Predicting the Outcome of Appeal Decisions in Germany's Tax Law

Bernhard WALTL [a,1], Georg BONCZEK [a], Elena SCEPANKOVA [a],
Jörg LANDTHALER [a], and Florian MATTHES [a]

[a] *Software Engineering for Business Information Systems,
Technical University of Munich, Germany*

Abstract. Predicting the outcome or the probability of winning a legal case has always been highly attractive in legal sciences and practice. Hardly any attempt has been made to predict the outcome of German cases, although prior court decisions become more and more important in various legal domains of Germany's jurisdiction, e.g., tax law.

This paper summarizes our research on training a machine learning classifier to determine likelihood ratios and thus predict the outcome of a restricted set of cases from Germany's jurisdiction. Based on a data set of German tax law cases (44 285 documents from 1945 to 2016) we selected those cases which belong to an appeal decision (5 990 documents). We used the provided meta-data and natural language processing to extract 11 relevant features and trained a Naive Bayes classifier to predict whether an appeal is going to be successful or not.

The evaluation (10-fold cross validation) on the data set has shown a performance regarding F_1-score between 0.53 and 0.58. This score indicates that there is room for improvement. We expect that the high relevancy for legal practice, the availability of data, and advance machine learning techniques will foster more research in this area.

Introduction

The formal procedure of modern societies allows to take legal actions in order to claim someone's right. Thereby, courts and judges decide the case based on a given set of facts (evidence) and the applicable law. From an economical point of view, those cases can be resource intensive, as to time, money, and data. This does not only count for legislation, and consequently the society, but also for the claiming individual, i.e. the plaintiff. Therefore, predicting the result of a case or a probability approximation of whether a case is successful or not, is highly desirable. Within this paper we describe our approach and results of predicting the outcome of cases for a narrow but relevant set of cases within the German tax law, namely the success rate of appeal decisions of German Fiscal Courts.

The Federal Fiscal Court being one of the five highest courts in Germany, is a court of last resort responsible for the interpretation and application of German tax law (exempt

[1] Corresponding Author: Bernhard Waltl, Software Engineering for Business Information Systems, Boltzmannstr. 3, 85748 Garching bei München, Germany; E-mail: b.waltl@tum.de.

criminal tax law). In most cases, people refrain from going into appeal, as for non-legals it is extremely difficult to assess their success odds correctly and thus the financial risk if losing the case. As a result, many people do not even try to challenge the first instance court decisions, remaining ignorant and losing on their chances of getting their legitimate right. Only about 4-5 % of about 70 000 currently pending cases at financial courts go into appeal [1]. This seems particularly problematic from the view of the rule of law principle in Germany. The decision if to appeal or not, depends on a couple of factors from an individual's perspective. By helping to predict the outcome of an appeal, we aim to find a fair deal between seeking justice and the economic risks of legal proceedings.

1. German Judicial Procedures: Fiscal Courts and Appeal Decisions

The judicial procedures in the German fiscal domain follow a clear structure. The process is initiated by a plaintiff, who brings his case to one out of 18 different fiscal courts (Finanzgericht FG) in Germany. The FG collects and structures the evidence and decides on the case. In case the plaintiff does not agree with the outcome, he can initiate an appeal procedure, which directly goes to the Federal Fiscal Court (Bundesfinanzhof BFH), which is located in Munich, Bavaria. In contrast to different jurisdictions, the tax law system only consists of two instances, with the BFH being the second and last instance for tax law related cases. Now the BFH investigates the case and decides whether the decision of the FG was compliant with applicable laws. If European legislation is decisive for the case outcome, the BFH is obliged to consult the European Court of Justice (EuGH), and await its binding ruling. Finally, the BFH renders a judgment which either confirms or overrules the decision of the fiscal court as court of first instance. Under certain circumstances, the BFH has to refer the case back to the fiscal court which decides the case anew. Finally, the case is decided and the plaintiff is informed.

We analyzed and modeled fiscal court decisions (Step 1a) and trained machine learning algorithms to predict the outcome of future appeal decision. Thereby, we collected cases from FG and BFH (responses of Step 1a and 2a) (see Section 4), processed them, proposed a model and extracted eleven different features (see Section 5). Those features served as the base line for a multinomial Naive Bayes classifier (see Section 6). Finally, we evaluated the performance of the classifier and discussed steps for improvements (see Section 7).

