IDENTIFYING CARING COMMUNITIES WITHIN DUTCH CHAMBER OF COMMERCE DATA:

A CLASSIFIER COMPARISON

by

Lisa Tessels

0515257

Master Thesis

MASTER APPLIED DATA SCIENCE

Utrecht University

Vilans supervisor: Mariëlle Zondervan

UU supervisor: Dr. Dong Nguyen

Second UU supervisor: Dr. Pablo Mosteiro Romero

July 2024

Abstract

This paper aimed to determine the most effective classifier for identifying registered 'caring communities' using data from the Dutch Chamber of Commerce. I optimized and assessed the performance of four classifiers: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting Tree (GBDT). The results show that LR consistently outperformed the other models across 2022 and 2023 test sets, excelling across all evaluation metrics. While GBDT showed competitive performance, SVM and RF were less effective. Despite LR's strengths, improvements in recall and data quality are essential for better identification of caring communities. Without these improvements, the algorithm may underestimate the total number of caring communities, leading to an incomplete understanding of their prevalence.

Keywords – Machine learning, Classification, Caring Communities

Acknowledgements

This thesis marks the end of my Master's program at Utrecht University (UU), and I want to express my sincere gratitude to those who have supported me along the way.

First, my gratitude goes to Vilans for commissioning this research project. Their support, including access to valuable data and guidance from experts in the field, was instrumental in shaping this thesis. I am also grateful to my supervisor, Dr. Dong Nguyen from UU, for her insightful feedback, constant support, and expert guidance throughout the research process. Your suggestions significantly improved the quality of this work. I also extend my thanks to Kalee Said, a fellow student at UU and a good friend. Our discussions and teamwork throughout the research process were not only stimulating but also fostered a supportive and motivating environment. Finally, I would like to thank my family for their support and encouragement throughout my studies. Their willingness to proofread and offer a fresh perspective on my writing is greatly appreciated.

This research would not have been possible without the contributions of all of you. Thank you.

Contents

1 Introduction

An ageing population and growing demand for community-based care have led to the emergence of 'caring communities' as vital components of social care systems. These resident-led collectives take the initiative to improve their local living environments by offering various services and activities, ranging from health programs to social activities (NZVE, n.d.; Zoest, 2023). As the number of elderly and disabled people living at home increases (CBS, 2020; Daalhuizen et al., n.d.), so too does the reliance on these communities to fulfil local care needs (Zoest, 2023). However, due to the diverse nature of these initiatives, this movement is difficult to monitor (Movisie, 2020).

Identifying registered caring communities is important for several reasons. Firstly, it provides a better understanding of the landscape of these crucial care providers, particularly considering the growing pressures on formal care systems. Caring communities play a vital role in alleviating this pressure by delivering personalized, community-driven care (Zoest, 2023; Zoest et al., 2023). Secondly, identifying these communities allows for a clearer picture of their reach and enables the provision of targeted support to initiatives that significantly contribute to social well-being and cohesion. This aligns with the interests of organizations like Vilans, the Dutch knowledge organization for independent healthcare research, which is actively involved in evaluating strategies to improve care delivery (Vilans, 2024). Lastly, integrating caring communities into formal care systems represents a shift towards a more inclusive and sustainable care model (Zoest, 2023).

To tackle the challenge of identifying caring communities, I will use data from the Dutch Chamber of Commerce (KvK). According to the Monitor Caring Communities 2020, approximately 80% of community initiatives are registered within the KvK (Zoest et al., 2023). This high rate of registration suggests that the KvK data is a comprehensive source for identifying most of these initiatives, making it a suitable choice for this study. By analysing repeated samples of KvK data from multiple years (2022 and 2023), I aim to not only identify caring communities but also assess trends and changes in their size and types over time.

By building a classifier model using Machine Learning (ML), I can automate the otherwise, time-consuming, and resource-intensive, identification process. This model could analyse key features within KvK data, such as business descriptions, SBI codes, or names, to learn patterns that distinguish registered caring communities from other types of organizations.

Therefore, the research question guiding this study is:

'What is the most effective classifier for identifying registered 'caring communities' within repeated samples of Dutch Chamber of Commerce data?'

By finding the most effective classifier for identifying caring communities within KvK data, this study aims to contribute to a more integrated and effective social care infrastructure. Additionally, this study seeks to investigate the extent to which KvK data is useful for this task. This will provide valuable insights for future efforts aiming to leverage similar data sources and aid the decision-making of organizations like Vilans regarding data acquisition strategies, specifically whether it is worth continuing to invest in the purchase of KvK data.

This study aims to go beyond theoretical contributions by addressing practical needs in social planning and policymaking. By identifying caring communities, we can better understand their size and growth trends. Future studies can use this data to explore their distribution, characteristics, and impact. This knowledge can then contribute to the development of policies that support caring communities, so these communities can continue to complement the formal care system and play a vital role in making care future-proof.

The structure of this study is as follows. First, a comprehensive literature review (chapter 2) examines existing research on classification algorithms and caring communities. This review lays the foundation for the methodology. Next, chapter 3 describes the data, elaborating on its characteristics, preparation process, and ethical and legal considerations. Chapter 4 outlines the experimental method, including parameter tuning and evaluation mechanisms. After, the results are presented in chapter 5, followed by an analysis and interpretation in the discussion (chapter 6). Finally, the conclusion is presented in chapter 7.

2 Related Works

This related works chapter is divided into two main subsections. The first section (2.1 Caring Communities) explores existing research on caring communities and the challenges associated with their identification. The second section (2.2 Classifiers) delves into the application of machine learning for organizational classification and introduces the specific classifiers I will compare in this study.

2.1 Caring Communities

Despite the lack of scientific articles specifically on the Dutch concept of 'caring communities', the concept draws upon what is described in the scientific literature as simply 'communities'. These communities often exhibit similar characteristics to caring communities, such as high levels of volunteerism, strong social networks, and a presence of organizations focused on social well-being (Chaskin, 2001). Research on communities highlights their role in fostering social cohesion, mutual support, and collective action, which are central to the functioning of caring communities (Chaskin, 2001; McLeroy et al., 2003).

Although there is alignment with existing community concepts, the definition and identification of caring communities as a distinct concept have not been extensively researched in international peer-reviewed journals. However, there has been a growing interest in this area in recent years, particularly by knowledge centres in the Netherlands such as Movisie, Nederland Zorgt Voor Elkaar and Vilans (Movisie, 2020; NZVE, n.d.; NZVE et al., 2019; Zoest, 2023; Zoest et al., 2023). These organizations have played a significant role in advancing this research. In 2020, these three organizations collaborated to publish a comprehensive report titled "*Monitor Zorgzame Gemeenschappen 2020: Een groeiende beweging van bewonersinitiatieven in welzijn, wonen en zorg*" (in English: *Monitor Caring Communities 2020: A Growing Movement of Resident Initiatives in Well-being, Housing and Care*) (Smellik et al., 2020). This report provides a thorough and insightful exploration of caring communities in the Netherlands, shedding light on their characteristics, impact, and growth. It reveals that there is a growth of diverse caring communities and emphasizes their positive impact on social cohesion and support for vulnerable populations. Next to portraying the movement of caring communities, the report also explores essential questions regarding financial sustainability, impact measurement, and effective communication of their contributions. The study positions caring communities as key players in enhancing well-being and complements formal care systems. While the "Monitor Zorgzame Gemeenschappen 2020" provides a valuable foundation, a research gap remains, particularly regarding the development of robust methods for identifying these initiatives.

My research aims to bridge this gap by focusing specifically on identifying caring communities within Dutch KvK data. Identifying caring communities within datasets remains challenging due to several factors.

First, caring communities use a wide variety of organizational structures and names to describe themselves, such as healthcare cooperatives, 'noaberzorgpunten' (neighbour care points), city villages, 'lief-en-leed straten' (streets where residents share both joyful and sorrowful events), and many more variants. They are foundations, associations, cooperatives or even just a group of local residents who have started working without a formal structure (Movisie, 2020).

Besides, they facilitate a wide range of activities and services. For example, a caring community might facilitate social gatherings, drop-in centres, neighbourly assistance, cultural activities, meals, transport, informal care support, shopping services, etc. Some communities are also involved in professional services, for example, daycare and community nursing, or employ a village support worker (Movisie, 2020; NZVE, n.d.). Thus, caring communities come in a range of different shapes and sizes, often serving the entire neighbourhood but sometimes focusing on specific populations like the elderly or people with disabilities (Movisie, 2020). The diverse nature of these initiatives makes it difficult to develop a universal definition.

The second challenge associated with identifying caring communities is that indicators can be subjective and nuanced. For instance, the trade name and organizational descriptions may not always accurately reflect the community-oriented nature. An illustrative example is 'Czaar 51', registered with the KvK under the broad category of 'sociaal-cultureel werk' (social-cultural work). While this description does not explicitly mention its role as a caring community, Czaar 51 is a perfect example of a caring community. It functions as a communal living room, fostering a space for residents to connect and engage in diverse activities from shared meals and study sessions to creative activities like dancing, drawing and learning to sew (Netwerk DAK, n.d.).

Because of these challenges, caring communities are hard to identify, which has caused them to remain largely undetected. With no way to precisely quantify the number of instances, it is difficult to aid the development of policies that support these initiatives so they can continue to complement the formal care system and play a vital role in making care future-proof (Zoest, 2023).

2.2 Classifiers

Classification is a type of supervised learning where an algorithm learns to identify a target variable y that represents a characteristic within the data. In this research, the target variable would be 'caring community' or 'non-caring community', where 'non-caring community' refers to all other types of organizations and entities that do not fit the criteria of a caring community. Data classification has important applications that cover a wide spectrum including engineering, e-commerce, and medicine. ML has also been employed to classify organizations based on various characteristics. For example, Litofcenko et al. (2020) demonstrated the possibility of using a decision tree classifier to classify non-profit organizations according to organization name and field of activity. This study builds upon the well-established classification system known as the International Classification of Nonprofit Organizations (ICNPO), developed by Salamon and Anheier (1992). ICNPO is a standardized system for categorizing nonprofit organizations. This framework has been instrumental in advancing research in the nonprofit sector by providing a common language and criteria for categorization. My study zooms in on a specific subgroup within one of the ICNPO's categories (community and neighbourhood organizations within group 6: development and housing). This subgroup, like caring communities, focuses on improving social and economic well-being within their communities (Salamon & Anheier, 1996).

