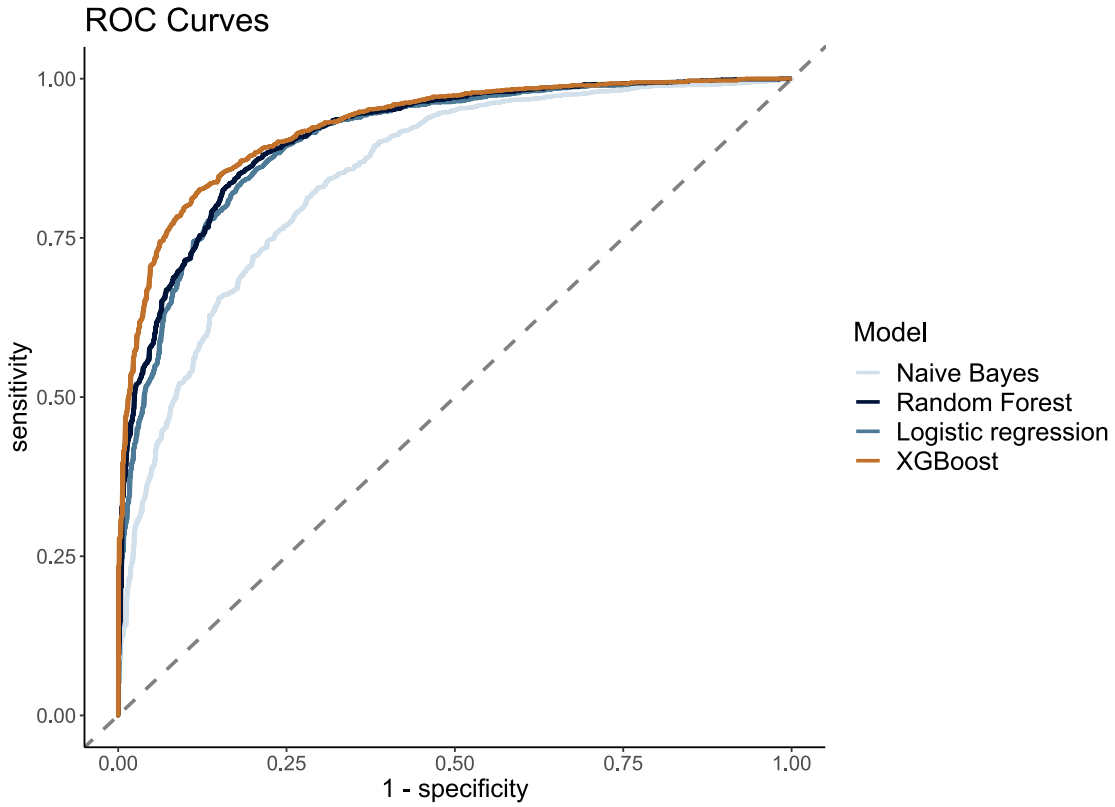

Figure 9: Graph of the ROC curve for the four models

The figure shows the classification abilities for various thresholds of the models, and it can be seen that XGBoost's curve arches most towards the upper left, which would indicate that this is the best model for classifying rejections at various classification thresholds. To be able to compare the models more easily, the area under each of the curves can be used, as this provides a numerical approach to the ROC. These ROC AUCs are presented in Table 13. Again, XGBoost achieves the highest score, logistic regression and Random Forest perform nearly identically, with Naive Bayes performing significantly worse.

Logistic regression	Naive Bayes	Random Forest	XGBoost
0.903	0.847	0.905	0.928

Table 13: The ROC AUC for each of the models

5.1.5 Metrics Summarization

Based on the chosen performance metrics, it seems evident that XGBoost has the best performance. Although there are clear differences in model performances, it should be noted that all models seem to perform well, indicating that the data is well suited for prediction purposes. Table 14 summarizes the performance metrics of the models.

Model	Cohen's Kappa	ROC AUC	Accuracy
Logistic regression	0.618	0.903	0.862
Naive Bayes	0.513	0.847	0.813
Random Forest	0.630	0.905	0.865
XGBoost	0.646	0.928	0.867

Table 14: Overview of the presented metrics for each model

5.2 Feature Importance

Next, the feature importance of the respective models is looked at, studying their interpretations of each variable's contribution to the predictions. It is important to note that the metrics for variable importance vary for each of the respective models, which means that the nominal value of importance for each variable is not comparable across models. However, it is still possible to get an understanding of whether the models identify the same variables as being influential.

Table 15 shows the ten most significant variables for each model, indicating which variables are important in deciding whether applications should be rejected. See Appendix 5 for the relative impact of each of these variables in the various models.

	Logistic regression	**Naive Bayes**	**Random Forest**	**XGBoost**
1.	Applicant organization unknown	Applicant organization	Applicant organization	Applicant organization unknown
2.	Topic 27	Biological material	Project period days	Project period days
3.	Topic 5	Year	Year	Year
4.	Topic 13	Cooperation abroad	Biological material	Biological material
5.	Processing organization REK North	Previous project research	Topic 34	Topic 27
6.	Topic 22	Previously applied	Topic 27	Topic 34
7.	Topic 7	Consent received	Processing organization	Processing organization REK West
8.	Topic 24	pid	Topic 5	Topic 5
9.	Topic 35	Consent received 2	Application title length	Processing organization REK North
10.	Topic 9	Top topic	Topic 22	Topic 22

Table 15: The top ten influential variables for each model

Table 15 shows that the most influential variable for all the models is the applicant organization, with overwhelming evidence that "unknown" applicant organizations are highly influential. Note that the requirement of converting categorical variables to numerical dummy variables makes it slightly easier to interpret the categorical variables for XGBoost and logistic regression compared to Random Forest and Naive Bayes, as XGBoost and logistic regression specifies that "Unknown" is the specific factor level that is most influential. Figure 10 offers a closer look at this variable and shows that there is indeed significant deviance in application outcomes for the various applicant organizations. Most notably, if the applicant organization is unknown, more than 75% of the applications are rejected.

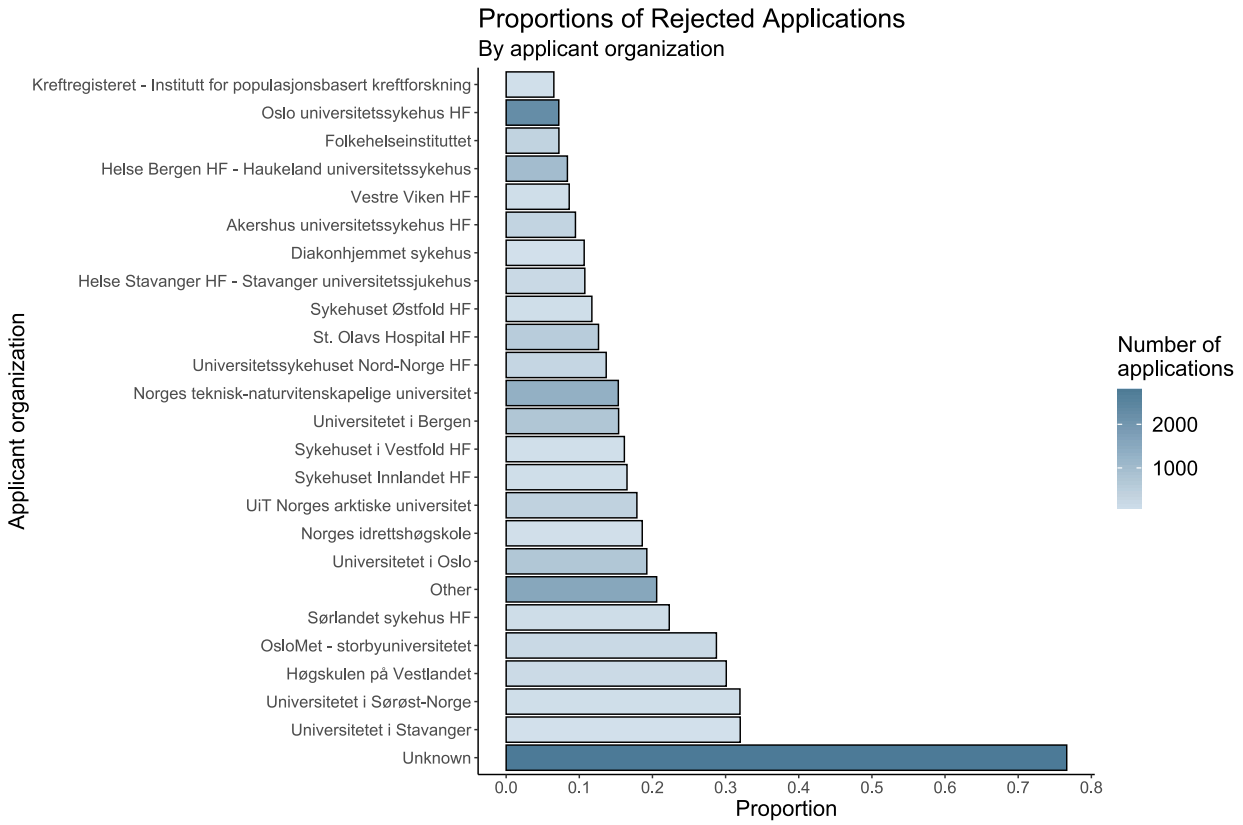

Figure 10: Applicant organizations and the proportion of rejected applications

According to REK, applications should only be rejected if REK does not consider the research project to fall under The Health Research Act. Thus, the applicant organization should in theory not affect the processing outcome. Yet, the data shown in Figure 10 indicates something different, signaling that the case officers might have degrees of bias towards some of the institutions. However, it may also be that Figure 10 merely signifies how good the various institutions are at producing applications. For instance, it might be that Oslo Universitetssykehus AF writes better or more relevant applications than Universitetet i Stavanger. Accordingly, further investigations into this potential bias might be sensible for REK to conduct.

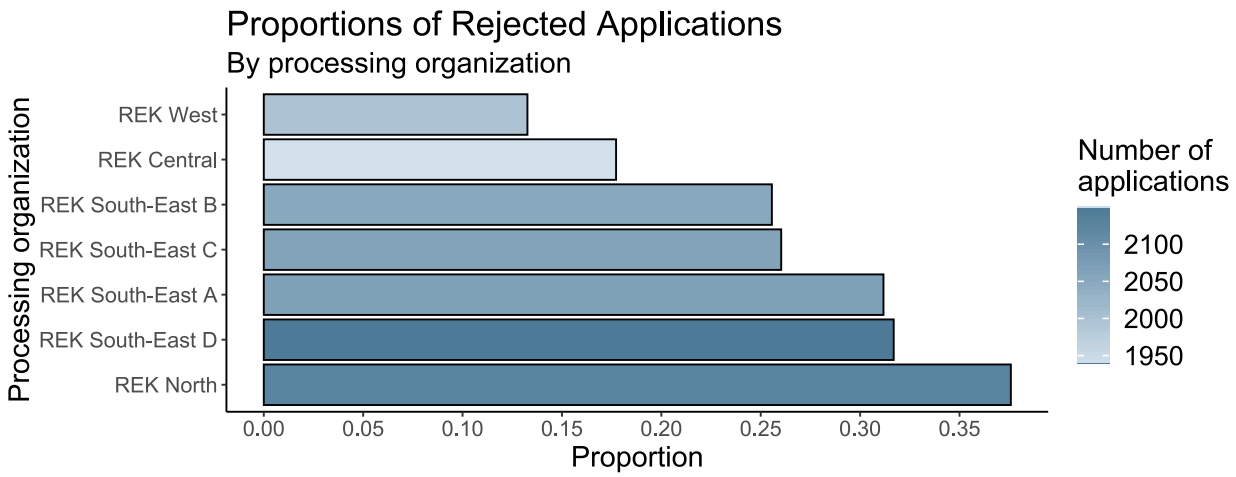

Figure 11: Proportion of rejected applications for each processing organization

Figure 11 shows that, like the applicant organization, there are large variations in the proportion of rejected applications for the processing organizations. 37.6% of the applications processed by REK North are rejected, which is almost three times as many as applications processed by REK West, where 13.3% are rejected. Considering the principle of equal treatment which is an important measure in case management, this seems like something that REK should investigate further, however such a discussion is outside the scope of this thesis.

