The legal prediction industry: meaningless hype or useful development?

Floris Bex[*] *& Henry Prakken*[*]

This is the second author's unofficial translation of Floris Bex & Henry Prakken, De juridische voorspelindustrie: onzinnige hype of nuttige ontwikkeling? Ars Aequi 69 (2020), pp. 255-259.

Recently there has been much discussion about the use of predictive algorithms in the law – some have even claimed that the robot judge is a matter of time. In this article we discuss whether, and if so how predictive algorithms can be of use for the legal world, in particular for the judiciary.

1. Inleiding

The legal prediction industry is on the rise: under the heading 'Artificial Intelligence (AI) for the judiciary'[1] there are many discussions about the use of AI algorithms for predicting outcomes of legal cases.[2] Despite the enthusiasm in the media about the opportunities offered by these algorithms as regards robot judges[3] and 'forum shopping',[4] there are also sceptical voices from the legal academic world.[5] An often-mentioned problem is that *predicting* a decision with the help of statistical correlations is not the same as *taking* the decision on the basis of legally valid reasons. Moreover, the reported results of the algorithms should not be overestimated. For example, also without an AI algorithm a very accurate prediction is possible for any given criminal case that it will result in a conviction, simply since more than 90% of the criminal cases decided in court result in conviction. Are predictive algorithms a

[*] Prof.dr. F.J. Bex is endowed professor in Data Science and the Judiciary at the Department of Law, Technology, Markets, and Society, Tilburg University, scientific director of the Dutch National Police-lab AI of the Innovation Centre for AI (ICAI), and lecturer in AI at the Department of Information and Computing Sciences, Utrecht University. Prof.dr.mr. H. Prakken is professor in legal informatics and legal argumentation at the Faculty of Law, University of Groningen and senior lecturer in AI at the Department of Information and Computing Sciences, Utrecht University.

[1] C. Prins & J. van der Roest, 'AI en de rechtspraak', *NJB* 2018/206; H. Prakken, 'Komt de robotrechter er aan?', *NJB* 2018/207; S. Verberk, M. Noordergraaf & C.E. du Perron (eds.) 'Algoritmes in de rechtspraak. Wat artificiële intelligentie kan betekenen voor de rechtspraak', *Rechtstreeks* 2019, issue 2.

[2] Publications about algorithms that predict outcomes of legal cases: D.M. Katz, M.J. Bommarito & J. Blackman, 'A general approach for predicting the behavior of the Supreme Court of the United States', *PLoS ONE* 2017, issue 4; N. Aletras et al., 'Predicting judicial decisions of the European Court of Human Rights', *PeerJ Computer Science* 2016-2; M. Medvedeva, M. Vols & M. Wieling, 'Using machine learning to predict decisions of the European Court of Human Rights', *Artificial Intelligence and Law* 2019, doi.org/10.1007/s10506-019-09255-y. Publications that discuss these predictive algorithms (critically): F. Pasquale & G. Cashwell. 'Prediction, persuasion, and the jurisprudence of behaviourism' *University of Toronto Law Journal* 2018-68.supplement 1, pp. 63-81; K.D. Ashley, 'A brief history of the changing roles of case prediction in AI and law', *Law in Context* 2019, issue 1, pp. 93-112; D.L. Chen, 'Machine learning and rule of law', in: M.A. Livermore & D.N. Rockmore (eds.), *Law as Data*, Santa Fe: Santa Fe Institute Press 2018, pp. 429-438.

[3] F. Jensma, 'Big data kunnen de rechter verdringen', *NRC* 28 oktober 2017; H.J. van den Herik, 'In 2030 zullen computers rechtspreken', *Mr. Online* 31 October 2016, www.mr-online.nl/in-2030-zullen-computers-rechtspreken/.

[4] C. Driessen, 'Wie personeel wil lozen na een ruzie moet bij rechtbank Den Haag zijn', *NRC* 11 september 2019; E. Kreulen, 'De rechtspraak is met deze nieuwe robot niet langer digibeet', *Trouw* 26 August 2019.

[5] Prakken 2018, Pasquale & Cashwell 2018.

meaningless and temporary hype or can they still be use for the legal world? In this article we try to answer this question.

