



FOM Hochschule für Oekonomie & Management

Hochschulzentrum Düsseldorf

Scientific Paper

part-time degree program

5th Semester

in the study course "Wirtschaftsinformatik"

as part of the course

Big Data & Data Science

on the subject

Predicting Music Genres based on Spotify Song Data using a Gradient Boosting Algorithm

by

Thomas Keiser

Martin Krüger

Jesper Wesemann

Luis Pflamminger

Advisor:	Prof. Dr. Adem Alparslan
Matriculation Number:	123456 (Krüger), 123456 (Keiser), 123456 (Wesemann), 123456 (Pflamminger)
Submission:	January 31st, 2022

Contents

List of Figures

List of Tables

List of Abbreviations

List of Symbols

1 Einleitung

Dies soll eine \LaTeX -Vorlage für den persönlichen Gebrauch werden. Sie hat weder einen Anspruch auf Richtigkeit, noch auf Vollständigkeit. Die Quellen liegen auf Github zur allgemeinen Verwendung. Verbesserungen sind jederzeit willkommen.

1.1 Problem Definition

1.2 Goal

1.3 Structure

2 Fundamentals

2.1 Basic Concepts of Big Data

Big Data is an umbrella term used to describe various technological but also organizational developments. Originally, Big Data refers to large sets of structured and unstructured data which must be stored and processed to gain business value. Today, Big Data is also often used as buzzword to outline countless modern use cases that deal with large amounts of data **Meier_2021**. Big Data is therefore often used in conjunction with other keywords like automatization, personalization or monitoring. This chapter presents the foundation of Big Data and gives an overview of technological and business standards. (QUELLE)

2.1.1 Relevance of Data

Data in combination with Business Intelligence has become increasingly important over the past decades and is closely associated with the advances of the internet itself **Chen_2012**. Looking back, Business Intelligence can be divided into three sub-categories, which follow another linearly. The first phase is centered around getting critical insights into operations from structured data gathered while running the business and interacting with customers. Examples would be transactions and sales. The second phase focuses increasingly on data mining and gathering customer-specific data. These insights can be used to identify customer needs, opinions and interests. The third phase, often referred as Big Data, enhances the focus set in phase two by more features and much deeper analysis possibilities. It allows organizations and researchers to gain critical information such as location, person, context often through mobile and sensor-based context **Chen_2012**.

In conclusion, organizations require Business Intelligence as it allows them to gain crucial insights which is needed to run the business and achieve an advantage over the competition. It is important to minimize the uncertainty of decisions and maximize the knowledge about the opportunity costs and derive their intended impacts. It is clearly noticeable that the insights and analysis possibilities become progressively deeper and much more detailed. Along this trend the amount of data required becomes larger and larger with increasingly complex data structures. Size, complexity of data and deep analysis form the foundation of Big Data and can be found again in the 5V matrix of Big Data.

2.1.2 The 5V Matrix for Big Data

When describing Data, a reference is often made to the five Vs, which highlight its main characteristics. The previous aspects of Big Data can again be recognized in averted form.

Volume: The size of the datasets is in the range of tera- and zettabyte. This massive volume is not only a challenge for storing but also extracting relevant information **Meier_2021**.

Variety: Variety refers to the diversity of the data itself. For modern Business Analytics almost every data format and type plays a vital role. They range from the more classical text and figures to images, audio and video **Meier_2021**. Whereas classical formats typically are stored in a structured way other formats rely on a semi- or unstructured database. The main differences between structured and unstructured data will be discussed in the following chapter as part of the different storage solutions for Big Data. Without much preface, unstructured data is more difficult to classify and further complicates the extraction of information but allows much deeper analysis possibilities **Tanwar2015**.

Velocity: Velocity is about the speed in which the data must be stored, and valuable information extracted. In a fast-paced environment, like the current globalized world, faster analysis can be a key advantage. Some special use cases, like malfunction detection, even require real-time processing of data **Fasel_2016**.

Value: The goal of Business Analytics and the extraction of information out of data is, as mentioned already in the previous, to create business value (Big Data Analytics, 6). Value that minimizes uncertainty of action or processes and gives the operators a key advantage **Meier_2021**.

Veracity: Veracity describes a challenge of analytics with data. The gathered data is often vague and not concise. Its information is not easily identifiable from the outside. Furthermore, some samples of the dataset are often of bad quality for multiple possible reasons and therefore hinder the algorithm and its training rather than supporting it. The predictions on the other hand must be precise. This conflict is a massive challenge when creating analysis models **Fasel_2016**.

2.2 Differentiation between Big Data and Big Data Analytics

The term Big Data Analytics is used in literature to describe the subcategory of Big Data that focusses primarily on the analytics of existing data, as the name already suggests. The analysis of data is of very high interest for many use-cases and organizations as its outcome is business value (comparison to the previous chapter: Why is data so important).

The term Analytics is used to describe a systematical analysis of data in some form. Analytics is about detecting hidden patterns, clusters and meaningful features ranging from simple detection over deep analysis and predictions. The goal is to process the data in such a way that value is created from it **Tanwar2015 Meier_2021**. The creation of value is closely linked to the use cases for which in-depth knowledge is required and in which form. Generally, literature distinguishes between four main groups of analytics by grouping them according to the methodology and their objectives starting with Descriptive Analytics. Descriptive Analytics is the simplest form of Analytics and is often the first step for further research. Descriptive Analytics gives information about what has happened in the past. Descriptive Analytics is often used for reporting on relevant topics for operation such as KPIs, sales or revenue. Typically, visual presentation methods are used for displaying the information. Diagnostic Analytics is based on Descriptive Analytics but serves a different goal, since its goal is to give insights on why something has happened in the past. Diagnostic Analytics evaluates the impact of features, detects correlations and dependencies. Predictive Analytics takes this approach one step further as it predicts what is most likely going to happen in the future based on the knowledge gained from past data. It creates models with the help of algorithms to determine the probability of outcomes. The final step of Analytics is Prescriptive Analytics, which again is based on its predecessor. Prescriptive Analytics predicts outcomes and recommends actions to avoid or support them respectively **Meier_2021**.

This project contains elements of descriptive, diagnostics and predictive analytics. Using the Crisp Dm process model (CHAPTER X), the data set is first analyzed and then a predictive model is generated to classify new data based on known data.

<https://www.sigmacomputing.com/blog/descriptive-predictive-prescriptive-and-diagnostic-analytics-a-quick-guide/>

2.3 Data Formats

As part of Big Data, the storage of data faces similar challenges as Big Data itself. Adequate storage solutions are key to providing data for the following analysis step. This chapter gives a quick overview on data types. Storage is a very important topic for Big Data but plays only a minor role for the project itself and therefore will not be discussed in more detail.

Variety already outlined the shift from structured to unstructured data formats. Structured data has a fixed format and fits into a predefined data model which can be stored in tabular form. Unstructured data, on the other hand, has no fixed format, schema or structure.

It comes in almost every form such as PDF, text, image, audio and many more. Basically, the whole internet and everything that is published on it is some form of unstructured data. Therefore, it is believed, that approximately 95 percent of all data is in unstructured form. In between structured and unstructured data there exists a subcategory called semi-structured data. Semi-structured data has no strict standard and can be read by machines since it often consists out of user-defined data tags. An example for semi-structured data is XML **Tanwar2015**.

Structured and semi-structured data is relatively easy to analyze compared to unstructured data because machines mostly rely on structural organization. Unstructured data, on the contrary, has great potential because the amount of information stored inside it is huge. To analyze unstructured data, it is often required to deconstruct it into metadata which again is comparable to semi-structured data. Often both structured and unstructured data is necessary to form well-founded business decisions and gain a competitive advantage **Tanwar2015**.

The dataset used for this project and described in more detail in chapter X is a structured dataset. It consists of classical numerical and categorical features which all have a predefined range of values.

2.3.1 Reinforcement Learning

2.3.2 Categories of Machine Learning

Machine learning refers to the ability of a computer to learn on its own. The field of machine learning includes a wide range of algorithms that learn from data and make predictions with variable quality. These predictions are not made programmatically but by data-driven predictions that are "learned" by generating knowledge. Basically, Machine Learning can be divided into 3 different methods which are discussed below.**2018VDMAQuick**

Supervised Learning In supervised learning, a function is determined by mapping input values to known target values using examples. Supervised learning is also called learning from examples for this reason.**schacht2019blockchain** The system learns based on a training data set that already contains the correct answers. Using the already given data sets, the algorithm learns to set up rules and patterns to reach the known target variable. The process is repeated until the prediction matches the desired quality. Experiences from each iteration are in turn included in the learning process. If the trained model fulfills the desired results, it can also be applied to unknown data. The goal of this learning method is therefore to make predictions and recommendations.**schacht2019blockchain**

It is also important to mention that target values in machine learning are called labels. Here again, one must distinguish between two use cases. If the label is nominal, it is a so-called classification. So, the model should assign data to specific classes. A regression is present if the label has metric target values.**Paass2020** The aim regression is to use data to make forecasts of future values or to identify trends. Problems that can occur when using supervised learning and should be avoided if possible are either overfitting or underfitting. Overfitting is when the algorithm is adapted too much to the training data. This means that it delivers better results when applied to training data than to unknown data. The opposite is called underfitting which means the models of the learning procedure are not complex enough. Therefore, the algorithm cannot deliver sufficient performance to make predictions.**buxmann2018kÄijnstliche**

Unsupervised Learning The opposite of supervised learning is unsupervised learning. Here, the algorithm itself acquires patterns and correlations without explicitly predefining target values. Which is why the algorithm is also called learning from observations.**schacht2019blockchain** This is also done using sample data, but without labeled output data as in supervised learning. The algorithm thus focuses on deriving or recognizing common structures and patterns in the sample data.**2018VDMAQuick** The search includes the complete training data and not only relations with concrete target values as in supervised learning.**ernst2016grundkurs** Unsupervised learning can again be divided into several types. Clustering, for example, deals with finding frequency structures, patterns and grouping the data into a few sets.**Ertel2021** Another type of learning are associations which search for rules that map connections between data points. Finally, there is also dimensionality reduction. This kind of learning tries to reduce all available variables to the most important ones.**FraunhoferMasch2018** The problem with unsupervised learning, however, is that no conclusions can be drawn about the quality of the algorithm due to the non-existence of target variables. Thus, there is no real right or wrong in the learning process and so the algorithm has the possibility to fail completely.**schacht2019blockchain**

Semi Supervised Learning Semi-supervised learning is a learning method that includes elements of supervised and unsupervised learning. Because of this combination, it is often not considered a separate form of learning, but it is important to mention it. This method uses training data that are only partially labeled.**schacht2019blockchain** This training data serves as a representative label of discovered structures if unsupervised learning is used. If supervised methods are used, the unlabeled data serves to make statistical accumulations more estimable. The main advantage of semi-supervised learning is that training of algorithms is already possible with little data. This reduces the high costs and the effort that often have to be spent on target values.**WuttkeDatasolutMachine**

Reinforcement Learning Reinforcement learning is fundamentally different from the learning methods already mentioned. Unlike the learning methods already discussed, no training data is available at the beginning.**Ertel2021** The algorithm acquires data only by interacting with the environment, which is why this learning method can also be called learning by interaction. Due to the few Requirements, this learning method is optimally suited**schacht2019blockchain** Models train by tackling challenging problems and having their decisions immediately rewarded if successful or punished if unsuccessful through feedback signals. This reward or punishment occurs, for example, in chess when a game is won or lost. Moves that lead to victory are saved, as are moves that lead to a possible loss.**schacht2019blockchain** In contrast to other learning methods, the algorithm does not know whether a decision is right or wrong before it decides what to do next. In addition, it does not know whether the decision it is currently making is the best one for the situation at hand, since its wealth of experience only grows with increasing runtime. In order to cover all possible situations, the algorithm must initially also cover previously unknown possibilities and not only act on the basis of actions that have led to reward or punishment in the past. However, if the algorithm is trained well enough, it can solve problems very well on this basis. Overall goal is to develop a strategy to maximize rewards received and get better as fast as possible.**Ertel2021**

2.4 Decision Trees

Decision Trees are one of the most widely used supervised Machine Learning Algorithms either as standalone solutions or in combination with enhancement approaches like Boosting. They are furthermore very flexible in their construction and can be used for various Machine Learning Problems such as Classification and Regression.

Decision Trees “predict an unknown value of a target variable by learning decision rules from data features” to reconstruct the dependence between the features and the respective labels for each sample. To perform Classification or Regression, Decision Trees rely on recursive splitting of the dataset into multiple subgroups. As the number of iterations increases, the subgroups become more and more homogeneous **James2021**. The ideal result is that each subgroup is fully homogeneous and therefore only represents a single category (in case of Classification). However, this is often only a theoretical best condition, as multiple risks, such as overfitting, are associated with the increasing depth of Decision Trees.

Trees consist out of four main components. A Node is a discrete Decision Function that takes samples as its input and splits them based on features into subgroups. The aim

of each split, as previously discussed, is to create a split that results in the overall most homogeneous distribution for all subgroups **lewis2000introduction**. Nodes can be sub-classified into three kinds. The Top-Node, from which the classification starts, is called a Root Node. Nodes that are located at the very end of a Decision Tree are referred to as Leaves. Leaves do not split data any further and only mark the end of a Decision Tree. When reached, Leaves categorize or predict a final output value depending on the prediction task. Nodes in-between the Root Node and Leaves are called Internal Leaves. Like the Root, Internal Leaves are responsible for the recursive splitting of the data. Branches connect Nodes with another. For classical Trees, information only flows from the top to the bottom of the Tree.

In practice, there exist various Algorithms for computing Decision Trees with the most common ones being: ID3, C4.5, C5.0 and CART. Each algorithm follows the same principle of regressively finding perfect splits to separate data but utilizes different methods to find the ideal splitting criteria, which strongly influences the structure of the Tree, its accuracy and performance. Additionally, each Algorithm has its benefits and constraints. Therefore, it is important to determine the best Algorithm before implementing a Decision Tree based on the prediction task and dataset. This project uses the Python library Sklearn to implement a Classification Tree. Sklearn is based on the CART Algorithm **sklearn Decision Trees**.

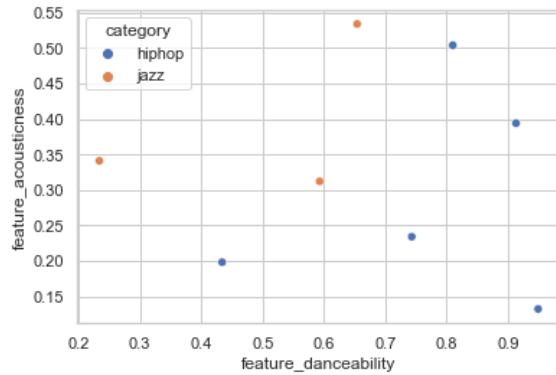
2.4.1 Decision Tree Algorithm

Initial Dataset:

To better visualize the procedure of the Decision Tree Algorithm, a simplified dataset is used, on which the individual steps are explained. For this example, a Classification Problem is chosen. The dataset consists out of actual features and labels from the project implementation phase. The features are "acousticness" and "danceability". Both features are numerical with a value range in-between 0 and 1. The Classification Problem is binary with "Hip hop" and "Jazz" representing the classes k for which the samples of the dataset are classified. Mathematically, the dataset is represented in the following form: x_i present the explanatory features while y_i represents the corresponding label for one data point of the input dataset N with a total number of n samples.

Figure 1: Initial Dataset

	category	track	feature_danceability	feature_acousticness	label
0	hiphop	h1	0.949	0.132	1
1	hiphop	h2	0.743	0.234	1
2	hiphop	h3	0.913	0.394	1
3	hiphop	h4	0.810	0.504	1
4	hiphop	h5	0.434	0.198	1
5	jazz	j1	0.654	0.534	0
6	jazz	j2	0.593	0.312	0
7	jazz	j3	0.234	0.341	0

Figure 2: Coordinate System

Recursive Tree Construction:

The Decision Tree Algorithms splits the Dataset recursively into subsets. At the beginning, the whole dataset is contained within the Root-Node from which the division into subtrees start. Each Node is mathematically described as Q_m and contains a subgroup N_m of the input dataset. Nodes split their respective groups of data further or act as a final Classifier as from of a Leaf (in case of a Classification Problem). The split can be binary or multiway. CART only utilizes binary splits and divides the Node Q_m into two repetitive subgroups labeled as Q_m^{left} and Q_m^{right} .