2. Related Work

One of the earliest approaches regarding predictions applied a nearest neighbor approach, where the cases closest to a problem are determined in terms of similarity measures and an outcome is assigned with regard to the majority of those cases [2]. Popple, in 1996, using a nearest neighbor algorithm, added more complexity to the similarity measures by assigning weights to different fact descriptors [3]. In our view, a nearest neighbor approach is limited by its definition to the circle of identifiable neighbors and does not allow for precise predictions outside this scope.

The IBP (Issue-Based Prediction Model) integrates case-based reasoning with a model of abstract legal issues associated with a legal claim of trade secret misappropriation [4]. The model's restriction to cases concerning trade secret misappropriation reflects the difficulty of a transfer to other fields of law. When the legal issues and relationships in the IBP Domain Model are "a distillation and interpretation of two authoritative sources on the law of trade secret misappropriation (a statute and a Restatement provision)" [4], this shows this model's strong connection to the legal content of cases. The identification of relevant issues in this model is thus time- and knowledge-intensive and has to be done anew for any other field of law, hindering the development of a universal prediction model.

Katz's prediction model leverages the random forest method together with feature engineering for the prediction of Supreme Court decisions [5]. Based on the extensive Supreme Court's database, where each case is assigned with around 240 variables, many of which are categorical, a number of formal features is derived. Except for the lack of a comparably extensive database and the information about judges "behavior" who don't play a dominant role in civil law jurisdictions as Germany, the use of formal features sets the possibility of creating a universal prediction model in a way we are aiming at.

3. Approach

This section briefly describes the steps performed within our approach, which follow a classical machine learning approach by beginning with a data preparation and preprocessing step. Subsequently, we came up with a model (features and priorities) which serves as the base line for the prediction algorithm. Based on that, we extracted the required features and trained a classifier, which we tested afterwards.

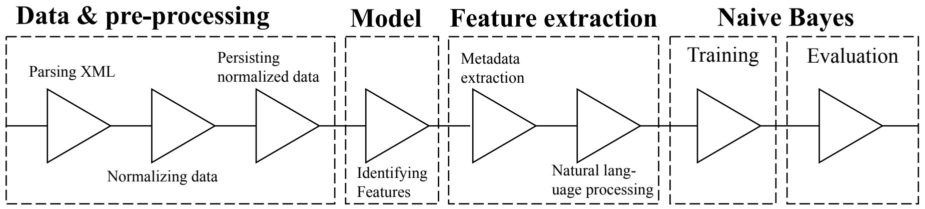

Figure 1. Stepwise and subsequent pre-processing, feature extraction, and training with evaluation of a Naive Bayes classifier to predict the outcome of fiscal court appeal decisions.

Data & Pre-processing The available data (see Section 4) needs to be processed. Therefore, it was necessary to develop specified importing routines and normalized the data such that it fits to one common data scheme, which is persisted in a database to easily enable data-intensive machine learning procedures.

Modeling During the model we have defined parameters that potentially indicate the outcome of an appeal case and that are available in the data at hand. Thereby, we have identified different variables, so-called features, and summarized them within a table (see Section 5). In addition, we assigned a priority to each feature indicating its suspected importance.

Feature selection Based on the collection of features, we have developed several routines extracting those from the data set. Thereby, we analyzed the metadata, such as author, publishing date, etc. and created the desired set of features for each of the document (see Section 5). We mainly used regular expressions for this step.

Naive Bayes classifier Using an existing machine learning framework, we trained and tested a common and simple probabilistic classifier, namely Naive Bayes. We have compared different classifiers and found that Naive Bayes is performing best. We split up the available data into a training and a test data set. Thereby, we used a common strategy, namely 10-fold cross validation (see Section 6).

Figure 1 shows the subsequent steps but it does not reflect the workload that was spent on each individual task. Especially data & pre-processing, modeling and the feature selection parts require lots of time and different implementations. Compared to that, training and testing the classifier can be done straight-forward. Existing machine learning libraries and frameworks can easily be integrated and used once the data is pre-processed, the modeling part done, and the required features extracted.

