Institute of Mathematics
Faculty II – Mathematics and Natural Sciences
Technische Universität Berlin

Internship Report

Studies: M. Sc. Scientific Computing

Company:
MMM Consulting GmbH

Duration:
01/03/2023-28/04/2023

Contents

List of Figures

List of Tables

1 Description of the Company

1.1 The Company

MMM Consulting GmbH is a consulting firm that has been operating since 2006, serving both international corporations and medium-sized businesses. The company specializes in providing consulting services in the areas of business intelligence, event management, and customer segmentation. With a strong focus on the pharmaceutical industry, it develops effective concepts and strategies in partnership with clients, and provides practical support for their implementation. Moreover, it offers cooperative outsourcing solutions to clients, allowing them to focus on core business activities and create flexible structures that can quickly adapt to changing market requirements. [2]

The company consists of several teams that specialize in different areas. The Consulting Team is responsible for providing clients with high-quality consulting services in business intelligence. It works closely with clients to understand their business needs and provide insights and recommendations based on data analysis. The Innovation Team is dedicated to staying ahead of the curve and identifying new and emerging trends focusing on cutting-edge ideas, and create new project concepts. The Marketing Team is responsible for promoting the company's services and expertise to potential clients, as well as for the promotion of ideas and projects developed by other teams within the company, while the Marktforschung Team conducts primary market research to gather data and information directly from the target group.

The Operation Team is comprised of several specialized sub-teams, including DevOps, Sales Operations, and Tools. DevOps is responsible for ensuring that data is collected, stored, and processed efficiently and accurately by maintaining the company's databases. The Sales Operations team creates reports and visualizations to assist clients understand complex data and make informed decisions. The Tools Team is responsible for developing and enhancing software. Their primary objective is to ensure the stability and reliability of the company's products. They achieve this by implementing the latest software development practices such as continuous integration and continuous deployment (CI/CD). By streamlining the software development and deployment process, the Tools Team helps to reduce errors and increase efficiency, such that tools and products continue to meet the evolving needs of clients in a rapidly changing business landscape. Lastly, the HR Team is responsible for handling various administrative tasks within the company, including recruitment, employee training and development, payroll and benefits administration, and ensuring compliance with labor laws and regulations. Additionally, they are responsible for organizing and planning bonding activities to promote team building and employee engagement.

1.2 My Role at the Company

From March 1st, 2023, to April 28th, 2023, I had the opportunity to work at MMM Consulting as a member of the Operations Team, specifically the Tools Team. As March was a lecture-free month, I was able to work full-time and completed a total of 168 hours during that month. In April, I worked 72 hours, which was split into a 20-hour per week workload. In total I worked 240 hours.

During my time at Tools Team, I participated in multiple projects with a primary focus on workflow processes in areas such as machine learning, Natural Language Processing, and web scraping. I gained a comprehensive understanding of the methods and techniques applied in these areas, including data preprocessing, feature extraction, model building, and evaluation.

Moreover, I acquired practical experience in programming languages such as Python, R, and SQL, as well as software tools such as TensorFlow, Keras, NLTK, Pandas, SMOTE, BeautifulSoup, Scrapy, and Selenium [3]. I learned about the basics of data management and analysis, including data cleaning, preparation, and visualization. Working alongside my team members, I gained experience in project management and collaboration, including communicating progress and challenges, prioritizing tasks, and adhering to deadlines.

2 Tasks and Results

"A program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T as measured by P, improves with experience E." [10]

There are two main types of machine learning: supervised and unsupervised [8]. In supervised learning, the algorithm is trained on a labeled dataset, where the correct output is provided for each input example. The algorithm then uses this labeled data to make predictions on new, unseen data [14]. In unsupervised learning, the algorithm is given an unlabeled dataset and must identify patterns and relationships on its own [9].

Natural language processing (NLP) is a subfield of machine learning that focuses on enabling computers to understand and interpret human language. NLP techniques are used in a variety of applications, including chatbots, virtual assistants, and automated text summaries. [12].