The Algorithm is greedy and thus determines for each split an optimal division of the samples according to the impurity of both subgroups. The goal for every split is to divide the group of data into two subgroups which are more homogeneous as the origin and maximize the overall homogeneity. The division of the data is always performed using a feature, on the basis of which the samples can be divided either by a threshold for nominal values or an is-equal-to query for categorical values. The split is mathematically described as the following (1). It consists out of the the feature j and a criteria t_m .

$$(1) \theta = (j, t_m)$$

Splitting Criteria and Information Gain:

The optimal split is determined by selecting the feature that has the highest Information Gain. Information Gain stands for the knowledge and information about the dependence between features and labels gained by splitting a group of samples by a feature. Ideally, the Information Gain is maximized. Mathematically spoken, the split is optimal if the Information Gain $G(Q_m, \theta)$ (2) is minimized as shown in (3).

$$(2) G(Q_m, \theta) = \frac{N_m^{left}}{N_m} H(Q_m^{left}, \theta) + \frac{N_m^{right}}{N_m} H(Q_m^{right}, \theta)$$

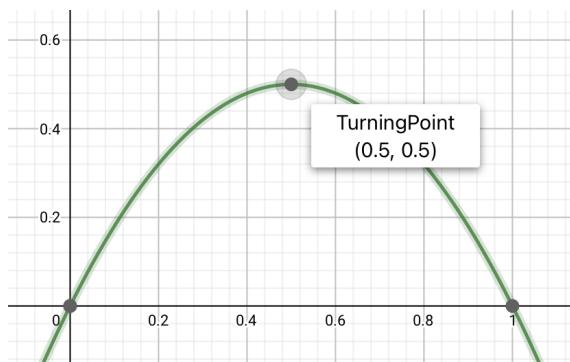
$$(3) \theta^* = \operatorname{argmin}_{\theta} G(Q_m, \theta)$$

To determine the overall best $G(Q_m, \theta)$ the respective Impurities H for both subgroups Q_m^{left} and Q_m^{right} are required. Each Impurity is weighted according to its overall impact by its relative size $\frac{N_m^{left}}{N_m}$.

The Impurity H for CART is calculated by the Gini Index. Gini index measures the probability that a sample does not belong to the category that represents the majority of the subgroup **James2021**. If both categories of a subgroup are identical in size, the Gini Index reaches its maximum point at 0,5 (5). The maximum of the Gini index means the worst possible data constellation for a subgroup with maximum heterogeneity. Gini index equal to 0 on the other hand represents the best possible result with the subgroup being fully homogenous. p_j represents that probability that a sample belongs to the class j .

$$(4) Gini = 1 - \sum_{j=1}^k (p_j)^2$$

Figure 3: Coordinate System



The optimal splits for the first iteration of the dataset look like the following. For numerical features the determination of the optimal split as for categorical ones as the optimal splitting criteria for a feature is inside its value range and must be determined by trial and error. In this example only the overall best splitting criteria is shown. The trial and error solution is implemented in python and be looked up ()�.

Danceability:

$$Gini^{left} = 1 - \left(\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2 \right) = 0,48$$

$$Gini^{right} = 1 - \left(\left(\frac{3}{3}\right)^2 + \left(\frac{0}{3}\right)^2 \right) = 0$$

$$G(Q_0, \theta) = \frac{5}{8} * 0,48 + \frac{3}{8} * 0 = 0,30$$

Acousticness:

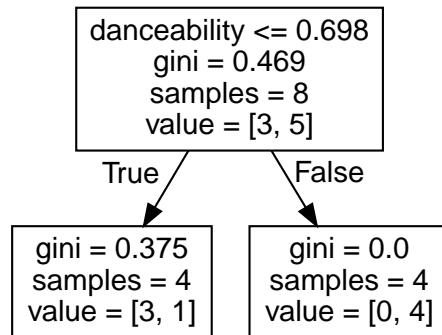
$$Gini^{left} = 1 - \left(\left(\frac{4}{4}\right)^2 + \left(\frac{0}{4}\right)^2 \right) = 0$$

$$Gini^{right} = 1 - \left(\left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2 \right) = 0,36$$

$$G(Q_0, \theta) = \frac{4}{8} * 0 + \frac{4}{8} * 0,36 = 0,18$$

The conclusion is that the Feature Danceability creates a better initial prediction than Acousticness does. It is therefore determined to be the initial splitting criteria for the Classification Tree. With the first iteration completed, the next iteration will start with each leaf as its input. The second iteration will split the Nodes further if no stop criteria is met.

Figure 4: Decision Tree: First split



Stop Criteria:

The natural stop criterion is a completely homogeneous data-group for a Node. If a Node only contains one single Class of samples, no further splitting is possible. This is the case for the right node of the example. The samples are completely homogeneous and only consist out of the category "Hip hop". The Node therefore automatically becomes a Leaf and represents the category of samples. Other stop criteria can be predefined depending on own preference. The most relevant criterion is the definition of a maximum depth of the Decision Tree. A maximum depth describes a maximum amount of recurrent iteration before the Algorithm automatically stops and treats the last Nodes as Leaves. Other hyperparameters are discussed in the implementation chapter **lewis2000introduction**.

Completion of the example:

The second iteration has both Nodes created by the first iteration as its input but only the left Node requires further splitting as the right Node is already fully classified. For the second iteration only the feature Acousticness is relevant as Danceability cannot split this subgroup of the dataset further. The input data for the second split looks like the following (X) and the same calculation takes place again.

Figure 5: Data for second split

category	track	feature_danceability	feature_acousticness	label
4	hiphop	h5	0.434	0.198 1
5	jazz	j1	0.654	0.534 0
6	jazz	j2	0.593	0.312 0
7	jazz	j3	0.234	0.341 0

$$Gini^{left} = 1 - ((\frac{0}{1})^2 + (\frac{1}{1})^2) = 0$$

$$Gini^{right} = 1 - ((\frac{3}{3})^2 + (\frac{0}{3})^2) = 0$$

$$G(Q_1, \theta) = \frac{1}{4} * 0 + \frac{3}{4} * 0 = 0$$

This time Acousticness leads to two fully homogeneous subgroups and thus marks the end of the Classification Tree as all samples are classified correctly. The second split looks like the following with the total Classification Tree displayed in figure X.

Figure 6: Decision Tree: Second split

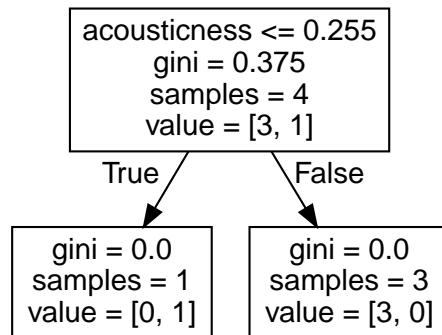
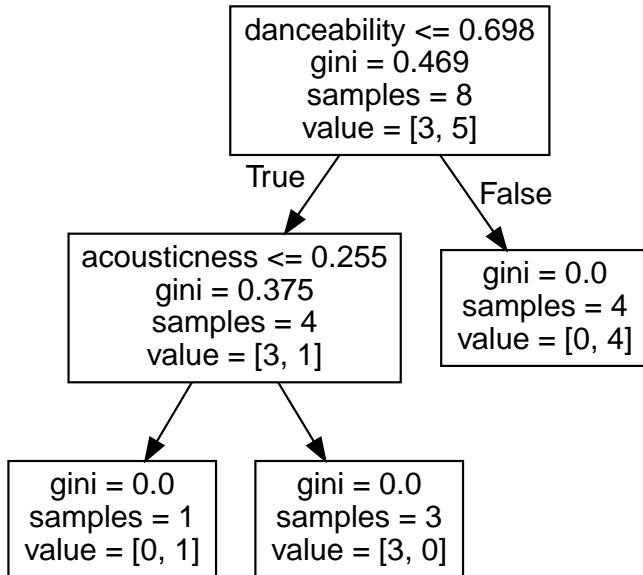


Figure 7: Decision Tree: Final Classification Tree



Tree optimization:

Tree optimization plays a very relevant role because Trees often overclassify training data without countermeasures **lewis2000introduction**. Although the training data is well classified, overfitted Decision Trees often produce very poor results for test data. Too accurate classification of training data can negatively affect Decision Trees, as they are less able to generalize the learned knowledge. Pruning is a technique used to overcome overfitting problems by reducing the size of Decision Trees **James2021**. Sections that provide little to no classification benefit are removed or not constructed during the recursive splitting process. In essence, worse results for training data are traded for better results for unknown data.

2.4.2 Evaluation of Decision Trees

In conclusion, Decision Trees can be assessed as follows. Starting with the advantages, the main benefit is the overall simplicity of Decision Trees, both from a technical and business point of view. For researchers and developers, Trees are easy to construct, require little to no data preparation, are almost universally applicable with a possibility of validation. However, the simplicity for business should not be underestimated either. When comparing Machine Learning Algorithms, the main comparison is often the accuracy of a model. The areas in which Decision Trees stand out include visualization and comprehensibility. The Decision Tree Algorithm is a white box model that allows complete transparency and explainability **James2021 sklearn Decision Trees**.

The disadvantages of Decision Trees are again closely related to its simplicity. Overfitting and the relative instability of Decision Trees are the main drawbacks and result in good memorization but a comparatively weak generalization ability **James2021** **sklearn Decision Trees**.

2.5 Gradient Boosting

The Gradient Boosting Algorithm is derived from Gradient Boosting Machines, which are a family of powerful Machine Learning Algorithms with a certain procedure pattern for the creation of models. In general, GBMs are very flexible in their characteristics with the possibility of utilizing multiple different Machine Learning Algorithms as their foundation.

Boosting differs from classical approaches as it does not consist out of a single predictive model but an ensemble approach. Ensemble Algorithms contain multiple Weak Learners that form a committee to create a strong prediction. Weak Learners are often very simple forms of traditional Algorithms, like Decision Trees, and must just be able to predict parts of the dataset correctly. Only the combination of many Weak Learners allows the model to perform overall accurate predictions. The most common form of Ensemble Algorithms are Bagging Algorithms with Random Forests as an example. Bagging, in essence, is the combination of multiple unique models. The prediction is formed by aggregating the outputs from all models into a single representative value. Typically, all models are derived from a single Algorithm, like Decision Trees for Random Forests, but technically there is no limitation to aggregate outputs from different Algorithms. This is also the case for Boosting.

Boosting, on the other hand, follows a different principle and does not rely on independent models with an aggregation function. Boosting fits new models sequentially and can thereby use earlier acquired knowledge for further iterations. This allows GBMs to train specific areas of the dataset where it has previously performed poorly **James2021**.

2.5.1 Gradient Boosting Algorithm

The generic Gradient Boosting Algorithm follows a sequence of three steps beginning with the initialization of the dataset, over the sequential training and ending with a final output. At the beginning, an additional initiation of the dataset and a Loss Function is necessary. The mathematical representation of the dataset is like the one used for Decision Trees. A summary: x_i present the explanatory features while y_i represents the corresponding label for one data point of the input dataset N with a total number of n samples.

The mathematical goal of the Algorithm is to reconstruct the unknown functional dependence f between x_i and y_i with an estimate F^* for every data point, such that the specific Loss Function $L(y, \gamma)$ is minimized (1) **Friedman_2001 Natekin2013**.

$$(1) F^* = \arg \min_{\gamma} L(y, \gamma)$$

Loss Function and Weak Learners:

The Loss Function is an indicator for the quality of the model. A small Loss for a data point means that the prediction is close or identical to the observed Label and the model therefore categorizes the sample correctly whereas a high Loss implies that the model could not predict the sample well. Given a particular learning task and dataset, different Loss Functions must be considered as Loss Functions are only suitable for specific data and task constellations. The most common Loss Function for Binary Classification is the so-called Bernoulli Loss (2). The Bernoulli Loss can be transformed into a log(odds)-prediction (3) as it is better suited for further calculations. Variations of (3) will be used in the following section to demonstrate the Gradient Boosting Procedure **Natekin2013**.

$$(2) L(y_m, \gamma) = -\log(\text{likelihood})$$

$$(3) L(y_m, \gamma) = -y_i * \log(\text{odds}) - \log(1 - p)$$

Additionally, a Machine Learning Algorithm must be defined as a Weak Learner. For GMB there are multiple Leaners to choose from, again the choice is mostly depending on the prediction task and available data **Natekin2013**. A classical approach is the use of Decision Trees, which was also chosen as the Weak Learner for this project. Decision Trees used for Gradient Boosting are always Regression Trees, regardless of whether they are used for Regression or Classification Problems. The optimization parameters are almost identical as for standalone Decision Trees, but the Trees often look very different because they are specifically created as Weak Learners. As a result, the Decision Trees often only consist out of very few layers with only 8-32 Leaves. (QUELLE)

1. Initialization:

To showcase the Gradient Boosting Algorithm the same sample dataset is used as for Decision Trees. It again consists out of 8 samples with two features and two categories as target Labels.

Figure 8: Initial Dataset

	category	track	feature_danceability	feature_acousticness	label
0	hiphop	h1	0.949	0.132	1
1	hiphop	h2	0.743	0.234	1
2	hiphop	h3	0.913	0.394	1
3	hiphop	h4	0.810	0.504	1
4	hiphop	h5	0.434	0.198	1
5	jazz	j1	0.654	0.534	0
6	jazz	j2	0.593	0.312	0
7	jazz	j3	0.234	0.341	0

The first step is to set an initial prediction for all samples of the dataset. The initial prediction is not unique for individual samples but a uniform value. The optimal initial prediction can be calculated using the following equation (4). For $F_0(x)$, representing the initial prediction, a minimum of γ is searched for. The right-hand side of the equation only consists out of a sum for each sample i (of the total dataset N) of the known Loss Function with the respective Label y_i and γ as its input. To find the low point of the equation, the derivate of the Loss Function is required. The final calculation of the overall $\log(\text{odds})$ that a song is classified as “hiphop” is the \log_e of the sum of songs of the category “hiphop” divided by the sum of the songs of the category “jazz” (4). The result can be checked graphically as is equal to the x_1 -coordinate value of the low point of the $\log(\text{odds})$ -prediction (6).

$$(4) F_0(x) = \arg \min_{\gamma} \sum_{i=1}^n L(y_i, \gamma)$$

$$(5) F_0(x) = \log_e\left(\frac{5}{3}\right) = 0.51$$

(6) COORDINATE SYSTEM

2. Sequential Processing:

With the completion of the initialization of the dataset the modelling can begin. The training is a sequential process of constructing the Regression Trees with a total of M iterations. The first iteration starts with $m = 1$. The modelling consists out of four sub-steps which are numbered consecutively.

First, the $\log(\text{odds})$ -prediction must be converted back into a probability p with the help of a Logistic Function as probability is easier to use for Classification (6). The result is that all songs have the probability of 0,63 to belong to the category “hiphop”.

$$(6) p = \frac{e^{\log_e(\text{odds})}}{1+e^{\log_e(\text{odds})}} = \frac{e^{\log_e\left(\frac{5}{3}\right)}}{1+e^{\log_e\left(\frac{5}{3}\right)}} = 0,63$$

In the **first step** the Pseudo Residuals r_{im} for each sample i of the dataset are created. The equation (7) for calculating the PR consists out of known fragments. For every sample

i a PR r_{im} is calculated using the derivative of the $\log(odd)$ -prediction with the Label y_i and the Prediction of the last iteration $F = F_{m-1}$ as its input. Again, the equation can be transformed into a very simple equation (8). For each sample the PR can be calculated by only subtracting the previously calculated probability p from the observed Label y . Ideally, an additional column is created in which the PRs are temporarily stored (figure 9).

$$(7) r_{im} = -\left(\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right)_{F=F_{m-1}}$$

$$(8) r_{im} = (y_i - p_i)$$

This and every following calculation is explained for the track h1 (9) of the dataset while the result for every other sample will be shown in form of a table.