The biological studies variable is also shown to be influential, being the second most important variable for Naive Bayes and fourth most important variable for the tree-based models. From Table 16, a clear correlation between this variable and actual rejected applications is demonstrated. Only 6% of the biological material studies are rejected, while 34% of non-biological material studies are rejected.

Biological material study	**FALSE**	**TRUE**
Proportion rejected	34%	6%

Table 16: Proportion of rejected applications for biological material studies

A correlation like this seems reasonable as applications sent to REK should be about "health-related research projects conducted on human beings, human biological material, or personal health data" (Regional Committees for Medical and Health Research Ethics, 2022). If this is not the case, the chance of rejection is naturally higher.

Another interesting finding is that the topic variables are shown to be influential for all models. Plotting the topic variables against the proportion of rejected applications, as in Figure 12, clear correlations can be seen, and it is for instance evident that topic 27 and topic 34 are associated with rejected applications, while topics 36 and 32 are associated with non-rejected applications.

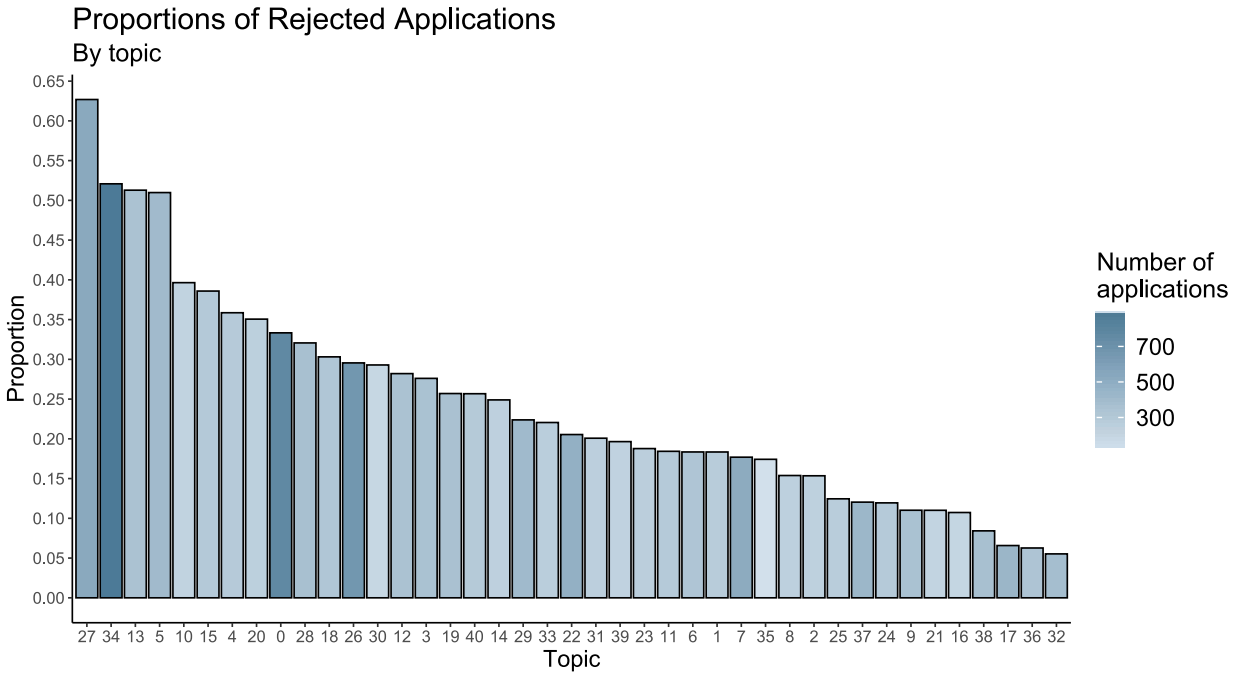

Figure 12: Proportion of rejected applications for each topic

By doing a qualitative study of the topics, an understanding of what types of project descriptions lead to rejections can be gained. Table 17 shows the top ten words for two of the topics most associated with rejections (27 and 34) and two of the topics least associated with rejections (32 and 9). The table shows that topic 27 is related to some issues that municipalities face, with words such as “municipality”, “health service” and “suicide”. Topic 34 captures descriptions of project type and study type, with some of the most frequently used words being “qualitative”, “parent”, “family” and “health personnel”. Both topics have in common that they do not contain medical terminology but rather describe aspects surrounding a project. The topics related to non-rejections on the other hand contain more medical terminology. Topic 32 is related to biochemistry, with words such as “cell”, “blood”, “cancer cell” and “stem cell”, and topic 9 captures research related to pregnancy and birth, with frequently used words being “pregnancy”,

"mother", "birth", "foster" and "newborn". The words in Table 17 are translated to English, see Appendix 6, Table 24, for the original words in Norwegian.

Topic 27	**Topic 34**	**Topic 32**	**Topic 9**
municipality	qualitative	cell	pregnancy
health service	parent	blood	mother
municipal	mentally	protein	birth
specialist health service	family	antibody	pregnant
practice	health personnel	cancer cell	giving birth
correct	semi-structuring	isolate	pregnancy
interaction	social	tissue	moba[2]
health personnel	quantitative	stem cell	newborn
qualitative	individual	human	fetus
suicide	relatives	body	premature

Table 17: The 10 most influential words in the four topics 27, 34, 31 and 9

From these findings, it seems like applications where the project descriptions regard specific subjects within the field of medicine are unlikely to be rejected, while applications with project descriptions that do not describe a specific medical field, but rather the surrounding factors of a project, are more often rejected.

Furthermore, the variable for project period is identified as influential by the tree-based models. Shorter projects are more often rejected than longer projects, as shown in Figure 13. Interestingly, the most common project length that is applied for is one year, and almost 50% of these projects are rejected.

[2] MoBa refers to the Norwegian Mother, Father and Child Cohort Study: https://www.fhi.no/en/studies/moba/

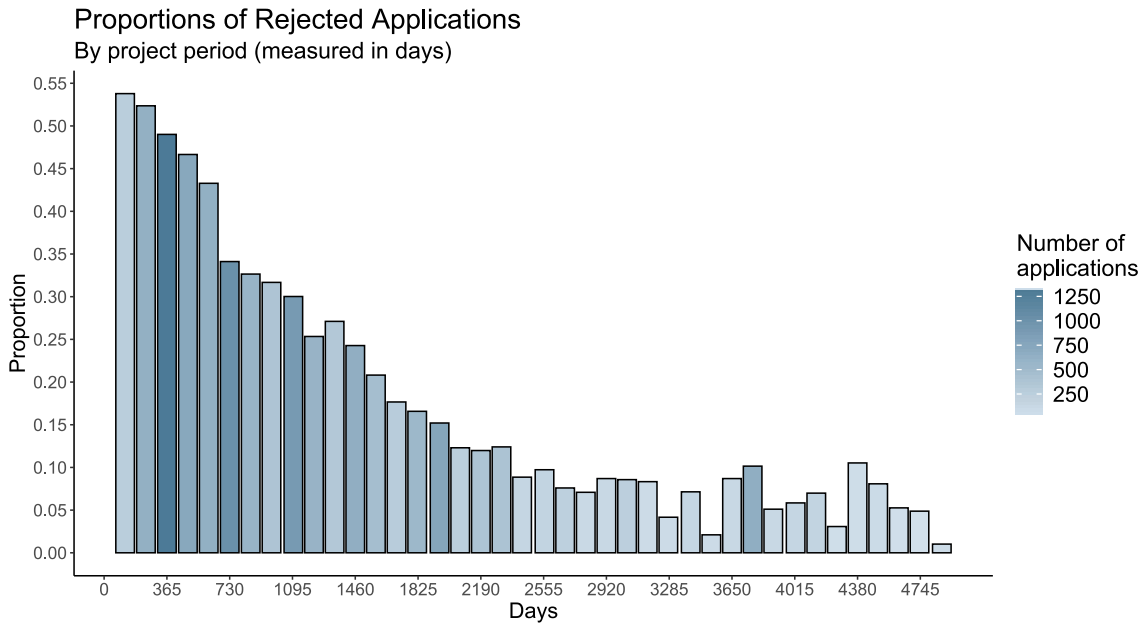

Figure 13: The number of days a project lasts versus the proportion of rejections

In sum, all four models identify similar variables as being influential when predicting rejections for applications that are sent to REK. From the most influential variables it seems evident that, all else held equal, applications are more likely to be rejected if:

1. The applicant organization is unknown
2. The project is not a study on human biological material
3. The project description is not about a specific subject within the field of medicine
4. The project period is short
5. The processing organization is REK North

When interpreting these results, one should keep in mind that about one third of the data set was filtered out from the data set initially due to missing variables, most importantly missing decision variables. Therefore, there is a risk regarding whether there exist relationships between the which decisions have been made and whether the decisions are present in the data set. For instance, there might have been procedures in REK for registration of decision outcomes of applications that are approved, but not for applications that are rejected. A systematic flaw like this would mean that there could be a high degree of bias in the data set which in turn would be represented in the results of this thesis. The authors of this thesis do not believe that such a systematic flaw in the data is present, as it was shown in section 3.3 that the missing variables rather are correlated

with the year applied. However, caution regarding the data quality in this case should be considered before a real-world implementation of the methodology, and it is also advised that REK analyze why there is missing data, especially missing decision variables, in their data set.

5.2.1 Model Comparison

Summarizing the results, it is evident that XGBoost has the best performance metric values. Although it is reasonable to consider that logistic regression is vastly less computational complex, and that its variable importance provides useful insights because it also includes whether the influence for each variable is positive or negative as opposed to only absolute influence. XGBoost is simple to implement but is computationally demanding due to the number of hyperparameters that require tuning. However, given the circumstances it can be concluded that XGBoost is the overall best model out of the four.

5.3 Predicting Case Officers' Assessments

As XGBoost is concluded to be the best performing model, the predictions of this model are used for the rest of this thesis.

XGBoost	**Predictions**	
Truth	False	True
False	1,975	151
True	232	528

Table 18: Confusion matrix of XGBoost's predictions

Table 18 shows that close to all non-rejected applications are correctly classified as not being rejected by XGBoost, at 92.9%. For rejected applications, 69.5% are correctly classified as such, which also indicates a robust model. However, as the purpose of this thesis is predicting rejected applications, the ability to correctly classify a substantial proportion of the rejected applications is more important than correctly classifying non-rejections. At the same time, there is a tradeoff between correctly identifying as many rejections as possible and making as few mistakes as possible regarding the predicted rejections. That is, if the threshold for predicting rejections is lowered, the model will correctly identify more of the actual rejected applications, but at the same time incorrectly classify more non-rejections. By adjusting the threshold of the model, it is

possible to influence how strict the model is in classifying applications and decide what the threshold should be based on this tradeoff. This is one of the subjects for discussion in the next section.

6 Discussion

Now that the results of implementing the machine learning methods on the data have been presented, potential use cases and applicability of the methodology is discussed. Furthermore, some reflections are made regarding weaknesses of the analysis as well as some important ethical considerations.

