

Predicting the Admission into Medical Studies in Germany: A Data Mining approach

Open Access at KIT

by Dominik Jung¹, Lorenz Kemper², Benedikt Kaempgen¹, Achim Rettinger¹

KIT SCIENTIFIC WORKING PAPERS 27



¹Institute of Applied Informatics and Formal Description Methods (AIFB)
Kaiserstraße 89
76128 Karlsruhe
Germany

²Chair of Public Finance and Public Management
Kronenstraße 34
76128 Karlsruhe
Germany

Impressum

Karlsruher Institut für Technologie (KIT)
www.kit.edu



Diese Veröffentlichung ist im Internet unter folgender Creative Commons-Lizenz
publiziert: <http://creativecommons.org/licenses/by-nc-nd/3.0/de>

2015

ISSN: 2194-1629

Predicting the Admission into Medical Studies in Germany: A Data Mining approach

Dominik Jung¹, Lorenz Kemper², Achim Rettinger¹ and Benedikt Kmpgen¹

¹ Institute of Applied Informatics and Formal Description Methods (AIFB) -
Karlsruhe Institute of Technology (KIT), 76128 Karlsruhe, Germany,

² Chair of Public Finance and Public Management - Karlsruhe Institute of
Technology (KIT), 76128 Karlsruhe, Germany,
dominik.jung@student.kit.edu, lorenz.kemper@kit.edu, rettinger@kit.edu,
benedikt.kaempgen@kit.edu

Abstract. In Germany, the placement into medical studies programs is highly competitive. Even for excellent applicants success is uncertain. Tackling this uncertainty, the aim of this paper was to investigate the success of an application from a data perspective. For this purpose, applicant data in the time frame from 2009 to 2012 was analyzed and various Data Mining approaches were developed and compared. The characteristics of the statistical patterns lead us to simple recommendations how to alter an application in order to succeed. Also, our results indicate that Data Mining can outperform personal predictions.

Keywords: Predicting Admission, Medical Studies, Data Mining, Knowledge Discovery, Data Exploration

1 Introduction

Each year ten-thousands of German high school graduates apply for a placement into a medical studies program. However, the supply of study opportunities is scarce. In fact, on average, German universities provide only one spot per five applicants [4]. In order to handle this excessive demand certain selection procedures have been put in place. As medical education in Germany is largely provided by public institutions and as the numbers of applications have become increasingly overwhelming the selection procedures mostly rest upon objective, easily-comparable criteria. This comes in handy to data science as hard criteria allow for a convenient collection, aggregation and analysis of relevant applicant attributes.

Hochschulstart.de, formerly known as the Central Office for the Allocation of Study Places (ZVS), is the responsible agency to administer the placement of applicants into medical studies for more than 30 universities. Thankfully, they have provided us the respective dataset of all applicant data between the summer semesters of 2009 and 2013. Each semester the agency carries out a selection process that is for one part determined by the government and for another by the individual universities. Legal regulation constitutes that 20%

of the college spots for medical studies have to be reserved for the best high school graduates regarding the Abitur grade (ABV: best-graduate-procedure). Another 20% are allocated to the applicants who have waited the longest to study (WZV: waiting-time-procedure). The remaining 60% are allocated according to university-specific rules (AdH: university procedure). While some universities are content with picking their potential students only depending on their graduation marks, others award bonuses for a finished apprenticeship or even decide on the grounds of softer criteria such as applicant interviews.

The university-specific selection procedures and criteria are always published on hochschulstart.de or the universities websites. Also, the admission results of former years are accessible in form of numerus clausus (NC) figures. The NC simply states the characteristics of the worst successful applicant. Therefore, the parameter implies that all applicants with better characteristics, such as better grades, have been accepted in that particular year. Consequently, applicants can have a grasp of their chances to be admitted. However, each year the number of applicants varies. Moreover, the performance levels of applicants and the number of supplied places to study fluctuates between years.

Therefore, applicants face uncertainty. As the general procedure merely allows electing up to six preferred institutions, an applicant must come to a decision under incomplete information and uncertainty.