NLP algorithms typically involve three main steps: tokenization, part-of-speech tagging, and parsing. Tokenization involves breaking a sentence or document into individual words or tokens. Part-of-speech tagging involves identifying the grammatical role of each word in the sentence (e.g., noun, verb, adjective). Parsing involves analyzing the sentence structure and identifying relationships between the words (e.g., subject-verb-object).[8]

Some common machine learning techniques used in NLP include:

- Supervised learning algorithms, such as logistic regression and support vector machines, are often used for tasks like text classification and sentiment analysis.[8]

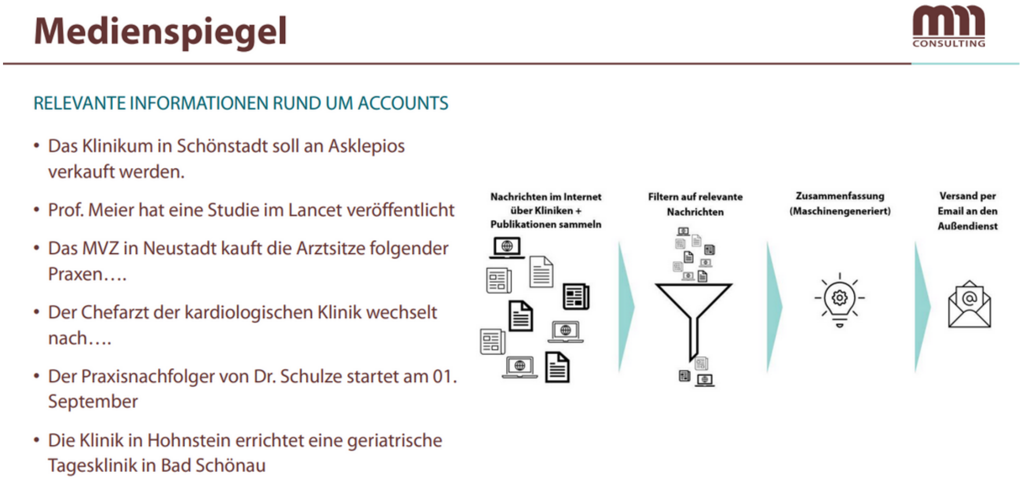
- Unsupervised learning algorithms, such as clustering and topic modeling, can be used to identify patterns and group similar documents together.[8]

- Deep learning techniques, such as recurrent neural networks and convolutional neural networks, have been very successful in recent years for tasks like language translation and speech recognition [8].

To train and evaluate machine learning models, it is common to split the data into a training set and a test set. The training set is used to train the model, while the test set is used to evaluate its performance on new, unseen data. It is important to ensure that the data is representative and unbiased, and that the model is not overfitting the training data.[14]

2.1 Medienspiegel Project

One of the projects that I participated in at MMM Consulting was the "Medienspiegel" project. The main goal of this project was to create a software solution providing customers with relevant news and publications from the medical world through a newsletter.



Medienspiegel

RELEVANTE INFORMATIONEN RUND UM ACCOUNTS

- Das Klinikum in Schönstadt soll an Asklepios verkauft werden.
- Prof. Meier hat eine Studie im Lancet veröffentlicht
- Das MVZ in Neustadt kauft die Arztsitze folgender Praxen....
- Der Chefarzt der kardiologischen Klinik wechselt nach....
- Der Praxisnachfolger von Dr. Schulze startet am 01. September
- Die Klinik in Hohnstein errichtet eine geriatrische Tagesklinik in Bad Schönau

Nachrichten im Internet über Kliniken + Publikationen sammeln

Filtern auf relevante Nachrichten

Zusammenfassung (Maschinengeneriert)

Versand per Email an den Außendienst

Figure 1: Process of Medienspiegel [1]

One of the essential parts of the "Medienspiegel" project was to automate the collection and classification of relevant data from various online sources. To achieve this, I used web scraping techniques to extract information from different websites, such as medical news websites and publications of doctors.

For web scraping, I utilized packages such as BeautifulSoup and Scrapy, which are powerful tools in Python for extracting and parsing data from HTML and XML documents [4]. I also used Selenium to automate web browsing activities, allowing us to navigate through websites and collect data more efficiently.