6.1 Use Cases of the Methodology

The Norwegian Ministry of Local Government and Regional Development proclaim in their strategy that case management in the public sector is highly driven by rule sets, with elements of discretionary assessment from the case officers (2020, p. 26). Consequently, they present the notion that the processes do not necessarily need to be either fully manual or automatic, but that a mix of methods can be used, where the manual handling of anomalies is used as an example (Norwegian Ministry of Local Government and Regional Development 2020, p. 26). This section of the thesis discusses the possibility of using the prediction methods to detect applications that are likely to be rejected, tagging them with a warning flag for the case officers and potentially rejecting them automatically.

6.2.1 Flagging of Applications

To present applications to the case officer that the model identifies as likely to be rejected, a flagging system is proposed. The system assigns a flag to applications that are above a certain threshold of probability of being rejected. Doing this would help the case officers identify such applications, potentially reducing the time spent in the initial assessment of applications.

To be able to flag applications it is necessary to look at the results and decide a threshold for when an application is flagged. The threshold should be decided based on the organization's tolerance of correctly versus wrongly flagged applications, which is measured in number of true and false positives, respectively. A low threshold can be beneficial for flagging as many of the rejected applications as possible, which gives the case officers certainty that most of the applications to be rejected are in the category of flagged applications. Inversely, a high threshold can be beneficial for achieving a high proportion of correctly flagged applications relative to

falsely flagged applications, leading to a high certainty for the case officers that the flagged applications in reality should be rejected. Thus, the trade-off regarding thresholds and flagging is finding the balance of flagging most of the rejected applications while simultaneously making as few errors as possible.

By plotting the true and false positive rate against various thresholds, as done in Figure 14, the tradeoff can more easily be studied. The horizontal line in the figure represents a criterion of 5% false positives, which is achieved at a threshold of 58.6%.

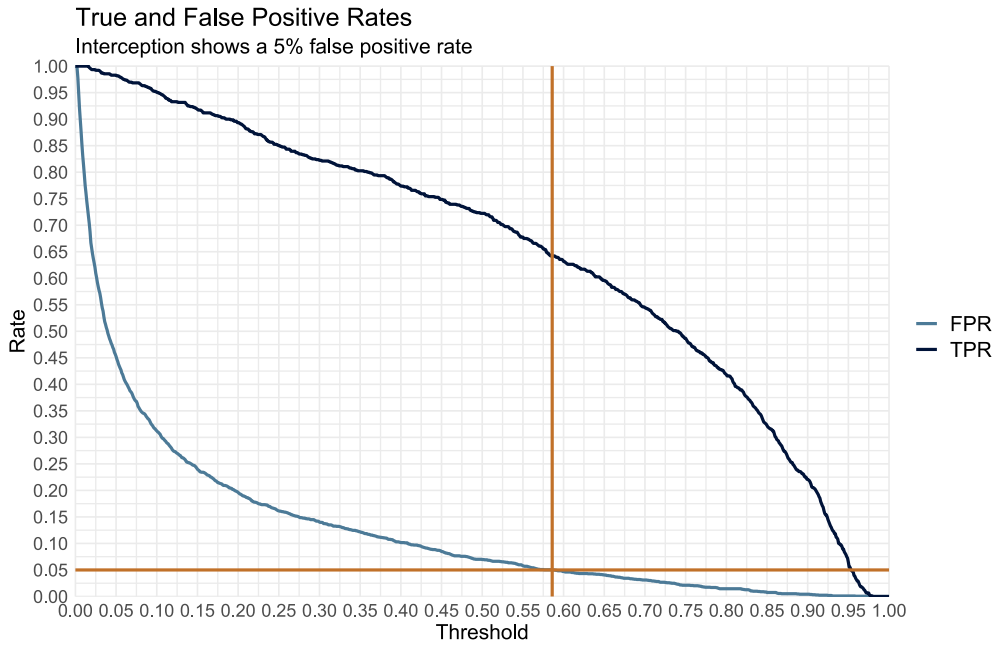

Figure 14: False positive rate criteria of 5%

This threshold results in a true positive rate of 64.2% meaning that when 5% of the flagged applications are non-rejections, 64% of all the actual rejected applications are correctly flagged. Table 19 displays the confusion matrix for this scenario, showing that this leads to 594 flagged applications, where 488 of these are actual rejections. Out of the rejected applications, 272 are not flagged in this case.

5% FPR		Predicted	
		False	**True**
Actual	**False**	2020	106
	True	272	488

Table 19: Confusion matrix of the XGBoost predictions at a 5% FPR criteria

To identify a larger proportion of the rejected applications, the threshold would have to be lowered, which would also lead to a scenario where more applications would be wrongfully flagged. For instance, if it is desirable that at least 90% of the rejected applications are flagged, corresponding to a 90% true positive rate, it can be seen in Figure 15 that a threshold of 18.4% should instead be used. Although this would mean that more of the rejected applications are correctly identified as such, it would also lead to 445 wrongfully flagged applications. By further increasing the threshold, one could imagine that REK could avoid the initial assessment of the non-flagged applications, as almost all applications that are not satisfactory would be flagged.

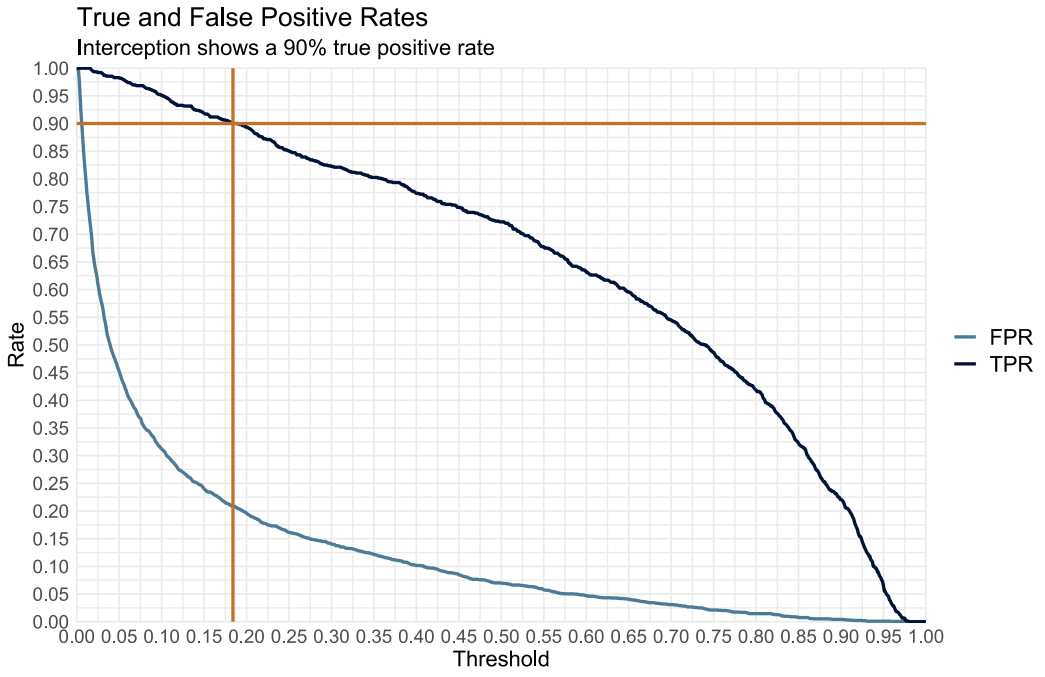

Figure 15: True positive rate criteria of 90%

The weakness of flagging applications is however the chance of creating a false support for the case officers, where the flags are used as reasoning for the decisions rather than guides. A

flagged application is an application that is likely to be rejected due to similarity to previously rejected applications and not a direct recommendation to reject. Hence, it is important to make sure that the involved parties who are presented with the potential flags are aware of their meaning and use case before an implementation.

6.2.2 Automatic Rejection of Applications

In line with the argument of automation from the Norwegian Ministry of Local Government and Regional Development (2020, p. 21) AI could potentially open the opportunity of automatic rejection of applications. In such a case the applications would be rejected without involvement from case officers, even furthering the reduction of their workload. Etscheid argues that "[in sensitive areas] where administrative decisions have a direct influence on people's livelihoods, systems should only be used when the error rate is acceptable." (2019, Conclusion, para. 4). For this use case, the false positive rate must be particularly low, where the rejection threshold should be set based on the processing organization's own judgment.

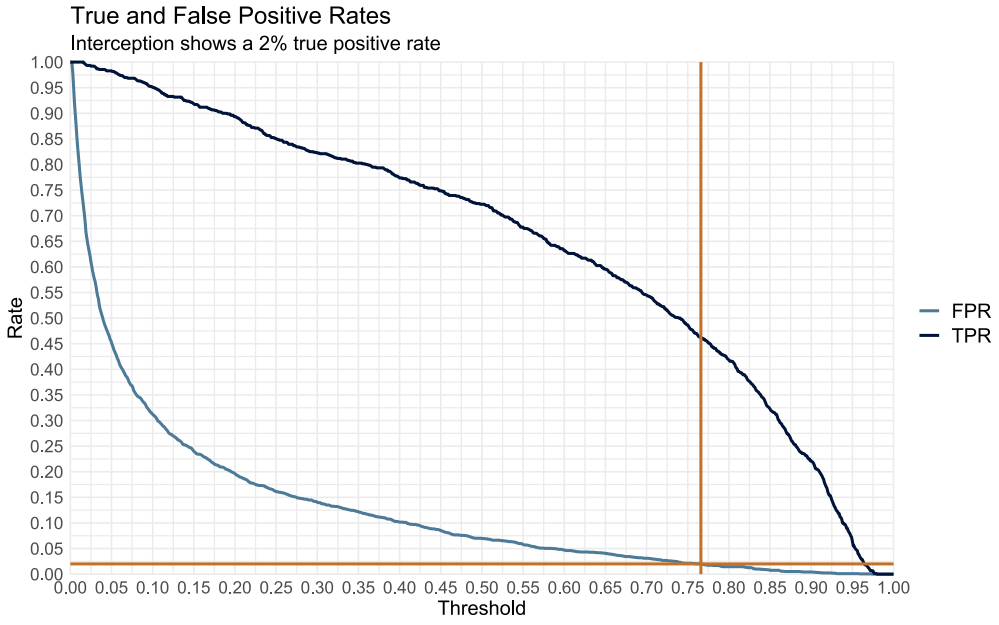

Figure 16: False positive rate criteria of 2%

For instance, Figure 16 shows that if a criterion of only allowing a two percent false positive rate is used, about half of the rejected applications are still identified correctly. Table 20 displays the confusion matrix of using a threshold of 76.6%, resulting in 351 of 760 true positives.

		Predicted			
		2% FPR		0.5% FPR	
		False	**True**	**False**	**True**
Actual	**False**	2084	42	2116	10
	True	409	351	579	181

Table 20: Confusion matrices of the XGBoost predictions at 2% and 0.5% FPR criteria

At an even stricter false positive criteria of 0.5%, 181 applications are automatically rejected, where 171 of these are correct rejections. This illustrates that the organization could control their own tolerance of false negatives if they desire to automatically reject the most acute cases.

When discussing the subject of automatically rejecting an application, it should be noted that the consequence of an application being rejected is that the applicant does not receive an approval from REK to conduct the research for which they have applied. An important question is then what to communicate to the applicant in the case of an automatic rejection. Should the applicant of the automatically rejected application be told that the application does not need REK's approval, or simply that REK has not processed the application? If the true positive rate is close to 100% and more importantly the false positive rate is close to zero, the first option of declaring that the project is not regulated by REK, can be defended. In the instance where this is not the case, communicating that the application has not been processed due to high likelihood of it being rejected, can be considered. In the case that the predictions are so poor that the false rejections cause more work for the case officers than it reduces, then automatic application rejections should be avoided altogether.