We first discuss several types of legal predictive algorithms and the distinction between 'algorithmic experts' and 'algorithmic outcome predictors'. Then we discuss some important issues concerning determining the quality of predictive algorithms – for example, several ways to evaluate their quality, and conditions on the data that have to be fulfilled in order to obtain valid predictions. Then we discuss our main question, how the various types of predictive algorithms can be useful for legal academic research, for justice-seeking parties and for the judiciary. We will argue that there is definitely a hype surrounding predictive algorithms but that they cannot be disqualified as meaningless, since they can be useful in the law in various ways. So-called *algorithmic outcome predictors* can support quantitative analysis of the law and case law, and so-called *algorithmic experts* can support judges in answering some factual issues in a legal case.

2. Different legal predictive algorithms

Predictive algorithms usually are typical examples of *supervised machine-learning* algorithms. Such an algorithm is first presented with a large number of historic cases – the training data - with the features and outcome of these cases, for example, in case of predicting recidivism, whether the person reoffended, or in case of predicting outcomes of legal cases, what was the outcome). From the training data the algorithm can learn the possibly very complex relations between these features and the possible outcomes, and use these relations to 'predict' the outcome of unseen cases.[6] A thus trained algorithm is evaluated with test data: historic cases of which only the features (so not the outcome, even though it is known) are shown to the algorithm.

2.1 Algorithmic experts: predicting facts that are relevant for a decision

Some legal predictive algorithms make estimates on issues that are relevant for a judicial decision and which estimates would otherwise have to be made by the judge or by a human expert. Well-known examples are algorithms that predict the probability of recidivism,[7] and algorithms that estimate the expected environmental impact of activities for which an environmental permit is requested, such as the AERIUS-system, which estimates emission of nitrogen.[8] We will call this kind of system *algorithmic experts*.

2.2 Algorithmic outcome predictors: predicting outcomes of legal cases

There are also *algorithmic outcome predictors*, algorithms that predict outcomes of legal cases. They come in, roughly, three types: predictors on the basis of features unrelated to the merits of a case, predictors on the basis of the textual description of a case, and predictors on the basis of legally relevant factors.

Predicting on the basis of features unrelated to the merits of a case.

Some algorithms base their predictions on features of a case that are not related to the merits of the case. An example is the algorithm that predicts decisions of the American Supreme

[6] 'Predict' is here between quotation marks since the cases for wich the algorithm predicts the outcome can also be cases from the past, of which the outcome is alread known.
[7] See, for example, R. Berk et al., 'Fairness in criminal justice risk assessments: the state of the art', *Sociological Methods & Research* 2018, https://doi.org/10.1177/0049124118782533.
[8] Aerius, Rekeninstrument voor de leefomgeving, www.aerius.nl/nl.

Court[9] on the basis of information that is available in a database[10] about the Supreme Court, such as the kind of case, the date at which it was decided and which lower court decided the original case. This algorithm, which correctly predicted 70% of the decisions, cannot explain the predicted decisions in a legally meaningful way, since the features on the basis of which it makes its predictions are not related to the merits of the case: a – strongly simplified – example of an explanation of the algorithm would be 'I predict that the court will in this case affirm the decision of the lower court, since this is what it usually does in economics-related cases in which mr. Roberts is the *chief justice*'.

Predicting on the basis of the textual description of a case.

Other algorithms predict outcomes with the help of a statistical analysis of the text of case law, where statistical correlations are identified between the frequency of word combinations and the outcome of a case. An example is the algorithm that predicts whether the European Court of Human Rights (ECHR) will for a specific article from the Convention with the same name decide whether that article was violated[11] on the basis of descriptions by the ECHR of the procedural history and the facts that gave rise to the appeal with the Court.[12] The algorithm correctly predicted 75% of the cases. Although it would seem that the algorithm looks at the legal aspects of the case (procedural history, facts), the identified statistical correlations do not say anything about the legally relevant reasons for the outcome of a case. Therefore this algorithm can also not explain its predicted outcomes in a legally meaningful way. For example, the three word combinations with the highest predictive value for 'violation' were, respectively, 'district prosecution office', 'the district prosecutor' and 'the first applicant'. This is legally not very informative.