Another study conducted by Allozi and Abbod (2022) employed three classification models to classify firms into healthy or failed, namely Logistic Regression, Artificial Neural Network and Support Vector Machine. Many other studies explored organization classification, particularly

in the financial and non-profit sectors, e.g. Aljawazneh et al. (2021), LePere-Schloop (2022), and Ma (2021). However, their focus and data source differed from this research. Nevertheless, these studies highlight the potential of using organizational characteristics for classification tasks.

In ML, classification methods like Support Vector Machines (SVM) (Cortes & Vapnik, 1995) and Random Forests (RF) (Breiman, 2001) are commonly used due to their high accuracy (Wu et al., 2008; Zhang et al., 2017). Although not as popular as SVM and RF, Stochastic Gradient Boosting Decision Trees (GBDT) (Friedman, 2002) can achieve excellent prediction performance. This algorithm is often overlooked in the literature of classification benchmarking, as seen in studies by Macià and Bernadó-Mansilla (2014) and Fernández-Delgado et al. (2014). These studies failed to include GBDT for comparison. Another established algorithm is Logistic Regression (LR) (Cabrera, 1994). This linear model is commonly used for practical binary classification tasks because of its simplicity, interpretability, and computational efficiency.

Zhang et al. (2017) provide a comprehensive comparison of 11 state-of-the-art and common classification algorithms, highlighting their respective strengths and weaknesses. The algorithms studied include established classifiers like RF, LR and K Nearest Neighbours classifier, and newer classifiers such as GBDT, Extreme Learning Machine and Deep Learning. Zhang et al. (2017) found that GBDT matches or exceeds the prediction performance of SVM and RF, even with non-exhaustive hyperparameter tuning. GBDT and RF both showed the best total average classification accuracy and mean rank across all 71 data sets they assessed, followed by SVM. This means that these three algorithms generally perform well, in terms of prediction accuracy, regardless of the data sets. They utilized data sets with varying numbers of instances and classes, and features with varying distributions, allowing for a comprehensive comparison. For my research, which involves classifying caring communities within KvK data, the characteristics of my data sets are crucial to consider. My data includes features like organizational descriptions and the total number of employees. Given the similarity in feature diversity (mixed), the number of features (8), the number of classes (2), and data set sizes (1000+), the insights from Zhang et al. (2017) are particularly relevant. Their findings suggest that GBDT and RF are likely to provide high classification accuracy for data with characteristics similar to mine, such as 2 classes and around 8 features.

Understanding the strengths, weaknesses, and overall behaviour of different algorithms across various domains is valuable. However, relying solely on this a priori knowledge to choose the most suitable classifier for the task at hand can be misleading (Harrell, 2015; Zhang et al., 2017). Classifier performance can vary significantly depending on the characteristics of the data. Therefore, the most effective way to determine the optimal classifier is to conduct an empirical experiment. This involves applying a selection of candidate classifiers to the specific dataset and comparing their performance metrics.

While previous studies have demonstrated the potential of machine learning for organizational classification, they do not focus on the unique attributes of caring communities. For this reason, I propose a comparison study between four models, the well-established LR, the widely adopted SVM and RF, and the commonly underutilized GBDT, as machine learning techniques to investigate their performance in identifying caring communities in KvK data. This list of classifiers is informed by the comparison of Zhang et al. (2017).

3 Data

This chapter dives into the data used in this study. It outlines the source of the data, provides an overview of the different subsets, and explores the various features within them. Additionally, it details the data preparation steps I undertook to clean, transform, and integrate the subsets. Finally, this chapter addresses the ethical and legal considerations associated with using data from the KvK.

3.1 Data Description

The datasets utilized in this study consist of information from the KvK, covering various aspects of businesses and organizations. The KvK is a government agency responsible for registering businesses and organizations in the Netherlands (KVK, n.d.). KvK data includes industry classifications, geographical locations, legal forms, textual descriptions, and various supplementary details, providing a comprehensive view of businesses and organizations.

3.1.1 Data Source

All data used in this research was requested from the KvK. Since it is not feasible to request the entire dataset from this repository due to excessive costs, Vilans requested the data selectively based on certain SBI (Standard Business Identifier) numbers. The numbers indicate the business activities of a company (KVK, 2023). To further minimize costs, they only requested specific columns of the data. For example, for the 2023 data, Vilans no longer requested telephone numbers, first and last name of the owner and many date features.

Which SBI numbers to include was guided by a monitor study conducted by Movisie, Nederland Zorgt Voor Elkaar and Vilans in 2020 (Smellik et al., 2020). This survey study aimed to make the scale and significance of the caring community movement visible. The survey targeted residents' initiatives through regional networks, which shared the survey invitation via newsletters, emails, and website announcements. Given the diversity within the population, the study did not prioritize a predefined representative sample but focused on mapping the scope and nature of the initiatives. The caring communities identified through this research were requested from the KvK registry and the corresponding SBI numbers were noted down. Zoest et al. (2023) determined a specific set of 12 SBI numbers to be the most relevant.

Thereafter, Vilans requested data again in 2022 using the same list of SBI numbers. In 2023, they repeated this data request, but this time also included instances that listed these SBI numbers as their secondary activities.

3.1.2 Data Overview

The dataset is comprised of four subsets, each serving a different function in the analysis (Table 1). An overview of which subset contains which features can be found in Appendix A.

The first subset contains data from 2022 and serves as the testing data for that year. This dataset, referred to as test_KVK2022, includes 8,035 rows and 55 columns, covering a wide

range of attributes such as dossier number, business address, municipality, province, various contact numbers, business descriptions, and registration details.

The second subset, referred to as test_KVK2023, contains data from 2023 and is used as testing data for that year. This dataset consists of 11,285 rows and 37 columns, featuring attributes similar to those in the 2022 dataset, though with fewer columns as those were deemed less useful. Since the test datasets (both 2022 and 2023) lacked pre-existing labels, I manually labelled a random sample from each set. For more information about the labelling process, see section 3.2.2.

The third subset originated from the Caring Community Monitor study done in 2020 and is used as training data. This dataset, referred to as train_monitor, consists of 448 rows and 57 columns, offering the same information as the 2022 dataset with two additional columns covering information about the date of deactivation ('DAT_UITSCH RP') and dissolution of the registration ('DAT_ONTB_RP').

The fourth subset contains a random selection of the 2022 dataset that is manually labelled by caring community experts at Vilans to determine whether they meet the definition of caring communities. The experts used a similar process to mine, which is described in section 3.2.2. Known as train_label, this dataset includes 500 rows and 58 columns and serves as training data. To ensure there is no overlap between the train and test sets, the instances included in train_label were removed from the 2022 test dataset (test_KvK2022).

Table 1: All used datasets, including their name, a brief description, function, and size.

Name	Description	Function	Size
test_KVK2022	Data 2022 based on KvK extract of requested SBI numbers	Testing data 2022	8035 rows
test_KVK2023	Data 2023 based on KvK extract of requested SBI numbers	Testing data 2023	11285 rows
train_monitor	Data based on KvK registrations from the Monitor study	Training data	448 rows
train_label	Random selection of the 2022 data for which Vilans employees manually checked whether they meet the definition of caring communities	Training data	500 rows

3.1.3 Example of the Data

To get a better idea about the format and information captured in the data, Table 2 shows two fictional registries from the labelled KvK 2023 dataset. This table showcases a selection of the key attributes that I used in the analysis, including the trade name in 45 positions ('HN45'), province ('PROV'), Standard Business Identifier ('SBI_CODE'), legal form ('RECHTSVORM'), registration date ('INSCHR_DAT'), total number of employees ('W_P_ TOTAAL'), textual description of the business activities ('Bedrijfsomschrijving'), and the target variable which indicates whether an organization is classified as a caring community (labelled as 1) or not (labelled as 0). A description of each attribute can be found in Appendix A.

Table 2: Example of fictional data from the KvK repository.

HN45	PROV	SBI_CODE	RECHTS-VORM	INSCHR_DAT	W_P_TOTAAL	Bedrijfsomschrijving	label
Stichting De Haan	A	88102	74	20130324	2	Aanbieden van dagbesteding	0
Het Buurt Huis	G	94997	61	20140231	0	Aanmoedigen van lokale welzijn	1

3.1.4 Data Exploration

As mentioned in section 3.1.2 Data Overview, the four subsets differ in size, considering the number of records and features (Table 1). After I merged the subsets that served as training data, it became clear that the target variable 'label' has a disproportionate distribution, with most records (67,9%) labelled as non-caring communities.

Furthermore, the various features have different relationships with the target variable as well as varying distributions. Appendix B illustrates a distribution analysis of the features and the target variable, which reveals several observations. First, there appears to be a negative relation between the number of employees and the possibility of a caring community (Figure B1). Second, Figure B2 shows that the legal forms most likely to represent a caring community are code 71 (association), 61 (cooperation), and 74 (foundation). Third, the SBI code for 'local social services' has the highest likelihood of representing a caring community (Figure B3). Fourth, the province of Flevoland contains the most caring communities and Overijssel and Zeeland the least, as shown in Figure B4. Last, examining the registration date shows a slight peak in caring communities' registrations in 1988 and 1989 and that most caring communities decided to register themselves after 2010 in general (Figure B5).

3.2 Data Preparation

Data preparation for this study involved several steps, including handling missing data, integrating different datasets, removing duplicates, and transforming the data.

3.2.1 Data Cleaning

Initially, the dataset contained a total of 20.268 records across all files. During the cleaning process, I had to discard several of these records because of missing, duplicate, or invalid values. As a result, I removed 1061 records, leaving a dataset containing 19.207 records.

Missing data I addressed various missing or invalid values across the datasets. I prioritized cleaning invalid or missing values in the features that I deemed useful for analysis (Appendix A). For example, for the target variable, I removed records with missing or illogical labels, such as '?'. Other variables such as dates also contained NaN values that I removed. Through this cleaning process, a total of 198 records were removed.

Data integration After cleaning both training datasets and ensuring they contained identical features, I combined them to create one larger set.

Duplicates I used the 'VGNUMMER' field to identify and remove duplicates. I used this feature, which contains unique establishment numbers, as the most suitable candidate for this purpose. This process eliminated one duplicate organization from the training data. For more information about the process, please refer to Appendix C.