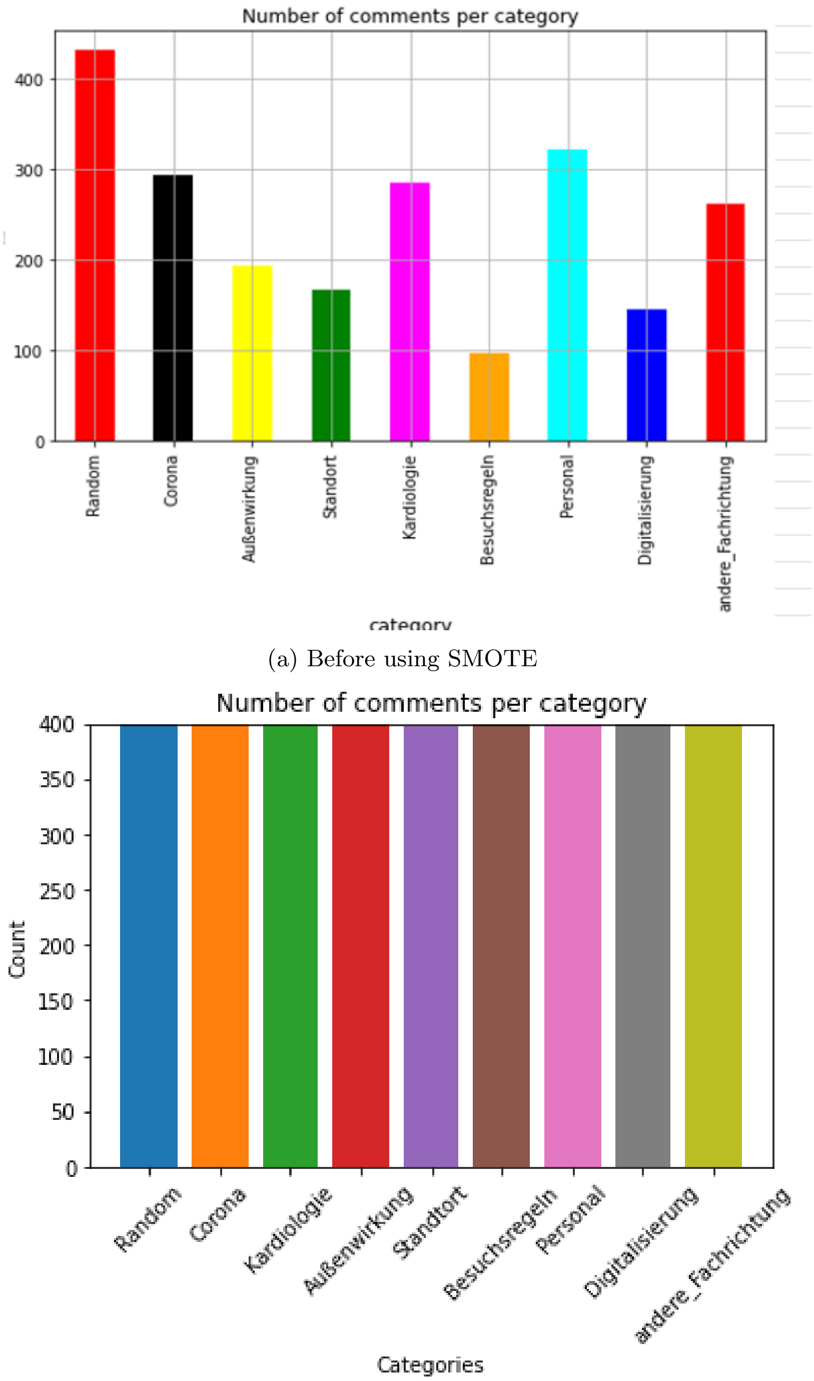
The collected data was then stored in a structured format for further analysis and processing. However, due to the nature of web scraping, the collected data often had imbalanced class distributions, which could cause issues when training machine learning models.

To tackle this problem, I used Synthetic Minority Over-sampling Technique (SMOTE) [5], which is a popular over-sampling technique for imbalanced datasets. I implemented SMOTE using the imblearn library in Python, which generates synthetic samples for the minority class to balance the dataset. In an imbalanced dataset, as can be seen in Figure 2(a), one class has significantly more data points than the others, leading to bias in the machine learning model trained on the data.