Predicting on the basis of legally relevant factors.

A third approach predicts outcomes on the basis of the legally relevant factors in a case. Well-known is the research of Ashley and colleagues on the case law concerning misuse of trade secrets in American law[13]. Both for previously decided cases and for the case to-be predicted the legally relevant factors are manually indicated in advance; for example, whether the used information gives a competitive advantage, whether a non-disclosure agreement was signed, whether the product was reverse-engineerable, i.e., whether it could be re-engineered on the basis of public information, and whether the information was disclosed during the negotiations. These factors are subsequently related to decisions in previously decided cases. Identifying these relations can happen manually but also with a *machine-learning* algorithm that automatically learns the (statistical) correlations between factors and outcomes. Such an algorithm can explain an outcome in a content-based way that is familiar to lawyers, since the prediction was made on the basis of legally relevant factors. For example: 'I predict that the plaintiff in this case will win, since in earlier cases in which the information gave a

---

[9] Katz et al. 2017. Similar commercial predictive algorithms have been developed in the United States for lower-court cases on the basis of data about, for instance, the judges, the solicitors and the process parties, for example, Lex Machina (https://lexmachina.com), LexPredict (www.lexpredict.com) and Premonition (https://premonition.ai/). In the Netherlands such kinds of data are currently also collected for training possible predictive algorithms. See reports in *NRC Handelsblad* and *Trouw*, footnote 4.

[10] The Supreme Court Database, http://scdb.wustl.edu/.

[11] Medvedeva et al. 2019.

[12] The texts are available in the HUDOC-database, https://hudoc.echr.coe.int/.

[13] K.D. Ashley & S. Brüninghaus, 'Automatically classifying case texts and predicting outcomes', *Artificial Intelligence and Law* 17 (2009): 125-165. A salient detail is that when the relations between factors and decisions were manually entered by humans, the accuracy of the prediction was the highest, namely, 91%.

competitive advantage the plaintiff also won, even though in both cases the product was reverse-engineerable'. Moreover, the algorithm can as further explanation refer to similar precedents with the same outcome. Such algorithms could correctly predict between 82% en 88% of the decisions in cases about the misuse of trade secrets.

A big disadvantage of this approach is that the manual encoding of legally relevant factors is very labour-intensive and, moreover, a form of legal interpretation. There is current research on natural-language processing algorithms for automatically recognising factors in texts, but this research is still preliminary.[14]

In sum, algorithmic outcome predictors have a considerable number of limitations when it comes to applying them in judicial decision-making. Either they cannot explain their predictions, or they require substantial pre-processing of the data with which they work. And in both cases their predictive power is for the time being still modest, as we will explain in more detail below.

3. General issues concerning predictive algorithms

How good are legal predictive algorithms nowadays in predicting legal cases? Because of several known issues from statistics and data science, an answer to his question is not easy: what are the criteria for measuring the quality of predictions and how are things with the quality and availability of the data on which legal predictive algorithms operate? Still it can be said that for the time being the results are not spectacular.

How is the quality of predictions assessed?
The first question that arises is which assessment metrics can best be used to measure the quality of predictions. Often the *accuracy* is mentioned: how many percent of the cases from the test set are correctly predicted by the algorithm? This accuracy measure does not say very much, since in reality the distribution of the outcomes is often uneven, or 'skewed': the ECHR decides on average in 76% of the cases that the article was violated. So if we always predict 'violation' we have an accuracy of 76%, higher than the 75% of the trained predictive algorithm.[15] Moreover, the above-discussed algorithms answer a yes/no question, so even tossing a coin already scores 50%. Finally, we have to be careful with interpreting the concept of accuracy. For example, an accuracy of 80% does not mean that the probability that any given judge would in the individual case take the same decision is 80%.