Overlap in train/test sets When comparing the training set to the test sets (2022 and 2023 KvK data), I identified several overlapping data points: 322 between the training set and the 2022 set, 192 between the training set and the 2023 set, and 4,930 between the 2022 and 2023 datasets. While overlap within the separate test sets is not critical, the overlap between the training and test sets is crucial to tackle to prevent data leakage. Data leakage occurs when information from the training data influences the test set, leading to overfitting. I prevented this by removing the overlapping records from the test datasets, which totalled 514 records.

Feature selection All datasets contained a different number of features prior to cleaning. To address this, I selected common features present in all datasets and carefully evaluated their relevance by only including those deemed important to the analysis. The rationale behind the inclusion/exclusion of each feature is documented in detail in Appendix A. I use the following features: trade name, province, SBI code, legal form, registration date, total number of employees, and organizational description. These features were selected based on the exploratory analysis and insights from the Monitor study (Smellik et al., 2020). For instance, province emerged as a potentially valuable predictor, as both the data exploration and the Monitor study indicated that certain areas in the Netherlands have a higher concentration of caring communities. Although this concentration varies by province, the validity of using this feature in classifying organizations remains an open question.

Data transformation During the data transformation phase, I created labels that indicate whether an instance is a caring community or not. For specific implication details about this process, please refer to Appendix C. The training data of the Monitor survey (train_monitor) were all labelled as caring communities since this subset solely contained caring communities.

3.2.2 Data Labelling

To assess the performance of the classifiers, it was necessary to manually label a sample of the 2022 and 2023 test data. This labelled data, consisting of 350 records for each year (700 total), provided a benchmark for evaluating the performance of the classification algorithms.

The labelling process was conducted in collaboration with a fellow student, with guidance from an expert in the field of caring communities at Vilans. Prior to labelling, we met with the expert, who provided us with pointers to guide our assessment. For instance, she emphasized that community centres that focus solely on rental and playgrounds would not classify as caring communities.

When assessing the entries, we initially focused on specific features like 'bedrijfsomschrijving' (company description) as it contained the most detailed information. In instances of uncertainty, we conducted online research and visited organisation websites to gather further information, all while keeping in mind the provided definition. The definition used during the labelling of both the train_label dataset and samples of test_KVK2022 and test_KVK2023 can be found in Appendix D.

During the labelling process, we worked together closely and discussed any questionable instances to ensure consistency and accuracy. In cases where it was unclear, we reached out to the expert. Notably, we relied on external sources for about 90% of the instances when determining if a data point could be positively classified as a caring community. Without the additional information, these instances were unclassifiable. Conversely, labelling instances that were non-caring communities required almost no further investigation.

From the labelled data, we identified 31 (8,8%) records in the 2022 sample and 26 (7,4%) records in the 2023 sample that met the criteria for caring communities.

3.3 Ethical and Legal Considerations

This study utilizes data that were legally purchased from the KvK, with few usage restrictions. The purchase was conducted by Vilans according to the procedures set by the KvK, ensuring compliance with relevant regulations. However, it is important to note that while the data may be used for personal use and reuse, it is not allowed to resell the data or make it available in the same way the KvK does (KVK, 2021b). Furthermore, the personal data should be handled in compliance with the Trade Register Act, GDPR, and other applicable regulations (KVK, 2021a).

While the data is technically public, it includes sensitive information including the names of business owners. To protect the privacy of these individuals and adhere to the regulations associated with personal data, I removed all identifiable information during the data cleaning process, while only retaining business names. However, due to the public nature of the data, individuals can still access it through online searches.

The data were securely stored on a company laptop belonging to Vilans. Access to the data was strictly controlled, with only authorized personnel able to view the files. I maintained transparency throughout the research process through detailed documentation of the methodology, including data collection, cleaning, and analysis procedures. This ensures that the research can be replicated and verified by other researchers, upholding the principles of transparency and accountability.

4 Methods

This methods chapter outlines the process I undertook to address the research question: *'What is the most effective classifier for identifying registered 'caring communities' within repeated samples of Dutch Chamber of Commerce data?'*. The chapter covers data preprocessing, including one-hot encoding and text preprocessing. I also detail the classification algorithms used (LR, SVM, RF, GBT), their hyperparameter tuning, and evaluation metrics (accuracy, precision, recall, F1 score, ROC, AUC).

4.1 Data Preprocessing

After preparing the data, I preprocessed it by transforming categorical features, preparing text data for natural language processing (NLP) tasks, and employing LDA for topic modelling.

One-hot encoding I used one-hot encoding to represent all categorical features: province ('PROV'), legal form ('RECHTSVORM'), and SBI codes ('SBI_CODES'). This approach avoids potential issues with ordinal encoding, where assigning numerical values might introduce unintended ordering to inherently categorical variables. To address the challenge of sparsity in the data, I combined categories that appeared less frequently. For legal forms, I grouped categories appearing less than 10 times into an 'other' category. Similarly, for SBI codes, I combined categories occurring less than 100 times.

Text preprocessing I applied text preprocessing to three text features: 'HN45', 'H_NAAM_VOL', and 'Bedrijfsomschrijving'. The 'HN45' feature represents the trade name in 45 characters. 'H_NAAM_VOL' is the full name of the organization, and 'Bedrijfsomschrijving' is the business description. I tokenized, normalized, and cleaned the text by removing URLs, punctuation, and non-alphabetic characters. Additionally, I filtered out Dutch stopwords using the word_tokenize and stopwords libraries from the Natural Language Toolkit (NLTK) package. By using unicode normalization I ensured that the encoding was consistent. Finally, I applied lemmatization using the WordNetLemmatizer available in the NLTK package to reduce words to their base forms while preserving their true meaning, which is more effective than stemming for maintaining word integrity. I then used the preprocessed text to create a Bag-of-Words (BoW) representation for further analysis.

Text representation To capture the thematic structure within the textual data, I employed Latent Dirichlet Allocation (LDA) (Blei et al., 2003) for topic modelling. To determine the optimal number of topics for LDA, I utilized coherence score (C_V) as the evaluation metric since it correlates best with human judgement of topic quality (Röder et al., 2015). Coherence measures the semantic similarity between words within a topic, and higher scores indicate better topic quality. I experimented with different topic numbers, specifically in the range of 1 to 50, increasing by steps of 10 (k=1, 10, 20, 30, 40, 50). Initially, I planned to select the number of topics that yielded the highest score before flattening out. However, I observed that the coherence score kept increasing with the number of topics, making it impractical to select the one with the highest score. After analysing the trends, I determined the optimal number of topics (k) = 11, as this seemed to yield the highest score before a major drop.

4.2 Classification Algorithms

After preprocessing the data, I trained, tuned, and evaluated four machine learning models: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting Tree (GBDT). The aim was to identify the best-performing model for the given classification task of identifying caring communities. I employed a systematic approach to hyperparameter tuning and model evaluation.

4.2.1 Train, Validation and Test Data

For tuning, I used the entire training set, which consisted of the combined train_label and train_monitor sets. To split this data into training and validation data I used K-fold cross-validation. The training data is used to train models with different hyperparameter combinations during GridSearchCV. The validation set is used to evaluate these models and select the best hyperparameters. The 2022 and 2023 data served as held-out test data for the final model evaluation. For this final evaluation, the entire training set was used for training the models with the best hyperparameters.

This approach offers several advantages. First, the train-validation split ensures that the data used for tuning hyperparameters (training set) is separate from the data used to evaluate the model (validation set). This helps in avoiding bias and selecting the best model configuration. Second, using separate held-out test data prevents overfitting, as the final model is evaluated on completely unseen data. Last, it provides a robust evaluation of the model's performance on new, unseen data, ensuring it generalizes well.

4.2.2 Hyperparameter Tuning

Each model's hyperparameters were optimized using GridSearchCV from scikit-learn with cross-validation. I chose GridSearchCV because it performs an exhaustive search over a specified parameter grid, exploring all combinations of hyperparameters within the grid. This method ensures the identification of the best possible set of hyperparameters for the specific dataset and model combination. While RandomizedGridSearchCV is a faster alternative as it samples a fixed number of parameter settings from the specified grid, it might skip over the most optimal combination, making GridSearchCV a more robust choice for thorough optimization. Furthermore, given the dataset's manageable size, I considered GridSearchCV computationally feasible and effective.

Within GridSearchCV, I employed 10-fold cross-validation for each hyperparameter combination, as illustrated in Figure 1. This method involves splitting the training data into 10 equally sized folds. For each iteration (i.e. hyperparameter combination), a model is trained on nine folds, and evaluated on the remaining held-out validation fold. This process is repeated for all 10 folds. The results from all folds are then averaged to increase the model adaptability and reduce overfitting (Shatnawi et al., 2022). To ensure the model is particularly sensitive to identifying caring communities, I used recall as the scoring metric. GridSearchCV ultimately selects the combination that achieves the highest average score across all 10 folds of the cross-validation process. This way, the model is trained and validated on different subsets of the data, providing a more reliable evaluation of its performance. The choice of 10 folds balances

computational efficiency with robust model evaluation, offering a good compromise between bias and variance.

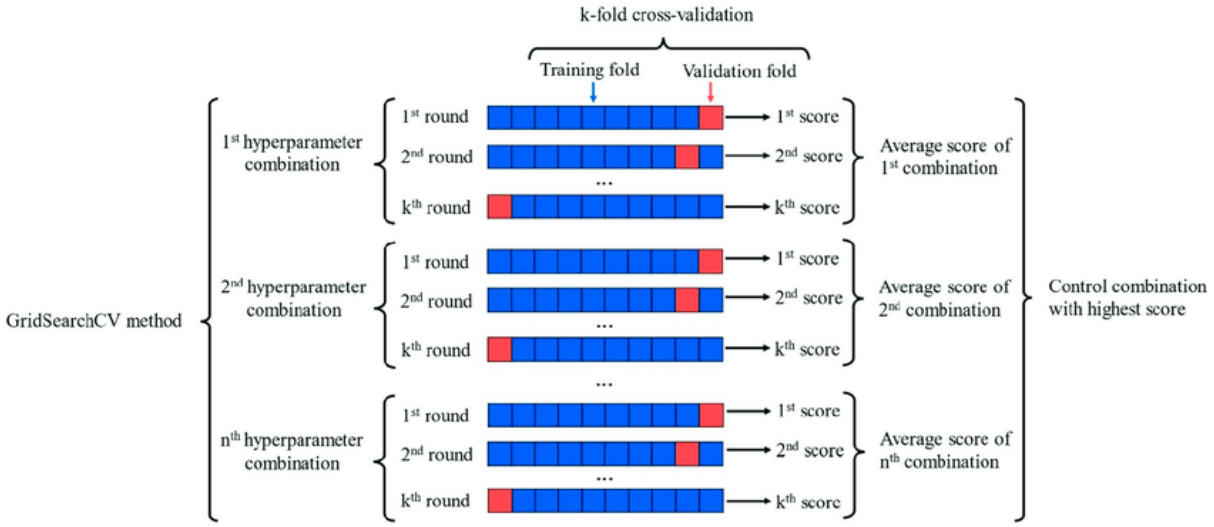

Figure 1: Hyperparameter tuning using 10-fold cross-validation (GridSearchCV) (Shatnawi et al., 2022).