We approach this environment on the grounds of knowledge discovery and data science. Our aim is to observe whether an analysis of the dataset will grant a significant edge to come to good decisions and predict outcomes.

As there is a great range of possible questions we constrain ourselves and focus on one particular, namely, we attempt to predict the success or failure of an individual applicant?

In the following, we wish to present our approach to this problem. We start by outlining our methodology and the features of our dataset. Subsequently, we summarize our data handling and data analysis before we conclude by evaluating the potential of our prediction results.

2 Methodology and Related Work

The discussions about admission into medical studies in Germany has a long history. The graduation of the first woman, Dorothea Erxleben, in 1754 in medical studies, was followed by numerous investigations, discussions and studies how, when and why woman and men will succeed in medical studies from then[1] until today[3].

But one relevant questions was never asked: With which likelihood someone is somebody admitted to his favorite studies?

So we see that work as an explorative approach to determine and detect relevant factors in the admission process. We also want to give potential students a possibility for forecasting and decision-support. As an explorative example for other areas we chose medical studies as in Germany this field is particularly

competitive. Because there are no other approaches in the literature at the moment, we had to build up an own research process developing different approaches to predict the admission success.

2.1 Research Design

We have oriented our approach based on the methodology for Data Mining process of Ester and Sander[2]. They divide the process of Knowledge Discovery into separate steps:

- **Focussing:** The objective of this step is to define the objective of the research and the data source
- **Preprocessing:** In this step you complete and integrate the data in your database. Additionally, you need to check for mistakes or missing values.
- **Transformation:** The third step is the process of converting data from the raw format to fit data analysis. You can also derive new attributes from the raw data.
- **Data Mining:** The fourth step covers all methods and techniques for pattern recognition.
- **Evaluation:** Finally you conclude by evaluating your results and present them in a comprehensive way.

In addition, we applied some standard descriptive statistical methods on the raw data for a better understanding of our data and the problem. Because we have

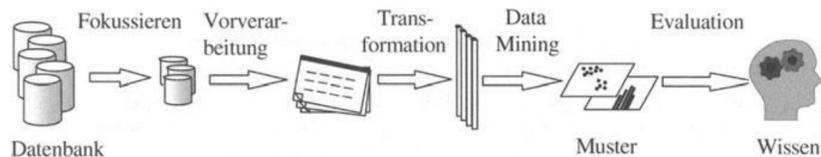


Fig. 1. Different steps of Knowledge Discovery by Ester and Sander [2]

comprehensively presented the process of "Focussing" in the first part of this work, we continue directly with preprocessing and data transformation.

2.2 Data Preprocessing and Transformation

One central step of our research was the "Data Preprocessing and Transformation". In this step we used KNIME and R to increase the quality of our data. Before our preprocessing and transformation the given attributes realizations within our dataset were:

Table 1. The attributes of the raw data

Attribute	Data type	Value range
Admission best graduates procedure (ABV)	binary	Yes/No
Admission waiting time procedure (WZV)	ordinal	Yes/No
Admission procedure of universities (AdH)	binary	Yes/No
University of admission	nominal	Among 36 universities
State of high school graduation	nominal	16 states
Graduation mark	ordinal	1.0-4.0
Waiting time	ordinal	0-100 semesters
Civilian service	binary	Yes/No
Test grade (TMS)	ordinal	1.0- 4.0
Medical apprenticeship	binary	Yes/No
University preferences 1-6 (ABV)	nominal	Among 36 universities
University preferences 1-6 (AdH)	nominal	Among 36 universities

While attributes such as a completed medical apprenticeship are increasingly incorporated in the universities selection procedures, the decisive factor for admission remains the graduation mark.