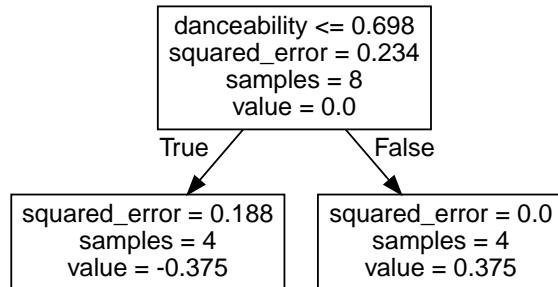
$$(9) r_{1,1} = 1 - 0.625 = 0.375$$

Figure 9: Gradient Boosting: Pseudo Residuals for first iteration

	category	track	feature_danceability	feature_acousticness	label	f0x	probability_f0x	pseudo_residuals_1
0	hiphop	h1	0.949	0.132	1	0.510826	0.624988	0.375012
1	hiphop	h2	0.743	0.234	1	0.510826	0.624988	0.375012
2	hiphop	h3	0.913	0.394	1	0.510826	0.624988	0.375012
3	hiphop	h4	0.810	0.504	1	0.510826	0.624988	0.375012
4	hiphop	h5	0.434	0.198	1	0.510826	0.624988	0.375012
5	jazz	j1	0.654	0.534	0	0.510826	0.624988	-0.624988
6	jazz	j2	0.593	0.312	0	0.510826	0.624988	-0.624988
7	jazz	j3	0.234	0.341	0	0.510826	0.624988	-0.624988

The **second step** constructs the Regression Trees out of the features of the samples with the corresponding PR as the label. For this example, only tree stumps are created to simplify the implementation. The Regression Tree for the first iteration is shown in Figure X. After the completion of the Tree, Terminal Regions R_{jm} must be defined for every Leaf. j starts with 1 and is increased for every Leaf **Friedman_2001**.

Figure 10: Gradient Boosting: First Regression Tree



In **step three**, following the completion of the Regression Tree, Output Values $\gamma_{j,m}$ are calculated by using the equation presented in (10). For each Leaf in the Tree $\gamma_{j,m}$ is computed by finding $\gamma_{j,m}$ that minimizes the Loss Function (10). Like for the initialization step,

the derivative has to be created and must set equal to 0. And again, after a complicated transformation, a very simple equation remains (11). The $\gamma_{j,m}$ can be calculated using only the PR and the most recent predicted probabilities p for all samples in the Leaf (11).

$$(10) \gamma_{jm} = \arg \min_{\gamma} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i + \gamma))$$

$$(11) \gamma_{jm} = \frac{\sum r_{im}}{\sum p_i * (1 - p_i)}$$

For $R_{2,1}$, which contains track h1, the calculation looks like the following (12). The Output Values of the complete dataset are displayed in figure X.

$$(12) \gamma_{2,1} = \frac{0.375 + 0.375 + 0.375 + 0.375}{0.625 * (1 - 0.625) + 0.625 * (1 - 0.625) + 0.625 * (1 - 0.625) + 0.625 * (1 - 0.625)} = 1.6$$

Figure 11: Gradient Boosting: First Output Values

	category	track	f0x	probability_f0x	pseudo_residuals_1	output_value_1
0	hiphop	h1	0.510826	0.624988	0.375012	1.600032
1	hiphop	h2	0.510826	0.624988	0.375012	1.600032
2	hiphop	h3	0.510826	0.624988	0.375012	1.600032
3	hiphop	h4	0.510826	0.624988	0.375012	1.600032
4	hiphop	h5	0.510826	0.624988	0.375012	-1.599926
5	jazz	j1	0.510826	0.624988	-0.624988	-1.599926
6	jazz	j2	0.510826	0.624988	-0.624988	-1.599926
7	jazz	j3	0.510826	0.624988	-0.624988	-1.599926

Step four marks the end of the first iteration and creates a new prediction $F_m(x)$ for each sample. The new $\log(odds)$ -prediction is based on the last $\log(odds)$ -prediction plus the Learning Rate v multiplied by the Output Value(s) for the sample of the last Regression Tree (13) **Friedman_2001**. Normally, there is only one Output Value for a sample which makes the summation sign obsolete. The Learning Rate v is a hyperparameter for Gradient Boosting. For this example, a high v of 0.6 is used to better visualize the changes. In practice a v in the order of 0.1 is common as a small Learning Rate tends to give better results **Friedman_2001**.

$$(13) F_m(x) = F_m(x - 1) + v * \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$$

The prediction $F_1(x)$ for h1 looks as follows (14) while the complete dataset is again displayed in figure X. In figure X the next probabilities are also shown. As one can see, the probabilities for most samples got closer to the Label. However, for track h5 the prediction got worse. This is why multiple iterations are necessary to form a well-founded prediction.

$$(14) F_1(x) = 0.51 + 0.6 * 1.6 = 1.47$$

Figure 12: Gradient Boosting: Predictions for $F_1(x)$

category	track	f0x	probability_f0x	output_value_1	f1x	probability_f1x
0	hiphop	h1	0.510826	0.624988	1.600032	1.470845
1	hiphop	h2	0.510826	0.624988	1.600032	1.470845
2	hiphop	h3	0.510826	0.624988	1.600032	1.470845
3	hiphop	h4	0.510826	0.624988	1.600032	1.470845
4	hiphop	h5	0.510826	0.624988	-1.599926	-0.449130
5	jazz	j1	0.510826	0.624988	-1.599926	-0.449130
6	jazz	j2	0.510826	0.624988	-1.599926	-0.449130
7	jazz	j3	0.510826	0.624988	-1.599926	-0.449130

The second iteration of the modelling phase follows the exact same principles and will again be showcased for track h1. It marks the end for this example.

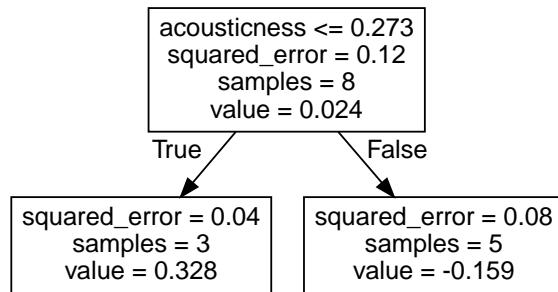
- New Probability:

$$p = \frac{e^{1.47}}{1+e^{1.47}} = 0,81$$

- New Pseudo Residual:

$$r_{1,2} = 1 - 0.81 = 0.19$$

- New Decision Tree with Terminal Regions:

Figure 13: Gradient Boosting: Second Regression Tree

- New Output Value:

$$\gamma_{2,2} = \frac{0.98}{0.54} = 1.81$$

- New Prediction:

$$F_2(x) = 1.47 + 0.6 * 1.81 = 2.56$$

- New Probability:

$$p = \frac{e^{2.56}}{1+e^{2.56}} = 0.93$$

The iterative modelling is repeated until M is reached. M marks the completion of the training of the Gradient Boosting Model. The $F_M(x)$ is the final prediction for every sample. Lastly, the predictions must again be transformed to probabilities like for the calculation of the PR. For the final probabilities thresholds are used to compute the category to which the samples belong **Friedman_2001**. A typical threshold for binary Classification is 0.5 as it splits $F_M(x)$ into two equal classes.

3. Output:

Sample h1 has a final probability of 0.93 of belonging to the category "Hiphop". 0.93 is by far greater than the predefined threshold of 0.5 and thus h1 is classified as a track that belongs to the category "Hiphop". Figure X shows that all samples are classified correctly but some less clear than others. A real-world implementation would consist out of more iterative models with different Regression Trees and a different Learning Rate to enhance the classification potential. Also the dataset would be very different with more samples that consist out of more features.

(FIGURE)

An unknown sample gets initialized by $F_0(x)$ and is sequentially routed through every model. For each iteration the Prediction is updated using the Output Value of the Regression Tree. Finally, the last probability is used to classify the sample using the predefined threshold.

2.5.2 Evaluation of Gradient Boosting

Gradient Boosting is a very powerful method as it can effectively capture complex dependencies for various Machine Learning Problems. GBMs improve many existing algorithms, such as decision trees, because many problems associated with single large models are (partially) solved by the iterative ensemble approach.

The main benefit over Decision Trees is the stability of Gradient Boosting. While large trees always have to make tradeoffs between detail and overcategorization, gradient boosting can gradually get to a deeper and deeper level of detail thanks to small trees with overall better generalization. Furthermore the flexibility of Gradient Boosting and Boosting in general is massive as it only represents the framework with many parameters to adapt the algorithm very specifically to the usecase.

The drawbacks of Gradient Boosting often arise in practice. Gradient Boosting has a significantly higher memory consumption and build time as the model must be constructed

sequentially. Also the evaluation is more time consuming as the sample must be processed by each model. From a business perspective Gradient Boosting also has its disadvantages. While the prediction is better, it is much more complex to evaluate the model and explain the results **Natekin2013**.

2.6 Cross Industry Standard Process for Data Mining

The CRISP-DM Reference Model is an organizational model for data mining projects and provides an overview of the life cycle of such projects. It includes the phases of a project, their respective tasks, and their deliverables. The life cycle of a data mining project is divided into six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment. The order of the phases is not strictly adhered to. In certain projects, the outcome of each phase determines which phase or which specific task in a phase must be performed next. However, the process of data mining is to be understood more as a part of the process. This means that data mining is not necessarily complete once a solution is implemented. The lessons that emerge during the process and from the solution can raise new, often more focused business questions. Subsequent data mining processes can benefit from, and often build on, the lessons learned from previous processes. The chapters that now follow deal with the respective phases and explain the steps to be completed in more detail.

2.6.1 Business Understanding

The first prerequisite for a successful data science project is business understanding. Without an exchange of information between the data science experts and the business experts, it will be difficult to achieve the desired result. It is necessary to define exactly what is to be achieved by the data analysis. This phase of the project defines the requirement and identifies the problem to be solved. The source of the data and the problem definition are discussed. Data, project scope, available resources and required procedures are defined. Only from this point it is possible to plan concrete steps.

2.6.2 Data Understanding

For the step of Data Understanding, Data is collected in a targeted or untargeted manner. The collected Data can be table entries, answers from questionnaires, sequence recordings from machines and so on. As a result, the relevant data must be pulled together and

made accessible. In the case of data, it is also crucial how it is collected. An insufficiently thought-out survey for example can very easily lead to bias or incorrect answers. This means that the design of the data collection must be representative. Large, i.e. sufficient, data volumes are a prerequisite for the subsequent formation of stable models and reliable forecasts. However, it is not impossible that even moderate amounts of data lead to good results. For example, a carefully selected sample may show better results than large amounts of data, which are not verifiable. Ideally, a data evaluation should be based on access to data that is sufficiently realistic for the complex processes of the company and the question to be investigated. This is the prerequisite for models that are to produce meaningful forecasts.

2.6.3 Data Preparation

In Data Preparation, data is first extracted from databases, tables and systems and transformed into the desired form (e.g. time series vectors, multidimensional matrices). This step allows the data to be further processed and analyzed using mathematical (statistical) based software or Data-Algorithms. The next step of this project phase is the verification, cleaning and, if necessary, the further transformation of the data. These processes are essential because data often cannot be further processed in its raw state. The algorithms used for analysis are often very demanding. In other words, they are not very tolerant. Even zero values (empty cells of a table) can become a problem. It is examined which values are plausible and which are rather extraordinary. For Example, it needs to be clear what empty Cell means, or the extreme Value of a cell. If it is possible to explain these values, what the measuring points are and so on. Zero values, no data, extreme values, multiple notations, or different scales can possibly be interpreted quickly by humans. With algorithms, the situation is somewhat different. They capture data as pure values, which can lead to different results depending on the situation. Data cleaning and quality assurance of the data within the data exploration is the most costly phase of the project. It will usually also require the largest share of the available project time, but it is crucial. Here, possible anomalies in the data structure as well as data errors can be identified in advance. This phase has a decisive influence on the quality of the overall result.

2.6.4 Modeling

This is the heart of any data analysis. The appropriate analytical procedure is chosen to generate predictions and groupings. Connections and structures in the data are also

visualized. It is important to present these in such a way that the most meaningful visualizations possible are created, which allow the company to make decisions based on evidence-based data. If good preparatory work has been done in the preceding project phases. That is, if the data has been checked, incorrect entries filtered out and business structures understood. It is possible to save a lot of time and costs in this phase.

2.6.5 Evaluation

In the iterative processes of model building, several models with different approaches and sometimes objectives are often formed. In the process, models are also reworked that have led to very good predictions but have become too complex at best. The intention is to simplify the models until the best possible results are obtained. At the end of the evaluation, there are many models that differ not only in the application of different methods, but also in their robustness, complexity and results. Discussing and assessing the intermediate results with company representatives leads to the optimization of the results. Furthermore, visual insights into the results (data) help to better understand the subject matter. By evaluating established criteria, results are assessed, models are compared, and the best approach or approaches are selected.

2.6.6 Deployment

The creation of the model is generally not the end of the project. Typically, the knowledge gained must be organized and presented in a way that the customer can use. Depending on the requirements, the deployment phase can be as simple as creating a report or as complex as implementing a repeatable data mining process. In many cases, it is the user, not the data analyst, who performs the implementation steps. In either case, it is important to understand in advance what actions need to be performed in order to use the models that have been created.

2.7 Application Programming Interfaces

This section gives an overview over the basic concepts and technologies behind Web **API!**s (**API!**s). An **API!** in general is an interface between two pieces of software.**reddy2011api** These might run on the same machine and communicate locally, in the case of a desktop application for example, or on separate machines that are connected via some network, e.g. in a client/server application. An API is a construct, which

abstracts complex functionality away and provides a simple interface to interact with and benefit from the complex structures behind the API.**Mozilla** This section focuses specifically on Web APIs, which is a form of API that can be accessed over the internet using the HTTP protocol.**StoplightAPITypes**

APIs are used in this project in the context of data collection and the fundamentals explained in this section are needed to understand section ??.

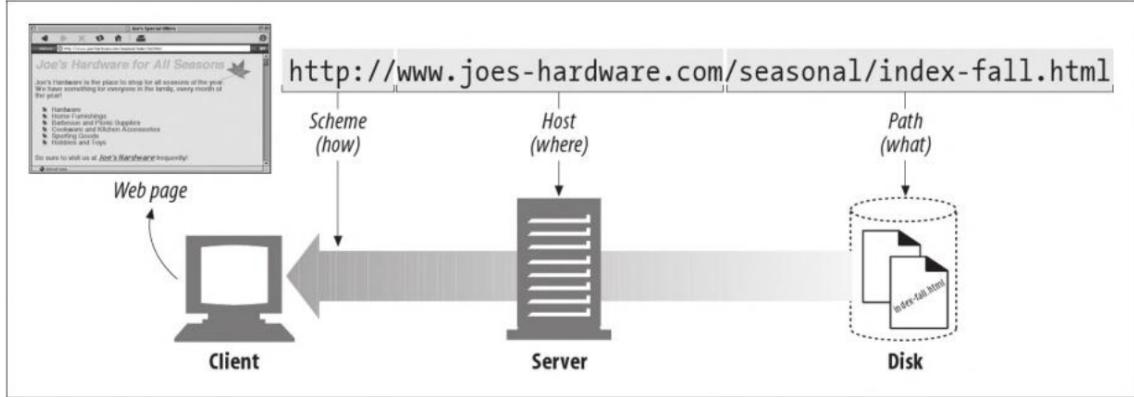
2.7.1 The HTTP Protocol

All communication on the world wide web is conducted through **HTTP!** (**HTTP!**), because it is reliable and guarantees data integrity and prevents the destruction or distortion of data during transit**gourley2002http**

Most communication on the Web is conducted between a client and a server. The client usually requests a resource from the server and the server responds with that resource, which the client can then use**gourley2002http**. As these resources could have a number of datatypes, e.g. an HTML page, a JPEG image or JSON objects, every message has a specific MIME type, which is used to inform the receiver, which type of data is contained in the message. The receiver (usually the client) can then handle the data appropriately, depending on its type.**gourley2002http**

When a client requests a resource, it must tell the server what specific resource it wants to have. A resource can be uniquely identified using a **URL!** (**URL!**). Figure ?? shows how URLs are structured. It starts with the protocol to be used, which tells the server, how the communication should commence. Next the host is specified, usually in form of a domain. It refers to a specific server or a group of servers which have the resource or know where to find it. After the domain, the path to the requested resource is given.**gourley2002http** If the server knows where to find the resource and the client is authorized to see it, the server will send it to the client. In the context of APIs, a URL pointing to a specific resource is also called an "API endpoint".**Cooksey2014** This interaction between client and server is usually done in an HTTP transaction.