4. Data

The data we base our research on is a corpus, maintained by professional editors, consisting of 44 285 judgments of German fiscal courts, which date back to 1945, whereas the most recent documents were issued in 2016. Out of these 44 285 documents, 27 055 depict first instance cases (FG), the remaining 17 230 are judgments ruled by the BFH. Ultimately, after cleaning documents which lack important data for feature extraction, our dataset contained 5 990 complete proceedings.

Each data point consists of a tuple: A first instance case, and a corresponding appeal decision, i.e. revision. The effectively used dataset consists of judgments from 1990 until 2015. Our data is relatively up-to-date, but there is a significant drop of cases from 2012, since those cases have not been decided yet. This might cause a so-called cold start issue during the training phase of machine learning algorithms. An analysis of the temporal distribution of the data set implies that the dataset does not cover many major changes in German fiscal legislation. One can expect however the dataset to be representative for the German fiscal legislation of the last years. As stated above, although the German tax law is part of a civil law jurisdiction and the main acts, e.g., EStG, have statutory character, the case law is particularly important for legal practice, e.g., tax consultants, auditors, etc.

The data is structured in XML documents collection, whereas each XML file represents one judgment. Each XML file contains a variety of different metadata such as referenced legal norms, decision date, filing numbers, years of dispute, the ruling court and a general markup for structuring the judgment text itself into different sections, e.g. statement of facts, reasoning, etc. Advanced information of the decision results such as the information whether the court ruled in favor of the plaintiff, what kind of juristic person the plaintiff constitutes etc. are not explicitly given. After its extraction, this data, in combination with the meta-data, is used as features (see Section 5.2).

In addition, we have access to a manually created and editorially maintained thesaurus containing numerous terms of the German tax law. The thesaurus is available in JSON format, can easily be accessed, and provides information about synonyms, hyponyms, abbreviations and similar terms to a given term. This thesaurus in its entirety includes 16 019 of such groups (synsets) and overall 42 598 tokens, i.e. terms.

5. Processing and Feature Extraction

5.1. Pre-processing

The pre-processing consists of two main parts. The first one constitutes the simple extraction of meta data of the concerning documents, whereas the second one contains several text mining tasks in order to extract features that are not already given.

We extracted four features: The references within the factual findings, since not all existing references are also stated in meta-data, the factual findings themselves and reasons given in the judgment as well as the type of juristic person that represents the plaintiff (if applicable). In the process of determining those features we acted on the assumption that legal texts often follow certain patterns of formulation. This approach allows us to extract the desired features with standard natural language processing techniques.

5.1.1. Dataset generation

In order to ascertain the result of the appeal, we needed to label our testing and training data. Thereby, we used the circumstance that the information, if the appeal got rejected or sustained, could be at the very beginning of the reasons part within the ruling. Also, the wording is carefully chosen, so the dismissal of a case is formulated with just a few adjectives. By means of several selected terms, it is possible to classify this first sentence and therefore determine the outcome of the judgment. Despite the small feature space of 8 different terms indicating the outcome, this method works reasonably good for all documents.

5.2. Modeling and feature selection

All information for our model was derived and is knowable prior to the date of the estimated decision (out-of-sample applicability). Consequently, the model allows to generate ex ante predictions, i.e. predicting in the real sense. Another characteristic of our model is generality and consistency. This means that our model generates predictions irrespective of changes in the composition of the courts (e.g., retirement, recusal, etc.) and not limited to specific time periods.

We considered a number of features, e.g. the year of dispute, the specific courts, the nature of the petitioner, the duration of a case, the decisive legal norms, the overall cited norms, the guiding principles and the heading. The different grade of impact each one of those features might have on the decision result, we are expressing in different weights manually attributed to them.

Considering the year of dispute the assumption is that different time periods correspond to different legal amendments with specific grades of legal complexity which influences the probability of reversals. Compared to other fields of law, tax law is immensely important for the state budget and thus highly influenced by political considerations, which result in more legal changes and amendments than in any other legal area. The more those amendments intervene with the overall tax law system, the more careful they have to be drafted in order to guarantee the application consistency within the tax law system itself.