There are many (statistical) metrics for algorithms that can work with skewed distributions, or with which probabilities for an individual case can be determined. A problem is that their application and interpretation are not trivial, especially not for legal professionals. Moreover, the quality of an algorithm often depends on the purpose for which it is used and how serious particular errors are: not recognising a tumour is more serious than missing a spam mail.[16]

---

[14] See Chapter 10 in K.D. Ashley, *Artificial Intelligence and Legal Analytics. New Tools for Law Practice in the Digital Age*, Cambridge: Cambridge University Press 2017 and M. Schraagen et al., 'Argumentation-driven information extraction for online crime reports', *CKIM 2018 International Workshop on Legal Data Analysis and Mining (LeDAM 2018), CEUR Workshop Proceedings* 2019.

[15] Incidentally, this does not mean that predictive algorithms are meaningless in case of such skewed distributions, or would not have learned anything, since an 'always violation' prediction will never recognise a 'no violation case' when the algorithm does recognise it.

[16] See for such a discussion about assessment metrics in the legal domain J. Bijlsma, F.J. Bex & G. Meynen, 'Artificiële intelligentie en risicotaxatie: drie kernvragen voor strafrechtjuristen', *NJB* 2019/2778.

Also relevant is how well humans perform the same task. An imperfect algorithm can still be useful if it performs better than humans on the same task.

Which data is available to train the algorithm?

The quality of predictions among other things depends on the quality and availability of the data. With predicting judicial decisions a problem is that only a small part of the case law is publicly available.[17] One can also wonder whether one can really speak of prediction if the algorithm is applied to the text of the decision to-be predicted, since this text is written after the decision has been taken and the judge therefore writes the text 'towards' the decision.[18] It would be better to predict on the basis of the case files that the court sees, but these case files are even less publicly available.

Another well-known problem from statistics is *overfitting* an algorithm on the data, where the algorithm is too much focused on specific elements in the data. For example, if the training set contains many cases about a specific state from a specific period, and during that period that state faced an uprising or civil war, then the algorithm could incorrectly conclude that cases against that state have a high probability of success.

Moreover, an algorithm that learns from past cases does not always correctly generalise to the future, since types of cases but also legal, moral or societal opinions can change. This can be seen in the experiments with the ECHR algorithm:[19] when the algorithm was only trained on cases from before the case to-be predicted, the accuracy decreased to between 58% and 68%, depending on how recent the training cases were.

4 Possible usefulness of outcome predictions

We have seen that the current outcome predictors predict far from perfectly and that it is not easy to assess and guarantee their quality. We have also seen that either they cannot explain their predictions in legally meaningful terms or they require substantial manual pre-processing of the data. Can these algorithms still be useful? We discuss this for three groups of potential users, legal academic researchers, justice-seeking parties and the judiciary, where our main focus will be on the judiciary.

4.1 Legal academic research

Algorithms and statistics can give legal academic researchers insight in the influences on judicial decision-making and in how judicial decisions change over time.[20] For example, the algorithm that predicts decisions of the US Supreme Court can help obtain insight in the

[17] In 2018 only 4,6 procent of all Dutch court cases was published at rechtspraak.nl, see *NRC Handelsblad*, footnote 4.

[18] Medvedeva et al. 2019 use, contrary to Aletras et al. 2016, only information from the case to-be predicted that is in principle available before the decision in the case is known. However, the case descriptions in the case to-be predicted are not fully identical to the desciptions that are sent to the process parties some years before the decision.

[19] Medvedeva et al. 2019.

[20] The (quantitative) study of judicial decision making is not new; the field of jurimetrics (recently renamed to *empirical legal studies*) has existed for decades; see e.g. J. Jacobs & M. Vols, 'Juristen als rekenmeesters: Over de kwantitatieve analyse van jurisprudentie', in: P.A.J. van den Berg & G. Molier (eds.), *In dienst van het recht: Opstellen aangeboden aan prof. mr. J.G. Brouwer ter gelegenheid van zijn afscheid als hoogleraar Algemene Rechtswetenschap aan de Rijksuniversiteit Groningen (Brouwer bundel),* Den Haag: Boom Juridisch 2017, pp. 89-104. What is new is the use of modern methods from machine-learning and natural-language processing technology in addition to more conventional statistical methods.