6.2.3 Summary of the Use cases

The results prove that, although there are some barriers and potential issues, there are definite possibilities with further model tuning for creating a system for automatic rejection. In addition,

it seems evident that flagging applications that are suspected by the model of being rejected could be highly beneficial for case officers in quickly identifying unsatisfactory applications. In a real-world scenario, it would be reasonable to include case officers and leaders in REK to decide a satisfactory false positive rate criteria for both approaches.

6.2 Limitations and Further Research

From the analysis and discussion there seems to be a significant ability and use case in predicting whether an application will be rejected using machine learning. Yet, there are several reasons for caution regarding the findings which should be highlighted before the conclusion.

One weakness of the results originates from the various model assumptions that do not fit with the realities of the data set. For instance, both logistic regression and Naive Bayes assumes that the predictor variables are uncorrelated, which is not the case for the data set in this thesis, for instance because the topic variables by nature are correlated due to the sum of the probabilities being 1. In addition, the models do not consider the fact that case officers' judgements might be influenced by other applications they are processing simultaneously. To increase the performance of these models, one could first try to apply decorrelation techniques on the variables. In this thesis, backwards stepwise variable selection is performed to reduce model dimensionality for Logistic regression. This partly deals with the correlation issue, but other decorrelation techniques such as principal component analysis would be highly relevant for further studies. However, XGBoost is well suited for dealing with correlated variables due to the nature of the decision tree splits. The splits are done sequentially, meaning that variables are in fact assumed to be correlated, i.e., each split is affected by previous splits. Therefore, the weakness regarding the correlation assumptions does not significantly impact the final results of the thesis, as these results are based on the XGBoost model. Consequently, the weakness is the missed potential of Logistic regression and Naive Bayes not being utilized to their fullest extent because of the correlation between the variables. From the results Logistic regression comes close to XGBoost and a further study decorrelating the variables might consequently affect Logistic regressions prediction capacity, making it superior.

Apart from the model's inability to perfectly interpret the data, it is also important to understand that the response variable is a result of human decision making, which introduces the risk of the data being biased. For instance, two case officers might not make the same decision given the exact same case. Such cognitive bias naturally disturbs the results of the models, as the predictions might be correct for a certain case officer's assessment and at the same time could be wrong for another case officer's assessment. There is a need for caution regarding this, as it has been shown that the processing organizations reject vastly different proportions of applications. This also relates to the issue of the missing decision variables, where there is a potential risk that the missing decision variables in the data set is related to a specific decision outcome.

Another reason for caution is due to the nature of the LDA topic model, which assumes that all the documents in a corpus to some degree contain all topics. If a new application is sent in that does not contain any of the predefined topics, it will still be assigned topics falsely and will act as noise in the topic distribution. Here it would be beneficial to detect that the application does not belong to any of the predefined topics and filter it out before it is of a nuisance. Such a filter should be taken into consideration for future research. However, this could also be solved through training new topic models often or training topic models only on the most recent applications. Furthermore, the topic model is not optimized for prediction purposes. The number of topics, k, is decided based on performance metrics that are used to assess the quality of a topic model in addition to a qualitative assessment of the semantics captured by the topics. By tuning the number of topics based on the performance metrics of the supervised prediction methods, the prediction results would be expected to increase.

Regarding future applications there is also another weakness in the methodology. The current methods lack the ability to consider changing administrative guidelines or political policy. In institutions such as REK, which has a public mandate, there is always a possibility that changes in the political climate might affect how they work and operate. An example of this is the Covid-19 pandemic, where studies regarding a covid vaccine were approved much more hastily than they normally would (Rahman & Islam, 2021). Updating the models regularly will reduce the impact of this weakness, but a prediction method solely based on application processing up to a point, will not sufficiently be able to adapt to rapid changes and will produce inaccurate results.

Any future developments of the methodology could therefore benefit from further studying how the methods can be developed in a way that such occurrences can be handled.

6.3 Ethical Considerations

In creating a decision support system, one should also do some ethical considerations before a potential recommendation of the implementation of the methodology. Implementing machine learning techniques in a real-world environment may have unforeseen or unwanted consequences (Dhar, 2016, p. 6) and some of these are hence discussed.

The first of such being whether all variables are ethical to use. Dhar makes the case of problematic AI inference regarding someone's race, gender and political beliefs and states that implementing AI "requires addressing the issue of 'ethics of inference' associated with systems that use data that may not be congruent with the best interests of individuals" (2016, p. 6). In the case of this thesis, some of the metadata variables should be put under such scrutiny. More specifically, the *applicant organization* and *processing organization* which turned out to be the most impactful variables for several of the models, might be problematic.

One might intuitively think that the organizations that are involved should not be of importance in ethically conducted application processing and the outcome should be based on the application content and decided by a uniform ruleset. The conclusion of such an argument would be to remove the variable of *applicant organization* from the modeling. On the other hand, it is possible to argue that an applicant organization's trust and credibility may affect a case officer's assessment, although this does not seem to conform with the rejection criteria that REK follows. Before concluding whether the variables should be included in the modeling or not, a further look into what the goal of the case management is should be conducted. For the variable *processing organization, s*uch a case is harder to defend. The laws regulating each of the REK branches are the same, indicating that they should conduct a uniform processing of the applications. It is hence easier to argue that this variable should not be used for the modeling and removed.

From Table 21 it is possible to derive that removing the variables under critique lowers the benchmarking metrics quite significantly. Before implementing this, it should be discovered why this is the case and to what extent removing the variables will affect the potential use cases of the models. It can also be of interest to collaborate with REK, to try to find ways to improve this aspect of the application processing.

Model	Cohen's Kappa	ROC AUC	Accuracy
Logistic regression	0.399	0.827	0.796
Naive Bayes	0.108	0.768	0.760
Random Forest	0.388	0.839	0.799
XGBoost	0.447	0.860	0.808

Table 21: The benchmark metrics after removing applicant and processing organization

Besides the question of whether a variable should be used in the model, there is also the important aspect of what the effect of it being there is. It is easy to imagine that a model using processing organization as a variable will experience a feedback loop, where the application is recommended “rejected” due to the processing organization variable and consequently is rejected. The rejection will then strengthen the relationship in the model between the processing organization and the chance of rejection, deepening the issue.

7 Conclusion

This thesis has investigated the potential of using machine learning to predict decision-making in an application-based case management process. Specifically, it studied whether supervised prediction methods could predict if medical research project applications sent to REK will be rejected or not. The study results are promising, showing that all the models can predict rejections well. The best-performing model is XGBoost, based on its performance metrics and simple implementation procedure. The results also show that using LDA topic modeling to construct structured features from text is valuable to the models as many topic variables are highly influential for the predictions. The thesis shows that for the practical use cases of machine learning, flagging applications predicted as having a high probability of being rejected is plausible. Subsequently, automatic rejections based on the predicted probabilities seems realistic but requires further clarifications with REK regarding potential consequences before real-world implementation.

Although the results are promising, it is necessary to make some critical considerations before making practical use of the methods. Most notably, the methods are not well suited to adapt to sudden changes in the environment, and following the results without caution makes the model prone to self-fulfilling feedback loops. Also, some of the influential variables in the models are metadata variables that one could argue should not be significant to a case officer's decision. Removing such variables reduces some of the models' predictive abilities. Therefore, it is important to consider the consequences and ethical implications of keeping or removing such variables before the practical use of the methods.

Looking at the thesis altogether, it seems evident that there is exciting potential for machine learning in applications-based case management processing. XGBoost proves to have tremendous predictive power in predicting rejections, and LDA topic modeling helps extract valuable insights from the project description. These insights enable a better understanding of the data and improve the predictions made by the supervised models.

References

Benoit, K., Muhr, D., & Watanabe, K. (2021). *Stopwords: Multilingual Stopword Lists*. CRAN. https://CRAN.R-project.org/package=stopwords

Berrar, D. (2019). Bayes' theorem and naive Bayes classifier. In *Encyclopedia of Bioinformatics and Computational Biology* (pp. 403–412). Elsevier. http://dx.doi.org/10.1016/b978-0-12-809633-8.20473-1

Biau, & Scornet. (2016). A random forest guided tour. *TEST*, *25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(1), 993–1022. https://doi.org/https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf?ref=https://githubhelp.com

Breiman. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Broomfield, H., & Reutter, L. M. (2019). Kunstig intelligens/data science: En kartlegging av status, utfordringer og behov i norsk offentlig sektor - første resultater. *Institutt for Sosiologi Og Statsvitenskap*. Publikasjoner fra CRIStin - NTNU. http://hdl.handle.net/11250/2634733

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, *72*(7–9), 1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011

Chen, T., & Guestrin, C. (2016, August 13). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. http://dx.doi.org/10.1145/2939672.2939785

Chu, P. C., & Spires, E. E. (2001). Does time constraint on users negate the efficacy of decision support systems? *Organizational Behavior and Human Decision Processes*, *85*(2), 226–249. https://doi.org/10.1006/obhd.2000.2940

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

D'Agostini, G. (1995). A multidimensional unfolding method based on Bayes' theorem. *Nuclear*

Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, *362*(2–3), 487–498. https://doi.org/10.1016/0168-9002(95)00274-x

Dhar, V. (2016). The Future of Artificial Intelligence. *Big Data*, *4*(1). https://doi.org/10.1089/big.2016.29004.vda5

Efron, B. (2013). Bayes' theorem in the 21st century. *Science*, *340*(6137), 1177–1178. https://doi.org/10.1126/science.1236536

Esposito, C., Landrum, G. A., Schneider, N., Stiefl, N., & Riniker, S. (2021). GHOST: Adjusting the decision threshold to handle imbalanced data in machine learning. *Journal of Chemical Information and Modeling*, *61*(6), 2623–2640. https://doi.org/10.1021/acs.jcim.1c00160

Etscheid, J. (2019). Artificial Intelligence in Public Administration. *18th International Conference on Electronic Government (EGOV)*, 248–261. https://hal.inria.fr/hal-02445801/document

Fradkov, A. L. (2020). Early history of machine learning. *IFAC-PapersOnLine*, *53*(2), 1385–1390. https://doi.org/10.1016/j.ifacol.2020.12.1888

Geletta, S., Follett, L., & Laugerman, M. (2019). Latent Dirichlet Allocation in predicting clinical trial terminations. *BMC Medical Informatics and Decision Making*, *19*(1). https://doi.org/10.1186/s12911-019-0973-y

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(suppl_1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Hardman, D., & Macchi, L. (2004). *Thinking: Psychological perspectives on reasoning, judgment and decision making*. John Wiley & Sons.

Hornik, K., Mair, P., Rauch, J., Geiger, W., Buchta, C., & Feinerer, I. (2013). ThetextcatPackage forn-Gram Based Text Categorization inR. *Journal of Statistical Software*, *52*(6). https://doi.org/10.18637/jss.v052.i06

Hosmer, D. W., Jr., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

Hubl, M., & Merkert, J. (2015, June). A survey of the application of machine learning in decision support systems. *AIS Electronic Library (AISeL)*. European Conference on

Information Systems 2015, Münster, Germany. http://aisel.aisnet.org/ecis2015_cr/133

Ihaka, R. (1998). *R : Past and Future History*. Statistics Department, The University of Auckland . https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.331.299&rep=rep1&type=pdf

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. Springer Science & Business Media. https://hastie.su.domains/ISLR2/ISLRv2_website.pdf

Karimollah, H.-T. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian J Intern Med.*, *4*(2), 627–634.