In the following sections, I will detail for each model which hyperparameters I tuned, which hyperparameter grid I explored, and what hyperparameter combination yielded the best result. The repository containing the code for this study is available on GitHub. However, the data is not openly available and therefore not included in this repository.

4.2.3 Logistic Regression

LR is a linear model used for binary classification that estimates the probability of a binary outcome based on one or more predictor variables. It works by fitting a logistic function to the data, thus transforming a linear combination of the predictors to fall within the range of [0, 1]. For this task, I utilized the LogisticRegression function from the scikit-learn library. The hyperparameters I tuned and the options I explored were:

- **penalty**: Type of regularization used.
  - Options: 'l1' (Lasso), 'l2' (Ridge), 'elasticnet', and None.
- **C:** Regularization parameter that controls the strength of regularization, with smaller values specifying stronger regularization.
  - Options: 100, 10, 1.0, 0.1, 0.01
- **solver**: Algorithm to use for optimization.
  - Options: 'lbfgs', 'newton-cg', 'liblinear', 'saga', and 'sag'.

Using GridSearchCV and evaluating all combinations using 10-fold cross-validation, I found the following hyperparameter combination to yield the best performance: {C: 100, penalty: 'l1', solver: 'liblinear'}.

I chose these parameters and grid to balance between different types of regularization and optimization algorithms, ensuring a comprehensive search across regularization strengths and solvers to find the most effective combination.

4.2.4 Support Vector Machine

SVM is a supervised learning model used for classification and regression. It works by finding the hyperplane that best divides a dataset into classes. The SVM can handle non-linearly separable data by using kernel functions to map the data into higher dimensions. I used the SVC function from scikit-learn. For SVM, the hyperparameters I tuned and the choices I explored were:

- **C**: regularization parameter that controls the trade-off between achieving a low training error and a low testing error.
    - Options: 0.1, 1, 10, 100, 1000, 1900, 2000, 2100, 3000
- **gamma**: kernel coefficient for 'rbf', 'poly', and 'sigmoid'. It determines the influence of a single training example.
    - Options: 10, 1, 0.1, 0.01, 0.001, 0.0001
- **kernel**: type of kernel.
    - Options: 'rbf', (‘linear’, ‘polynomial’, ‘sigmoid’)

After exploring different kernel functions during testing, I found that only the 'rbf' kernel yielded results on this dataset. The other kernel functions (such as linear, polynomial, and sigmoid) did not show any progress during training, which I observed in verbose mode. This lack of progress indicated that these kernels were not suitable for the data, failing to fit the model effectively.

To decide on the range for the C parameter, I started with a wide range of values. The theory for determining how to set C is not very well developed, so I chose this range based on exploratory trials.

The best hyperparameter combination found was: {C: 1900, gamma: 0.001, kernel: 'rbf'}.

4.2.5 Random Forest

RF is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the class which is the mode of the classes of the individual trees. It improves the predictive accuracy and controls overfitting by averaging the results of various decision trees. I used the RandomForestClassifier from scikit-learn. For RF, I optimized the following hyperparameters and considered these options:

- **n_estimators**: number of trees in the forest.
    - Options included a range with 10 evenly spaced values between 50 and 100.
- **max_features**: number of features to consider when looking for the best split.
    - Options: None, 'log2', 'sqrt'
- **max_depth**: maximum depth of the tree.
    - Options: 2, 4, 6, 10, 12, 15, 20, None

- **min_samples_split**: minimum number of samples required to split an internal node.
    - Options: 2, 5, 10
- **min_samples_leaf**: minimum number of samples required to be at a leaf node.
    - Options: 1, 2, 5
- **bootstrap**: method for sampling data points when building trees (with or without replacement).
    - Options: True, False

I found the best combination of hyperparameters to be: {bootstrap: False, max_depth: 12, max_features: None, min_samples_leaf: 2, min_samples_split: 10, n_estimators: 50}.

4.2.6 Gradient Boosting Tree

GBDT is an ensemble technique that builds trees sequentially, each new tree correcting errors made by the previous ones. It combines the predictions of several base estimators to improve robustness and accuracy. I utilized the GradientBoostingClassifier from scikit-learn for this task.

For GBDT, I tuned fewer hyperparameters compared to, for example, RF, since Zhang et al. (2017) found that GBDT matches or exceeds the prediction performance of most classifiers, even with non-exhaustive hyperparameter tuning. The hyperparameters I adjusted and the options I experimented with were:

- **n_estimators**: number of boosting stages to be run.
    - Options included a range with 10 evenly spaced values between 50 and 200.
- **learning_rate**: how much each tree contributes to the overall model.
    - Options: 0.01, 0.05, 0.1, 0.2
- **max_depth**: maximum depth of the individual trees.
    - Options: 2, 4, 6, 10, 12, 15, 20, None

The best hyperparameter combination I found was: {learning_rate: 0.2, max_depth: 2, n_estimators: 133}.

4.3 Evaluation

After tuning the hyperparameters, I retrained each model on the entire training set using the best hyperparameters identified through GridSearchCV. This ensured that the models were optimized for performance before being evaluated on the testing sets. The testing sets contained the manually coded 2022 and 2023 KvK subsets. For evaluation, I leveraged the manually labelled samples of the held-out test sets. To assess the strength of each classifier I utilized the following performance metrics: accuracy, precision, recall, F1 score, Receiver Operating Characteristic (ROC) curve, and Area Under the Curve (AUC).

- **Accuracy** is the most frequently used metric that measures the proportion of correctly classified instances among the total instances. While included, I used this measure with caution since accuracy can yield misleading high performances in imbalanced data (He & Garcia, 2009; Thölke et al., 2023).

- **Precision** indicates the proportion of true positive predictions among all positive predictions made by the model. It reflects the accuracy of positive classifications.
- **Recall** also known as sensitivity or true positive rate, measures the proportion of actual positive cases correctly identified by the model.
- The **F1 score** provides a harmonic mean of precision and recall, providing a single metric that balances both. It is particularly useful for this dataset as it contains imbalanced classes and considers both true positives and true negatives.
- **ROC curve** is a graphical representation of the model's performance across different threshold values. It plots the true positive rate (recall) against the false positive rate.
- The **AUC** measures the area under the ROC curve and provides a summarized measure of the model's performance. AUC values range from 0 to 1, with higher values indicating better performance. This evaluation measure, together with ROC, provides a more reliable performance evaluation for imbalanced data (He & Garcia, 2009; Thölke et al., 2023).

To determine the best overall model, I calculated the average of each performance metric across the two test datasets (KvK 2022 and KvK 2023). This approach allowed me to obtain a consolidated view of each model's performance, considering variations and consistency across different time periods. Specifically, the average performance metrics were computed as follows:

$$Average\ Metric = \frac{Metric_{2022} + Metric_{2023}}{2}$$

By averaging the performance metrics, I ensured that the chosen model is robust and generalizable across different temporal datasets since the impact of any anomalies or outliers present in a single test dataset is mitigated.

5 Results

In this chapter, I examine the performance of the machine learning models used to identify caring communities. In section 5.1 I will explore the strengths and weaknesses of each model by analysing various evaluation metrics like accuracy, precision, recall, F1 score, and AUC. I will also compare the performance of the models across two years (2022 and 2023) to assess their consistency and robustness. Besides, in section 5.2 I will analyse the average performance to determine the best overall model and discuss the number of caring communities this model identified.

In section 5.3, I will conduct an error analysis to understand why the models struggled to identify caring communities. This analysis focuses on false negatives, which are caring communities misclassified as non-caring. By examining consistently misclassified instances, I will hypothesize common weaknesses in the model's ability to accurately classify caring communities.

5.1 Performance Analysis

After evaluating the models, I compared the performance metrics for each model to identify the best-performing model. I specifically focussed on the recall, F1 score and AUC. Recall is important to consider since identifying the positive records (caring communities) is the specific focus of this research. The F1 score balances precision and recall, and AUC provides a comprehensive measure of model performance.

Each model's performance metrics from 2022 and 2023 are presented in Table 3 and 4, respectively.

Table 3: Performance metrics for each model on the 2022 test set. The highest score for each metric is highlighted in bold.

Model	Accuracy	Precision	Recall	F1 score	AUC
Logistic Regression	0.85	0.27	**0.42**	**0.33**	**0.83**
Support Vector Machine	0.86	0.23	0.26	0.24	0.72
Random Forest	0.84	0.16	0.19	0.17	0.74
Gradient Boosting Tree	**0.88**	**0.33**	0.32	**0.33**	0.79

In the 2022 test set, all models achieved relatively high accuracy (above 0.8). However, other performance metrics revealed their limitations in identifying caring communities (as measured by precision, recall and F1 score). SVM had moderate recall (around 0.3) but suffered from low recall (around 0.2). This indicates it missed a substantial number of true positives (actual caring communities) while also misclassifying some non-caring communities as caring. RF exhibited even lower precision and recall rates (below 0.2). This shows that this model frequently misclassified both caring and non-caring communities. LR achieved the highest recall, F1 score and AUC. GBDT had the highest accuracy, precision and tied F1 score. However, its recall was lower than LR. Based on the combined analysis of all metrics**, both LR and GBDT emerged as the top performers**.