In the example given, the "random" category has 400 data points, while the other categories have significantly less. If we were to train a machine learning model on this imbalanced dataset, it may overfit to the majority class (random), resulting in poor performance on the minority classes.

With SMOTE the dataset transforms such that each minority class has the same number of data points as the majority class (in this case, 400).The result is shown in Figure 2 (b).

With this balanced dataset, we can train our machine learning model without worrying about bias towards the majority class. The model will be able to learn from the minority classes as well, leading to better performance overall.



(a) Before using SMOTE

(b) After using SMOTE

Figure 2: Balancing the imbalanced data

After data collection and preparation, I used a supervised learning approach to classify the collected data into relevant categories. I followed a multi-class classification approach for the news article categorization task. Each category was represented by a label, and the goal was to develop a machine learning model that could accurately predict the appropriate label for each news article. I experimented with different algorithms to improve the performance of the model, including the support vector machine (SVM) algorithm [13] and deep learning techniques using the BERT model.

Initially, I used the SVM algorithm as it is well-suited for handling imbalanced datasets.[7] This algorithm works by assigning different weights to the classes in order to address the imbalance, and I observed that this algorithm provided good results. However, I also explored deep learning techniques to further improve the performance of the model. In particular, I used the BERT model [6], which is a powerful deep learning model that has shown remarkable success in natural language processing (NLP) tasks. I fine-tuned the pre-trained BERT model on the dataset and observed significant improvements in classification accuracy compared to the imbalanced SVM algorithm.

To implement these techniques, I utilized the TensorFlow framework [3], which allowed us to efficiently train and evaluate different models. Overall, the approach using both the imbalanced SVM algorithm and the BERT model demonstrated promising results for the news article categorization task [6].

The frequency of words in text is an important aspect of natural language processing (NLP) because it can provide insight into the meaning and context of a text. By analyzing the frequency of different words, we can identify important topics, recurring themes, and even detect sentiment.

In NLP, the frequency of words is often represented using a term frequency-inverse document frequency (TF-IDF) score. This score measures the importance of a word in a particular document (or corpus of documents) by weighing its frequency against how often it appears in the entire corpus. This allows us to identify words that are important to a particular document, but not necessarily to the corpus as a whole.

To visualize the frequency of words in a text, we can create a word cloud (see Figure 3). A word cloud is a visual representation of the most frequently occurring words in a text, where the size of each word is proportional to its frequency. This allows us to quickly identify the most important words in a text and can provide insight into the overall theme or topic.