For further data analysis we decided to derive the following additional attributes:

Table 2. New derived attributes

Attribute	Data type	Value range
Average graduation mark per year (AG)	ordinal	1.0-4.0
Average graduation mark per year per state (AGS)	ordinal	1.0-4.0
Deviation of the student from AG	ordinal	1.0-4.0
Deviation of the student from AGS	ordinal	1.0-4.0
State of the University in ABV	nominal	16 states
State of the University in AdH	nominal	16 states

3 Exploration of the raw data

From summer 2009 to summer 2013 about 250.000 applications have been received. Each of these has been stored as a data tuple including a list of up to 22 relevant attributes. Firstly, the attributes include information about the applicant such as his graduation mark, whether he has already finished a medical apprenticeship or how long he has already waited for a placement. Second, his preferences regarding university locations are listed. Finally, it is noted if and where the candidate was eventually admitted.

3.1 The meaning of the grade:

As we will show in our analysis part, plainly looking at the grade will enable us to predict the admission of 84% of candidates correctly. The constant medial discussion about the numerus clausus and the hardship of not-excelling graduates to gain a placement in medical studies already suggests that the importance of the grade has been apprehended by the public. The dataset gives evidence of this assumption as it reveals clear selection effects. As you can see in Fig. 2,

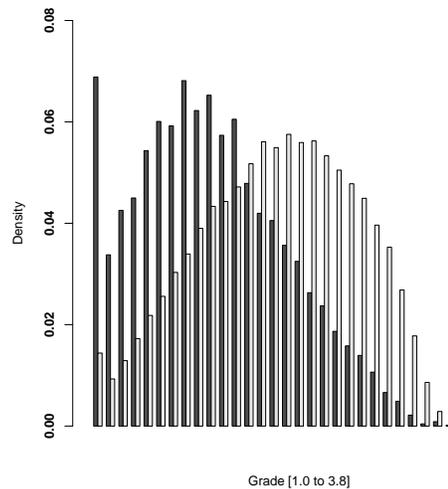


Fig. 2. Comparing grade densities of high school graduates in 2012 (light) and medical studies applicants (dark) from Baden-Württemberg (1 of 16 regions in Germany)

excelling high school graduates are clearly overrepresented in the set of medical studies applicants. The graph presents comparison values from the state of Baden-Württemberg. Applicants from the state (dark bars) clearly outdo the 2012 cohort of high school graduates (light bars) in respect to their Abitur performance [7]. While strong students might also be simply more prone to become a doctor, the magnitude of the numbers nurtures the hypothesis that many interested graduates with bad or average grades are discouraged to apply at all. Certainly, they do so due to a good reason, since their expectation not to be accepted is congruent with reality.

However, the graph also shows that there are students with bad and average grades that apply, at all. This can be explained in two ways. First, they simply may have false expectations about their chances; second, they have waited a long time to study and hope to be accepted through the 20% waiting-time quota

(WZV). Not surprisingly, as the WZV only considers the grade as a secondary factor, Fig. 3 shows that these hopes are very valid.

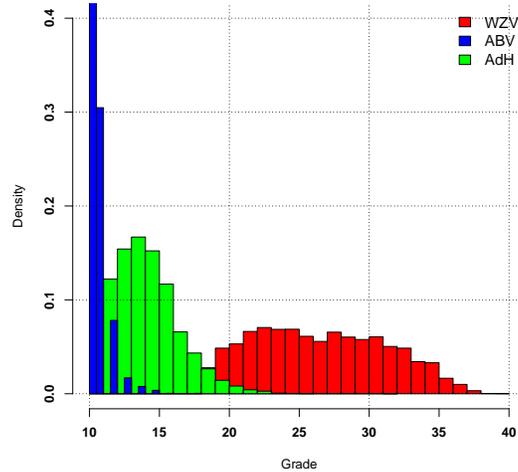


Fig. 3. Comparing grade densities of admitted applicants depending on the respective selection procedure

Comparing the marks of candidates accepted through the WZV (red bars) with those admitted by AdH (black) or ABV (blue) the diminishing weight of the graduating mark as decisive factor becomes apparent.

In fact, the distribution of accepted candidates through the WZV is almost uniform and the slight curving can easily be explained by the likewise-curved distribution of all High School graduates. It does not matter if one graduates with 2.0 or 4.0, students are apparently equally prone to study medicine.