We distinguished geographically between different courts and the specific Chambers deciding the case (German: Gerichtskammer). Courts having jurisdiction ratione loci and ratione materiae decide autonomously within their circuit, which leads to inconsistency between the different court circuits. In a comparable way, Chambers as parts of the same court are autonomous in deciding cases, often dominated by the concrete personal composition. The observed autonomous deciding is grounded in the principle of the judge being bound only by law and his own consciousness. We assume that there is some correlation between the outcomes of the case and case durations on the one hand, and court locations, including Chamber specifications, on the other hand.

Selecting legal norms is motivated by the fact that legal norms are the decisive factor when adjudicating a case. Moreover, our feature selection considered norms not just as a whole, but - following its specific citation in the case - splits it into paragraphs, articles, sentences, numbers, letters etc. Certain norms, or rather elements of a norm are more controversial in their application than others, i.e., creating more scope for different interpretations. This is why the splitting is necessary for more precise predictions. We

Feature	**Description & rational**	**Priority ↓**
Courts	Courts having jurisdiction ratione loci and ratione materiae, decide autonomously in their geographically assigned circuit, leading to inconsistency within the circuits.	High
Court chambers	Chambers of the same court may and do decide autonomously, leading to inconsistency within the same court.	High
Decisive legal norms	Those have the function of legally justifying the outcome of the case.	High
Guiding principals	Those summarize the legal statement of the decided case in a few sentences.	High
Petitioner	The different groups of petitioners (individuals and corporate entities) incorporate different values with regard to the public law domain of tax law.	High
Cited legal norms	Those are necessary for legal reasoning, albeit not of decisive nature for the outcome of the case.	Middle
Duration of the case	This reflects either the complexity of a case or the workload in a specific court.	Middle
Keywords of statement of facts	The ’statement of facts’ section contains by law only the legally essential, resp. for the legal reasoning relevant facts of a case.	Middle
Keywords of the ’legal reasoning’ part	The legal reasoning part is dominated by legal language - extracted keywords thus support semantically the outcome of a case.	Middle
Year of dispute	This time period reflects the applicable law at the time of the dispute.	Middle
Heading	This serves as a quick classification of judgments without the aim to reflect the legal reasoning.	Low

Table 1. An overview of the selected features, description and corresponding priority we attributed to them.

distinguished between decisive legal norms, which are explicitly cited at the beginning of a case, and the overall cited norms in the judgment text. As the former are considered as primarily decisive for the overall outcome of the case, we weigh them stronger than the rest of the cited norms, having impact often on solely procedural matters or cited for reference reasons only.

Considering the petitioner as a selective feature we looked into the function he is acting in – as an individual or as a legal person. The assumption is that courts might be more willing to attribute rights to individuals than to legal entities, as the former are usually in an (economically) weaker position than the latter ones. Exerting influence on this imbalance of powers might be a factor on the subconscious level of judges as decision makers. We further grouped legal entities into two categories, the private entities (German: "Personengesellschaften") and the corporate entities (German: "Kapitalgesellschaften"). Whereas corporate entities are characterized by their strong economic purpose, limited liability and legal capacity, the focus of private entities lies usually not on an economic aim, but the people involved, expressed in a personal liability, which is reflected in the involved values having a possible impact on judges' decisions.

Another feature is the duration of cases as the time period from the year of the case filling to the actual decision date. The case duration may reflect both the complexity of a case or the workload at a particular court. By extracting the workload cases by way of comparisons, we filtered the factual or legally complex cases. Complexity itself increases the probability of different interpretations and thus the risk of reversal.

Considering the heading of the case and the guiding principles, we need to distinguish. Whereas the guiding principles summarize the main statement of the case and thus serve as the nucleus of the case, the heading's first purpose is classification of judgments without the requirement to capture the legal reasoning itself.

By selecting the keywords of the statements of facts as a selective feature (nouns and thesaurus) we are taking into account the fact, that the statement of facts-section by law should contain only the "decisive facts of a case", i.e. all the facts that are necessary to construct a legal reasoning and therefore the factual pillars of a case. A judgment should contain the essential facts of a case, keeping up with a tight and lean presentation of those (§ 313 s. 2 ZPO). Considering further the keywords of the legal reasoning part of the judgment is motivated by the fact, that this part is more than any other one dominated by legal language; the extracted legal knowledge supports semantically the outcome of a case.