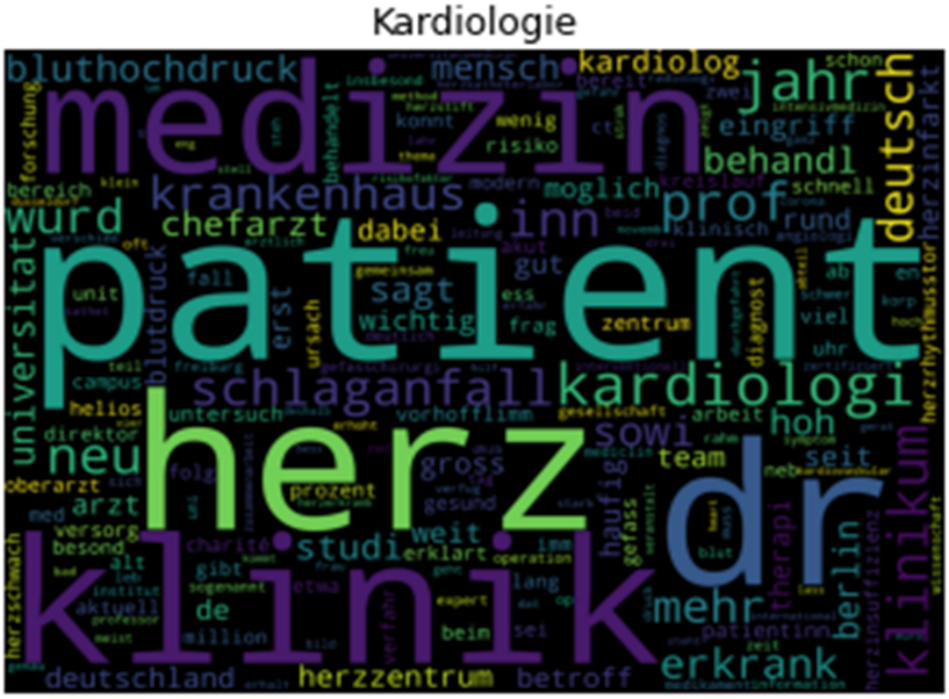


Figure 3: Frequency of Words in the data

It's worth noting that I chose to use precision as the main evaluation metric instead of accuracy because the data was imbalanced. In imbalanced datasets, accuracy is not a good measure of performance, since it can be misleading. For example, if I have 100 news articles and only 10 of them are relevant for sending to customers, then a model that simply labels all the articles as not send would achieve an accuracy of 88, even though it hasn't correctly identified any relevant articles. Precision, on the other hand, measures the proportion of relevant news articles among the total articles that the model labeled as *send*. This is a more suitable metric for the use case, since I want to make sure I are only sending relevant news to customers. In "Medienspiegel" project, I used a train-test split of 70-20-10. This means that 70 of the data was used for training the machine learning models, 20 was used for validation, and the remaining 10 was used for testing the performance of the model on unseen data.

The reason I use a train-test split is to evaluate how well the machine learning models are performing on new, unseen data. By training the model on a subset of the data, I can validate the model's performance on a separate subset of the data that the model has not seen before.

In this case, I used a 70-20-10 split to ensure that I have enough data to train the model effectively, while still having enough data to validate and test the model's performance. The validation set was used to tune the hyperparameters of the model, while the test set was used to evaluate the final performance of the model.The train-test split is an important technique in machine learning to ensure that the models are performing well on unseen data and to avoid overfitting. The final results of the model can be seen in the Table 1.

Category	Precision	Recall	F1-score	Support
Random	0.89	0.86	0.87	400
Corona	0.83	0.88	0.89	400
Kardiologie	0.90	0.90	0.91	400
Außenwirkung	0.80	0.79	0.83	400
Standort	0.77	0.80	0.80	400
Besuchsregeln	0.88	0.89	0.88	400
Personal	0.71	0.72	0.71	400
Digitalisierung	0.70	0.74	0.73	400
andere_Fachrichtung	0.80	0.80	0.80	400

Table 1: Classification report.