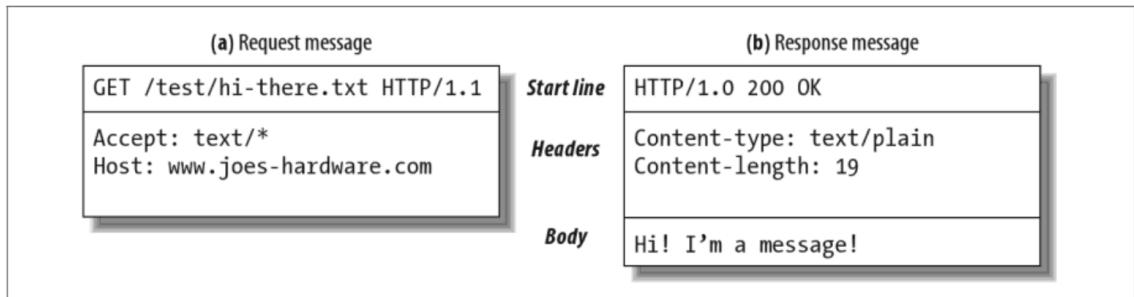
Figure 14: Description of different parts of a URL



Source: [gourley2002http](#)

A transaction consists of a request (sent from client to server) and a response in which the server answers the clients request in some way. In the context of Web APIs, a request made to an API is also called "API call". **StoplightAPITypes** Figure ?? shows the structure of an HTTP request and response message.

Figure 15: Example HTTP request and response messages



Source: [gourley2002http](#)

Every HTTP request message has a start line. It contains a method, which tells the server what action to perform on the resource, the URL and some protocol information. After the start line, some headers are given. These contain information about the request, e.g. what types of data the client can accept or credentials for authorization. Some requests also contain a body to give some data to the server. **gourley2002http**

A response message also starts with a start line, although it does not contain a method or URL. Instead it contains protocol information and status information in the form of a numerical status code and status text. **gourley2002http** A response also contains headers and might contain a body. **gourley2002http**

There are many HTTP methods defined, that can be used to delete a resource, store data on the server, get header information for a document and more.**gourley2002http** In this paper, only two HTTP methods are used, GET and POST. It is used to ask the server to send a resource to the client. A GET request message is used to ask the server to send a resource to the client. It does not contain a body.**gourley2002http** A POST request message is used to send data to the server for processing. The data to be processed is sent in the request body.**gourley2002http**

There are also many status codes defined, which tell the client the status of the transaction. This is necessary, because the client's request can not always be fulfilled successfully by the server.**gourley2002http** The server might not be able to find the server, he might not be able to store the data or the client might not have given the right authorization credentials to perform the requested action. For each of these problems, there are predefined status codes. For example, codes in the range 200-299 represent success, codes from 400-499 tell the client that his request contained an error and 500-599 are used to inform about an error which happened on the server.**gourley2002http** Table ?? shows some common status codes. There are many other status code defined in the standard, but they are not relevant for this paper.

Table 1: Common status codes

Status code	Reason phrase	Meaning
200	OK	Successful transaction. Any requested data is in the response
401	Unauthorized	The response needs to contain credentials to access the requested
404	Not Found	The server cannot find a resource for the requested URL.

Source: **gourley2002http**

There are also many different headers defined in the HTTP standard. Some that are important for this paper:

- Authorization: Request header, which contains the credentials the client is using to authenticate. This might be username and password or a token.
- Content-Type: The datatype of the data contained in the body

2.7.2 JSON

Most modern APIs transmit data to the client using **JSON!** (**JSON!**). It is a text-based format able to represent structured data. **MozillaJSON** It is easy for humans to read

and write JSON objects, but is also easily processed by computers. JSON is tightly integrated in the JavaScript programming language and can be used without any parsing or serialization. **OracleJSON** This is beneficial, as JavaScript is widely used both on the client and server side of web applications. JSON can also be used with many other programming languages. **JsonOrgIntroduction**

JSON supports only six very basic datatypes: String, Number, Boolean, Null, Object and Array. **OracleJSON** The following code snippet shows how each of the datatypes are defined in JSON.

```

1  {
2      "string": "Martin",
3      "number": 200,
4      "number2": 300,
5      "boolean": true,
6      "booleanFalse": false,
7      "nullValue": null,
8      "object": {
9          "objectAttribute": "string",
10         "objectAttribute2": "string2"
11     },
12     "array": [
13         "entry1",
14         "entry2",
15         {
16             "objectInArray": true
17         }
18     ]
19 }
```

The code snippet shows, that the whole JSON object is wrapped in curly braces. Each attribute is defined as a attribute-value-pair, where the attribute name is defined in double quotes, followed by a colon and the value definition. String values are defined in double quotes, numbers and booleans without. Objects are defined in curly braces and can again contain any datatype. Arrays are defined using brackets. They can contain any number of entries of any type. Types can also be mixed within the same array. All attributes-value-pairs or array items are separated by commas. Indentation and newline characters are not parsed in JSON as all syntax is given using the characters "[":'. **JsonOrgIntroduction**

The MIME type of a JSON object is called "application/json". An HTTP response containing JSON data will therefore always have the header "Content-Type: application/json". **richardson2008restful**

The text-based nature, good readability, simplicity and easy parsing and processing are reasons for the wide usage of the data format. **OracleJSON** JSON is also used by the Spotify API to transfer data to the client. **SpotifyWebAPI**

2.7.3 Types of Web APIs

As previously mentioned Web APIs are accessed through the HTTP protocol**StoplightAPITypes**, and offer a way for applications and data servers to communicate and exchange data over the web.**Lane2019WhatIs** They are used to power desktop, web and mobile applications, pass data between different services, systems and devices. They make automated data access and retrieval possible and enable the integration of internal and external systems into processes and data flows.**Lane2019WhatIs**

There are different types of Web APIs. Open or Public APIs can be used by anyone with minimal restrictions, such as logins or authentication. They are provided by a company, government or other organization and make it possible for any developer to access that organizations data or services.**StoplightAPITypes** Internal APIs are only used to share resources within one organization. They are meant to enable the integration of different tools, databases and teams with each other and are not exposed to the public.**StoplightAPITypes** Partner APIs are exposed to the public, but are only open to specific users, which might pay for the right to use the API. Composite APIs aggregate multiple endpoints into one to enable easy retrieval of data. This aggregation might contain data from multiple other internal and external APIs and services. It is used to simplify API usage and reduce server load by minimizing the number of API calls.**StoplightAPITypes**

Web APIs can follow different types of architectures, which define the structure, data types and commands used in the API. Some competing architectures are REST, RPC and SOAP. **StoplightAPITypes** As the Spotify implements a REST approach**SpotifyWebAPI**, the next subsection explains the REST architecture of Web APIs.

2.7.4 The REST Architecture

REST! (REST!) is a set of constraints applied to modern web services, that stems from a basic set of rules, which were defined to ensure the scalability of the web.**masse2011rest** A web service, which implements REST principles is also called a "RESTful service". Such an API consists of a set of interlinked resources called a resource model.**masse2011rest**

There are six basic rules to a RESTful design:**masse2011rest**

- Client-server: The API is designed to work in a client/server architecture.**masse2011rest**

- Uniform Interface: The API needs to conform to the established standards of the web.**masse2011rest** This includes
 - Identification of resources through a unique identifier, such as a URL.**masse2011rest**
 - The manipulation of resources through representations. This means that information about the resource is only exchanged using representations. As an example, a set of data could be stored in a database on the server. When a client requests the resource, a JSON representation of the data is sent, not the actual database. This ensures flexibility and interoperability, as the data could also be represented using other models, such as XML.**masse2011rest**
 - Self-descriptive messages, which means that a HTTP message should always completely describe the action to be performed on the server. No additional knowledge or state should be required to understand the message.**masse2011rest**
 - Hypermedia as the engine of application state. This means that links to other resources can be used in a resource's representation to describe its state or the state of other resources.**masse2011rest**
- Layered System: This enables the use of proxies, gateways to enforce security balance server load.**masse2011rest**
- Cache: The response from a web server should always include information about the cacheability of the response data. This ensures load reduction, availability and low latency.**masse2011rest**
- Stateless: A web server must not be required to save information about the state, which the client receiving the data is in. Clients should always communicate their state to the server so it can respond accordingly.**masse2011rest**
- Code-On-Demand: This means that servers must be able to transfer the execution of code to the client, e.g. in the form of client-side JavaScript.**masse2011rest**

All RESTful APIs should follow these rules and users of the API can expect it to handle resources and state in this way.

2.8 Basic Concepts of Music Theory

Since this paper aims to use Music streamed on Spotify to create a machine learning model it is also necessary to briefly discuss the Basics of Music. Therefore, first a definition for music is given and after that the basic characteristics of music are analyzed. Secondly methods are shown that help to classify music.

Music itself is a form of art that combines either vocal or instrumental sounds, sometimes both to express an emotion or convey an idea.**Becker2021WhatIsMusic** Music plays a key factor in many cultures of the world and is unique and different in each of them. But despite being able to define music it can be quite difficult even for artists itself to describe it and they struggle to put their perception into words. Famous Tenor and Saxophonist of the 1950s/1960s John Coltrane described Music as follows. "My music is the spiritual expression of what I am my faith, my knowledge, my being. When you begin to see the possibilities of music, your desire to do something really good for people, to help humanity free itself from its hang ups. I want to speak to their souls."**Havers2021Sax**

2.8.1 Melodical Components of Music

Based on John Coltrane's quote, music could colloquially be described as something, "that is used to trigger human sense".**Havers2021Sax** To achieve this music consists of various elements that work together harmoniously to form a pleasant-sounding melody. In music theory there are different opinions about the number of elements, which again only illustrates how complicated grasping is from the point of view of analysis. Many factors are highly interdependent and correlate with each other.

- **Pitch and Melody:** Pitch is the word used to describe the highness or lowness of a musical sound. A series of Pitches, also described as scales, is used to create a Melody. Melodies can be derived from various scales for e.g., the traditional major and minor scales of music or even more unusual ones like the whole tone scale.**Hemming2015** Furthermore melodies can be described in two ways. Conjunct melodies are smooth and easy to sing or play. Disjunct melodies however represent the opposite. They are ragged or jumpy and difficult to sing or play.
- **Harmony and Chords:** Harmony can be described as the sound created when two or more Pitches are performed at the same time to form a chord. Chord is simply the definition for three or more notes sounding at once. Chords themselves are used to create a musical mood. Harmony can be split up into two categories taking the sound of pitches into consideration. Harmony can sound Consonant,

which means the pitches sound pleasant together or Dissonant, meaning the pitches sound unpleasant together.**Hemming2015**

- **Rhythm:** Rhythm refers to the recurrence of notes and silences in time. A rhythmic pattern is formed by a series of notes and silence repeats. In addition to signify when notes are played, rhythm also defines how long notes are played and with what intensity. The difference in time creates varying note durations as well as different type of accents.**MilneMusicFundamentals**
- **Texture:** Texture indicates the number of instruments or voices that are used to contribute to the density of the music. Texture can be split up into 3 types of Monophonies, Homophony and Polyphony. Monophony describes a single layer of sound e.g. a solo voice. Homophony is a melody with an accompaniment which can be a lead singer with a band or a solo singer an a guitar or piano. Polyphony is a form of texture which consists of two or more independent voices. One voice forms the melody and the other forms a support role. This could be for example a lead singer with a choir.**2019ShawMusic**
- **Timbre:** Timbre, also known as tone colour or tone quality, can be described as the specific tone or quality an instrument or a voice has. Timbre helps to distinguish instruments from each other when playing the same melody or simple notes. For example, a “C” Note on the Piano and a sung “C” have the same pitch but different sound quality which gets differentiated by timbre. Timbre usually can be described with adjectives used to describe color, temperature, or the human voice e.g., warm, cold, metallic, harsh, dry.**Hemming2015**

2.8.2 Lyrics and Instruments

Another way to characterize music is Text. Lyrics are the definition for the linguistic part of a song. Lyrics usually consist of verses and choruses and can be implicit or explicit. More or less every song uses unique lyrics but nevertheless we can categorize lyrics by topics they address.**Shipman2014Analysis** Statistics have shown that the most common themes are Growing Up, Friendship, Statements of Discontent, heartbreak or Death.**2020AIMMListening** Another factor that must be taken into consideration when characterizing music are instruments. In the previous chapter we already talked about the fact that timbre is used to distinguish instruments from each other. The timbre focuses mainly on the sound of the instruments to distinguish them. However, we can distinguish instruments not only by the way they sound but also by how the sound is produced in the first place. This is done with a classification of the instruments into 5 categories.

- Idiophones: Sound gets produced by the body of the instrument vibrating e.g., xylophones
- Membranophones: Sound produced by vibration of a tightly stretched membrane such as drums
- Chordophones: Sound produced by vibrating strings e.g., piano or cello
- Aerophones: Produce sound by vibrating columns of air such as the pipe organ or oboe
- Electrophones: Produce sound by electronic means e.g., synthesizers or electronic instruments.