Or in other words, this supports the thesis that it is the self-selection in anticipation of failure of average to bad students that leads to the excellence of applicants on the whole.

3.2 Extracting distinctive features and other decisive attributes

Beside the grade (and secondarily the waiting time) which other attributes determine a successful application? Following, we present selected observations for all secondary determinants.

University preferences: For one, there are the university preferences. Even though the German university system is known to be comparatively undifferentiated there exist significant differences between groups of individuals electing

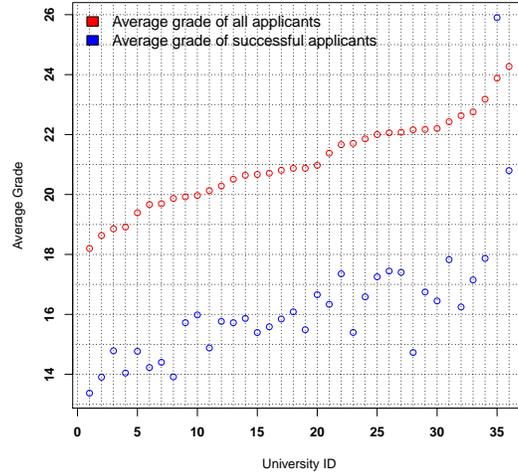


Fig. 4. Average grade of all applicants vs successful applicants

different universities. As shown in Fig. 4 (red dots) there is a gap of more than half a grade point in applicant average grades between the most competitive and the least competitive university.

While applicants for the Charite Berlin average a grade of 1.8, applicants for the University of Greifswald only aggregate to a mean of 2.4. The same trend can be observed looking at mean data of admitted applicants (blue dots). Consequently, choosing particular universities will have a crucial impact on application success.

Waiting time: Moreover, as already pointed-out the waiting time is important for the selection process as it is the primary factor regarding 20% of the placements. In addition, the procedure is global i.e. not split up into many university-specific procedures. As expected, the data simply reflects the procedure rules by having a clear-cut two-level distribution with the turnabout at about 12 semesters waited.

Medical apprenticeship and civilian service: Third, a medical apprenticeship or time spent for civilian service can be influencing factors. However, as the descriptive statistics show they are only able to tilt applicants on the edge of admission towards success. As Fig. 6 shows, the likelihood of admission for candidates with completed medical apprenticeship only increases positively for candidates with a grade within the low 2 range and only for candidates with grades below 1.9 it increases significantly.

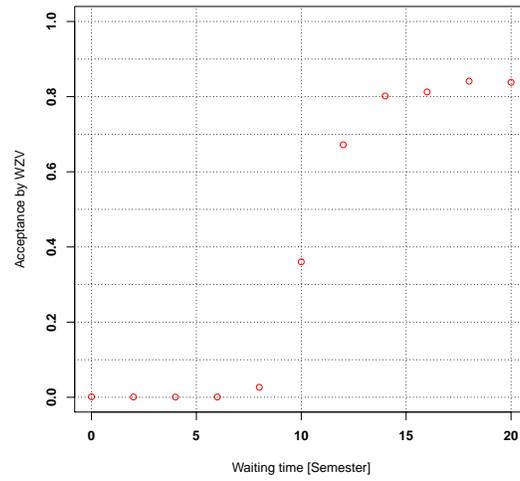


Fig. 5. Admitted through the WZV per time waited

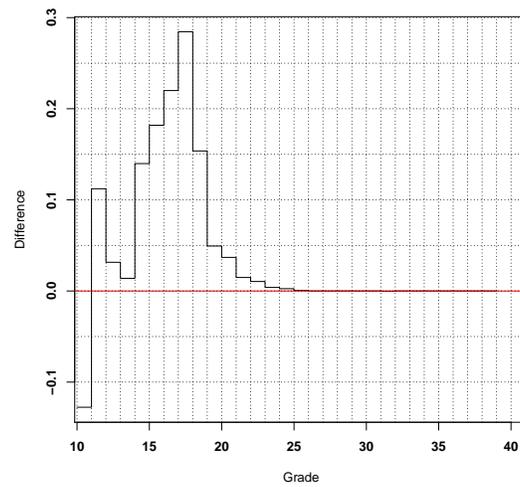


Fig. 6. Increased likelihood of admission due to a medical apprenticeship (per grade)

Fig. 7 uses the same method in order to assess the effect of a civilian service. Much like a medical apprenticeship, a civilian service appears to be able to tip the scales. This time, however, applicants need even better grades (less than 1.7) in order to benefit.