Liu, S. (2019, January 11). Dirichlet distribution. *Towards Data Science*. https://towardsdatascience.com/dirichlet-distribution-a82ab942a879

Maalouf, M. (2011). Logistic regression in data analysis: An overview. *International Journal of Data Analysis Techniques and Strategies*, *3*(3), 281. https://doi.org/10.1504/ijdats.2011.041335

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, *5*(9), 1315–1316. https://doi.org/10.1097/jto.0b013e3181ec173d

May, C., Cotterell, R., & Van Durme, B. (2019). *An Analysis of Lemmatization on Topic Models of Morphologically Rich Language*. arxiv. https://doi.org/10.48550/arXiv.1608.03995

McEnery, T., McEnery, A., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. Taylor & Francis.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282. https://doi.org/10.11613/bm.2012.031

Ministry of Local Government and Regional Development. (2020). Nasjonal strategi for kunstig intelligens. In *Regjeringen.no* (p. 1-1). Ministry of Local Government and Regional Development. https://www.regjeringen.no/no/dokumenter/nasjonal-strategi-for-kunstig-intelligens/id2685594/

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic Evaluation of Topic Coherence. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, *1*(2010). https://doi.org/https://aclanthology.org/N10-1012.pdf

Rahman, M. A., & Islam, M. S. (2021). Early approval of COVID-19 vaccines: Pros and cons. *Human Vaccines & Immunotherapeutics*, *17*(10), 3288–3296. https://doi.org/10.1080/21645515.2021.1944742

Regional Committees for Medical and Health Research Ethics. (2022). *Regional committees for medical and health research ethics*. Forskningsetikk. https://www.forskningsetikk.no/en/about-us/our-committees-and-commission/rek/

REK-Portalen. (2022, May 26). *About applying to REK*. https://rekportalen.no/#hjem/søke_REK

Rhys, H. I. (2020). *Machine Learning with R, the tidyverse, and mlr*. Manning Publications. https://livebook.manning.com/book/machine-learning-for-mortals-mere-and-otherwise/about-this-book/

Rish, I. (2001). An empirical study of the naive Bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, *3*(22), 41–46.

RStudio, PBC. (2022). *About RStudio*. RStudio. https://www.rstudio.com/about/

Schofield, A., Magnusson, M., & Mimno, D. (2017). Pulling Out the Stops: Rethinking stopword removal for topic models. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. http://dx.doi.org/10.18653/v1/e17-2069

Sessions, V., & Valtorta, M. (2006, January). The Effects of Data Quality on Machine Learning Algorithms. *International Conference on Information Quality*. Proceedings of the 11th International Conference on Information Quality, MIT, Cambridge, MA, USA. https://www.researchgate.net/profile/Marco-Valtorta-2/publication/220918649_The_Effects_of_Data_Quality_on_Machine_Learning_Algorithms/links/09e4150a506b15472b000000/The-Effects-of-Data-Quality-on-Machine-Learning-Algorithms.pdf

Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach*. "O'Reilly Media, Inc."

Slof, D., Frasincar, F., & Matsiiako, V. (2021). A competing risks model based on latent Dirichlet Allocation for predicting churn reasons. *Decision Support Systems*, *146*(2021), 113541. https://doi.org/10.1016/j.dss.2021.113541

Snipes, M., & Taylor, D. C. (2014). Model selection and Akaike Information Criteria: An example from wine ratings and prices. *Wine Economics and Policy*, *3*(1), 3–9. https://doi.org/10.1016/j.wep.2014.03.001

Taheri, S., & Mammadov, M. (2013). Learning the naive Bayes classifier with optimization models. *International Journal of Applied Mathematics and Computer Science*, *23*(4), 787–795. https://doi.org/doi:10.2478/amcs-2013-0059

United Nations Statistics Division. (2022). *UNSD — methodology*. Unstats; UN. https://unstats.un.org/unsd/methodology/m49/

Vayansky, I., & Kumar, S. A. P. (2020). A review of topic modeling methods. *Information Systems*, *94*, 101582. https://doi.org/10.1016/j.is.2020.101582

Weber, L. M., Saelens, W., Cannoodt, R., Soneson, C., Hapfelmeier, A., Gardner, P. P., Boulesteix, A.-L., Saeys, Y., & Robinson, M. D. (2019). Essential guidelines for computational method benchmarking. *Genome Biology*, *20*(1). https://doi.org/10.1186/s13059-019-1738-8

Wickham, H., & Grolemund, G. (2016). *R for data science*. O'Reilly.

Widmann, M. (2020, August 4). *Cohen's Kappa: What it is, when to use it, and how to avoid its pitfalls*. The New Stack. https://thenewstack.io/cohens-kappa-what-it-is-when-to-use-it-and-how-to-avoid-its-pitfalls/

Wijffels, J. (2021). *udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the "UDPipe" "NLP" Toolkit.* CRAN. https://CRAN.R-project.org/package=udpipe

Wijffels, J. (2022). *Package 'udpipe.'* CRAN. https://cran.r-project.org/web/packages/udpipe/udpipe.pdf

XGBoost Documentation — xgboost 1.6.0 documentation. (2021). XGBoost Documentation. https://xgboost.readthedocs.io/en/stable/

A Appendix

A.1: Language Labeling

To be able to label each application with what language they are written in, three different methods for language recognition are implemented. Each method results in a sub-label of what language the method believes the application is written in. After the three methods have had their say, the majority vote decides what language the application is labeled as e.g. if two methods label an application as English and one labels it as Norwegian, the final label on the application will be English. This way of labeling the applications is proven effective, giving better results than using any of the methods alone. The three language sub-labels are described in the following paragraphs.

A.1.1 Language Variable

The first of the three methods used is extracting the language variable from the application form. In the form there is a question asking if the application is written in Norwegian formulated in a binary manner. Looking at data it becomes clear that this variable is a good indicator of the language of the application, but not a foolproof one, as some applications written in English are marked as Norwegian and vice versa.

A.1.2 Language Detection Function

The second method used is the language recognizing function “textcat” first published and documented by Hornik et al. in 2013. This function labels each of the applications based on the content of the project description variable. The function is quite successful at labeling the applications, especially the English ones. The Norwegian applications were sometimes labeled as Danish, so they were manually labeled as Norwegian.

A.1.3 LDA Topic Modeling

The last method used for labeling the applications is using a LDA topic model with only two topics. When creating topic models before filtering out one language, it was discovered that the language that the application is written in heavily guides how the topics are distributed. This

resulted in binary topic models, where the topics consisted of either mostly English or Norwegian words, and the LDA language labels were set accordingly.

A.2: Inverse Document Frequency (IDF)

Inverse document frequency is a value used to identify words that are important to a document in the context of the corpus it is in *(Silge & Robinson, 2017, pp. 31-33)*.

The inverse document frequency is calculated using Equation 19.

$$IDF(term) = \ln\left(\frac{n_{documents}}{n_{documents\ containing\ term}}\right)$$

Equation 19: IDF

From the formula, one can derive that terms that appear in fewer documents overall get a higher value, and that words that appear in many documents are assigned a lower value.

In this thesis, the IDF value is used for the preprocessing of the project description, where it is used to identify words that might be noise or have little to no value for the topic model. To do this, the IDF score is calculated for each word in the corpus. Then the 100 words with the highest IDF score are removed, as these words are present in many of the documents and might therefore be overrepresented in the many of the topics. After that, the words with a score indicating that the word is only present in one document is removed, due to the nature of how LDA topic modeling works, where these words do not help assigning document topics. See Appendix 3 for the list of stop words resulting from this IDF stop word approach.

A.3: Stop-words

Four different lists of stop words are used:

1) The first set of stop words are the Norwegian words from the package "Stopwords", refer to Benoit et al. (2021) for the full list.

2) The second set of stop words are based on words with an IDF score of 9.59, which are words that are only present in one application. This amounts to a total of 57,626 words, making the list too extensive to display.

$$IDF(one\ document\ words) = \ln\left(\frac{14422}{1}\right) = 9.59$$

Equation 20: IDF of one-document words

3) The third set of stop words are based on the one hundred words with the highest IDF score. The words are displayed in Table 22 below, ranking the top 100 words according to their respective IDF scores. The top word, "patient" has an IDF of 0.66, which would imply that it is used in a 7454 of the 14422 documents.

1	0.66	pasient	**26**	1.57	formål	**51**	1.95	høy	**76**	2.23	studere
2	0.82	studie	**27**	1.58	samt	**52**	1.96	under	**77**	2.24	tid
3	0.92	prosjekt	**28**	1.58	vise	**53**	1.97	gruppe	**78**	2.25	hensikt
4	0.93	undersøke	**29**	1.62	klinisk	**54**	1.97	person	**79**	2.25	vurdere
5	0.98	behandling	**30**	1.63	finne	**55**	1.97	norsk	**80**	2.25	identifisere
6	1.02	hos	**31**	1.64	ulik	**56**	1.98	følge	**81**	2.26	ofte
7	1.02	gi	**32**	1.65	bruk	**57**	1.99	barn	**82**	2.26	hjelp
8	1.06	få	**33**	1.66	norge	**58**	1.99	undersøkelse	**83**	2.26	forhold
9	1.09	studium	**34**	1.68	mange	**59**	2.00	redusere	**84**	2.26	basere
10	1.15	god	**35**	1.69	viktig	**60**	2.02	kartlegge	**85**	2.26	forekomst
11	1.21	ønske	**36**	1.75	utvikle	**61**	2.02	helse	**86**	2.27	samle
12	1.24	ny	**37**	1.78	metode	**62**	2.03	blant	**87**	2.27	endring
13	1.26	kunnskap	**38**	1.79	bidra	**63**	2.04	utvikling	**88**	2.27	sykehus
14	1.26	år	**39**	1.79	sykdom	**64**	2.05	grad	**89**	2.28	del
15	1.27	øke	**40**	1.80	resultat	**65**	2.07	informasjon	**90**	2.28	livskvalitet
16	1.31	bruke	**41**	1.84	inkludere	**66**	2.08	liten	**91**	2.28	type
17	1.33	annen	**42**	1.84	dag	**67**	2.12	forskning	**92**	2.29	alder
18	1.34	mål	**43**	1.86	ta	**68**	2.13	vid	**93**	2.29	utføre
19	1.35	stor	**44**	1.88	gjennom	**69**	2.14	vanlig	**94**	2.29	alvorlig
20	1.41	mye	**45**	1.89	gjennomføre	**70**	2.14	faktor	**95**	2.29	lite
21	1.42	data	**46**	1.90	risiko	**71**	2.15	mulig	**96**	2.30	analyse
22	1.52	se	**47**	1.91	to	**72**	2.15	behov	**97**	2.34	teste
23	1.52	tidlig	**48**	1.91	påvirke	**73**	2.19	spørreskjema	**98**	2.36	årsak
24	1.53	effekt	**49**	1.91	tillegg	**74**	2.20	betydning	**99**	2.36	føre
25	1.53	gjøre	**50**	1.92	derfor	**75**	2.21	behandle	**100**	2.39	oppleve

Table 22: The top one hundred words with the highest IDF score

4) The final set of stop words are identified manually, capturing words that have been a disturbance in interpreting topics, either because they are over-represented in many topics, or because they do not provide any additional value when it comes to interpreting the semantics of the topics. These stop words are displayed below