5.2.1. Feature extraction

A part of the references used by the court are already contained in the meta-data of the document. The remaining norms were extracted by parsing the textual content of the case. Since we only considered a relatively small subset of German legal texts, we used regular expressions to detect those references. After finding such a reference we normalized it, such that it corresponds the format that is used throughout the corpus.

For the extraction of the information whether the plaintiff represents a certain type of juristic person, we again relied on certain structures in legal formulations. We analyzed the first few sentences of the facts which cover the basic traits of the plaintiff. Those also cover whether it is a juristic or natural person raising the claim. Afterwards, we searched for the terms referring to the plaintiff. We extracted common terms and phrases that occur in combination with the most relevant forms of juristic persons. In order to avoid false positives arising through formulations such as “the plaintiff works at X-GmbH”, we did not consider sentences that contain verbs indicating some form of employment. Despite this method obviously not being the most effective one, we consider it to be more efficient in comparison to more advanced techniques with respect to implementation efforts.

5.2.2. Processing of textual data

After extracting the textual features, we normalized them with respect to the thesaurus mentioned earlier. For each concept in this thesaurus, we chose one representative with which we replaced all occurrences that pose an abbreviation, synonym or similar term to this representative. Furthermore, for the facts and reasons we only kept a bag of words that contain the keywords (also their multiplicity) appearing in the thesaurus. This allowed us to preserve the legal terms, while removing terms and nouns that induce noise due to their irrelevance for the legal case. By replacing the synonyms etc., we expect an edge in efficiency when classifying, since the semantic relation between words is not taken into account. When unifying terms that are similar to each other, we might lose some nuance that differentiates them, but we consider the advantage in the classification step worth this hypothetical loss, since there is no other trivial way of creating a relation between them. After these steps, we also apply stop-word removal and stemming.

6. Predictive Analytics and Performance

For the training and classification we used the scikit-learn [6] machine learning framework. We passed different features through a pipeline, calculating TF-IDF vectors for textual features and count-based vectors for the remaining features. After trying different common estimators, the multinomial Naive Bayes classifier has performed best producing the most promising results. Using a 10-fold cross-validation, we achieved a F_1-score of 0.57 (see Table 2).

We see that both types of judgments, the ones in favor of the plaintiff and the ones in favor of the defendant, have been classified by our approach (precision). Also, 60% of the judgments with positive outcome have correctly been identified having no negative outcome. Since the overall precision and recall are both 57%, so is the F_1-score. In Section 7 we will interpret these values in this specific application, thereby we differentiate

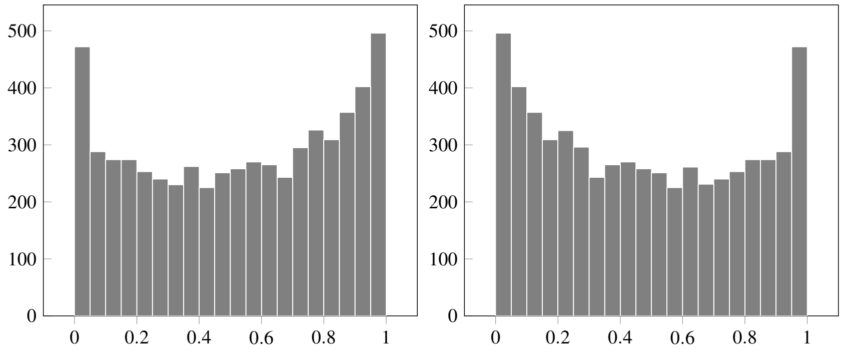

Figure 2. Histogram of predicted probabilities for positive (left chart) and negative outcomes (right chart).

	precision	recall	F_1-score	support	support (rel)
pos. outcome	0.57	0.60	0.58	3 012	50.28 %
neg. outcome	0.57	0.53	0.55	2 978	49.72 %
avg / total	0.57	0.57	0.57	5 990	100.00 %

Table 2. Confusion matrix summarizing the performance of the prediction using a multinomial naive Bayes estimator (evaluation using a 10-fold cross-validation).

between three different aspects: Quality of features and feature extraction, accuracy of predictions, and potential for improvements.

We also used feature weighting, but initial parameter studies have been of little success. We also observed loss in both precision and recall when lowering auxiliary feature weights such as court, plaintiff type or references. Based on this fact, we conclude that there is in fact potential for hyperparameter tuning since the likelihood of ideal parameters being the default ones is quite low.