Table 4: Performance metrics for each model on the 2023 test set. The highest score for each metric is highlighted in bold.

Model	Accuracy	Precision	Recall	F1 score	AUC
Logistic Regression	**0.92**	**0.44**	**0.42**	**0.43**	**0.88**
Support Vector Machine	0.86	0.19	0.27	0.23	0.76
Random Forest	0.76	0.07	0.19	0.11	0.67
Gradient Boosting Tree	0.84	0.16	0.27	0.20	0.77

Similar to the 2022 results, models in the 2023 test set achieved high accuracy scores. However, examining other metrics again revealed limitations. SVM and GBDT exhibited similar performance, with moderate recall (around 0.3) but concerningly low precision (around 0.2). Also, the performance of RF decreased even further compared to 2022, with the lowest performance rates across all metrics. This makes this model the least suitable for identifying caring communities. **LR on the other hand showed the best performance across all metrics**, with the highest accuracy, precision, recall, F1 score, and AUC.

Additionally, the average performance metrics across both test sets are summarized in Table 5.

Table 5: Average performance metrics for each model. The highest score for each metric is highlighted in bold.

Model	Accuracy	Precision	Recall	F1 score	AUC
Logistic Regression	**0.87**	**0.35**	**0.42**	**0.38**	**0.85**
Support Vector Machine	0.86	0.21	0.26	0.23	0.74
Random Forest	0.80	0.12	0.19	0.14	0.70
Gradient Boosting Tree	0.86	0.24	0.30	0.26	0.78

In general, the **LR model showed the highest overall performance** across all metrics. GBDT also demonstrates solid overall performance. However, its metrics were slightly lower compared to LR. This suggests GBDT might prioritize correctly identifying caring communities while keeping false positives low, potentially at the expense of missing some true positives. The RF model presents an entirely different picture with consistently low scores across all metrics (accuracy, precision, recall, F1 and AUC). This implies it struggles to accurately identify caring communities. Finally, the SVM model fell between the top and worst-performing models. While its recall is slightly better compared to RF, its overall precision remained low. This suggests SVM might struggle to identify a substantial number of true positives while also misclassifying some non-caring communities as caring.

Upon examining the ROC curves of the 2022 data (Figure 2), the **LR model demonstrated the best performance** with an AUC of 0.83. This indicates that the LR model is the most effective at distinguishing between positive and negative classes. The curve showed a consistent improvement in the true positive rate as the false positive rate increased, suggesting that this model maintained a reasonable balance between sensitivity (identifying true positives) and specificity (avoiding false positives). The GBDT model exhibited slightly lower performance compared to LR but still demonstrated good discriminative ability. SVM and RF performed similarly, with AUC scores around 0.73. This is substantially lower compared to LR

and GBDT which had values around 0.8. Besides, the ROC curve for RF showed a different trend compared to SVM, with less of a curve and more of a straight line. This suggests limitations in its ability to distinguish classes.

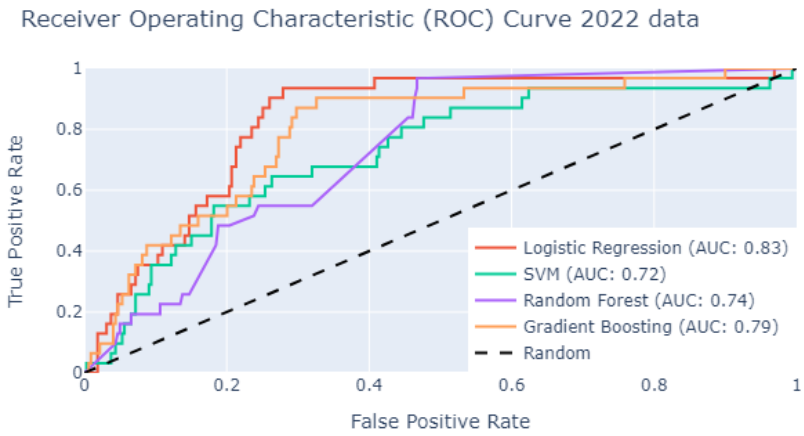

Figure 2: ROC curves for each model on the 2022 test set.

For the 2023 data, the ROC curves showed more variation but similar AUC trends. The **LR model again achieved the highest AUC score** (0.88), indicating its best overall performance in distinguishing between caring and non-caring communities. The SVM and GBDT models followed with an AUC of around 0.77. This demonstrates their robustness in classification tasks and indicates that these models performed reasonably well, although they were less discriminative compared to LR. RF remained the least discriminative model, with the lowest AUC score.

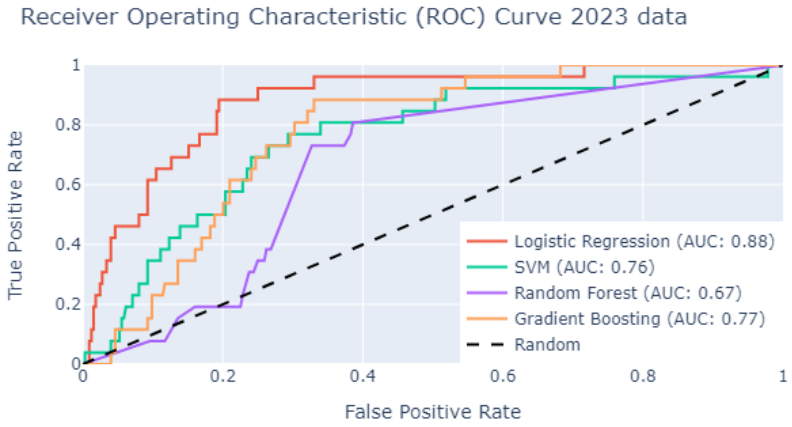

Figure 3: ROC curves for each model on the 2023 test set.

By comparing the two years, it is evident that the performance of the models was quite similar between 2022 and 2023. The **LR consistently performed well across both years**, highlighting its reliability in classification tasks. While GBDT offered a strong alternative, SVM and RF showed limitations in discriminative ability across both datasets.

5.2 Number of Predicted Caring Communities

In applying the best-performing classifier to the datasets, I identified a total of **1004 ± 58 caring communities** in the 2022 dataset and **947 ± 58 caring communities** in the 2023 dataset. The values are accompanied by 95% confidence intervals, which indicate the range within which the true number of caring communities likely falls.

5.3 Error Analysis

To understand why these models performed poorly overall and specifically failed to identify many caring communities, I conducted an error analysis. In this analysis, I specifically focused on false negatives. In this context, false negatives are caring communities incorrectly classified as non-caring communities. This analysis is crucial because high recall is essential for this study, as it prioritizes capturing all relevant caring communities.

The analysis revealed a steady pattern for the 2022 and 2023 datasets: a set of 12 unique data points in 2022 and 11 in 2023 consistently misclassified as non-caring by all models, with varying frequencies of misclassification. By examining these consistently misclassified instances I revealed shared weaknesses of the models in their ability to accurately identify caring communities. I provided four examples for which I noted the business description and hypothesized what could have possibly caused the classifiers to fail:

1. Corlian Mooibroek
    - Description: provides care and support for children and young people with disabilities.
    - Possible reasons for misclassification:
        - Personal name: the use of a personal name rather than a descriptive organization name might have confused the models.
        - Specific niche: the specific focus on children and young people with disabilities might not have aligned with the training data used for the models.
2. Senioren Advies Bureau
    - Description: offers advisory services to seniors.
    - Possible reasons for misclassification:
        - Service type: advisory services might not be as strongly associated with direct caregiving in the model's understanding.
3. NAH Coach
    - Description: provides coaching and guidance to individuals with acquired brain injuries.
    - Possible reasons for misclassification:
        - Specific niche: the specific focus on individuals with acquired brain injuries might not have aligned with the training data.
        - Ambiguous keywords: the description might not have emphasized caregiving-related keywords strongly enough.
4. Global Care Capacity BV
    - Description: provides home care and support to care recipients.
    - Possible reasons for misclassification:

- Business structure: the "BV" (limited company) designation might have led the model to misclassify it as a business.
- Generic keywords: the description might not have contained enough specific caregiving-related keywords.

6 Discussion

This chapter dives into the theoretical and practical contributions, the limitations of this research, potential future research possibilities, and the ethical implications and considerations associated with my study.

6.1 Contributions

This study has several theoretical, but mostly practical contributions. In terms of theoretical contributions, this study is among the first to explore the potential of machine learning models to identify registered caring communities using Chamber of Commerce data. While challenges remain, the findings provide valuable insights for future research in this area, which will be discussed in section 6.3. For instance, the error analysis revealed that models struggle to identify communities with specific niches (e.g., children with disabilities). This insight suggests that future research should explore developing methods to manage the diverse nature of caring communities. Second, my results highlight the importance of considering various performance metrics beyond just accuracy. By focusing on metrics like recall and F1 score, the limitations of some models in identifying true positives (caring communities) were revealed. This emphasizes the need for careful selection and evaluation of machine learning models in this context.

From a practical perspective, the findings of this study have important implications for policymakers and stakeholders involved in community welfare that are concerned with using the outputs of this classification algorithm to gain insights into caring communities in the Netherlands. The use of such algorithms is not yet good enough to put into practice since the recall rates do not suffice. Using this algorithm has the potential to get an inaccurate picture of the total number of caring communities, which can ultimately lead to wrongful conclusions drawn by policymakers who decide to trust these models. However, after addressing the challenges connected to this research, the algorithm has the potential to inform decision-making processes aimed at resource allocation and policy development regarding community support interventions.

6.2 Limitations

In conducting this study, several limitations emerged that should be acknowledged. Many of these limitations emerged from the data itself. The data sources used, while offering valuable insights, presented certain challenges that affected the overall analysis.

The first limitation concerns the data source, the KvK registration database. While the KvK offers a structured overview of registered organizations, the data does not contain many details on aspects such as the types of business activities carried out. Those details are crucial when it comes to identifying caring communities due to their diverse nature. As mentioned in chapter 2, caring communities facilitate a wide range of activities and services, from providing meals to professional services like community nursing (Movisie, 2020; NZVE, n.d.). While manually labelling the data, I noticed that additional information from external sources, like websites, was crucial to accurately classify them as caring or non-caring communities. This reliance on

supplementary data introduces constraints in the classification process since the algorithm does not have access to this data.