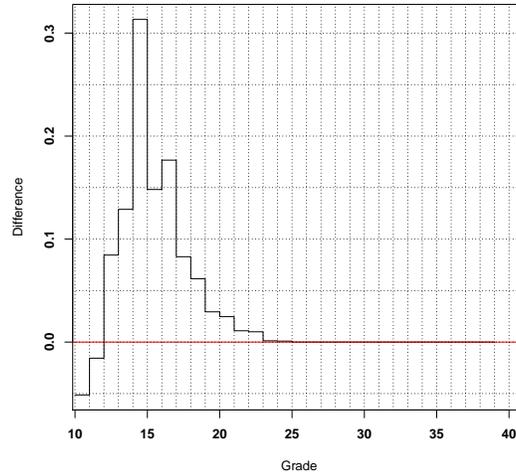


Fig. 7. Increased likelihood of admission due to a civilian service (per grade)

Test for medical students (TMS): Finally, we are left to analyze the relevance of the test for medical students (TMS). First, we were interested in how much the test results deviated from Abitur grades and therefore how much value they added to the selection procedure. Surprisingly, we found the test results to be less correlated to Abitur performance than expected (correlation: 0.34, see figure 8). Seemingly, they add, in fact, another notion to the selection process. So how much does the TMS participation and score contribute to a successful admission? Performing linear regression analysis (on the set of candidates that have taken the test) we find a significant impact (see table 3). While a large part of the influence of TMS scores on admission is confused for Abitur performance, integrating it into the equation still leaves us with a significant effect of TMS scores to predicting admission success.

Interim summary: Descriptive statistical analysis enables us to get a feeling for the relevance of the various attributes and therefore helps understanding what to expect from the dataset. We found that the Abitur grade is the major factor to determine admission into medical studies. But also a smart election

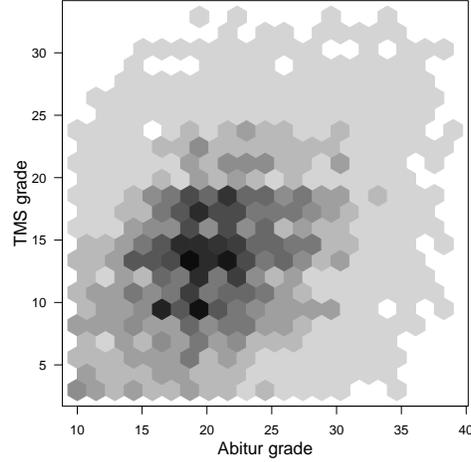


Fig. 8. Abitur grade vs TMS grade

Table 3. Linear regression - Admission Abitur grade + TMS score

Intercept	Abitur grade TMS Score	
0.64910***		-0.02331***
1.04306	-0.02825	-0.01408

**** significant at 5 % level

of universities, the waiting time, a medical apprenticeship or a strong TMS can contribute in certain scenarios.

In the following section we will test several inductive methods. Thereby, we attempt to bring the factors together in the right fashion - ultimately aiming to find a sound rule to predict admission success on an individual level.

4 Results and Data Analysis

In this section we start by presenting simple approaches like the One-Rule Classifier [9]. Then we proceed with more complex techniques like decision trees and an own model. By choosing the different Data Mining approaches, we follow the most commonly used algorithms in Data Mining [9]. Finally, we collect and compare our data analysis results.

4.1 Simple rule-based classifier

One very simple approach that can show in simple data very useful results is the One-Rule Classifier[9][8]. It has only one parameter, the minimum bucket size for discretization. The algorithm generates one rule for each value of the predictor in the data and then selects the rule with the smallest total error.