influence of the political preferences of the justices or of the presidents who appointed them. Moreover, undesirable trends or influences can be detected, such as that particular courts decide more strictly in comparable cases, or the influence of the time of the deliberations (just before lunch or not) on how strictly judges decide.[21] One might fear that this will undermine confidence in the judiciary, since it reveals that judges are also subject to typical human prejudices. However, in our opinion, such research can, by contrast, help improve the quality of judicial decision-making. Although it is important here to look further than the sometimes biased reports in the media.[22]

4.2 Justice-seeking parties

For justice-seeking citizens or companies information about the probability of success or the expected amount of damages is useful for, for instance, deciding whether to sue or to accept a settlement agreement. For decisions whether to sue the prediction does not have to be perfect: a company that is regularly involved in litigation can already be financially better off if the algorithm is just a little bit more accurate than a human employee who makes the same estimates. For the same reason a prediction does not have to be based on grounds related to the merits of a case – although justice-seeking parties will find predictions accompanied with an explanation on legally relevant grounds more useful, since they can then adduce these grounds in the case.[23] A sometimes-mentioned disadvantage of statistics is 'forum shopping': justice-seeking parties file their case with the court where they have the highest probability of success. However, forum shopping is what solicitors have always done for their clients, namely, estimating where they have the highest probability of success. Moreover, as regards legal equality it is important that the judiciary, possibly using statistics, makes sure that there are as few differences between courts as possible.

4.3 The judiciary

Can algorithmic outcome predictors in the future become algorithmic judges? Some think they can. Here sometimes the medical domain is mentioned, in which it is widely accepted that, for instance, a human oncologist has to consult a data-based predictive algorithm for recognising skin cancer if this algorithm has been proven to perform better than humans.[24] However, this analogy beaks down, since unlike in the medical example, a legal predictive algorithm and a judge perform different tasks.

In the medical example human and algorithm perform the same task, namely, recognising cancer in images of, for instance, birthmarks. Moreover, the estimates of human and algorithm are compared to the same (objective) truth: by examining the cells under a microscope it can be determined with certainty whether there is cancer. Thus a human expert

---

[21] S. Danziger, J. Levav & L. Avnaim-Pesso, 'Extraneous factors in judicial decisions', *Proceedings of the National Academy of Sciences of the Unites States of America* 2011, issue 17, pp. 6889–6892.

[22] For example, much has been written in the media about very strict 'hungry judges' (Danziger et al. 2011), but these results are contested by other researchers (K. Weinshall-Margel & J. Shapard, 'Overlooked factors in the analysis of parole decisions', *Proceedings of the National Academy of Sciences of the United States of America* 2011, issue 42; A. Glöckner, 'The irrational hungry judge effect revisited', *Judgment and Decision Making* 2016, no. 6, pp. 601-610).

[23] For example, compare the system of www.magontslag.nl, in which explicit legal rules are given concerning dismissal in labour law are given, with the more statistical predictions of LexIQ (*NRC Handelsblad* and *Trouw* 2019, footnote 4).

[24] J. Susskind, *Future Politics: Living Together in a World Transformed by Tech*, Oxford: Oxford University Press 2018. See also A. Karsemeijer, 'Zijn dit de langetermijneffecten van algoritmen?' *Rechtstreeks* 2019, issue 2, pp. 35-38.

and an algorithmic expert are compared in terms of the same standard.[25] In such a case a comparison between how humans and algorithms perform is meaningful and the algorithm can be said to perform better than the human doctor, namely, by recognising malign spots missed by the human doctor. However, in case of an algorithmic outcome predictor something else happens. Such a predictor predicts which diagnoses a human doctor (without microscope) would make, and then it is meaningless to say that the algorithm performs better than the human doctor. What is more, even a correct prediction of an incorrect diagnosis made by the human doctor counts as a success for such an algorithm. For the same reason it is meaningless to compare a legal outcome predictor with a judge, since here too a correct prediction of a legally incorrect decision would count as a success for the predictive algorithm. So the accuracy of an algorithmic outcome predictor cannot be a measure of the legal quality of the predicted decisions, since among the correctly predicted decisions there may well be legally incorrect or dubious decisions.