Independently of the tone production, a classification according to the type of player is also possible, i.e., wind instruments, percussion instruments, string instruments, keyboard instruments, plucked instruments.**GoshenInstrumentClass**

2.8.3 Chronological Classification

Not only compositional aspects can be used to distinguish music, but also the history of music provides a basis for this. This is done with the help of epochs. In music, an epoch is defined as a period in which stylistic similarities prevailed. However, epoch terms are a bit problematic because they give the impression that different styles were abruptly replaced. This is not the case, because with different styles, which often do not resemble each other at all or even contradict each other, there was often a smooth transition or simultaneous coexistence. Moreover, epochs are usually only generic terms for many undercurrents. The epochs that are considered relevant and for the most part include all styles since the 8th century are the Middle Ages, Renaissance, Baroque, Classicism, Romanticism, and the modern era.**LexikonMusikepochen** Although the music classifications of the classical and modern eras occur together in history, there are several differences in distinguishing music. While classical eras could often be delineated by, for example, the preferred use of instruments, the increased emergence and sheer diversity in music since the 20th century has made it almost impossible to bundle all styles under a single epochal term. For this reason, genres were developed for the classification of so-called "new music".**MusicfixRichtungen**

2.8.4 Genres

Musical genres assign different pieces of music with common characteristics to a unified tradition, history, or convention. Thus, a genre combines songs that sound the same or emit the same feelings, as well as songs that are based on the same characteristics of lyrics and instruments in songs. Dividing music into genres has become standard in today's music industry and due to the diverse number of possibilities in the creation of songs, there are now already over 1800 recognized specific genres with various subcategories. These subcategories or also called subgenres have again basic characteristics of the main genre as a basis but also their own characteristics to clearly distinguish themselves in their own genre and are often also often called style of the genre. These subgenres in turn have subcategories so that somewhat of a "genre tree" can be formed. The sheer number of genres and subgenres reflects the diversity of music. In this thesis the focus will be on the 3 main genres Hip Hop, Jazz and Rock which will be discussed in more detail below.**MusicfixRichtungen**

Hip Hop

Although Hip Hop is often widely considered as a synonym for rap music it can rather be defined as somewhat of a cultural movement or a form of lifestyle that includes a music genre. Hip Hop culture consists of various elements which are united under this umbrella term. Elements are djing/turntablism, rapping, beatboxing, breakdancing, and graffiti/visual art. **MusicalDictHipHop** These typical street style elements or activities created a cultural revolution and shaped music styles, fashion, technology, art, and more. **TateHipHop** Even to this day hip hop continues to develop new art forms that impact the lives of new and old generations. When it comes to the history of Hip Hop the culture has its origins in the 1970s where it first appeared in African American ghettos. Because of this heritage, hip hop still sees itself as street culture, and musicians who practice this form of music often have close ties to gangs, clans, and poverty. **Rory2019** Basic elements of hip hop come from the genres of funk and soul, which are also influenced by African Americans. Especially because of the distinctive rhythms, these genres offered themselves very well as a basis. Through the development of hip hop, however, it is also becoming more and more unique in type and form. The main characteristic of the music of hip hop is the interaction between a rapper and a beat. Rarely are real instruments used for the music and beats are usually created by electronic instruments or samples which is, the use of already finished or existing sound/music recordings. Beats can be very diverse and can represent many moods, e.g., aggressive, relaxed, harsh, etc. **MusicalDictHipHop** But the main focus in rap is usually on the rapper himself and even more on the lyrics. Rappers use rhythm, lyrics, and the timbre of their voice to express themselves. The Artists use their voice as

an instrument and more importantly different voice pitches to adapt to the intention of the lyrics. Rappers are often measured by the scene on their so-called "flow", the ability how quickly and how smoothly they can perform chants without errors. The themes of Hip Hop songs are often poverty, drugs, violence, struggles, or righteousness. Since many hip-hoppers come from ethnic and racial minorities, they often use personal life experiences for their lyrics. In addition, many songs also want to convey a message that often revolves around necessary political changes, social justice, or grievances in society.**Goodrich2017**

Jazz

Jazz is a musical style in which improvisation plays a central role. In many jazz performances, artists often play pieces they just made up from their heads and perform them on spot. Jazz has its historical roots in the early 19th century America more specifically in New Orleans. The city was an ideal breeding ground for jazz music because of given cultural diversity. The percentage of ethnic minorities was much higher in this city than elsewhere in America, which is why it was often called a melting pot of cultures. African Americans, people from the Caribbean, European immigrants, and sometimes even white Americans usually lived in the same neighborhoods and racially segregated ghettos that often existed in other American cities were not present here.**Beeck2021Jazz** However, the biggest influence on Jazz had the Afro-American culture as the music somewhat reflected the breaking away from previous rules of slavery and oppression. The improvisation of songs should act symbolically as a counter to the rules they had to deal with for a long time. Since its inception, jazz has gone through many different phases. The beginnings of jazz, from the 1910s to the 1920s, were characterized by small bands, often consisting of only a frontman and a few accompanists. The frontman often improvised pieces using a cornet or a trumpet, while the accompanists supported him with clarinets or trombones. In addition, often instruments like the banjo, piano, double bass, or drums were used to create a rhythm. In the 1930s and 1940s, the Swing and Big Band era emerged. Now, for the first time, singers appeared before big bands and bandleaders, and the clarinet was largely replaced by the saxophone. Moreover, the jazz epicenter shifted from New Orleans further and further to New York.**Wildridge2020** The 1950s and 1960s then introduced laid-back cool jazz in contrast to the more fast-paced songs of the previous decades. Here a jazz quartet often played soothing and slow songs. With the introduction of electronic instruments in jazz from 1970 onwards, many subgenres were formed, such as jazz-rock. Until today, there are many forms of jazz, which are mainly oriented to the New Orleans origins and the 3 mentioned eras. Jazz has its main musical roots in the blues, but there are also elements of rock and classical music in it.**JazzAmHistory** A distinctive rhythm is a key characteristic of jazz music, these are created mainly by "swinging" eighth notes. The so-called swinging is created by emphasizing one note of the eighth

note pair while the second note is lighter and "swings" to the next note. Jazz is also very polyphonic, which means that many sounds are played simultaneously and as a result, various layers of harmony are laid over an initial basic melody. In addition to the use of classical European instruments such as the saxophone or trumpet, instruments such as drums, bass, keyboard, guitar, and trombone are often used. Some types of jazz also have front singers, but often pieces are played without vocals. **2020 Masterclass Jazz** As already mentioned before. As already mentioned before the main characteristic of Jazz is the spirit of improvisation. This unifies almost all forms of jazz music. All members of a jazz band can be asked to improvise on a jazz tune when performing it. In addition, jazz artists attach great importance to imprinting their own sound and style on the music, so they usually even play their own songs slightly modified or with a distinct style. This leads to the fact that thousands of jazz recordings can be found for the same song, but they all sound different. Finally, it is important to mention that jazz can also reflect many different emotions. Everything from pain to joy is possible. For many People of Colour, it resembles the feeling of freedom, because for them, as mentioned above, jazz represents a strong voice against suffering, oppression, and injustice. **Musical Dict Jazz**

Rock

The musical genre Rock is a popular music genre that combines elements of rhythm and blues, jazz and country music while adding electric instruments. It originated as Rock 'n' Roll in the late 1940's and early 1950's and of course also is constantly changing and evolving. The basis of rock formed the music genres blues, gospel and country. Early rock 'n' roll came from cities like Memphis, Chicago, New Orleans or St. Louis. However, the genre spread very quickly throughout Western culture, and so it was mainly the British who liked the genre very much and who, in turn, took it to new heights. **2021 Masterclass Rock** British bands like the Beatles or the Rolling Stones emerged and became very popular in the USA as well. Rock dominated the music with its loud, dynamic, energetic and intense style for almost over 50 years until it was replaced by hip hop as the most mainstream genre. One of the main characteristics of rock is the infectious beat and rhythm. The music was primarily designed for dancing and was a clear distinction from other music genres popular at the time, such as jazz or swing. That's why from its start in the 50s the genre was very popular among young people as they felt they could break out of rigid traditions. Furthermore it is also important to mention that rock, like hip-hop later on, not only had an impact on music but also a cultural influence on the generations from the 1950s to the 1990s. **2021 Masterclass Rock** Rock especially supported the sense of rebellion and social justice in the western world of the 1960s and influenced clothing style, hair style and even attitude of its listeners for over 40 years. The older, more conservative generation at the time, which favored more quiet songs, rather detested rock. The energy already

mentioned above is a unique characteristic that sets rock apart from many other genres. Rock music is often very wild, impulsive and driving. But what defines rock the most is the use of electric instruments and especially the use of the electric guitar.**MusicalDictRock** The indispensable electric guitar makes rock probably the genre that is most influenced by a single instrument. Pioneers of rock like Elvis Presley, Jimi Hendrix or Chuck Berry experimented a lot with the electric guitar and used the unique sound to their advantage. Hendrix in particular often used the so-called "scream" of the electric guitar as his trademark, where he plucked the string in such a way that a shrill sound could be heard through the amplifier. This instrument allowed the artists to reach new melodic aspects and pitches that are not achievable with the use of pure acoustic instruments.**MusicalDictRock** Rock music is typically performed only in bands with a lead singer. He usually plays an electric guitar himself and is accompanied either by other electric guitars, normal guitars, other electric instruments and a drum kit. Lyrical texts of rock are also very diverse due to its division into many subgenres. Lyrics may not be very profound other lyrics by artists like Bob Dylan, however, are considered comparable to fine poetry. Rock music subgenres are very different and vary greatly in terms of rhythm and tempo. For example, the music genre of heavy metal is almost incomparable with soft rock.**Clark2021**

3 Implementation

3.1 Data Collection

In this section the approach and implementation of data collection for this project is examined. First, the basic requirements for the dataset are given and some existing datasets are evaluated. Then the resources used are presented and the general approach to data collection is explained, including authorization and how features and labels are collected. Lastly, the concrete implementation in Python is explained.

3.1.1 Requirements for the dataset

Basic requirements the dataset should fulfill are

- **Includes Spotify song features**

Spotify provides a set of song features that were generated using their own models. The dataset should include these features, as they are needed to train the model.

- **Includes genre as label**

The dataset needs to include the genre of the track to use as a label for the classifier

- **Has sufficient sample size per genre**

In order to train the model well, a sufficient sample size is needed per genre. It was not known before collecting the data, how many samples are enough to get a decent result, which is why the dataset with the highest sample size is preferred, without compromising too much on the other criteria.

- **Song and Artist name**

The best way to filter out duplicates is to use the song and artist names. Spotify does provide a track id for each song, however, if a song is released twice (e.g. as a single and later in an album), these track ids will differ which will lead to a duplicate entry.

Additional fields are not going to be used in this analysis, but might still be collected in order to publish the dataset and enable others to use it for different applications.

3.1.2 Existing Datasets

As this paper examines creating a model specifically on Spotify Song Data, a search on the internet was conducted first, to find a potential pre-made dataset, pulled from the Spotify **API!**, which could be used. Kaggle¹ lists an extensive catalogue of community provided datasets, so the main sources of this search were Kaggle and Google search for the term "Spotify Song Data". Kaggle lists a couple of datasets that could be applicable to the research question in this paper. Some examples of datasets listed are given and explained, why they could not be used for this project.

"Spotify music analysis" by user Aeryan² is a dataset of 2017 rows, which includes musical features like acousticness and tempo, the song title and artist, but lacks a genre field. Because of the small sample size and the missing genre field, this dataset could not be used.

There are multiple datasets which include songs that were featured in Spotify's "Top 50" Playlists, charts, or year in review, recorded at a single point in time or historically.^{3,4} These could not be used, as the sample size is again too small and the focus is specifically on the most popular tracks and not a wide variety of music in a genre.

"Dataset of songs in Spotify"⁵ is the most promising dataset examined, as it has a big sample size and includes genre data. However, the methodology of how the data was collected is not included and there could be multiple ways of how genre data for a given song is collected, as is explained later. Also the genres are limited to very specific directions of Electronic Dance Music and Hip-Hop.

As no optimal dataset for this research paper could be found using our search criteria, a dataset was specifically created for this paper using the Spotify Web API.

3.1.3 Ressources and Approach

Spotify provides extensive documentation for developers on their developer website. **SpotifyDev** This includes development and design guidelines for teams, that want to integrate Spotify's service into their own apps, documentation on iOS and

¹ Link to the Kaggle Website: <https://www.kaggle.com/>

² Link to the dataset: <https://www.kaggle.com/aeryan/spotify-music-analysis>

³ Link to the dataset: <https://www.kaggle.com/nadintamer/top-spotify-tracks-of-2018>

⁴ Link to the dataset: <https://www.kaggle.com/leonardopena/top50spotify2019>

⁵ Link to the dataset: <https://www.kaggle.com/mrmorj/dataset-of-songs-in-spotify>

Android development a community forum, a developer dashboard and the Web API documentation, which is the main resource for data collection from Spotify.

The **API!** is based on the **REST!** architecture. The different endpoints return **JSON!** meta-data directly from the Spotify Data Catalogue **SpotifyWebAPI**. There are also features to query for user related data using an authorization flow with the users Spotify account, but this is not relevant in this context. **SpotifyWebAPI** Requests to the **API!** are made via HTTP GET or POST methods. The **API!** can be used by anyone, but authorization via the OAuth protocol is required to access data from the **API!**. To explore the **API!** and find endpoints to use, Spotify provides a developer console, which can be used to send requests and see what kind of responses come back. This is not suitable for saving the data or making multiple requests programmatically, but is helpful for API exploration. As there is not one single endpoint that delivers all required fields, multiple queries that build on top of each other have to be made.

The approach began with using the **API!** reference to get an overview over the endpoints and their responses. The specific endpoints that might return interesting data were queried using the Spotify Web Console to see a response with live data and which exact fields are returned. Beyond the Web Console, the tool "Postman" was used to explore the **API!**. It is a platform that can be used to make HTTP requests to an API and store these requests in a collaborative environment. **Postman** **WhatIs** It supports the required authorization workflows and enabled the research team to explore endpoints together, easily make **API!** calls without having to authenticate by hand, and save the endpoints and required input parameters in a shared workspace. Once exploration was complete, the complete data collection was implemented in Python 3, mainly using the libraries http, json and requests. The final result was saved as a CSV file to be used for data exploration and further processing.

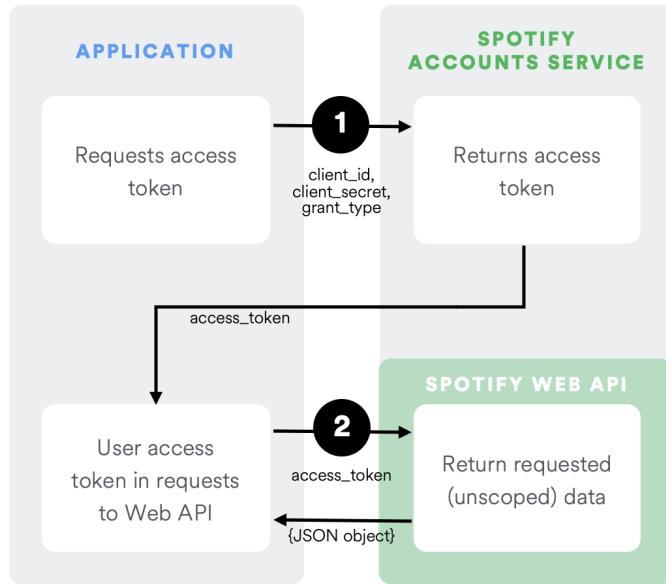
Many API endpoints give the option to specify a country and locale that the results should apply to. Whenever possible these values were explicitly specified to eliminate the risk of receiving data for multiple different countries or languages, which would have made the dataset incoherent. For country, the country tag *US* was used, which means that the API replies only with datapoints that are applicable to users in the United States of America. For locale, the value *en_US* was used in order to receive the data in the english language.

3.1.4 Authorization

Using the Spotify documentation for authorization workflows **SpotifyAuth**, authorization was first tested using Postman and then implemented in Python. Using a Spotify account to log in, the developer dashboard can be accessed. Here an application was registered

for the research project. On the API Dashboard, a "Client ID" and "Client Secret" can be retrieved. These credentials are used to start the authorization flow, as described by Spotify in Figure ??.

Figure 16: Spotify Authorization Flow



Source: [SpotifyAuth](#)

Step one is a post request to the Spotify Account Service containing the client id and client secret in the body, which returns an access token valid for one hour. This token can be used to access any endpoint of the actual **API!** that does not require user specific data. When the token expires, a new one has to be requested before querying the **API!** again. Figure ?? shows the full request and response to acquire the token.

Figure 17: Access Token Request

POST	https://accounts.spotify.com/api/token
<i>request access token</i>	
Body	application/x-www-form-urlencoded
	<pre> 1 { 2 "grant_type": "client_credentials", 3 "client_id": client id from application dashboard, 4 "client_secret": client secret from dashboard 5 } </pre>
Response	application/json
200	ok
	<pre> 1 { 2 "access_token": "BQDgQCSx-tIMD9LfVeZxm6Yml2p_WbEU3Q 3 9ENsVl7e--6d_vockTsfzMVUhPWiHSSnFUuHvm_9POA1kYEw" 4 , 5 "token_type": "Bearer", 6 "expires_in": 3600 7 } </pre>

3.1.5 Getting Features

In order to predict the genre of a track based on audio features, these features have to be requested for every track. Spotify provides an endpoint to get audio features for up to 99 tracks at a time. The typical request/response pattern for the audio-feature request endpoint is shown in figure ??.