Moreover, not all caring communities in the Netherlands are registered with the KvK. According to the Monitor Caring Communities 2020 (Smellik et al., 2020), approximately 20% of community initiatives operate without formal registration (Zoest et al., 2023). These unregistered initiatives often include informal groups that function outside legal frameworks or perceive registration as unnecessary. Therefore, by focusing solely on KvK data, this study overlooks a substantial portion of the caring community landscape, limiting the comprehensiveness of its findings.

The quality and consistency of business descriptions within the KvK database most likely also had a significant impact on the performance of the classifiers. As highlighted by Litofcenko et al. (2020), the accuracy of machine learning models highly depends on the clarity and uniformity of input texts. Their experience has shown that the quality of input texts is key. From the error analysis, I suspect that variations in terminology, abbreviations, or incomplete descriptions were part of the reason the classifiers struggled to accurately identify and categorise caring community initiatives.

Additionally, during manual labelling, for which I often had to consult external sources, I discovered many initiatives to be inactive. This highlights a crucial consideration: the operational status of caring communities. Policymakers are primarily interested in the number of active communities, unfortunately, some of the subsets I used lacked a specific "activity status" column, namely 'OPHEFF_DAT' which contained the dissolution date. This feature was not included in the purchase of the 2023 data by Vilans.

Consequently, these limitations prevented me from providing an accurate number of total (active) caring communities in the Netherlands.

6.3 Future Research

This study establishes a foundation for future research on identifying caring communities using machine learning. Several avenues for future research should be considered to enhance the methodology and broaden the scope of this research.

Data quality and training From the insights from the error analysis, I identified multiple areas for improvement. First, ensuring descriptions and organization names indicate caregiving roles can significantly improve the relevance and accuracy of input data. Second, incorporating a wider variety of activities and organization types into the training data is crucial. This step ensures the models are trained on a representative sample of the caring community landscape and can therefore better capture the nuances and diversity within caring communities.

Text representations Future research could also address the mentioned limitations by exploring other text representations and data enrichment strategies. As noted by Litofcenko et al. (2020), improving the quality of input texts could significantly enhance performance. They managed to substantially improve the quality by preselecting the relevant features of the input texts. Investigating how this strategy can be adapted to this study's context could provide

valuable insights into improving classification accuracy. Other techniques could also be explored to enhance the text inputs.

Expanding the data Exploring alternative data sources beyond the KvK and/or integrating external datasets could enhance the understanding of caring communities, including informal and unregistered initiatives. This may involve incorporating sources such as community surveys, social media analytics, or qualitative interviews alongside registration data. By enriching the dataset with supplementary information that is not captured by registration data alone, the broader spectrum of caring communities could be captured better, which in turn enhances classification accuracy. These approaches could help mitigate challenges associated with data quality and completeness.

Algorithm refinement Additionally, continual refinement of machine learning algorithms, including ensemble methods and deep learning architectures, could enhance the robustness and predictive power of classification models. Future work may involve further tuning of hyperparameters and exploring additional classification techniques to enhance model performance.

Abstain classifiers Exploring the concept of 'abstain' classifiers, which explicitly abstain from making a classification when confidence is low, could also be a promising approach (Xin et al., 2021). These classifiers could help mitigate misclassification risks by identifying instances where data quality or classifier confidence is insufficient to make a reliable prediction.

Collaboration with KvK An alternative, possibly unworkable, approach to get a better answer to the question of how many caring communities the Netherlands contains, could involve collaborating with the KvK to create a dedicated category for caring communities within their registration system. This new category would allow caring communities to self-identify during registration, improving the accuracy and completeness of data on caring communities within the KvK database.

Incentivizing registration It is important to acknowledge that some initiatives, particularly those with a more informal structure or those who view registration as unnecessary, might remain outside the scope. In combination with collaborating with the KvK, policymakers or stakeholders could incentivise caring communities to register themselves. This would provide the perfect solution to the question that this research is ultimately trying to answer: ‘How many caring communities are there in the Netherlands?’. Policymakers could consider initiatives to streamline registration processes and enhance public awareness of the benefits of formal registration.

In summary, while this study represents a significant step in utilizing KvK data for classifying caring communities, it also highlights the critical need for methodological advancements and broader data considerations to improve the identification of caring communities. By addressing these limitations, future research can contribute to a more comprehensive understanding of the caring community landscape and inform effective policy interventions that support these vital social initiatives.

6.4 Ethical Considerations

Developing a machine learning classifier to identify registered caring communities raises several ethical concerns. First of all, a key concern is that not all caring communities are registered with the KvK (Zoest et al., 2023). This means that the classifier unintentionally excludes many unregistered caring communities that operate outside formal frameworks. This exclusion could skew policy due to the underrepresentation of certain caring communities, potentially leading to unintended consequences like underfunding or lack of recognition. Another limitation arises from using only a portion of the available KvK data due to cost constraints. This potentially creates a biased sample that does not reflect the full diversity of caring communities. Selecting data based on SBI codes systematically excludes specific "categories" of communities. Consequently, the classifier might not recognize or accurately classify these omitted categories, overlooking significant parts of the caring community landscape. Furthermore, even the best-performing classifier, chosen based on metrics like recall and precision, might still struggle to identify caring communities adequately due to their highly diverse nature and lack of training data. This shortcoming could mislead policymakers about the number and types of caring communities, further hindering deserving communities from receiving policy consideration or resource allocation.

The limitations have the potential to lead to unintended consequences, including unequal resource distribution, overlooking uncommon communities, and reinforcing inequities in community support and recognition. These consequences in turn might impact the caring communities' ability to thrive and serve their residents effectively.

One way to address these ethical concerns is by enhancing data inclusivity. This entails exploring ways to incorporate data from unregistered communities and expanding the training data, so the classifier becomes more familiar with the diverse nature of caring communities.

By acknowledging these ethical implications and considerations, my research aims to contribute positively to the understanding and classification of caring communities. I recognize the importance of continuing to refine methods to better capture the diverse and complex nature of caring communities, thereby promoting fairer policy-making.

7 Conclusion

This study aimed to answer the following research question: *'What is the most effective classifier for identifying registered 'caring communities' within repeated samples of Dutch Chamber of Commerce data?.'* To answer this question, I cleaned and preprocessed the data, and performed hyperparameter tuning to optimize classifier performance. Thereafter I evaluated the performance of Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting Tree (GBDT) models based on key metrics including accuracy, precision, recall, F1 score, AUC, and ROC. This provided a comprehensive assessment of their strengths and weaknesses.

The results have shown that across 2022 and 2023, LR consistently demonstrated superior performance among the models across all metrics. Although GBDT showed competitive performance, it was slightly inferior to LR. GBDT not being the top-performing was somewhat surprising since Zhang et al. (2017) found this model to exceed the prediction performance of LR and the other two classifiers, even with non-exhaustive hyperparameter tuning. However, the difference with LR is marginal and therefore negligible. In contrast, SVM and RF did not prove to be effective, specifically demonstrated by their low recall rates.

Based on these findings, I conclude that the LR model emerged as the most effective classifier for identifying registered caring communities within the Dutch Chamber of Commerce data. Its high performance across all metrics indicates its suitability for this task, though enhancements in recall are necessary for more reliable identification.

Using LR to investigate the number of caring communities revealed a total of 1004 ± 58 caring communities in the 2022 dataset and 947 ± 58 caring communities in the 2023 dataset. The values are accompanied by 95% confidence intervals, indicating the range within which the true number of caring communities likely falls.

Despite LR's strengths, I revealed significant limitations that must be addressed. The error analysis, focused on false negatives, highlighted challenges such as the misclassification of communities with specific niches and generic or ambiguous business descriptions. Other limitations included the reliance on external data for accurate labelling and the lack of diverse training data. These issues underscore the need for further refinement in data quality, algorithmic robustness, inclusivity of diverse community types, and the exploration of alternative data sources. Policymakers and stakeholders should consider the nuanced challenges highlighted in this study when using classification algorithms to inform resource allocation and policy decisions.

In conclusion, while LR demonstrates promising performance in identifying registered caring communities, ongoing research and methodological advancements are essential to enhance the accuracy of classification models. Using this algorithm has the potential to get an inaccurate picture of the total number of caring communities, which can ultimately lead to wrongful conclusions drawn by policymakers who decide to trust these models. However, by addressing the challenges, future studies can contribute to a more comprehensive understanding of community welfare initiatives and support fair policy interventions.

References

Aljawazneh, H., Mora, A. M., García-Sánchez, P., & Castillo-Valdivieso, P. A. (2021). Comparing the Performance of Deep Learning Methods to Predict Companies' Financial Failure. *IEEE Access*, *9*, 97010–97038. https://doi.org/10.1109/ACCESS.2021.3093461

Allozi, Y., & Abbod, M. (2022). Predicting Business Failure Using Neural Networks: An Empirical Comparison with Statistical Methods and Data Mining Method. In L. Troiano, A. Vaccaro, N. Kesswani, I. Díaz Rodriguez, & I. Brigui (Eds.), *Progresses in Artificial Intelligence & Robotics: Algorithms & Applications* (pp. 146–156). Springer International Publishing. https://doi.org/10.1007/978-3-030-98531-8_15

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*(null), 993–1022.

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

Cabrera, A. (1994). "Logistic Regression Analysis in Higher Education: An Applied Perspective. In *Higher Education: Handbook of Theory and Research(225-256)* (Vol. 10, pp. 225–256).

CBS. (2020, July 13). *2. Woonsituatie* [Webpagina]. Centraal Bureau voor de Statistiek. https://www.cbs.nl/nl-nl/longread/statistische-trends/2020/laatste-levensjaren-tachtigplussers/2-woonsituatie

Chaskin, R. J. (2001). Building Community Capacity: A Definitional Framework and Case Studies from a Comprehensive Community Initiative. *Urban Affairs Review*, *36*(3), 291–323. https://doi.org/10.1177/10780870122184876

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*(3), 273–297. Scopus. https://doi.org/10.1023/A:1022627411411

Daalhuizen, F., Dam, F. van, Groot, C. de, Schilder, F., & van der Staak, M. (n.d.). *Zelfstandig thuis op hoge leeftijd*. Retrieved 19 May 2024, from https://themasites.pbl.nl/zelfstandig-thuis-hoge-leeftijd

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, *15*, 3133–3181. Scopus.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Harrell, F. E. (2015). Introduction. In *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis* (pp. 1–11). Springer International Publishing. https://doi.org/10.1007/978-3-319-19425-7_1

He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

KVK. (n.d.). *Over KVK*. KVK. Retrieved 27 May 2024, from https://www.kvk.nl/over-kvk/

KVK. (2021a). *Regels rond het gebruik van Handelsregistergegevens*.