A fundamental problem for legal outcome predictors is that judges do not predict on the basis of statistical correlations, but decide on the basis of legally relevant reasons. The human doctors and medical predictive algorithms in the above-discussed examples look for *statistical* relations between the features and the outcomes. By contrast, judges do not look for statistical but for *legally relevant* relations. For example, suppose that a criminal judge finds it legally relevant whether the accused would lose his job in case of an unconditional prison sentence. Unemployment statistically correlates with other factors, such as residence and level of education, so a data-driven predictive algorithm would find a statistical correlation between someone's residence and whether that person received an unconditional prison sentence. But for a judge someone's residence is, of course, not legally relevant. A justification like 'you receive an unconditional prison sentence, since you live in the Bronx, but your accomplice receives a conditional prison sentence since he lives in Manhattan Upper East Side' will in general not be regarded as acceptable.

In addition, in the law there is often no clear objective truth – precisely because of this it is important that judges extensively justify their decisions, so that the decision can be assessed on its *legally relevant* content. So predictive algorithms would also have to be able to justify their predictions on legally relevant grounds. However, earlier we saw that algorithms based on machine learning can often not explain or justify their outcomes in legally meaningful terms. Only algorithms that base their predictions on relevant legal factors can do so. However, these algorithms require, as we saw in Section 2.2, that all legally relevant factors in the case files are manually indicated and classified. To do so, judges have to think about the case in legal terms, as they have always done. This restricts the usefulness of this kind of algorithm, although it is conceivable that they can still provide judges with useful information, especially if in the future it will become possible to learn relevant factors from the text of case law.

In addition, there are no fundamental objections to the use of algorithmic experts in the court room. Since algorithmic experts give advice in domains in which the judge is no expert, critically examining the grounds of a prediction makes less sense. What makes more sense is to let (technical) experts in the domain determine whether the algorithm is in general of sufficient quality.

[25] For the same reason human and algorithmic experts that predict recidivism can be meaningfully compared, since we know of persons who were released whether they reoffended after their release. See Bijlsma et al. 2019.

5 Conclusions and recommendations

In this article we have discussed what predictive algorithms can mean for the law, in particular for judicial decision-making. It turned out that there is a crucial distinction between algorithmic experts and algorithmic outcome predictors. Regarding a prediction of *algorithmic outcome predictors* – in particular data-driven outcome predictors – as legally relevant for a case decision is fundamentally flawed; it confuses prediction with decision-making, and in particular statistical correlations with legally relevant relations between a case's features and its outcome. However, there are no fundamental objections between the use of *algorithmic experts* in court cases. The same holds for the use of statistics and algorithmic outcome predictors in legal academic research and by justice-seeking parties. Among other things, predictive algorithms can give legal academic researchers insight in what influences judicial decision-making and how decisions change over time, and they can be useful for citizens and solicitors for estimating the probability of success in court. The only problem that remains is the practical problem of determining the quality of such algorithms. This is far from trivial, witness the many pitfalls about which data science teaches us.

The enormous attention for predictive algorithms is definitely a hype, but we do not want to call it meaningless. In our article we have discussed several ways in which they can be useful in the law. In addition, research on such algorithms, and more generally on AI algorithms for the law, is strongly needed. The legal world benefits from solid quantitative analyses supported by AI, and is in this respect far behind, for example, medical science. Moreover, the extremely rapid developments in natural-language processing make it possible to better digitalise aspects of judicial decision-making, for example, by providing support for searching or summarising relevant documents or finding similar cases. It is important that the legal world teams up with AI developers and researchers in the development, validation and explanation of algorithms.

To conclude, we make the following recommendations with respect to the use of predictive algorithms by the judiciary.

Data-driven outcome predictors have no role in court cases. 'Content-based' outcome predictors may have a role in court cases, provided they can justify their outcomes on legally relevant grounds and their quality is sufficiently ensured. Moreover, the judge may only be guided by these legally relevant grounds and not by statistical metrics like accuracy. It is not accuracy percentages or probability estimates but only content-based arguments that can be legally relevant reasons for a decision.

Predictions of algorithmic experts can be useful for judges, but they should only be used if the judge has sufficient insight into the quality of the algorithm. Therefore it is important to have rules and procedures concerning the admissibility of algorithmic experts, just as there are rules and procedures concerning the admissibility of human experts.