Figure 18: Audio Feature Request

GET	https://api.spotify.com/v1/audio-features?ids={ids}
<i>request audio features for id</i>	
Parameter	
ids	comma seperated list of up to 99 song ids
Response	application/json
200	ok
	<pre> 1 { 2 "audio_features": [3 { 4 "danceability": 0.677, 5 "energy": 0.638, 6 "key": 8, 7 "loudness": -8.631, 8 "mode": 1, 9 "speechiness": 0.333, 10 "acousticness": 0.589, 11 "instrumentalness": 0, 12 "liveness": 0.193, 13 "valence": 0.435, 14 "tempo": 82.810, 15 "type": "audio_features", 16 "id": "2e3Ea0o241ReQFR4FA7yXH", 17 "uri": "spotify:track:2e3Ea0o241ReQFR4FA7yXH 18 ", 19 "track_href": "https://api.spotify.com/v1/ 20 tracks/2e3Ea0o241ReQFR4FA7yXH", 21 "analysis_url": "https://api.spotify.com/v1/ 22 audio-analysis/2e3Ea0o241ReQFR4FA7yXH", 23 "duration_ms": 211497, 24 "time_signature": 4 25 }, 26 ... 27] 28 }</pre>

With the exception of *type*, *id*, *uri*, *track_href* and *analysis_url*, all of the fields included in this response can be used as features in the dataset. However, this api call expects a track id, which we need to get using other api calls first. This could be a search endpoint, getting all tracks in a playlist, etc. Also, it does not give the track or artist names and doesn't include a genre.

3.1.6 Getting Track IDs and Labels

There is no simple endpoint that takes one or more track ids and returns a "genre" field in its response. The exploration of the API using the reference, web console and Postman only revealed two ways of getting the genre of a track.

The first way is using the artist of a track. Given a track id, the artists of the track and their corresponding ids can be requested by using the `/tracks/{id}` endpoint. Then, using the artist id, the genres that an artist is known for are returned, as shown in figure ??.

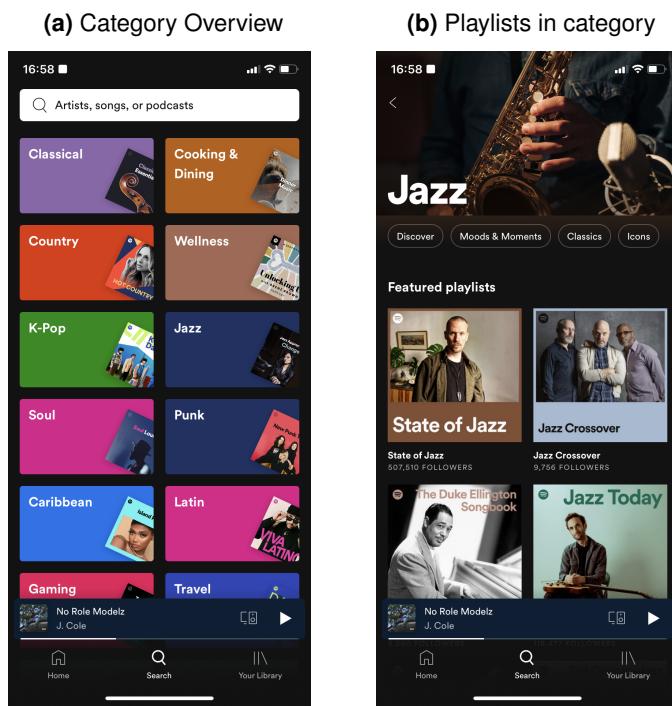
Figure 19: Artist Request

GET	https://api.spotify.com/v1/artists/{id}
request information about an artist by their id	
Parameter	
id	id of the artist
Response	application/json
200	ok
1	{
2	"external_urls": {
3	"spotify": "link to resource..."
4	},
5	"followers": {
6	"href": null,
7	"total": 15554811
8	},
9	"genres": [
10	"conscious hip hop",
11	"hip hop",
12	"north carolina hip hop",
13	"rap"
14],
15	"href": "link to resource...",
16	"id": "613HvQ5sa6mXTsMTB19r05",
17	"images": [links to album covers...],
18	"name": "J. Cole",
19	"popularity": 89,
20	"type": "artist",
21	"uri": "spotify:artist:613Hv..."
22	}

The exemplary API response shows a problem with this approach. One artist can be sorted into multiple genres. A given track might be associated with either of the artists genres, but the data does not show, which one exactly. Additionally a track might have multiple artists which further complicates this. Given these circumstances, this approach is problematic.

The second way is Spotify's "categories" feature. The app's search tab provides a number of categories that a user can browse to find new music in their preferred genre or style. In figure ??a an overview over some of the categories that are available in the Spotify app is shown. There are categories of multiple types, e.g. specific activities, like "Cooking" or "Gaming", or places, like "At Home" or "In the Car". But there are also categories for nearly all major genres. In the app screenshot there for example "Classical", "Jazz" or "Soul". These categories can be used to get tracks that belong in each specific category. When a user taps on one of the categories, playlists that contain tracks of the respective category are shown to the user, like shown in figure ??b.

Figure 20: Categories and Playlists in Spotify App



The API mirrors the app's behaviour and provides an endpoint to get a list of categories and their ids, one to get all playlists and playlist ids in a category, and one to get all tracks and track ids in a playlist. This chain of API calls is used to request every track in every playlist in a certain category.

3.1.7 Python Implementation

This section describes, how the previously explained API calls are used to request data from the API and save the complete dataset as a CSV file. The script was implemented in a way that lets anyone with a registered Spotify developer application run it to gather their own data. It supports filtering for certain genres to speed up the process and exporting the current data after each step to be able to pause the data collection and check the current results or continue later. The details of these functions are explained in this section using code snippets.

The Python requests library is used to execute the requests, as it supports the features needed to send the GET and POST requests we need without having to write complicated code. The dotenv library is used to read the client id and secret from a seperate .env file, rather than writing it into the code. This prevents these sensitive credentials from being committed into version control, which is hosted in a public GitHub repository and would therefore make the credentials public. The *json* library is used for transforming Python dictionaries into json data, *math* for access to rounding functions, *os* for managing system paths and access to files, *http* to define retry and timeout behaviour and *csv* to finally transform the json data into a CSV file.

First, the credentials are read using *load_dotenv*. Then a method is defined which takes the client id and secret as input, executes a POST request to the accounts API, as shown in ??, and return the access token from the *access_token* field in the response.

```
1 #Get environment variables from ".env" file and read credentials
2 load_dotenv('.env')
3 client_id = os.environ.get('CLIENT_ID')
4 client_secret = os.environ.get('CLIENT_SECRET')
5
6 # Authenticate and get an API Token from Spotify using a Client ID and secret
7 def getAuthTokenFromCredentials(id, secret):
8
9     url = "https://accounts.spotify.com/api/token"
10
11     payload = f'grant_type=client_credentials&client_id={id}&client_secret={secret}'
12     headers = {
13         'Content-Type': 'application/x-www-form-urlencoded',
14     }
15
16     response = requests.request("POST", url, headers=headers, data=payload)
17
18     return response.json()["access_token"]
19
20 auth_token = getAuthTokenFromCredentials(client_id, client_secret)
```

Next the http library is configured to use a ten second timeout and attempt five retries. If a request to the API is made and no response is given from Spotify's servers after ten seconds, the request will be sent again. If there is still no answer from the server after five attempts, an exception is thrown and the script ends. Timeouts are crucial as short disconnections from the Internet could result in the whole data collection step failing. The http setup code is found in appendix ??.

The next part of the script implements the filter mechanism

```
1 category_filter = None
2
3 # comment this out if you don't want to set a filter
4 category_filter = ["hiphop", "jazz", "rock"]
5
6 # tells data collection functions to not use filter if it's not set
7 if category_filter is not None:
8     use_filter = True
9 else:
10    use_filter = False
```

The option to filter the categories that are to be taken into account when collecting the data is given. Possible values are all valid category ids, which can be requested using the API call shown in figure ??

Figure 21: Categories Request

GET	https://api.spotify.com/v1/categories?country={country}&locale={locale}&limit={lim
<i>Get a list of categories used on Spotify's browse tab</i>	
Parameter	
id	id of the artist
country	only display categories available in a specific country
locale	the language in which the categories should be returned
limit	the maximum number of items to be returned
offset	index of the first item to return
Response	application/json
200	ok
<pre> 1 { 2 "categories": { 3 "href": "...", 4 "items": [5 { 6 ... 7 "id": "hiphop", 8 "name": "Hip-Hop" 9 }, 10 { 11 ... 12 "id": "pop", 13 "name": "Pop" 14 }, 15 ... 16], 17 ... 18 } 19 }</pre>	

The response shows an *id* field for each item in the *items* array. These ids can be used in the filter list. If the *category_filter* variable is undefined, no filter is set.

Next, the option to import existing data from a file is given. If a user already has the json data containing all playlists for all categories they want to use, they can load the json file here and skip executing the collection of categories and playlists and continue right at the collection of tracks for each playlist.

```

1  data = {}
2
3  #To load data_object from file instead of rerunning the scripts, uncomment this:
4
```

```

5     file = open(os.path.join("data_collection", "json", "tracks_full.json"))
6     data = json.load(file)

```

Next, the four main steps of data collection will be defined and executed:

1. Getting categories
2. Getting playlists
3. Getting tracks
4. Getting audio features

Each step is defined as a function with some input parameters and the resulting json data as output, which can then be used as input for the next step.

The function definition for getting the categories looks like this:

```

1 def getAllCategories(
2     # takes the requests session, which was set up in the beginning
3     requests_session,
4     # this is the token which was retrieved in the authorization step
5     auth_token,
6     # set to True if the category_filter should be used
7     use_category_filter=False,
8     # this takes the list which is used as a category filter
9     category_filter=None,
10    # set to True if the result of this step should be written into a json file
11    write_to_file=False,
12    # give the file path at which the json file should be stored
13    path_to_file=''):

```

The function first makes a simple API call with *limit=1* to get the total amount of categories available. This is necessary, because the API returns a maximum number of 50 entries per request, so if the total number of items exceeds 50, multiple calls have to be made.

```

1 # Establishing the requests session
2 http = requests.Session()
3
4 # First API call used to get the total amount of categories
5 headers = { 'Authorization': f'Bearer {auth_token}' }
6 url = "https://api.spotify.com/v1/browse/categories?country=US&locale=en_US&limit"
7     =1
8
9 try:
10     response = http.request("GET", url, headers=headers, data={})
11     if response.status_code != requests.codes.ok:
12         raise Exception
13 except Exception as e:
14     raise SystemExit(e)

```

`auth_token` is wrapped in an Authorization header conforming to the OAuth Bearer Token standard, which is used by the API for authorization. The URL is given with the right country and locale settings and the request is sent inside of a try block. If the response doesn't contain an HTTP status code 200, an Exception is raised and the program is stopped, as there is possibly an error in the request or response which could mean corrupted data.

Next, the total amount of categories is extracted from the response and the number of pages is calculated.

```

1   categoryAmount = response.json()["categories"]["total"]
2
3   # API only returns 50 items at a time. Offset can be used to gradually get all
4   # items
5   # Calculate number of pages with 50 items
6   pages = int(math.ceil(categoryAmount/50))
7   data = {"categories": []}

```

By dividing the category amount by 50 and rounding up to the next integer, the number of API calls necessary to get all entries is calculated. Each call returns a "page" of items, e.g. page one contains items 0 to 49 and is retrieved by using offset 0 and limit 50, page two contains items 50 to 99 and is retrieved by using offset 50 and limit 50.

A for loop is used to execute the request for every page as shown before. Each category's id and name are read from the response and appended to the data object:

```

1   for x in range(pages):
2       url = f"https://api.spotify.com/v1/browse/categories?country=US&locale=en_US&
3           limit=50&offset={x * 50}"
4
5       try:
6           response = http.request("GET", url, headers=headers, data={})
7           if response.status_code != requests.codes.ok:
8               raise Exception
9       except Exception as e:
10          raise SystemExit(e)
11
12       # categories are stored in the data dictionary
13       for el in response.json()["categories"]["items"]:
14           if (use_category_filter == True and el["id"] in category_filter) or
15               use_category_filter == False:
16               data["categories"].append({
17                   "id": el["id"],
18                   "name": el["name"]
19               })

```

Outside of the for loop, the resulting data is written to a file if `write_to_file` was set to `True` and the data object is returned:

```

1   if write_to_file == True:
2       with open(path_to_file, 'w') as outfile:

```

```

3         json.dump(data, outfile, indent=2)
4
5     return data

```

The data collected after this step looks like this:

```

1 {
2     "categories": [
3         {
4             "id": "hiphop",
5             "name": "Hip-Hop"
6         },
7         {
8             "id": "pop",
9             "name": "Pop"
10        },
11        ...
12    ]
13 }

```

This concludes the function *getAllCategories*. The data retrieved from this step is used for the next step, getting the playlists. The function definition here is slightly different:

```

1 def getPlaylistsForCategories(
2     requests_session,
3     auth_token,
4     data_object,
5     write_to_file=False,
6     path_to_file=''):

```

Instead of the filtering options, which only apply to the categories, this function takes the *data_object* from the previous step. This could be passed in directly from the previous function or read from a file. The function implements a nested for loop. The outer loop iterates over all categories in the data object. The inner loop uses an endpoint to get each category's playlists. The API request/response is shown in figure ??.

Figure 22: Get Category's Playlists Request and Response

GET	https://api.spotify.com/v1/categories/{category_id}/playlists?country={country}&limit={limit}&offset={offset}
<i>Get Category's Playlists</i>	
Parameter	
category_id	category id
country	only display categories available in a specific country
limit	the maximum number of items to be returned
offset	index of the first item to return
Response	application/json
200	ok
<pre> 1 { 2 "playlists": { 3 "href": "...", 4 "items": [5 { 6 "collaborative": false, 7 "description": "New music from YoungBoy...", 8 "external_urls": ... 9 "href": "...", 10 "id": "37i9dQZF1DX0XusuxWHRQd", 11 "images": ... 12 "name": "RapCaviar", 13 "owner": ... 14 "primary_color": null, 15 "public": null, 16 "snapshot_id": "...", 17 "tracks": { 18 "href": "...", 19 "total": 50 20 }, 21 "type": "playlist", 22 "uri": "spotify:playlist:37i9dQZF1DX..." 23 }, 24 ... 25], 26 "limit": 50, 27 "next": "...", 28 "offset": 0, 29 "previous": null, 30 "total": 50 31 } 32 }</pre>	

As there is an upper limit of 50 playlists per request, the same paging method is used as in the category function. If an items *type* is not *playlist* it is filtered out, as it doesn't fit the schema. Then each playlists name and id is stored in the data object, which is returned from the function. The JSON data after this step looks like this:

```

1  {
2      "categories": [
3          {
4              "id": "hiphop",
5              "name": "Hip-Hop",
6              "playlists": [
7                  {
8                      "id": "37i9dQZF1DX0XUsuxWHRQd",
9                      "name": "RapCaviar"
10                     },
11                     {
12                         "id": "37i9dQZF1DX6GwdWRQMpq",
13                         "name": "Feelin' Myself"
14                     },
15                     ...
16                 },
17                 ...
18             }
19 }
```

The third step gets all tracks in each playlist. The function definition is the same as before, as this function also takes data from the previous step. This time, there are three nested for loops. One for looping through the categories, then another one for looping through the playlists. In this one the number of pages is calculated. The third loop requests the pages and appends the tracks to the playlist data. The request parameters for this request differ from the other steps. Instead of country, there is a market parameter, which essentially does the same thing. There is also a fields parameter, which takes a string that can be used to define which fields the API should return. This is done to save bandwidth, as the response of this endpoint is very large and most of the datapoints might not be needed by the client. In this case, the value *items(track(name, id, album(name, id), artists))*, *total* is used to return only items of type track, their id and name, their albums name and id, and their artist information. Also the total amount of tracks in the playlist is returned. The parameter *additional_types* is set to *track*, as podcasts are irrelevant to this analysis. An exemplary API request/response is shown in figure ???. Because a track can have multiple artists, each artist's name and id are extracted and saved with the track. The function code can be found in the appendix.