"hun", "th", "and", "of", "frå", "gang", "tre", "ca.", "en", "ad", "vår", "la", "ii", "iii", "mm", "smite", "tiltak", "lang", "sist", "assistere", "feil", "etterlevelse", "stadig", "effektiv", "la", "forebygge", "hendelse", "utsette", "tilstand", "land", "antall", "sammenhengen", "rolle", "delstudie", "mm", "spiller", "venstre", "pasientgrupp", "mann", "avdeling", "pasientane", "personar", "meir", "desse", "vert", "pasientar", "funksjon", "deltager", "evaluere", "fylle", "oppfølging", "uke", "deltaker", "måned", "in", "mekanisme", "tilstand", "stille", "utredning", "samarbeid", "tilbud", "intervju", "erfaring", "bruker", "tjeneste", "primær", "deltaker", "åpen", "måned", "evaluere", "kobinasjon", "fase", "uke", "nivå", "lav", "tilskudd", "veiledning", "f", "materiale", "in", "verdi", "vurdering", "funn", "forårsake", "oppmerksomhet", "deltaker", "vanske", "funksjone", "problem", "registere", "analysere", "hemte", "søke", "periode", "rek", "innhente", "registrere", "opplysning", "menneske", "utforsk", "møte", "tema", "deltaker", "forståelse", "egen", "opplevelse", "erfaring", "intervju", "tilbud", "arbeid", "uke", "så", "fall", "inntak", "vekst", "sen", "tilstand", "kjent", "risikofaktore", "motivasjon", "deltaker", "minutt", "enkel", "kvinne", "sammenheng", "retningslinje", "ansatt", "kvalitet", "avdeling", "modell", "gjennomgå", "fjerne", "stadium", "evt", "bakgrunn", "deltakere", "å", "kvinner", "menn", "tekst", "ord", "minutter", "runde"

A.4: Example of a REK Form

Figure 17 shows an example of a category of the application form that applicants fill out and send to REK for processing. Most of the data used in the analysis is gathered from this form, and the picture shows the category where the project field lies, which is one of the central fields in this thesis.

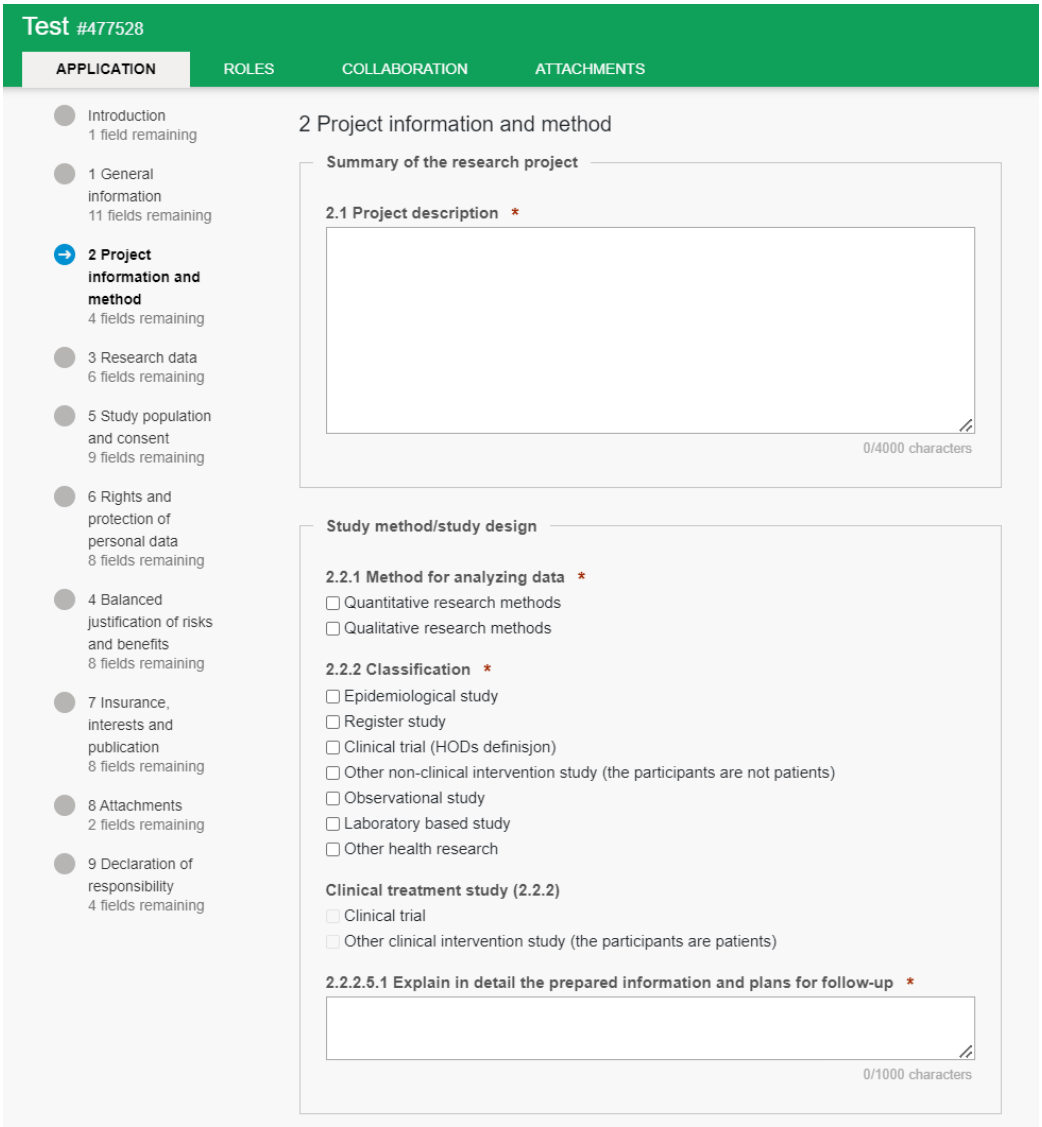

Figure 17: Example of an empty application form in Rekportalen

A.5: Feature Importance

Figures 18 to 21 show the feature importance of the various models. Only the twenty most influential variables are displayed. In the logistic regression model, it is shown whether the variable has a positive or negative impact on predicting rejections.