A detailed inspection of the classifiers outcome is shown in Figure 2. The figure shows two histograms for the classifiers performance on predicting positive (left chart), and negative outcomes (right chart). The histograms show the confidence with which the classifier predicts a certain outcome. Maybe one would expect the classifier to decide very confident on a subset of cases but this only holds for a small set of cases in which his prediction is above 80 or 90 %. Instead the distribution shows that the classifier's confidence is, with a few minor exceptions, equally distributed and covers the whole range from high confidence ($\geq 90\%$) to very low confidence ($\leq 10\%$).

7. Discussion

7.1. Quality of features and feature selection

The features we are currently using largely represent data about the legal process. When it comes to the content of the document, its title, the headnote, the types of plaintiffs as well as keywords of judicial relevance contained in the facts or the reasoning of the court are considered. These chosen features mainly constitute the factual basis of a judgment and are thus in our opinion essential for its efficient classification. However, the actual benefit of supporting features is to be put into question. The impacts of features such as the duration of the process are nominal and could turn out to be the source of overfitting. In addition, the extraction of metadata and especially of features using natural language processing (NLP) is — up to a certain degree — always vulnerable to errors. Hardly any technique from NLP can be performed without any error.

However, the formal nature of the features we selected for our model allows to build a prediction model across different legal areas. In contrast to successful, however predominantly issue-based prediction models (e.g. IBP [4]) our model bears the chance to create a universal prediction model, applicable across different legal areas.

7.2. Accuracy of predictions

Regarding the overall complexity, it is hard to define a "minimal" threshold for a F_1-score to be considered meaningful or valuable for legal practice. Due to the low precision and recall scores, it is currently not feasible to make any final statements about the ability to classify judgments of the fiscal courts. However, our results support the hypothesis that a classification of such judgments is principally possible. It also should be kept in mind that we use a rather small feature set, so adding more high quality features we expect a further increase both in precision and recall.

In addition, we only considered a very limited amount of data set with a restricted set of document types, as we haven't had the chance to access important documents that are relevant within a court's decision process. If more and more data is becoming publicly available within open data initiatives, the potential for making predictive analytics may be more attractive and more accurate.

8. Conclusion

This paper summarizes the results of an interdisciplinary research topic on using machine learning to predict the outcome of court decisions based on a huge set of prior cases. We restricted ourselves to predict the outcome of appeal decisions within the German tax law. Thereby, a plaintiff can appeal if he does not agree with the result of the fiscal court (first instance). The appeal goes directly to the German Federal Fiscal Court (BFH). This consumes a lot of time and monetary resources both of the plaintiff and the German State financing jurisdiction. Using the meta-data and natural language processing, we analyzed 5 990 documents and extracted 11 different features for each case. This served as the input for a multinomial Naive Bayes classifier. The evaluation has shown that the classifier's performance is limited (F_1-score between 0.53 and 0.58).

Although the overall performance of the classifier is not satisfying at the current stage, there is strong evidence that the performance could be improved by taking more features and additional data into account. Since more and more data is going to be publicly available, a synthesis of those combined with powerful machine learning algorithms could lead to better performing algorithms that could potentially be used by legal practitioners, e.g. judges and lawyers, or legislators to evaluate and improve the current legal situation.

References

[1] "The federal supreme finance court," https://www.bundesfinanzhof.de/sites/default/files/Booklet.pdf.

[2] E. Mackaay and P. Robillard, "Predicting judicial decisions: The nearest neighbour rule and visual representation of case patterns," *Datenverarbeitung im Recht*, 1974.

[3] J. Popple, "A pragmatic legal expert system," *Applied Legal Philosophy Series*, 1996.

[4] K. D. Ashley and S. Brüninghaus, "Automatically classifying case texts and predicting outcomes," *Artif. Intell. Law*, vol. 17, no. 2, pp. 125–165, Jun. 2009. [Online]. Available: http://dx.doi.org/10.1007/s10506-009-9077-9

[5] D. M. Katz, I. Bommarito, J. Michael, and J. Blackman, "A general approach for predicting the behavior of the supreme court of the united states," *arXiv preprint arXiv:1612.03473*, 2016.

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