One other simple rule-based classifier is the The Zero-Rule Classifier[9]. It chooses the class that is most likely within the data. The classifier is useful for having a reference baseline to which we can compare other approaches.

4.2 Decision Tree

A decision tree is a machine learning method that can be used for classification or for regression problems[5]. Like the One-Rule-Classifier, its structure is easy to understand and it can predict numerical as well as categorical classes.[5][9]

The building of such a tree normally consists of two steps: First, a tree is built utilizing some rules, that split the data into different subsets. Second, the tree gets pruned with the objective to not overfit and to remove parts that are unimportant with regard to accuracy.

The main difference between trees for classification or regression is, that the classification tree predicts a qualitative response rather than a quantitative response. [5] In this paper we are content with assessing if a student will be accepted or not. In this scenario a classification decision tree will be a good solution.

One of the most popular decision trees is the ID3 or the C4.5 from John Ross Quinlan.[6] The C4.5 algorithm is an improvement of the ID3, because it is, for one part, capable of handling missing values in the dataset (which is necessary in this scenario, as not every student had joint all accommodation processes). Also the C4.5 the decision tree is pruned after one run. In this research we used a Java implementation of the C4.5, the J48 decision tree.[9]

4.3 Modelling the selection process

At hochschulstart.de selection mechanisms as well as numerous clausus data of 21 universities is made easily available. For the next approach we used this information in an attempt to recreate a prognosis from the viewpoint of an individual applicant. We wrote a computer program comprising the universities' selection procedures as functions. The program would evaluate the data tuples regarding each selection procedure of individually elected universities. In the following, the program would compare the outcome with the previous year's numerous clausus. In other words, we simply took individual characteristics and assessed whether; depending on the selection procedures, an applicant would have made it in the previous year.

Thus, we recreated the approach, a real person would take to evaluate her or his admission chances. The method was not only useful as an end itself, but posed, in addition, a benchmark whether we could actually give better forecasts than a single person searching for information on the internet could.

4.4 Results

Table 4. Comparison of the different approaches

Approach	Parameters	Accuracy in %
Zero Rule	--	73.78
One Rule	minBucketSize=6	84.46
J48	Conf=0.25,folds=3	91.06
J48 pruned	Conf=0.25,folds=3	92.06
Modelled selection process	--	91.14

The Zero-Rule gives the percentage of applicants not admitted to medical studies. Consequently, predicting 'not admitted' for everyone would already yield 73% correct predictions. The One-Rule splits the data up into two subsets regarding their Abitur grade. The accuracy increases to 84%. The comparatively large precision gain of over 10% reflects the 'sure' candidates on the top of the scale and the 'basket cases' on the bottom of the scale. Growing more precise in our modelling of the selection process we manage to cross the 90% mark. Notice that the results of this estimator are probably overestimated, as the 21 universities that published their selection criteria on a central website are rather transparent and are therefore presumably easier to predict as the others. Bearing this in mind we can be quite confident that our Data Mining approach, the decision tree, yields convincing results.

In Data Mining different approaches exist to evaluate the fit of a model. The most common are the 10 fold cross-validation and the partitioning into testing and training sets.[9][2] In this work we choose to evaluate the different approaches on our newest dataset, the "Winter Semester 2012", which we consequently not used for training our classifiers. Evaluating on our newest dataset, gives us the most valid values for the accuracy for future years.

Learning the classifiers only with the use of former year datasets such as '2011' or '2011 and 2010' yielded only slightly less accuracy than learning on the whole dataset. This may show, that the underlying admission process did not change radically in the last years. More importantly it reassures us in believing that the data comprises valuable information, in order to predict future admission processes.

5 Conclusion

In this paper we aimed to reveal some aspects of what we can read from applicant data. We evaluated the individual factors of university selection to medical studies and pointed out their situational importance. Moreover, we were able to

show that methods of Data Mining can outdo individual persons in predicting their admission.

While we are content to have made a first step, 92% accuracy is not an overwhelming number. As additional data, such as graduate numbers, may be able to further increase precision; we are convinced that much of the remaining 8% is made up of random factors.