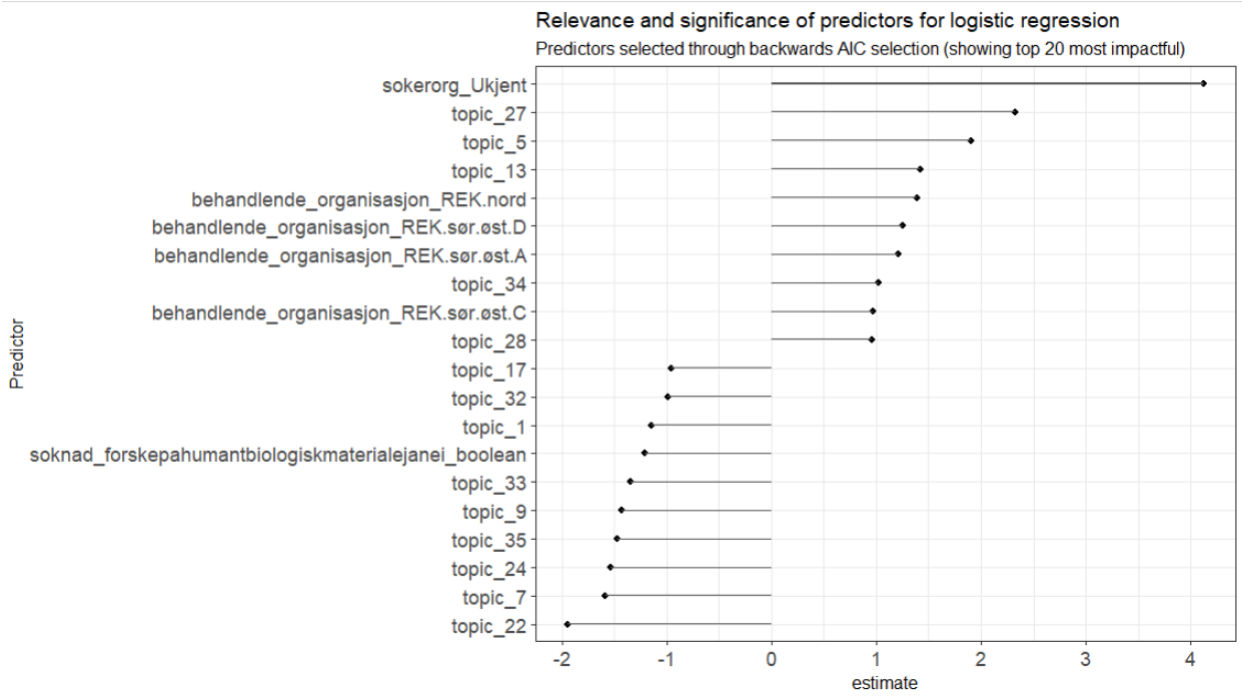

Figure 18:Variable importance for the Logistic regression model

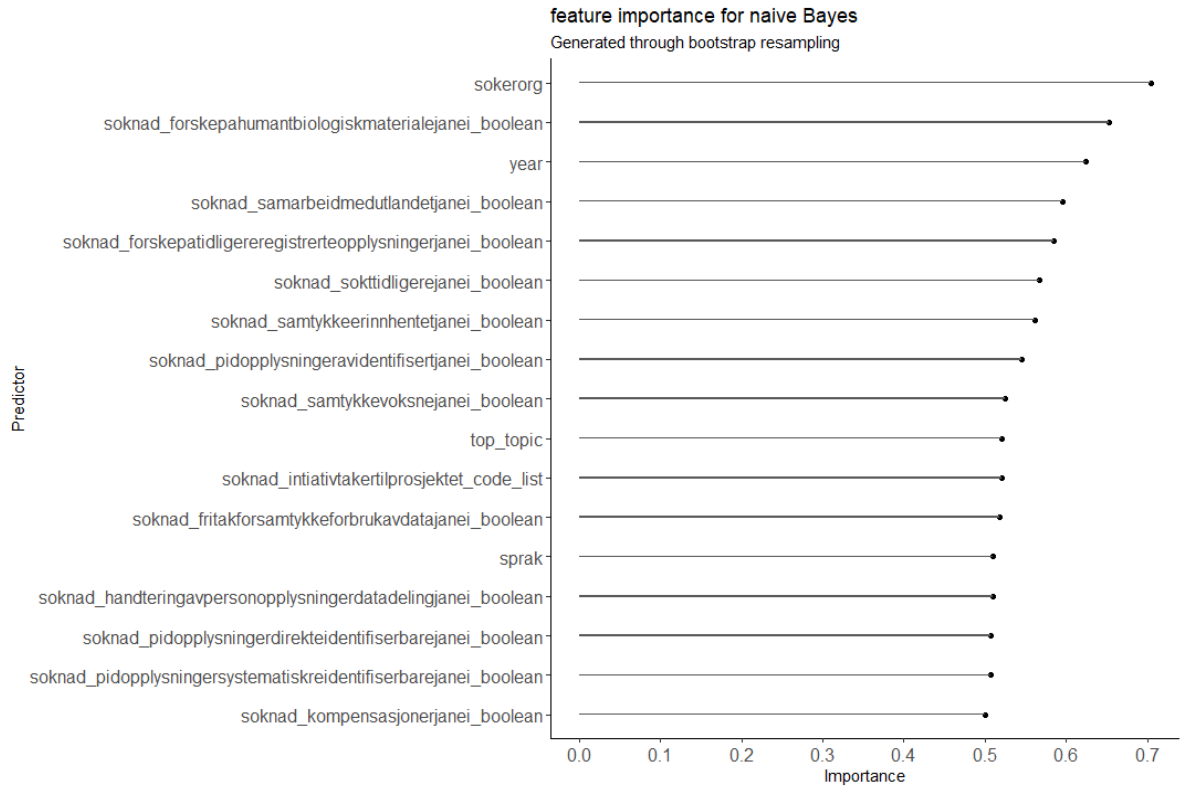

Figure 19: Variable importance for the Naive Bayes model

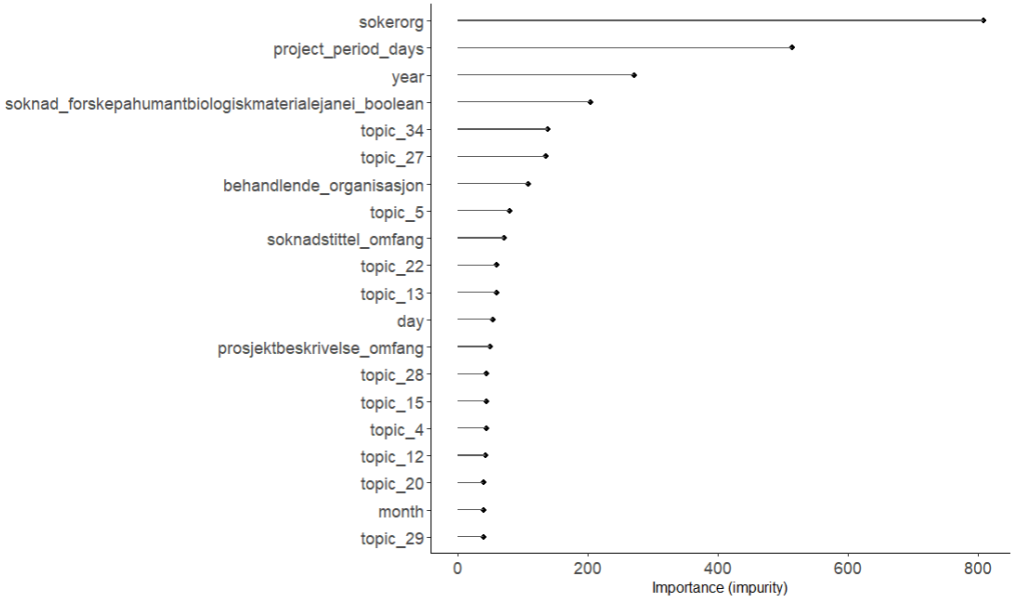

Figure 20: Variable importance for the Random Forest model

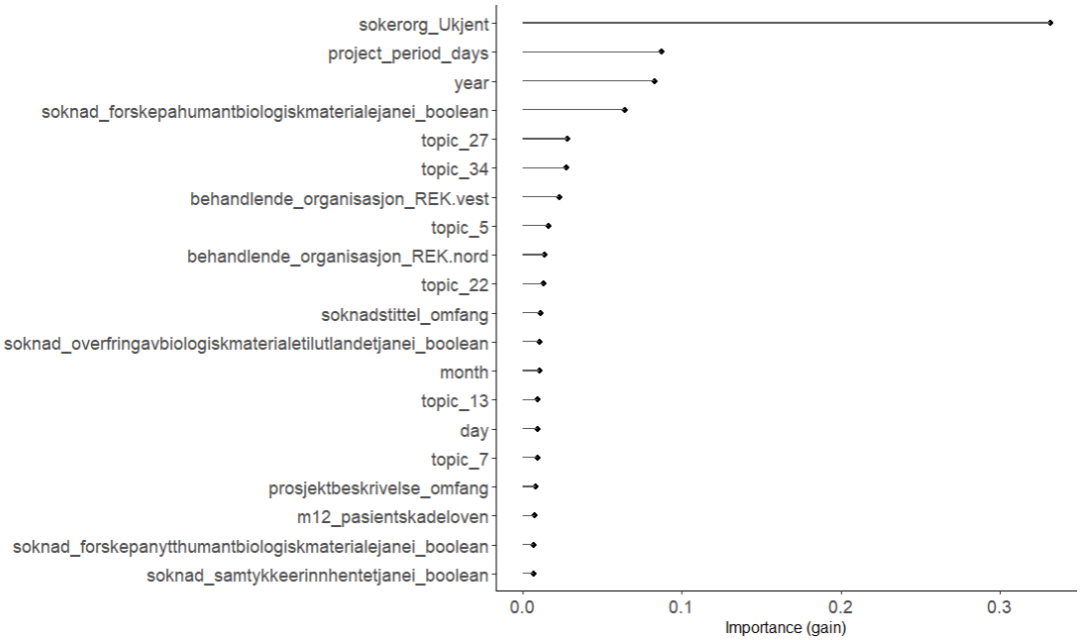

Figure 21: Variable importance for the XGBoost model

A.6: Norwegian Topic Words for Selected Topics

Top 10 Words for Topics 6, 11, 15, and 22

Topic 6	Topic 11	Topic 15	Topic 17	Topic 22
fedme	hjerte	gammel	overlevelse	psykisk
overvekt	variasjon	pårørende	tilbakefall	lidelse
vitamin	atrieflim	demens	kreft	depresjon
kosthold	hjertesvikt	sykehjem	strålebehandling	angst
metabolsk	karsykdom	omsorg	bivirkning	symptom
ernæring	hjertesykdom	palliativ	kombinasjon	søvn
mat	hjertefunksjon	nyre	cellegift	psykologisk
vekt	alkohol	hjemmeboende	kjemoterapi	terapi
diabete	genetisk	liv	svulst	psykiatrisk
spise	blodtrykk	delirium	lungekreft	stress

Table 23: Top 10 words for topics 6, 11, 15, and 22 (Norwegian)

Top 10 Words for Topics 9, 27, 32, and 34

Topic 27	Topic 34	Topic 32	Topic 9
kommune	kvalitativ	celle	svangerskap
helsetjeneste	forelder	blod	mor
kommunal	psykisk	protein	fødsel
spesialisthelsetjenest	familie	antistoff	gravid
praksis	helsepersonell	kreftcelle	føde
rett	semistrukturere	isolere	graviditet
samhandling	sosial	vev	moba
helsepersonell	kvantitativ	stamcelle	nyfødt
kvalitativ	individuell	human	foster
selvmord	pårørende	kropp	premature

Table 24: Top 10 words for topics 9, 27, 32, and 34 (Norwegian)

A.7: Backwards AIC variable selection

AIC, or Akaike Information Criterion, is a criterion for measuring a model's goodness of fit, and is used "to compare different models on a given outcome" (Snipes & Taylor, 2014, p. 2). The criterion considers both the log likelihood of the model, *L*, as well as the number of parameters that must be estimated, *K*, and is calculated as follows:

$$AIC = 2K - 2\log\left(L\left(\hat{\theta}\middle|y\right)\right)$$

Equation 21: The Akaike Information Criterion

The best model is the model with the lowest AIC score, and it can therefore be seen that the criterion penalizes models with many parameters to estimate (K). The AIC is ordinal and means nothing on its own, but can be used to compare models that are fit on the same data (Snipes & Taylor, 2014, p. 3).

In the logistic regression method for this thesis, the AIC is used as a performance measure of a large number of models through a backwards variable selection method. Backwards selection entails iteratively removing the least useful predictor from the formula that is used in a method (James et al., 2013, p. 231). The first step of the procedure is to fit a model on all predictors. Then, new models are fit on all possible variations of p-1 predictors, that is, models with one less predictor included in the formula. Out of these models, the best one in terms of the selected performance measure, in this case AIC, is selected. This procedure is repeated until the AIC no longer improves when removing predictors. James et al. (2013, p.231), points out that the procedure is not guaranteed to yield the best model, as not all possible predictor combinations possible is tried out. However, it provides a simple approach for automatic variable selection, which has been useful for this thesis as there are many predictors in the data set. The final formula resulting from the backwards AIC variable selection method is displayed below:

Logistic regression formula *= avvist ~ topic_1 + topic_4 + topic_5 + topic_7 + topic_8 + topic_9 + topic_13 + topic_16 + topic_17 + topic_18 + topic_21 + topic_22 + topic_24 + topic_27 + topic_28 + topic_31 + topic_32 + topic_33 + topic_34 + topic_35 + topic_36 + topic_38 + topic_39 + month + year + project_period_days + prosjektbeskrivelse_omfang +*

soknad_forskepahumantbiologiskmaterialejanei_boolean +
soknad_forskepatidligereregistrerteopplysningerjanei_boolean +
soknad_handteringavpersonopplysningerdatadelingjanei_boolean +
soknad_samarbeidmedutlandetjanei_boolean +
soknad_samtykkeerinnhentetjanei_boolean +
soknad_sokttidligerejanei_boolean + m11_na + m11_legemiddelstudie +
m10_annet_samarbeid +
m10_skal_det_gjennomfores_en_selvstendig_datainnsamling_i_utlandet +
m5_na + m5_kliniske_undersokelser + m5_sporreskjema + m5_fysiske_inngrep +
m2_na + m2_16_18_ar + m1_na + behandlende_organisasjon_REK.nord +
behandlende_organisasjon_REK.sør.øst.A +
behandlende_organisasjon_REK.sør.øst.B +
behandlende_organisasjon_REK.sør.øst.C +
behandlende_organisasjon_REK.sør.øst.D +
behandlende_organisasjon_REK.vest + behandlende_organisasjon_other +
sokerorg_Helse.Bergen.HF...Haukeland.universitetssykehus +
sokerorg_Norges.teknisk.naturvitenskapelige.universitet +
sokerorg_St..Olavs.Hospital.HF + sokerorg_UiT.Norges.arktiske.universitet +
sokerorg_Ukjent + sokerorg_Universitetet.i.Bergen + sokerorg_Universitetet.i.Oslo +
sokerorg_other

A.8: Variables

Variable	Type
soknadsid	character
topic_1	numeric
topic_2	numeric
topic_3	numeric
topic_4	numeric
topic_5	numeric
topic_6	numeric
topic_7	numeric
topic_8	numeric
topic_9	numeric
topic_10	numeric
topic_11	numeric
topic_12	numeric
topic_13	numeric
topic_14	numeric
topic_15	numeric
topic_16	numeric
topic_17	numeric
topic_18	numeric
topic_19	numeric
topic_20	numeric
topic_21	numeric
topic_22	numeric
topic_23	numeric
topic_24	numeric
topic_25	numeric
topic_26	numeric
topic_27	numeric
topic_28	numeric
topic_29	numeric
topic_30	numeric
topic_31	numeric

topic_32	numeric
topic_33	numeric
topic_34	numeric
topic_35	numeric
topic_36	numeric
topic_37	numeric
topic_38	numeric
topic_39	numeric
topic_40	numeric
top_topic	factor
behandlende_organisasjon	character
month	numeric
year	numeric
day	integer
project_period_days	numeric
sprak	factor
sokerorg	factor
prosjektbeskrivelse_omfang	integer
soknadstittel_omfang	integer
avvist	factor
soknad_andreopplysningerrelevantforbehandlingjanei_boolean	numeric
soknad_antallforskningsdeltakereinorge_numeric	numeric
soknad_antallforskningsdeltakeretotalt_numeric	numeric
soknad_biologiskmaterialeanonymisertjanei_boolean	numeric
soknad_biologiskmaterialeavidentifisertjanei_boolean	numeric
soknad_biologiskmaterialedirekteidentifiserbartjanei_boolean	numeric
soknad_forskepahumantbiologiskmaterialejanei_boolean	numeric
soknad_forskepainnsamlethumantbiologiskmaterialejanei_boolean	numeric
soknad_forskepanytthumantbiologiskmaterialejanei_boolean	numeric
soknad_forskepatidligereregistrerteopplysningerjanei_boolean	numeric
soknad_forskningsansvarligorganisasjon_picker	factor
soknad_forskningsdesignkontrollgrupperjanei_boolean	numeric
soknad_fritakforsamtykkeforbrukavbiologiskmatjanei_boolean	numeric
soknad_fritakforsamtykkeforbrukavdatajanei_boolean	numeric
soknad_fritaksamtykkedataalleredeinnsamlet_boolean	numeric