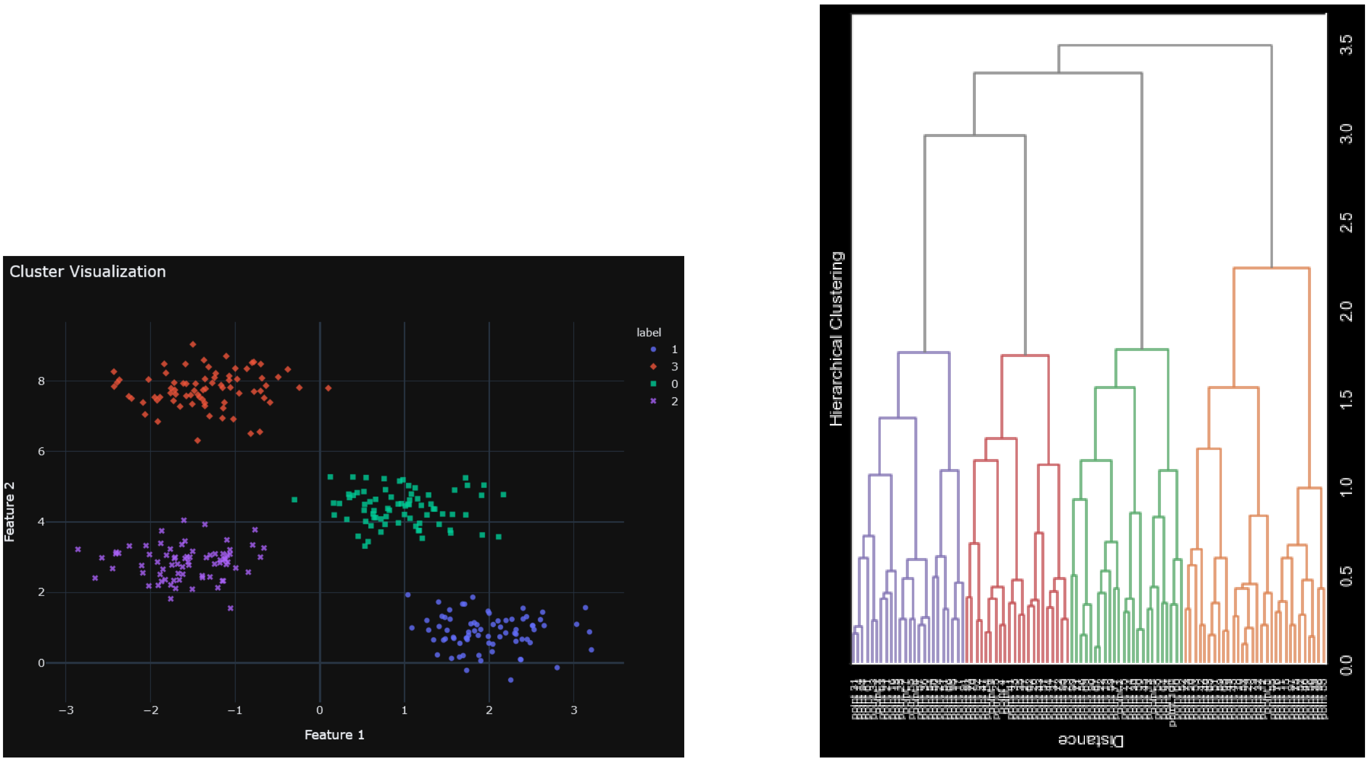
The table shows the results of a classification model for different categories. The model has been evaluated on a test set and the table shows the performance of the model in terms of precision, recall, and F1-score for each category. The precision score for a category indicates the percentage of predicted positives that are actually positive. In other words, precision measures how accurately the model predicts positive instances for a given category. The recall score for a category indicates the percentage of actual positives that are correctly identified by the model. In other words, recall measures how well the model is able to identify positive instances for a given category. The F1-score for a category is the harmonic mean of precision and recall. It provides a single score that takes both precision and recall into account. In the table, I can see that the model has performed well for all categories, indicating a good overall performance. I can also see that the model performs particularly well for the 'Kardiologie' category, with a precision, recall, and F1-score of 0.90, indicating almost perfect performance for this category. The 'Personal' category has the lowest recall score of 0.7, indicating that the model is not as good at identifying positive instances for this category.
Overall, the "Medienspiegel" project involved various stages of data collection, preparation, and analysis using a range of machine learning techniques and tools.

2.2 Customer Segmentation

Another project I worked on aimed at segmenting customers based on their activities and language usage to use targeting techniques to help customers find doctors who might be interested in the product being sold. This was achieved using machine learning techniques, specifically a combination of unsupervised and supervised machine learning.

1. Data Collection: The first step involved collecting data on the customers. This was done by analyzing their activities and language usage.

2. Unsupervised Machine Learning: Next, unsupervised machine learning was used, specifically bottom-up clustering, to classify the doctors into categories based on their activities, specializations, and other characteristics. A total of 4 clusters were developed.[9]