Figure 23: Get Playlist's Tracks

GET	https://api.spotify.com/v1/playlists/{playlist_id}/tracks?market={market}&fields={fi}
<i>Get Playlist's Tracks</i>	
Parameter	
playlist_id	playlist id
market	only display items available in this market
fields	a string describing which fields to return
additional_types	supported item types. valid types are "track" and "episode"
limit	the maximum number of items to be returned
offset	index of the first item to return
Response	
200	ok application/json
1	{
2	"items": [
3	{
4	"track": {
5	"album": {
6	"id": "4oxmme6i4mypSt2DDzPTsW",
7	"name": "DS4EVER"
8	},
9	"artists": [
10	{
11	"external_urls": {
12	"spotify": "..."
13	},
14	"href": "...",
15	"id": "2hlmm7s2ICUX0LVIhVF1ZQ",
16	"name": "Gunna",
17	"type": "artist",
18	"uri": "spotify:artist:2hlmm7s2ICUX0LVIhVF1ZQ"
19	},
20	...
21]
22	}
23	},
24	...
25],
26	"total": 50
27	}

The fourth step is getting the audio features for each track. The request that is used was already shown in ?? The function used has the same definition as before but implements

one more for loop to iterate through the tracks. Again, the code for this step is found in the appendix.

The data collected after this step looks like this:

```

1 {
2   "categories": [
3     {
4       "id": "hiphop",
5       "name": "Hip-Hop",
6       "playlists": [
7         {
8           "id": "37i9dQZF1DX0XUsuxWHRQd",
9           "name": "RapCaviar",
10          "tracks": [
11            {
12              "id": "2AaJeBEq3WLcfFWly8svDf",
13              "name": "By Your Side",
14              "album": {
15                "id": "2RrZgDND03MLu6pRJdTkz5",
16                "name": "By Your Side"
17              },
18              "artists": [
19                {
20                  "id": "45TgXXqMDdF8BkjA83OM7z",
21                  "name": "Rod Wave"
22                }
23              ],
24              "features": {
25                "danceability": 0.649,
26                "energy": 0.508,
27                "key": 8,
28                "loudness": -10.232,
29                "mode": 1,
30                "speechiness": 0.0959,
31                "acousticness": 0.0345,
32                "instrumentalness": 3.59e-05,
33                "liveness": 0.0736,
34                "valence": 0.405,
35                "tempo": 157.975,
36                "duration_ms": 194051,
37                "time_signature": 4
38              }
39            },
40            ...
41          ]
42        }
43      ]
44    }
45  ]
46}
```

This JSON contains all the necessary fields needed for continuing with the CRISP DM process. To prepare the dataset for data analysis, the JSON data needs to be transformed into a flat structure to be stored as a CSV file. Another Python function is used to flatten the data. It takes the resulting data from the last step and transforms it into the following format.

```

1 [
2   {
3 }
```

```

3   "categories.id": "hiphop",
4   "categories.name": "Hip-Hop",
5   "categories.playlists.id": "37i9dQZF1DX0XUsuxWHRQd",
6   "categories.playlists.name": "RapCaviar",
7   "categories.playlists.tracks.id": "2AaJeBEq3WLcfFW1y8svDf",
8   "categories.playlists.tracks.name": "By Your Side",
9   "categories.playlists.tracks.album.id": "2RrZgDND03MLu6pRJdTkz5",
10  "categories.playlists.tracks.album.name": "By Your Side",
11  "categories.playlists.tracks.artists": "Rod Wave",
12  "categories.playlists.tracks.features.danceability": "0.649",
13  "categories.playlists.tracks.features.energy": "0.508",
14  "categories.playlists.tracks.features.key": "8",
15  "categories.playlists.tracks.features.loudness": "-10.232",
16  "categories.playlists.tracks.features.mode": "1",
17  "categories.playlists.tracks.features.speechiness": "0.0959",
18  "categories.playlists.tracks.features.acousticness": "0.0345",
19  "categories.playlists.tracks.features.instrumentalness": "3.59e-05",
20  "categories.playlists.tracks.features.liveness": "0.0736",
21  "categories.playlists.tracks.features.valence": "0.405",
22  "categories.playlists.tracks.features.tempo": "157.975",
23  "categories.playlists.tracks.features.duration_ms": "194051",
24  "categories.playlists.tracks.features.time_signature": "4"
25 },
26 ...
27 ]

```

The function is also found in the appendix. Notice, that the nested fields and arrays have been flattened to represent a two dimensional data structure, containing only one array to represent rows which contains JSON objects with 22 fields. Each field represents a column. In the original JSON data, *artists* is a list containing one or more entries. If the data was flattened in a way that groups the entries by each individual artist, there could be multiple rows for the same track. If for example *track1* had two artists *artist1* and *artist2*, there would be two datapoints, both containing the same song name, id and features, but different artists. As the dataset should be grouped by individual tracks, not by artists, which results in multiple entries per track, multiple artists are condensed into a comma-separated list during flattening. This list is stored in the *categories.playlists.tracks.artists* field. This means, that the individual artist data is lost in favor of grouping the dataset by tracks.

After this process, the resulting two-dimensional structure is converted to a CSV file and saved for further processing.

3.2 Data Understanding

3.2.1 Spotify

Spotify is a music streaming service and has established itself as the Number one in this market. Since their launch in 2006, the company has gained over 365 million users, of which nearly 45 percent are subscribed to the chargeable premium service. This Premium service comes at a subscription cost of nowadays 12,99€ per month, giving the opportunity to listen to every song that is available at Spotify - this being ca. 70 million tracks from over 1,2 million Artists.

Similar to other streaming services, Spotify also generates a lot of data. The question here is how to filter out data from music in order to personalize user accounts. The founders evaluated millions of songs and taught an AI to independently recognize patterns and certain structures within different genres. And how does this concept translate to Spotify? The results of the algorithm can be seen above all in the "Discovery Weekly Playlist". Users are presented with 30 songs that match their tastes here every Monday. And the algorithm is amazingly successful at this. At the moment, the Spotify algorithm, which to this day goes by the name Echo Nest, works because of three main components.

First of all, it creates an individual Taste Profile for each Spotify user. There, it is recorded which artists, songs, albums, playlists and also podcast and audio books the respective user listens to. And also how long he listens to the music tracks, where he does it and how often. This data is then compared with the Taste Profiles of other users. For example, if another user has a large overlap in musical taste with your own playlists, you will be shown songs in the Discovery Weekly playlist that the other user already listens to regularly. Playlists with more followers end up in the Discovery Weekly folders more often than those with fewer followers.

In addition, the algorithm analyzes the individual components of the music tracks. It breaks down each song available on Spotify by tempo, instruments, duration, highs and lows, timbre, and so on. A highly complicated analysis process is used to classify these categories, which Spotify itself published in a detailed presentation in 2015. Based on the large amount of data, the Spotify algorithm divides the songs into more than 1500 different genres so far. The classification can be viewed online on the website Every Noise At Once. However, in addition to many regional differences in music genres, "nonsensical" genres keep coming out in the AI's analysis. While you might be able to imagine something under deep power pop punk, genres like djent are not really close to reality. But the Echo Nest algorithm is always learning, after all.

By the way, this analysis step also includes the classification of songs into happy, sad, melancholic or other types of music. So the classification according to feelings. However, this feature is also controversial among experts. What defines sad or happy music? That always depends on the individual listener and his mood. To support the AI at this point, it reads the titles of the playlists of the users. If a song often lands in playlists with titles like love songs or romantic music, the AI assumes that the song is to be rated as a love song. At the moment, this is still the algorithm's biggest hurdle. After all, the classification based on playlists is naturally prone to error as well.

At the latest, the feature to search the world wide web allows Spotify to be much more precise in rating music than comparable competitors. The Echo Nest algorithm searches the entire web daily for blog posts, comments on websites, articles, YouTube videos, Facebook postings and comments and tweets, evaluates them, and uses them to create an opinion about a particular song. Thus, within a few days, the algorithm notices how a song is received overall. Whether the reviews are good or bad. Whether people recommend the songs to others or advise against them, etc. This quickly creates a specific rating for each track, which in turn influences the Discover function and other recommendation formats within the platform.

3.2.2 Feature analysis

In the course of the analysis carried out here, three different genres were selected in order to subsequently assign songs to the correct genre on the basis of the underlying data. The genres selected here are "rock", "hip-hop" and jazz. The selection was made under the original, subjective assumption that these genres differ particularly strongly in their characteristics and style. The assumption focused primarily on characteristics such as vocals, the general use of instruments, the instruments used, or the energy of the songs of the respective genre.

The important insight for further consideration of the analysis is here: Each genre that is considered, each song feature that is evaluated does not reflect reality, but only reflects the results of the Spotify algorithm. The analysis can, according to subjective opinion, assign a Hip-Hop song to the category, or the genre Rock without being wrong - since the assignment makes sense based on the Spotify evaluations. Since the categories in the selected genres are very accurate due to their popularity, the analysis is performed under the assumption that the two are congruent.

The Data Understanding step starts with filtering out the required songs of the selected genres. How the song data in Spotify is structured can be explained very well. Each

individual song has a so-called "Song ID". This is unique and gives the possibility to identify the respective song. In addition, each song has thirteen features, which capture and reflect various properties. Genres do not exist directly in the data structures. The closest to these are the so-called "Categories". These "Categories" are, as described in the previous part, a product of the algorithm used to create them based on their listenership and the songs with their features. These "Categories" are not always necessarily congruent with today's common genres, so there are, for example, categories with the name "x, y", but they represent, according to subjective observation, the most popular genres of reality to a large extent. Songs are not directly, but only indirectly assigned to a category. The connection runs over playlists, which contain the songs. These playlists also have a unique ID, the so-called "playlist ID", and are created by the algorithm. Similar to the categories, the algorithm uses the features of the songs, user stories and the like to create suitable playlists. Playlists are then assigned to the categories via the "Category ID". The path to the songs of a genre, or the song ID's begins with the selection of the appropriate category, or the appropriate categories, via the playlists assigned to the categories to be able to access the songs from there.

The features are decisive in the classification of the songs to the genres. For a better presentation, they are divided into three clusters below, which can be used to better capture them. These clusters are Mood, Properties and Context.

Musical standards

However, at the beginning the features that capture and reflect the standard properties of music will be discussed. These features are "Duration", "Key", "Mode" and Time Signature. The feature "Duration", as the name suggests, holds the information of how long the Track is. This is measured in seconds. The feature "Key" is measured with an Integer between -1 and 11. It holds information about the key the track is in. Integers map to pitches using standard Pitch Class notation. E.g., 0 = C, 1 = C/D, 2 = D, and so on. If no key was detected, the value is -1. "Mode" is an Integer (either 0 or 1) and indicates the modality (major or minor) of a track. The type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0. The feature "Time Signature" is measured in an integer between 3 and 7. It is a notational convention to specify how many beats are in each bar (or measure). The time signature ranges from 3 to 7 indicating time signatures of "3/4", to "7/4".

Mood

Following these features is the Mood cluster. This cluster lists those features that measure the emotive values in songs, i.e. whether the song encourages dancing, spreads a positive

or rather negative mood, etc. These include: "Danceability", "Valence", "Energy" and "Tempo".

Danceability, Valence and Energy are all measured in a float between 0 and 1. Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable. Valence as a feature describes the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry). Energy represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy. A value near 1.0 indicates high energy, while tracks near 0.0 indicates low Values in this Feature. At last, Tempo holds the information about the overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.

Properties

This cluster includes all features that capture the audio characteristics of the tracks, such as "Loudness", "Speechiness" and "Instrumentalness".

Loudness is measured in a float number between -60 and 0 and measures the overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Speechiness, measured in a float number between 0 and 1, detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.

Instrumentalness, measured in a float between 0 and 1 as well, plays the counterpart to Speechiness. It predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the Instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.

Context

This Cluster unions the Features "Liveness" and "Acousticness". These capture everything around the Songs, for example if it is played in front of a live audience.

Both Features are measured in a float number between 0 and 1. Liveness detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live. Acousticness is a confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.

3.2.3 Overall correlation

After these explanations, the characteristics of the features in the individual categories are evaluated. A correlation matrix is used for this purpose. This shows on a scale between 1 and -1 how features correlate with each other. Since every song uses every feature, a high correlation here does not mean that the features often appear together, but how similar the values per song are. To start with an overall look over the data, fig. 1 shows a genre independent matrix. This means that the features of all songs that could be gathered in the step of data collection were analyzed to create this figure.

Generally, the features of the "Musical Standards" cluster can be ignored in this representation, as they do not measure but only reflect characteristics (e.g., time signature or key). It can be seen that "Energy" and "Loudness" have the highest correlation of the features, with a value of approx. 0.65. So, energetic songs often seem to be loud as well. This can be easily explained, since, as explained in the definition of the features: "Typically, energetic tracks feel fast, loud, and noisy". In addition, there is a comparatively high correlation between "Valence" and "Danceability" of about 0.5, which means that positive songs, i.e., songs with a high value for Valence, encourage more dancing than rather negative ones. The correlation between "Energy" and "Acousticness" is particularly negative, with a value of approx. -0.7. Acoustic tracks therefore often do not seem to be very energetic and vice versa. The same can be said for "Acousticness" and "Loudness", with a correlation value of about -0.5. Outside of these extreme values, the remaining correlation values settle between -0.2 and 0.2. Worth mentioning here are the correlations between "Instrumentalness" and "loudness" at approx. -0.3 as well as "Instrumentalness" and Valence also at approx. -0.3. Furthermore, in the positive "Speechiness" and "Energy" at a value of approx. 0.3.

If these correlations are now considered on a category-specific basis, it can be seen that the patterns described above also appear there. However, the average correlation changes depending on the category. For example, hip-hop has the same characteristics as the

independent matrix category but shows fundamentally more negative correlations (cf. Fig. 1 with Fig. 2).

3.2.4 Correlation Music theorie - Data

In the final step of Data Understanding, the music theory described in the basic section is compared to the data used in the analysis. This is necessary to develop a deeper understanding of the data. As mentioned at the beginning of this chapter, theory and Spotify are not always necessarily congruent, which is why a clear delineation is important here. For this purpose, the properties of the genres in the music theory are tried to be compared with the help of the features (e.g., "You can't dance well to rock" - What is the overall value for Danceability?). To do this comparison, three features are chosen and discussed more deeply. Subsequently, a better picture can be obtained with the help of further representations.

Starting with rock, one feature can be extracted here: The songs are mostly loud. The loudness feature captures this property very well. It measures the actual, average loudness of a song in decibels, with most tracks falling in the -60 to 0 range, where -60 is very quiet and 0 is very high. Overall, across all three categories, this feature has an average of -8.2544, with a minimum value of -32.06 and a maximum value of 2.044. Interestingly, the loudest song is not found under rock, but under hip-hop. The quietest song is found in jazz.

Figure 1 shows a so-called distribution plot. The X-axis contains the scale of the feature used (in this case "loudness"). The Y-axis reflects the density of the respective category. This means that the more songs are found in the respective value, the higher the density. As can be seen in the figure, all categories offer a very wide range of volume levels. Rock and Hiphop turn out to be significantly louder than Jazz. Jazz spreads its songs over an area from -32.06 (smallest value overall) to -2.695 but locates most songs at about -10. Hip hop and rock go less wide but can accommodate more songs in the upper quarter of the scale. Rock places its point furthest to the right of the scale. Thus, it can be confirmed that rock can generally be classified as loud, but also shows as the loudest of the three categories. However, it must also be emphasized here that hip-hop comes shortly after. In addition, jazz is not as loud as the other two categories, but it serves a much wider range.

The next feature that can be scrutinized here is the large occurrence of instruments in jazz. This can be checked in the Spotify data via the "Instrumentalness" feature. It measures the sole presence of instruments in songs with values between 0 and 1. The closer the value approaches 1, the higher the probability that there are no vocals or other elements

in the song. The overall average for this feature is 0.174. The lowest value overall is 0.0 while the highest goes up to 0.989.

For a better understanding, Figure 2 also shows a distribution plot, but adapted to the current feature. It shows that both hip-hop and rock are almost exclusively found in the 0.0 - 0.1 range. This means that songs in these categories almost exclusively contain vocals or similar. However, slight increases in the range between 0.8 and 1.0 can also be seen. Jazz, on the other hand, has two equally strong expressions, on the one hand in the range around 0.0, on the other hand also between 0.8 and 1.0. Since jazz often contains only musical tracks without any vocals, this reflects the reality quite well. Once again, it shows that rock and hip-hop are very similar in their high points. In addition, jazz again does not have such a high peak, but covers the scale much more broadly.