soknad_fritaksamtykkeinformasjonsfritakjanei_boolean	numeric
soknad_fritaksamtykkematerialealleredeinnsamlet_boolean	numeric
soknad_fritaksamtykkematerialealleredeinnsamletinformert_boolean	numeric
soknad_fritaksamtykkematerialelik_boolean	numeric
soknad_fritaksamtykkematerialendssituasjoner_boolean	numeric
soknad_fritaksamtykkematerialeutensamtykkekompetanse_boolean	numeric
soknad_gammeltbiologiskmaterialedestrueres2mndetterprv_boolean	numeric
soknad_gammeltbiologiskmaterialelagresinyspesifikkbioba_boolean	numeric
soknad_genetiskeunderskelseravbiomaterialejanei_boolean	numeric
soknad_genetiskeunderskelseravbiomathelsegevinstjanei_boolean	numeric
soknad_genetiskeunderskelseravbiomattilbakefring_boolean	numeric
soknad_handteringavpersonopplysningerbiologiskmateriale_boolean	numeric
soknad_handteringavpersonopplysningerbiologisksamtykke_boolean	numeric
soknad_handteringavpersonopplysningerdatadelingjanei_boolean	numeric
soknad_handteringavpersonopplysningerviderebruksamtykke_boolean	numeric
soknad_innhentingavdataioniserendestralingdoseestimatr_code_list	factor
soknad_innhentingavdataioniserendestralingjanei_boolean	numeric
soknad_innhentingavdatasprreskjemavalidert_boolean	numeric
soknad_innhenting_av_data_utproving_utstyr_bekreftelse_1_boolean	numeric
soknad_innhenting_av_data_utproving_utstyr_ce_merket_nei_code_list	factor
soknad_innhenting_av_data_utproving_utstyr_innenfor_eufor_boolean	numeric
soknad_innhentingavdatautprvingutstyrcemerketjanei_boolean	numeric
soknad_innhentingavdatautprvingutstyrjanei_boolean	numeric
soknad_innhentingavnyedatajanei_boolean	numeric
soknad_intiativtakertilprosjektet_code_list	factor
soknad_kompensasjonerjanei_boolean	numeric
soknad_legemiddelstudieavbruddbehandlingjanei_boolean	numeric
soknad_legemiddelstudiefase_code_list	factor
soknad_nyttbiologiskmaterialedestrueres2mndetterprve_boolean	numeric
soknad_nyttbiologiskmaterialelagresigodkjentbiobank_boolean	numeric
soknad_nyttbiologiskmaterialelagresigodkjentbiobankhviken_external_lookup	factor
soknad_nyttbiologiskmaterialelagresinyspesifikkbiobank_boolean	numeric
soknad_opplysningerfraandreregistrejanei_boolean	numeric
soknad_opplysningerfraandretyperregistrejanei_boolean	numeric
soknad_opplysningerfraannetregisterjanei_boolean	numeric

soknad_opplysningerfrabefolkningsbaserthelseundjanei_boolean	numeric
soknad_opplysningerfralokalthelseregjanei_boolean	numeric
soknad_opplysningerfranasjonaltkvalitetsregisterjanei_boolean	numeric
soknad_opplysningerfrapasientjournaljanei_boolean	numeric
soknad_opplysningerfrasentralthelseregisterjanei_boolean	numeric
soknad_opplysningerfratidligeregodkjentprosjektjanei_boolean	numeric
soknad_opplysningerfrautelandskeregistrejanei_boolean	numeric
soknad_overfringavbiologiskmaterialefrautlandetjanei_boolean	numeric
soknad_overfringavbiologiskmaterialefrautlandetland_picker_multi	character
soknad_overfringavbiologiskmaterialefrautlandetlandeu_picker_multi	character
soknad_overfringavbiologiskmaterialetilutlandetavtaler_boolean	numeric
soknad_overfringavbiologiskmaterialetilutlandetbeskytt_boolean	numeric
soknad_overfringavbiologiskmaterialetilutlandetjanei_boolean	numeric
soknad_overfringavbiologiskmaterialetilutlandetland_picker_multi	character
soknad_overfringavbiologiskmaterialetilutlandetlandeu_picker_multi	character
soknad_overfringavbiologiskmaterialetilutlandetrest_code_list	factor
soknad_overfringavbiologiskmaterialetilutlandetsamtykke_boolean	numeric
soknad_overfringavhelseoppltilutlandetbeskyttetinfo_boolean	numeric
soknad_overfringavhelseoppltilutlandetsikret_boolean	numeric
soknad_overfringavhelseopplysningerfrautlandetjanei_boolean	numeric
soknad_overfringavhelseopplysningerfrautlandetland_picker_multi	character
soknad_overfringavhelseopplysningerfrautlandetlandeu_picker_multi	character
soknad_overfringavhelseopplysningerfrautlandetsamtykke_boolean	numeric
soknad_overfringavhelseopplysningertilutlandetjanei_boolean	numeric
soknad_overfringavhelseopplysningertilutlandetland_picker_multi	character
soknad_overfringavhelseopplysningertilutlandetlandeu_picker_multi	character
soknad_overfringavhelseopplysningertilutlandetsamtykke_boolean	numeric
soknad_pidopplysningeravidentifisertjanei_boolean	numeric
soknad_pidopplysningerdirekteidentifiserbarejanei_boolean	numeric
soknad_pidopplysningersystematiskreidentifiserbarejanei_boolean	numeric
soknad_publiseringrestriksjonerjanei_boolean	numeric
soknad_registrertesrettigheteroppdatertinformasjonjanei_boolean	numeric
soknad_samarbeidmedutlandetannetland_picker_multi	character
soknad_samarbeidmedutlandetdatainnsamlingland_picker_multi	character
soknad_samarbeidmedutlandetjanei_boolean	numeric

soknad_samarbeidmedutlandetmultisenterstudieland_picker_multi	character
soknad_samiskjanei_boolean	numeric
soknad_sammenstillingavopplysningerjanei_boolean	numeric
soknad_samtykkebarnjanei_boolean	numeric
soknad_samtykkeerinnhentetjanei_boolean	numeric
soknad_samtykke_vil_bli_innhentet_begge_foreldre_boolean	numeric
soknad_samtykkevoksnejanei_boolean	numeric
soknad_sokttidligerejanei_boolean	numeric
soknad_utdanningsprosjektjanei_boolean	numeric
soknad_utdanningsprosjektstudieniva_code_list	factor
soknad_utproving_av_medisinsk_utstyr_sokt_slv_boolean	numeric
soknad_utsattoffentliggjoringjanei_boolean	numeric
soknad_utsattoffentliggjringtidspunkt_date_only	POSIXct
m12_na	numeric
m12_for_a_beskytte_legitime_patentrettslige_eller_konkurransemessige_interesser	numeric
m12_av_hensyn_til_et_lopende_forskningsarbeid	numeric
m11_na	numeric
m11_legemiddelstudie	numeric
m11_annen_klinisk_intervensjonsstudie_deltakerne_er_pasienter	numeric
m10_er_studien_en_del_av_en_internasjonal_multisenterstudie	numeric
m10_annet_samarbeid	numeric
m10_skal_det_gjennomfores_en_selvstendig_datainnsamling_i_utlandet	numeric
m10_na	numeric
m9_na	numeric
m9_nav	numeric
m9_statistisk_sentralbyra	numeric
m9_folkeregisteret	numeric
m9_strafferegisteret	numeric
m8_na	numeric
m8_fullblod	numeric
m8_serum	numeric
m8_plasma	numeric
m8_celler	numeric
m8_slimhinneavstryk	numeric
m8_ekspektorat	numeric

m8_avforing	numeric
m8_urin	numeric
m8_kroppsvaesker	numeric
m8_biopsimateriale	numeric
m8_beinmarg	numeric
m8_dna_ekstrahert	numeric
m8_rna_ekstrahert	numeric
m8_annet_materiale	numeric
m8_cytologier	numeric
m8_morsmelk	numeric
m8_bakterieisolat	numeric
m8_autopsimateriale	numeric
m8_har_og_negler	numeric
m8_cerebrospinalvaeske	numeric
m8_bein	numeric
m8_tannmateriale	numeric
m8_fostervann	numeric
m8_eggceller	numeric
m8_saedceller	numeric
m7_na	numeric
m7_i_human_farmakologi	numeric
m7_ii_terapeutisk_utproving	numeric
m7_iii_terapeutisk_bekreftelse	numeric
m7_iv_terapeutisk_bruk_intervensjonsstudie	numeric
m7_iv_terapeutisk_bruk_ikke_intervensjonsstudie	numeric
m6_na	numeric
m6_pasienter_som_ikke_vil_ha_direkte_fordel_av_prosedyren	numeric
m6_pasienter_som_potensielt_kan_ha_direkte_medisinsk_fordel_av_prosedyren	numeric
m6_voksne_yngre_enn_50_ar	numeric
m6_voksne_50_ar_eller_eldre	numeric
m6_mindrearige_under_18_ar	numeric
m6_friske_frivillige	numeric
m6_gravide	numeric
m5_na	numeric
m5_observasjoner_uten_opptak	numeric

m5_kliniske_undersokelser	numeric
m5_sporreskjema	numeric
m5_annet	numeric
m5_intervjuer_uten_opptak	numeric
m5_fysiske_inngrep	numeric
m5_intervjuer_med_opptak_lyd_video	numeric
m5_observasjoner_med_opptak_lyd_video_foto	numeric
m5_dagboker	numeric
m4_na	numeric
m4_konvensjonell_rontgen	numeric
m4_computertomografi_ct	numeric
m4_radiofarmaka	numeric
m4_akselerator	numeric
m4_kapslet_radioaktiv_kilde	numeric
m3_na	numeric
m3_ekspektorat	numeric
m3_fullblod	numeric
m3_celler	numeric
m3_beinmarg	numeric
m3_avforing	numeric
m3_slimhinneavstryk	numeric
m3_plasma	numeric
m3_cerebrospinalvaeske	numeric
m3_urin	numeric
m3_dna_ekstrahert	numeric
m3_serum	numeric
m3_biopsimateriale	numeric
m3_annet_materiale	numeric
m3_cytologier	numeric
m3_rna_ekstrahert	numeric
m3_autopsimateriale	numeric
m3_kroppsvaesker	numeric
m3_bakterieisolat	numeric
m3_fostervann	numeric
m3_har_og_negler	numeric

m3_eggceller	numeric
m3_saedceller	numeric
m3_tannmateriale	numeric
m3_bein	numeric
m3_morsmelk	numeric
m2_na	numeric
m2_under_12_ar	numeric
m2_12_15_ar	numeric
m2_16_18_ar	numeric
m1_pasientskadeloven	numeric
m1_produktansvarsloven	numeric
m1_saerskilt_forsikring	numeric
m1_na	numeric
northern_europe	numeric
southern_asia	numeric
northern_america	numeric
western_europe	numeric
southern_europe	numeric
western_asia	numeric
northern_africa	numeric
south_eastern_asia	numeric
australia_and_new_zealand	numeric
eastern_europe	numeric
eastern_africa	numeric
western_africa	numeric
eastern_asia	numeric
middle_africa	numeric
central_america	numeric
south_america	numeric
southern_africa	numeric
caribbean	numeric
central_asia	numeric
melanesia	numeric
micronesia	numeric

Table 25: All the variables in the data set