KVK. (2021b). *Wat mag wel/niet met data uit het Handelsregister*.

KVK. (2023, February 14). *Overview Standard Business Categories (SBI codes)*. KVK. https://www.kvk.nl/en/about-the-business-register/overview-standard-business-categories-sbi-codes/

LePere-Schloop, M. (2022). Nonprofit Role Classification Using Mission Descriptions and Supervised Machine Learning. *Nonprofit and Voluntary Sector Quarterly*, *51*(5), 1207–1222. https://doi.org/10.1177/08997640211057393

Litofcenko, J., Karner, D., & Maier, F. (2020). Methods for Classifying Nonprofit Organizations According to their Field of Activity: A Report on Semi-automated Methods Based on Text. *Voluntas: International Journal of Voluntary and Nonprofit Organizations*, *31*(1), 227–237.

Ma, J. (2021). Automated Coding Using Machine Learning and Remapping the U.S. Nonprofit Sector: A Guide and Benchmark. *Nonprofit and Voluntary Sector Quarterly*, *50*(3), 662–687. https://doi.org/10.1177/0899764020968153

Macià, N., & Bernadó-Mansilla, E. (2014). Towards UCI+: A mindful repository design. *Information Sciences*, *261*, 237–262. https://doi.org/10.1016/j.ins.2013.08.059

McLeroy, K. R., Norton, B. L., Kegler, M. C., Burdine, J. N., & Sumaya, C. V. (2003). Community-Based Interventions. *American Journal of Public Health*, *93*(4), 529–533. https://doi.org/10.2105/AJPH.93.4.529

Movisie. (2020, December 14). *Explosieve groei van zorgzame gemeenschappen | Movisie*. https://www.movisie.nl/artikel/explosieve-groei-zorgzame-gemeenschappen

Netwerk DAK. (n.d.). *Czaar 51*. Retrieved 21 June 2024, from https://netwerkdak.nl/organisaties/czaar-51/

NZVE. (n.d.). *Definitie*.

NZVE, Vilans, & Movisie. (2019). *De organiserende burger: Leerprogramma en monitor bewonerscollectieven*.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399–408. https://doi.org/10.1145/2684822.2685324

Salamon, L., & Anheier, H. (1992). In search of the non-profit sector II: The problem of classification. *Voluntas*, *3*, 267–309. https://doi.org/10.1007/BF01397460

Salamon, L., & Anheier, H. (1996). *THE INTERNATIONAL CLASSIFICATION OF NONPROFIT ORGANIZATIONS:*

Shatnawi, A., Alkassar, H., Moneem, N., A. Al-Hamdany, E., Bernardo, L., & Imran, H. (2022). Shear Strength Prediction of Slender Steel Fiber Reinforced Concrete Beams Using a Gradient Boosting Regression Tree Method. *Buildings*. https://doi.org/10.3390/buildings12050550

Smellik, J., Heijden, N. van der, Sok, K., Schaijk, R. van, Stouthard, L., & Zoest, F. van. (2020). *Monitor Zorgzame Gemeenschappen* (p. 60).

Thölke, P., Mantilla-Ramos, Y.-J., Abdelhedi, H., Maschke, C., Dehgan, A., Harel, Y., Kemtur, A., Mekki Berrada, L., Sahraoui, M., Young, T., Bellemare Pépin, A., El Khantour, C., Landry, M., Pascarella, A., Hadid, V., Combrisson, E., O'Byrne, J., & Jerbi, K. (2023). Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. *NeuroImage*, *277*, 120253. https://doi.org/10.1016/j.neuroimage.2023.120253

Vilans. (2024, June 18). *Wie zijn we?* vilans_nl. https://www.vilans.nl/wie-zijn-we

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., & Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1–37. https://doi.org/10.1007/s10115-007-0114-2

Xin, J., Tang, R., Yu, Y., & Lin, J. (2021). The Art of Abstention: Selective Prediction and Error Regularization for Natural Language Processing. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1040–1051). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.84

Zhang, C., Liu, C., Zhang, X., & Almpanidis, G. (2017). An up-to-date comparison of state-of-the-art classification algorithms. *Expert Systems with Applications*, *82*, 128–150. https://doi.org/10.1016/j.eswa.2017.04.003

Zoest, F. van. (2023). *Projectplan: KCS Werkprogramma 2024*.

Zoest, F. van, Stouthard, L., & Zondervan, M. (2023). *Verslag analyse 2023*.

Appendices

Appendix A - Description of Features

Table A1: Overview of the features in the datasets, including the name, a description, which dataset the feature is in, and the reason for including or excluding it from further analysis.

Name	Description	2022 data	2023 data	Monitor data	Label data	Reason for inclusion or exclusion
RGL	Register letter	X	X	X	X	No, same value for all entries
DOSSIER	Dossier number	X	X	X	X	No, this attribute is not relevant for classification
VGNUMMER	Establishment number	X	X	X	X	No, this unique identifier is not relevant for classification
HN1X30	Trade name 1 x 30 positions	X	X	X	X	No, HN45 provides a more complete version of the trade name
STRVA	Street /house number/addition of the establishment address	X	X	X	X	No, it is a proxy for address. PROV provides a more general and useful version of the address
PCPLVA	Postal code and city of the establishment address	X	X	X	X	No, it is a proxy for address. PROV provides a more general and useful version of the address
STRCA	Street/house number/addition of the correspondence address	X	X	X	X	No, it is a proxy for address. PROV provides a more general and useful version of the address
PCPLCA	Postal code and city of the correspondence address	X	X	X	X	No, it is a proxy for address. PROV provides a more general and useful version of the address
HN1X2X30	Trade name 1st line 2 x 30 positions	X	X	X	X	No, HN45 provides a more complete version of the trade name
HN2X2X30	Trade name 2nd line 2 x 30 positions	X	X	X	X	No, HN45 provides a more complete version of the trade name
HN45	Trade name 45 positions	X	X	X	X	Yes, most complete in representing the trade name
PCVA_CIJF	Postal code of the establishment address		X			No, it is a proxy for address. PROV provides a more general and useful version of the address
PCVA_LTRS	Postal letters of the establishment address		X			No, it is a proxy for address. PROV provides a more general and useful version of the address
PCCA_CIJF	Postal code of the correspondence address		X			No, it is a proxy for address. PROV provides a more general and useful version of the address
PCCA_LTRS	Postal letters of the correspondence address		X			No, it is a proxy for address. PROV provides a more general and useful version of the address
BEHKN	Managing chamber number		X			No, it is a proxy for address. PROV provides a more general and useful version of the address

Name	Description	2022 data	2023 data	Monitor data	Label data	Reason for inclusion or exclusion
GEOKN	Geographical chamber number		X			No, it is a proxy for address. PROV provides a more general and useful version of the address
GEMK_VA	Municipality code of the establishment address	X	X	X	X	No, it is a proxy for address. PROV provides a more general and useful version of the address
GEMK_CA	Municipality code of the correspondence address	X	X	X	X	No, it is a proxy for address. PROV provides a more general and useful version of the address
GEMNAAM	Municipality name	X	X	X	X	No, it is a proxy for address. PROV provides a more general and useful version of the address
PROV	Province	X	X	X	X	Yes, used as a location indicator
TEL_NRS	Telephone number	X		X	X	No, contains no information useful for classifying caring communities
MOB_TEL_NR	Mobile phone number	X		X	X	No, contains no information useful for classifying caring communities
FUNCTIE	Function	X		X	X	No, only a few companies have this (too much missing data)
VOORLETTER	First letter	X		X	X	No, contains no information useful for classifying caring communities
VOORVOEGSE	Prefix	X		X	X	No, contains no information useful for classifying caring communities
ACHTERNAAM	Last name	X		X	X	No, contains no information useful for classifying caring communities
SBI_CODE	Standard Business Classification code	X	X	X	X	Yes, contains useful information related to the business activities
SBI_OMSCHR	Standard Business Classification description	X	X	X	X	No, proxy for SBI code
NEVENACT_1	Secondary activity code (1st)	X	X	X	X	No, SBI_CODE is more important
NEVENACT_2	Secondary activity code (2nd)	X	X	X	X	No, SBI_CODE is more important
HFD_N_VEST	Head/branch office indication	X		X	X	No, contains no information useful for classifying caring communities
CD_EC_ACT	Economically active	X	X	X	X	No, contains no information useful for classifying caring communities
KL_WP_TOT	Classes of total employees	X	X	X	X	No, W_P_TOTAAL provides a more complete version of the employee count
KL_WP_FULL	Classes of full-time employees	X	X	X	X	No, W_P_TOTAAL provides a more complete version of the employee count
PEILDAT_WP	Reference date of employees at the entity	X		X	X	No, contains no information useful for classifying caring communities
PEILDAT_WP_OND	Reference date employees at the company			X		No, contains no information useful for classifying caring communities
P_DAT_WP_O	Reference date employees at the company	X			X	No, not useful and is identical to attribute mentioned above
RECHTSVORM	Registered legal form	X	X	X	X	Yes, caring communities are registered under certain legal forms
INS_REDEN	Reason for registration	X		X	X	No, overwhelming majority are not deregistered or dissolved