How many graduates of a year decide to go for medicine? How good are they? Which universities will they choose? A large part of these yearly variations will be unpredictable noise. However, we can enhance our predictions in quantifying how much individual admission chances may be influenced by this noise. So, instead of having a binary prediction outcome of 'will or not will be admitted' we could create a percent estimation of admission chances - 'you will be admitted with a ??% chance'. This were to reflect our varying level of certainty and serve applicants the required information to make an informed decision.

This leads to the concluding objective. We believe that applicants striving to go into medicine can make use of the information of data analysis. They gain a realistic view of their options and are now able to make career decisions more informed. Thus we aim at publishing our predictions in an accessible fashion for the applicants. For this purpose we have created a prototype recommendation program, which we will be released on the website www.zulassungsorakel.de in January 2015 (see 9).

The screenshot shows the web application interface for 'Zulassungsorakel.de'. At the top, there is a navigation bar with 'Medizin' selected and links for 'Tipps zur Bewerbung', 'Mehr über Zulassungsorakel', and 'Nutzungsbedingungen'. Below this is a section titled 'Angaben zur Hochschulzulassung' with dropdown menus for 'Bundesland: Baden-Württemberg', 'Abschlussnote: 2,0', and 'Jahr: 2013'. The main content area is divided into two columns. The left column, 'Priorität im WZV/AbV', contains six rows, each with a 'Priorität' label and a dropdown menu set to 'keine Angabe'. The right column, 'Sonstige Qualifikationen', includes dropdowns for 'Wartesemester: keines' and 'Medizinertest: keinen', and checkboxes for 'Zivildienst absolviert' (checked), 'Ausbildung im WZV' (unchecked), and 'Ausbildung im AdH' (unchecked). Below these is a 'HINWEISE' box with text explaining alternative admission procedures. At the bottom left, there is a 'Bewerbungssemester' dropdown set to 'SS2012' and a blue button labeled 'Chance bestimmen »'.

Fig. 9. Servlet Zulassungsorakel - the prediction program for students

References

1. Brinkschulte, E.: "... das Weib lässt sich nicht gegenüber dem Gelehrten vergessen" - Zur Aufrechterhaltung der Virilität des Hochschulbetriebes um 1900 (2005), in Stahnisch, F., Steger, F. (Hg.): *Medizin, Geschichte und Geschlecht - Körperhistorische Rekonstruktionen von Identitäten und Differenzen* (Geschichte und Philosophie der Medizin). Stuttgart: Franz Steiner Verlag.
2. Ester, M., Sander, J. (2000): *Knowledge Discovery in Databases - Techniken und Anwendungen*. Berlin/Heidelberg: Springer-Verlag.
3. Glocker, D. (2010): *Enrollment and success in higher education - Structural estimation and simulation evidence for Germany*. Diss., Freie Universität Berlin.
4. Hochschulstart - Stiftung für Hochschulzulassung (2013): *Daten der bundesweit zulassungsbeschränkten Studiengänge an Universitäten - WS 2013/2014*, available online at: hochschulstart.de/fileadmin/downloads/NC/WiSe2013_14/bew_medizin_ws13.pdf
5. James, G., Witten, D., Hastie, T., Tibshirani, R. (2013): *An Introduction to Statistical Learning - with Applications in R*. Berlin/Heidelberg: Springer-Verlag.
6. Quinlan, J. R. (1993): *Programs for Machine Learning*. Morgan Kaufmann Series in Machine Learning.
7. Statistisches Landesamt Baden-Württemberg (2013): *Pressemitteilung Nr. 92/2013*, available online at: statistik-bw.de/Pressemit/2013092.asp?BildungKultur
8. Weka Toolkit API Documentation, version august 2014, available online at: cs.waikato.ac.nz/ml/weka/documentation.html
9. Witten, I. H., Frank, E., Hall, M. A. (2011): *Data Mining - Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems.

KIT Scientific Working Papers
ISSN 2194-1629

www.kit.edu