Finally, a comparison of the categories in the "Energy" feature. Typically, energetic tracks feel fast, loud, and noisy. The higher the value approaches 1, the more energetic it is. If we now compare this with music theory, we can say that, as already noticed in the previous comparisons, rock and hip-hop are similar in their characteristics here and clearly distinguish jazz. Basically, it is to be noted that there is a high average value of 0.65 overall. This is clearer for rock with an average value of 0.7531.

If we now look at the data in a distribution plot (Figure 3), we see clear differences between the categories. Jazz spreads over the whole graph but has no high point. This can be attributed to the different types of jazz (slow jazz, free jazz), which means that different subtypes have different levels of energy. Surprisingly, HipHop is also broadly positioned and sets the high point at about 0.61. HipHop also has many facets as a genre, which is why this expression also makes sense. The curve of Rock builds up early but rises very steeply in the last third of the graph and marks its high point at about 0.9. So, Rock also contains quieter songs, but offers by far the most strongly energetic songs.

After these three sample features, some more distributions and boxplots are added to better explain the characteristics of the individual genres.

3.3 Data Preparation

In the following section the data collected during data collection is prepared and cleaned up for training the model. At the end of data collection, the result was saved in a CSV file. This file is imported and converted into a Pandas Dataframe using the Pandas Python library and its "read_csv" method:

```

1 import os
2 import pandas as pd
3
4 # reading data from csv
5 df = pd.read_csv(os.path.join('..', 'data_collection', 'final_result.csv'))

```

Next, duplicates are removed. If there are two or more datapoints, which have the same value in their artist and name column, only the first one is kept and all subsequent entries are removed. This deduplication could also be done by using track ids. The drawback of this method is, that many artists release their tracks multiple times, e.g. as a single and later in an album. These duplicates would not be caught using the id, as different releases have different id values. As the artist and title doesn't change, almost all duplicates are caught using artist and name.

```

1 df = df.drop_duplicates(subset=['categories.playlists.tracks.artists', 'categories
    .playlists.tracks.name'])

```

Next, all genres that are not to be used for training the model are filtered out.

```

1 genre_filter = ['hiphop', 'jazz', 'rock']
2 df = df[df['categories.id'].isin(genre_filter)]

```

Then columns that are not needed for training are removed, including category name, all playlist information and all track and album information. The category id is kept, as it will serve as the label. The remaining columns are renamed, as the long JSON tree names are no longer needed. The field "category.id" is renamed to "category" and each audio feature is renamed for example the danceability column is now called "feature_danceability".

```

1 columns_to_drop = [
2     "categories.name",
3     "categories.playlists.id",
4     "categories.playlists.name",
5     "categories.playlists.tracks.id",
6     "categories.playlists.tracks.name",
7     "categories.playlists.tracks.album.id",
8     "categories.playlists.tracks.album.name",
9     "categories.playlists.tracks.artists"
10    ]
11 df = df.drop(columns=columns_to_drop)
12
13 df = df.rename(columns={
14     "categories.id": "category",
15     "categories.playlists.tracks.features.danceability": "feature_danceability",
16     "categories.playlists.tracks.features.energy": "feature_energy",
17     "categories.playlists.tracks.features.key": "feature_key",
18     "categories.playlists.tracks.features.loudness": "feature_loudness",
19     "categories.playlists.tracks.features.mode": "feature_mode",
20     "categories.playlists.tracks.features.speechiness": "feature_speechiness",
21     "categories.playlists.tracks.features.acousticness": "feature_acousticness",

```

```

22     "categories.playlists.tracks.features.instrumentalness": "feature_instrumentalness",
23     "categories.playlists.tracks.features.liveness": "feature_liveness",
24     "categories.playlists.tracks.features.valence": "feature_valence",
25     "categories.playlists.tracks.features.tempo": "feature_tempo",
26     "categories.playlists.tracks.features.duration_ms": "feature_duration_ms",
27     "categories.playlists.tracks.features.time_signature": "feature_time_signature"
28   })
29 df

```

With unnecessary columns removed, a check is done to show any remaining null values.

```
1 df.isnull().sum()
```

In this dataset, there are no null values present. As the dataset is now clean, it can now be prepared for input into sklearn's Gradient Boosting Classifier, which is used for training. The classifier only supports integer values as labels, which means that the category data needs to be encoded. This is done by defining a function "encode_target", which takes the dataframe and the name of the column containing the target. It then maps an integer to each unique value in the target column, adds a new "target" column in the dataframe and stores the mapped integer in it.

```

1 # sklearn takes the features and labels as separate lists
2 # df needs to be split
3 def encode_target(df, target_column):
4
5     df_mod = df.copy()
6     map_to_int = {name: n for n, name in enumerate(df_mod["category"].unique())}
7     df_mod["target"] = df_mod[target_column].replace(map_to_int)
8
9     return (df_mod)
10
11 df_target = encode_target(df, "category")

```

The head of the resulting dataframe is shown in table ???. The category to target integer mapping is shown in table ??.

Table 2: Dataframe after cleanup

category	feature_danceability	feature_energy	feature_key	feature_loudness	feature_mode
hiphop	0.649	0.508	8.0	-10.232	1.0

Table 3: Category to target integer mapping

target	category
0	hiphop
1	rock
2	jazz

Now that all columns needed for training the model are prepared, the dataset needs to be split in test and training data. This is necessary to be able to calculate an accuracy score for the model after training. In order to calculate this score, data is needed that the model has never seen before, which ensures that the score is not just a result of overfitting, but the accuracy would be the same on real world data. For splitting the data, the function `train_test_split` from `sklearn` is used. First the dataset is shuffled so each set contains a near equal amount of entries for each genre. Then the shuffled dataset is split, with an 80% of samples going into the training set and 20% going into the test set. This results in 10.701 rows of training data and 2.676 rows of test data.

```

1  from sklearn.model_selection import train_test_split
2  df_target_shuffled = df_target.sample(frac=1, random_state=45)
3  train, test = train_test_split(df_target_shuffled, test_size=0.2, random_state=45,
        shuffle=False)

```

Next, features and labels are split into separate datasets. This is necessary, because the `GradientBoostingClassifiers` fit method takes features and labels as separate arguments called X for input and y for label. Datapoints are matched by their row numbers. Labels are stored in a list containing all values from the target column like so.

```

1  # y contains list of target values
2  y_train = train["target"]
3  y_test = test["target"]
4  y_all = df_target_shuffled["target"]

```

The features are stored in dataframes. The columns are selected using a list of all column names which contain features and then creating a new dataset only containing those columns:

```

1  # columns 1 to 14 contain the features, column 0 is the category and 15 the target
2  features = list(train.columns[1:14])
3
4  # create datasets only containing feature columns
5  X_train = train[features]
6  X_test = test[features]
7  X_all = df_target_shuffled[features]

```

At this point, all data has been cleaned up, targets have been generated, it has been split into training and test data and features have been separated from labels. The data is now ready for training the model.

3.4 Modeling

3.4.1 The Classifier

For training the Gradient Boosting Algorithm, the GradientBoostingClassifier class from scikit learn is used. It is part of a group of classes offering all sorts of ensemble methods. As explained in {Grundlagenteil Schwabe} ensemble methods combine the predictions of several weak learners to generate a more robust model and reduce overfitting issues. sklearn supports averaging methods like Bagging and Random Forests, which take the average of each learners prediction as their output. It also supports Boosting Methods like AdaBoost or Gradient Boosting, which build base estimators sequentially improving the output each time. Sklearn also provides a classifier and regressor model for each method.

The Gradient Boosting Classifier supports binary and multi-class classification and uses 20 hyperparameters to control the size of each regression tree, the number of trees, the learning rate and many more. The most impactful of these parameters are explained here.

- **learning_rate**

As explained in section ??, the learning rate is used to control how much each tree contributes to the result, by multiplying it with the output values of the previous tree. The default value here is 0.1

- **n_estimators**

The number of weak learners to be used while boosting. The default is 100.

- **max_depth**

The maximum depth of each regression tree. This also impacts its number of nodes. The default value is 3.

- **loss**

The type of loss function to be used. Possible values are deviance and exponential. Deviance describes how well a logistic regression model, in our case the regression tree, fits the data. A deviance of 0 would be a perfect fit, the higher the number, the worse the fit.

- **criterion**
- **min_samples_split**
- **min_samples_leaf**
- **max_features**

3.4.2 Cross Validation

Cross Validation is a method structure the dataset for modelling. It allows the use of one single dataset for both training and validation as the dataset is randomly split into N sections. Each section contains an equal distribution of label data as the original dataset. One subset is reserved as a validation dataset while the other subsets are used for the modelling of the dataset. The modelling takes place N times and for every iteration another subset is used for validation. Therefore N models are created in total. Each model can again be tested against the subset that is reserved for validation with the best model chosen **lewis2000introduction**.

3.4.3 Hyperparameter Tuning

Sklearn also provides methods for Hyperparameter tuning. In this project, Exhaustive Grid Search was used to tune hyperparameters. To perform this search, the class GridSearchCV can be used. It takes an estimator, in our case the Gradient Boosting Classifier, and a grid of parameters. In this case, the hyperparameters are the parameters of GradientBoostingClassifier as explained above. This codeblock shows a parameter grid, as it is defined in this project.

```
1 param_grid = {  
2     "loss": ["deviance", "exponential"],  
3     "n_estimators": [97, 98, 99, 100, 101, 102, 103],  
4     "learning_rate": [0.8, 0.9, 0.1, 0.11, 0.12],  
5     "criterion": ['friedman_mse', 'squared_error'],  
6     "min_samples_split": [1, 1.5, 2, 2.5, 3],  
7     "min_samples_leaf": [1],  
8     "max_depth": [1, 2, 3, 4, 5],  
9     "random_state": [42],  
10    "max_features": [None, 0.95, 0.90]  
11 }
```

The parameter grid is given as a Python dictionary containing the parameter name as keys and an array of values for each key. Additionally a GradientBoostingClassifier object is defined and both are passed into the GridSearchObject.

```

1     gbc = GradientBoostingClassifier(random_state=45)
2     search = GridSearchCV(gbc, param_grid)

```

The search object's fit method combines cross-validation, hyperparameter tuning and classifier fitting to find the best possible combination of hyperparameters to the estimator for the given dataset. This is done in the following way:

1. An estimator is created using the first combination of hyperparameters. In the current example, this is equivalent to creating the following estimator.

```

1     estimator = GradientBosstingClassifier(loss="deviance", n_estimators=97,
                                              learning_rate=0.8, criterion='friedman_mse', min_samples_split=1,
                                              min_samples_leaf=1, max_depth=1, random_state=42, max_features=None)

```

2. The newly created classifier is fitted to the training data using this set of parameters using 5-fold cross validation. When this is done, an average of the five scores resulting from cross-validation is built. This average is now used as the score for this specific combination of parameters.
3. The score is saved together with the parameters
4. All of the above steps are repeated with every possible combination of parameters from the parameter grid. This results in a score for every set of parameters.
5. As every model was fitted using the same method, it is clear that the model with the highest score is using the best hyperparameters. A new model is trained without cross-validation so using all data, to create the final model using the best parameters.

The model created by grid search can then be evaluated using test data that was not used in cross validation. Executing this GridSearch in Python is done be calling the fit method of the search object defined in the code above using the train features and labels.

```

1     search.fit(X_train, y_train)

```

After fitting, the search object can be used as if it was a normal estimator built with the optimal hyperparameters. To see, which hyperparameters were optimal, the get_params can be used. It's output looks like this:

```

1     {'cv': 5,
2      'error_score': nan,
3      'estimator__ccp_alpha': 0.0,
4      'estimator__criterion': 'friedman_mse',
5      'estimator__init': None,
6      'estimator__learning_rate': 0.1,

```

```

7     'estimator__loss': 'deviance',
8     'estimator__max_depth': 3,
9     'estimator__max_features': None,
10    'estimator__max_leaf_nodes': None,
11    'estimator__min_impurity_decrease': 0.0,
12    'estimator__min_samples_leaf': 1,
13    'estimator__min_samples_split': 2,
14    'estimator__min_weight_fraction_leaf': 0.0,
15    'estimator__n_estimators': 100,
16    'estimator__n_iter_no_change': None,
17    'estimator__random_state': 45,
18    'estimator__subsample': 1.0,
19    'estimator__tol': 0.0001,
20    'estimator__validation_fraction': 0.1,
21    'estimator__verbose': 0,
22    'estimator__warm_start': False,
23    'estimator': GradientBoostingClassifier(random_state=45),
24    'n_jobs': -1,
25    'param_grid': {'loss': ['deviance', 'exponential'],
26                  'n_estimators': [97, 98, 99, 100, 101, 102, 103],
27                  'learning_rate': [0.8, 0.9, 0.1, 0.11, 0.12],
28                  'criterion': ['friedman_mse', 'squared_error'],
29                  'min_samples_split': [1, 1.5, 2, 2.5, 3],
30                  'min_samples_leaf': [1],
31                  'max_depth': [1, 2, 3, 4, 5],
32                  'random_state': [42],
33                  'max_features': [None, 0.95, 0.9]},
34    'pre_dispatch': '2*n_jobs',
35    'refit': True,
36    'return_train_score': False,
37    'scoring': None,
38    'verbose': 0}

```

All fields starting with "estimator__" contain the optimal hyperparameters for the data. Even though a broad spectrum of values was given in the search grid, most of the optimal values are equal to the default values given by sklearn. This might suggest, that the developers set the default values using a similar dataset to the one used in this project.

Although many combinations were evaluated, there is still a possibility, that this is a local maximum and a set of hyperparameters that is far different from the ones tested here are actually optimal. Trying out every possible combination of values would require a great amount of computing resources, as the Grid Search shown above already took more than 1.5 hours on a modern 16 core CPU.

Finally the model can be evaluated using the test set defined previously. This set has never been seen by the model even while doing cross validation.

```

1     search.score(X_test, y_test)

```

The estimators score function shows a 86.286% accuracy for the final model using the test set.

3.5 Evaluation

4 Fazit

Appendix

Appendix 1: Beispielanhang

Dieser Abschnitt dient nur dazu zu demonstrieren, wie ein Anhang aufgebaut seien kann.

Appendix 1.1: Weitere Gliederungsebene

Auch eine zweite Gliederungsebene ist möglich.

Appendix 2: Bilder

Auch mit Bildern. Diese tauchen nicht im Abbildungsverzeichnis auf.

Figure 24: Beispielbild

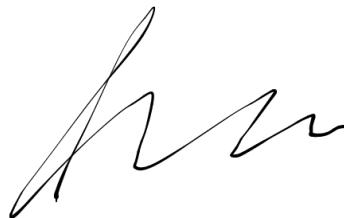
Name	Änderungsdatum	Typ	Größe
abbildungen	29.08.2013 01:25	Dateiordner	
kapitel	29.08.2013 00:55	Dateiordner	
literatur	31.08.2013 18:17	Dateiordner	
skripte	01.09.2013 00:10	Dateiordner	
compile.bat	31.08.2013 20:11	Windows-Batchda...	1 KB
thesis_main.tex	01.09.2013 00:25	LaTeX Document	5 KB

Declaration in lieu of oath

I hereby declare that I produced the submitted paper with no assistance from any other party and without the use of any unauthorized aids and, in particular, that I have marked as quotations all passages which are reproduced verbatim or near-verbatim from publications. Also, I declare that the submitted print version of this thesis is identical with its digital version. Further, I declare that this thesis has never been submitted before to any examination board in either its present form or in any other similar version. I herewith **agree/disagree** that this thesis may be published. I herewith consent that this thesis may be uploaded to the server of external contractors for the purpose of submitting it to the contractors' plagiarism detection systems. Uploading this thesis for the purpose of submitting it to plagiarism detection systems is not a form of publication.

Düsseldorf, 25.1.2022

(Location, Date)

A handwritten signature in black ink, appearing to read "Hans".

(handwritten signature)