Name	Description	2022 data	2023 data	Monitor data	Label data	Reason for inclusion or exclusion
UITS_REDEN	Reason for deregistration	X		X	X	No, overwhelming majority are not deregistered or dissolved
REDEN_OPH	Reason for discontinuation	X		X	X	No, overwhelming majority are not deregistered or dissolved
RSIN	Identification number for legal entities and partnerships	X		X	X	No, contains no information useful for classifying caring communities
VENN_NM_DM	Name of the entity	X	X	X	X	No, HN45 provides a more complete version of the trade name
NMI	Non-Mailing Indicator	X		X	X	No, contains no information useful for classifying caring communities
BOEKJAAR	Bookyear	X		X	X	No, contains no information useful for classifying caring communities
DAT_OPRI_A	Date on which the entity was officially established	X		X	X	No, INSCHR_DAT provides the most useful information and is present in all subsets
DAT_DEP_JS	Date of filing of annual accounts	X		X	X	No, contains no information useful for classifying caring communities
INSCHR_DAT	Registration date	X	X	X	X	Yes, most useful date attribute when classifying caring communities
OPHEFF_DAT	Dissolution date	X		X	X	No, INSCHR_DAT provides the most useful information and is present in all subsets
DAT_OPRICH	Date of establishment	X		X	X	No, INSCHR_DAT provides the most useful information and is present in all subsets
DAT_VEST	Starting date of establishment	X		X	X	No, INSCHR_DAT provides the most useful information and is present in all subsets
DAT_UITSCHRP	Date of deactivation of the registration			X		No, only present in one subset
DAT_ONTB_RP	Date of dissolution of the registration			X		No, only present in one subset
VEST_DATUM	Date the entity moved to its current establishment	X		X	X	No, contains no information useful for classifying caring communities
DAT_VOORTZ	Date of appointment as chair	X		X	X	No, contains no information useful for classifying caring communities
URL	Website URL	X		X	X	No, contains no information useful for classifying caring communities and is a proxy for trade name to some extent
DOMEIN	Website URL		X			No, contains no information useful for classifying caring communities is a proxy for trade name to some extent
P_W_FULLT	Full-time employees	X		X	X	No, W_P_TOTAAL provides a more complete version of the employee count
W_P_FULLT	Full-time employees		X			No, W_P_TOTAAL provides a more complete version of the employee count
W_P_TOTAAL	Total number of employees	X	X	X	X	Yes, most complete, and exact in representing the total number of employees

Name	Description	2022 data	2023 data	Monitor data	Label data	Reason for inclusion or exclusion
W_P_PARTT	Part-time employees	X		X	X	No, W_P_TOTAAL provides a more complete version of the employee count
WP_TOT_OND	Total number of employees at the company	X	X	X	X	No, W_P_TOTAAL provides a more complete version of the employee count
H_NAAM_VOL	Full registered trade name	X	X	X	X	Yes, complete in representing the trade name and is slightly different to HN45 in some cases
IND_OPHEFF	Indicator for discontinuation	X		X	X	No, contains no information useful for classifying caring communities
ZZG DEF	Label of caring communities				X	Yes, renamed to label and used as the target variable
Bedrijfsomschrijving	Textual description of business activities	X	X	X	X	Yes, useful for determining the activities of the entity

Appendix B – Data Exploration

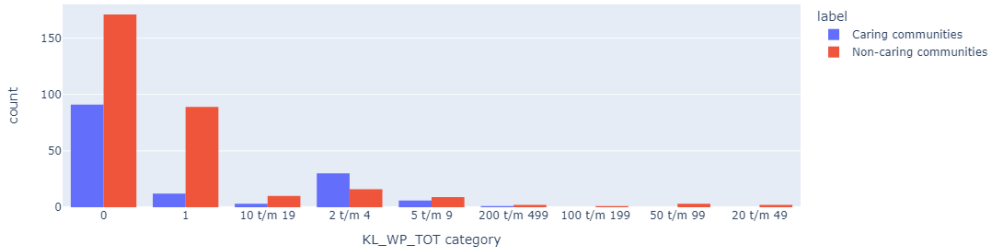

Figure B1: Relation between the number of total employees and the target variable 'label' in the training data. There seems to be a negative relation since the top 4 classes do not contain any caring communities.

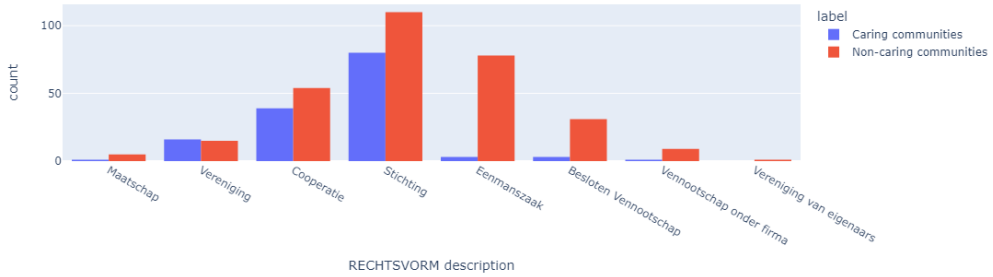

Figure B2: Relation between legal forms and the target variable 'label' in the training data. Code 71 (stichting aka association), 61 (cooperatie aka cooperation), and 74 (vereniging aka foundation) are most likely to contain caring communities.

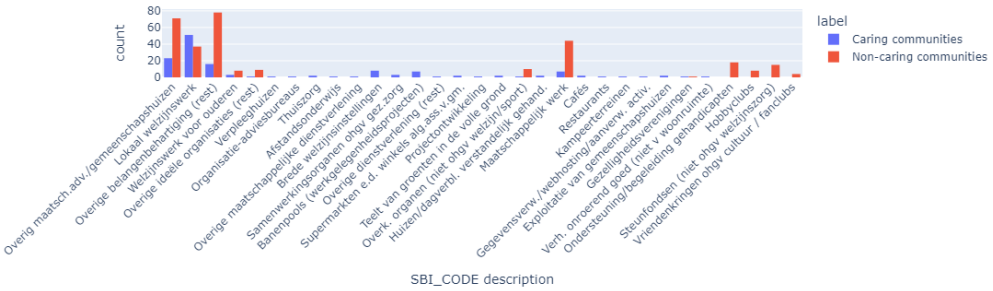

Figure B3: Relation between SBI codes and the target variable 'label' in the training data. The SBI code corresponding to local social services (Lokaal welzijnswerk in Dutch) has the highest likelihood of representing a caring community. Several other SBI codes only contain caring communities, indicating that this feature is of high importance to the classification algorithm.

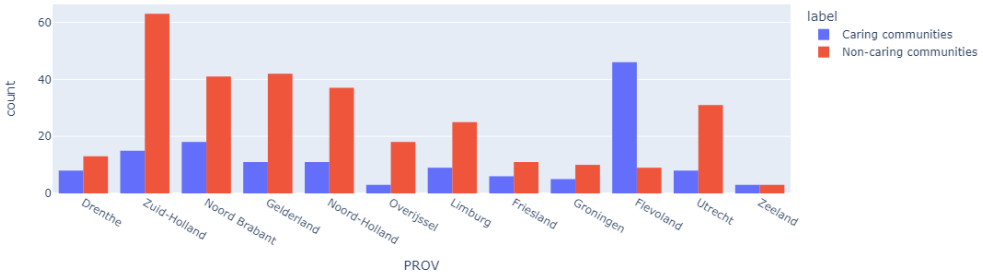

Figure B4: Relation between the province and the target variable 'label' in the training data. Flevoland has the most caring communities (46), and Overijssel and Zeeland the least (3).

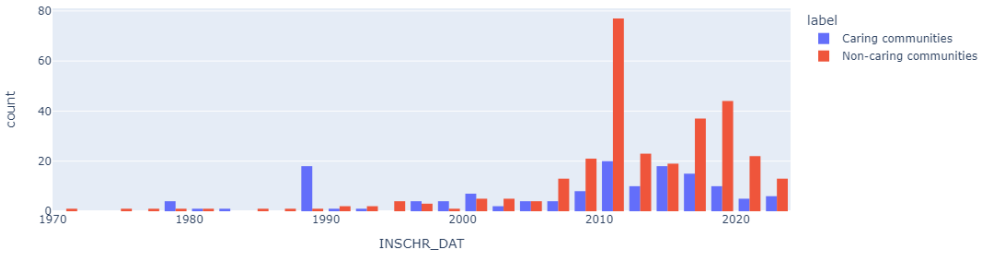

Figure B5: Relation between the registration date and the target variable 'label' in the training data. Most caring communities registered themselves after 2010, with an exception for 1988 and 1989 since there is a small peak there.

Appendix C – Data Cleaning

Duplicate removal

Before identifying duplicates within the dataset, it was essential for each record to have a unique identifier. I identified the feature 'VGNUMMER' as the most suitable candidate for this purpose; however, not all records contained a value in this field. To address this issue, I supplemented missing values by generating unique identifiers for the corresponding rows. Before doing so, I thoroughly checked for duplicates within this subset, yielding no instances of duplication. After, I attempted to utilize data from the years 2022 and 2023 to populate the 'VGNUMMER' field, but found no suitable matches, resulting in either mismatches or additional missing values. Given the absence of data to inform the filling of missing values, I decided to randomly generate identifiers. I opted for randomly generating identifiers over deleting records to maximize the use of available data. After assigning unique identifiers ('VGNUMMER'), I identified and removed duplicate entries. This process eliminated one duplicate organisation from the training data.

Data transformation

During the data transformation phase, I made one important adjustment to refine the dataset, namely the creation of a new label named 'label' in all subsets. This feature served as the target variable. One of the subsets, train_label, contained a pre-existing feature that included the necessary information and only had to be renamed. In the train_monitor set, I added a feature, with all entries assigned a value of 1 since this dataset solely contained caring communities.

Appendix D – Definition of Caring Communities

This definition was originally constructed by Nederland Zorgt Voor Elkaar (NZVE, n.d.) to be used to identify resident initiatives that received the survey during the Monitor study (Smellik et al., 2020). I also used this definition when manually labelling samples of the testing data.

Resident Collective / Caring Community:

- Residents who jointly take the initiative to improve their own living environment form a residents' collective. They do this from their own autonomy, have control over the initiative, and are responsible for it and ownership rests with them.
- The status of the initiative can be formal (legal entity) or informal.
- The residents' collective strives for a vital community in its own neighbourhood, district, village or district, is part of this community and offers services and products and/or organizes activities in the field of well-being, health, care and living as well as participation, poverty reduction, integration and shelter for the homeless. In short: community care is based on reciprocity.
- With this initiative, the residents' collective intends to be active for a long time (structural in nature, not incidental or one-off).
- The services and activities of the residents' collective are accessible to everyone (from the target group) who wants to use them and are inclusive in nature.

- Examples that do not belong to the definition:
  - collective private commissioning (CPO) without a community function
  - social entrepreneurs where the continuity of services/activities lies with the social entrepreneur and not with the residents
  - village or community center that focuses exclusively on rental and/ or catering