(a) The 4 clusters

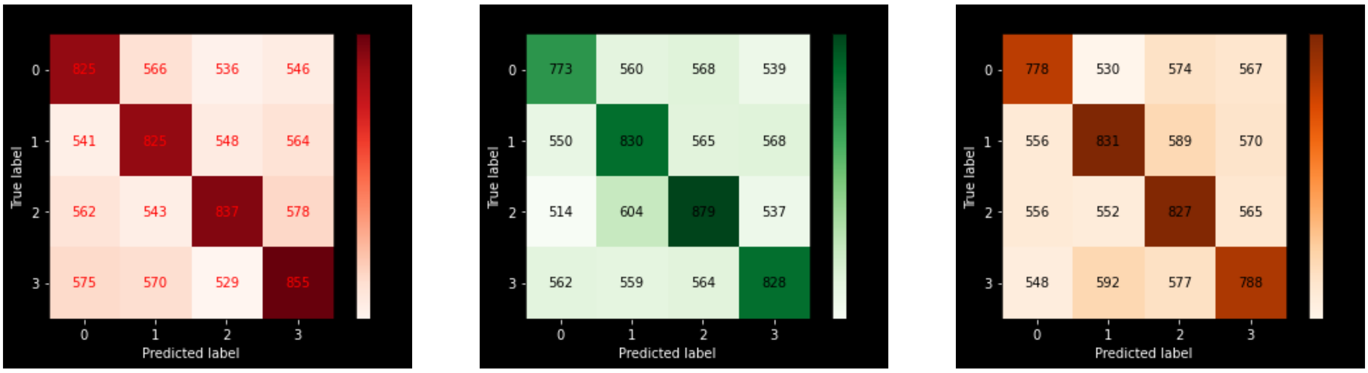
(b) Bottom up hierarchical Clustering

Figure 4: Unsupervised Machine Learning : Clustering

3. Segmentation: After the doctors were classified using unsupervised learning, segmentation was performed. This involved training a supervised machine learning model to be applied to future customers. For machine learning, I used SVM, Random Forest, and Naive Bayes.

4. Supervised Machine Learning Models:

   - SVM (Support Vector Machines): SVM is a popular machine learning algorithm used for classification tasks. It works by creating a hyperplane that separates the data into different classes. SVM has high accuracy, but it can suffer from overfitting, which leads to poor generalization performance. [11]
   - Random Forest: Random Forest is a decision tree-based algorithm that creates multiple decision trees and aggregates their predictions. It is an ensemble algorithm that can handle both categorical and numerical data. Random Forest has high accuracy and is less prone to overfitting compared to SVM.[14]
   - Naive Bayes: Naive Bayes is a probabilistic algorithm that uses Bayes' theorem to classify data. It works by assuming that the features are independent of each other. Naive Bayes is fast and efficient, but it can suffer from the "zero-frequency" problem, where a feature that does not appear in the training data causes the model to make incorrect predictions. [14]



(a) Random Forest

(b) SVM

(c) Naive Bayes

Figure 5: Confusion matrix of each method

5. Model Selection: After training the models, I found that Random Forest had the best performance, with a

precision of 86. Naive Bayes had a precision of 80, while SVM had a precision of 88. However, SVM was not used due to its low recall (63), that the model is not effective in identifying all relevant instances of that class.

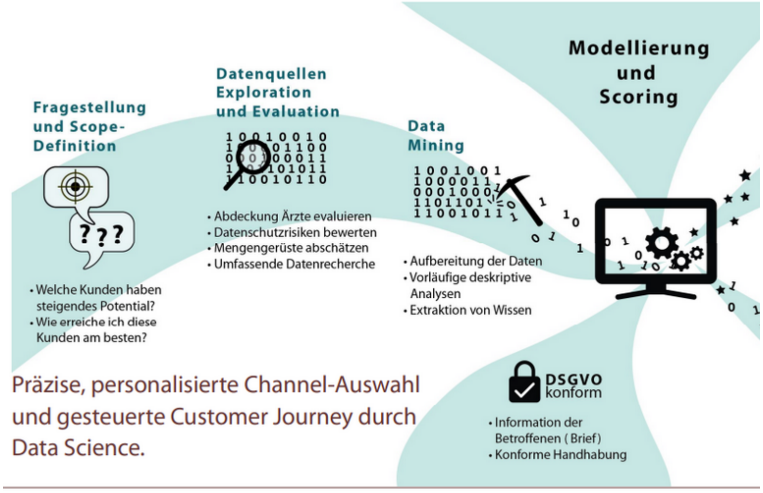

Fragestellung und Scope-Definition
???
• Welche Kunden haben steigendes Potential?
• Wie erreiche ich diese Kunden am besten?
Datenquellen Exploration und Evaluation
• Abdeckung Ärzte evaluieren
• Datenschutzrisiken bewerten
• Mengengerüste abschätzen
• Umfassende Datenrecherche
Data Mining
• Aufbereitung der Daten
• Vorläufige deskriptive Analysen
• Extraktion von Wissen
Modellierung und Scoring
DSGVO konform
• Information der Betroffenen (Brief)
• Konforme Handhabung
Präzise, personalisierte Channel-Auswahl und gesteuerte Customer Journey durch Data Science.

Figure 6: Data Collection and Prediction Process [1]

In conclusion, customer segmentation using machine learning techniques can be an effective way to target customers and improve sales. The use of unsupervised and supervised machine learning algorithms can help classify customers and develop targeted marketing strategies.

3 Conclusion and Outlook

3.1 Conclusion

During my internship, I gained valuable experience in the field of data analysis and machine learning. I had the opportunity to work on two exciting projects: the ”Medienspiegel” project and the customer segmentation project. In the ”Medienspiegel” project, I learned how to use natural language processing techniques to classify news articles and publications into relevant and irrelevant categories. I achieved an 87 precision rate in the classification, and I used the BERT model to train the classifier.

In the customer segmentation project, I learned how to use unsupervised learning techniques to group doctors into different categories based on their activities and specializations. I used bottom-up clustering to create four clusters, and then I used supervised learning techniques such as Random Forest, SVM and Naive Bayes to train a model that could predict the category of future customers. The Random Forest model had the highest precision rate at 86, and I decided to use this model for the segmentation.

Moreover, I had the chance to participate in team meetings, where I discussed project goals, progress, and challenges. I also presented my work to the team and received valuable feedback, which helped me improve my skills and achieve better results. Furthermore, I had the opportunity to work with a talented and supportive team, who taught me valuable skills and techniques throughout my internship. I learned how to use various tools and programming languages such as Python, R, and SQL. I also gained experience in data preprocessing, feature engineering, and model selection.

3.2 Outlook

The experience I gained during my internship has been invaluable, and I plan to continue to develop my skills in data analysis and machine learning. I hope to use these skills to work on more projects that can help companies make informed decisions based on data. I also plan to continue learning about new techniques and tools in this field, such as deep learning and reinforcement learning.

In addition to my personal development, I believe that the projects I worked on during my internship have the potential to make a significant impact on the healthcare industry. By using machine learning to analyze customer behavior and segment customers based on their interests, companies can more effectively market their products and services to the right audience. This has the potential to not only increase sales but also improve the overall customer experience.

Furthermore, I believe that there is still much to be explored in the field of healthcare data analysis and machine learning. As the healthcare industry continues to grow and evolve, there will be more opportunities to use these techniques to improve patient outcomes, streamline processes, and reduce costs. I look forward to seeing what the future holds and being a part of this exciting field.

References

[1] `https://www.mmm-consulting.de/`.

[2] `https://www.mmm-consulting.de/unternehmen/`.

[3] `https://www.tensorflow.org/`.

[4] `https://beautiful-soup-4.readthedocs.io/en/latest/`.

[5] `https://imbalanced-learn.org/dev/references/generated/imblearn.over_sampling.SMOTE.html`.

[6] `https://www.tensorflow.org/text/tutorials/classify_text_with_bert`.

[7] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[8] K. R. Chowdhary. *Natural Language Processing*, pages 603–649. Springer India, New Delhi, 2020.

[9] Katherine A. Heller and Zoubin Ghahramani. Bayesian hierarchical clustering. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, page 297–304, New York, NY, USA, 2005. Association for Computing Machinery.

[10] Rob Hierons. Machine learning. tom m. mitchell. published by mcgraw-hill, maidenhead, u.k., international student edition, 1997. isbn: 0-07-115467-1, 414 pages. price: U.k. £22.99, soft cover. *Software Testing, Verification and Reliability*, 9(3):191–193, 1999.

[11] Tasadduq Imam, Kai Ming Ting, and Joarder Kamruzzaman. z-svm: An svm for improved classification of imbalanced data. In Abdul Sattar and Byeong-ho Kang, editors, *AI 2006: Advances in Artificial Intelligence*, pages 264–273, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[12] Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551, 09 2011.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[14] Susmita Ray. A quick review of machine learning algorithms. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pages 35–39, 2019.