# RAPP: A Responsible Academic Performance Prediction Tool for Decision-Making in Educational Institutes

Manh Khoi Duong[1], Jannik Dunkelau[2], José Andrés Cordova[3], Stefan Conrad[4]

**Abstract:** Due to the increasing importance of educational data mining for the early intervention of at-risk students and the growth of performance data collected in educational institutes, it becomes natural to employ machine learning models to predict student's performances based off prior data. Although machine learning pipelines are often similar, developing one for a specific target prediction of academic success can become a daunting task. In this work, we present a graphical user interface which implements a customizable machine learning pipeline which allows the training and evaluation of machine learning models for different definitions of academic success, e. g., collected credits, average grade, number of passed exams, etc. The evaluation is exported in PDF format after finishing training. As this tool serves as a decision support system for socially responsible AI systems, fairness notions were included in the evaluation to detect potential discrimination in the data and prediction space.

**Keywords:** educational data mining; fairness; decision making; machine learning; academic performance prediction

## 1 Introduction

*Academic performance prediction* (APP) systems can be used to identify *at-risk students* in higher education early on, allowing the university to use resources in a targeted manner to prevent them from achieving poor academic performances. The definition of at-risk students varies as it depends on the context and the purpose of prevention. It can comprise of, e.g., higher chances of dropping out, longer study durations, and worse graduation grades. In this case, the APP system acts as a supporting *artificial intelligence* (AI) system for the university at the institutional level. However, given the impact of such systems onto the student body, social challenges arise. Marcinkowski et al. [Ma20] surveyed the perception of a student body of the use of such AI-based systems and show that APP is viewed as problematic by students as far as their own data and planning are concerned. Furthermore,

[1] Heinrich Heine University, Department of Computer Science, Universitätsstraße 1, 40225 Düsseldorf, Germany manh.khoi.duong@hhu.de

[2] Heinrich Heine University, Department of Computer Science, Universitätsstraße 1, 40225 Düsseldorf, Germany jannik.dunkelau@hhu.de

[3] Heinrich Heine University, Department of Computer Science, Universitätsstraße 1, 40225 Düsseldorf, Germany jose.cordova@hhu.de

[4] Heinrich Heine University, Department of Computer Science, Universitätsstraße 1, 40225 Düsseldorf, Germany stefan.conrad@hhu.de

doi:10.18420/BTW2023-29

the notion of *fairness-aware machine learning* (FairML) [DL19, Fr19, PS20] becomes an increasingly important topic and also found its way into educational data mining systems [LMZ19, KLM22, HR20, KL20, LQN21, AC19].

Acknowledging these issues, we developed a tool for *responsible academic performance prediction* (RAPP) which tackles two main tasks: it is a tool for (1) academic performance prediction and acts as a (2) decision support system for the social responsibility when employing AI in tertiary education. The first task deals with generating multiple prediction targets and datasets for the prediction of academic performances. The goal of the second task is to find socially acceptable machine learning (ML) models and justify their use from the extensive fairness and interpretability evaluation in the tool. For the full deployment of an AI system to identify at-risk students, ethical aspects and the perception by those affected have to be researched. The fairness and interpretability evaluation plays a supportive role to disregard or regard certain ML models by, e. g., checking whether they comply with student's perception of discrimination or do not discriminate through socio-demographic features.

The source code of the RAPP tool is published under the MIT License and available online at `https://github.com/hhu-rapp/rapp-tool`.

## 2 Related Systems

Our proposed tool combines functionalities from two different research communities: (educational) data mining and fairness assessment. In this section, we will briefly present selected tools already available from either community.

RapidMiner [HK16], Orange [De13], and WEKA [Ha09]—to name a few—are data mining tools with a graphical user interface (GUI) just as the proposed tool in this paper. The aforementioned tools mostly include data visualization, pre-processing, feature selection, clustering, classification, regression, and evaluation metrics. The tools are modular, meaning the pipeline and its specific configurations are highly modifiable. Their aim is to enable data mining practitioners the comparison of machine learning models on custom datasets without having to write code themselves.

Although not as comprehensive and powerful, tools that were explicitly developed for educational data exist as well. They predominantly focus on a specific dataset that was provided by a particular educational institute. Especially, they analyze and predict several students' data such as programming grades [Ba16], examinations of the final school year [LMP16], students' contributions in group programming [SA20], or students' written feedback [Gr20].

Fairness and transparency in machine learning have become more important in recent years due to the awareness of potential mistreatment of AI over different demographic groups [DL19, Fr19, PS20]. As a response, authors began developing tools to audit the fairness of an ML system and to produce bias reports, to guide the selection process of a fitting fairness metric, or to apply intervening methods to reduce exhibited bias. Examples for such tools are Aequitas [Sa18], FairSight [AL19], Fairlearn [Bi20], or Fairness Compass and Fairness Library [RD22]. These topics have also been recognized by the educational data mining (EDM) community lately. To name some, Hu and Rangwala [HR20] and Kizilcec and Lee [KL20] consider prejudice and unfairness where Le Quy and Ntoutsi [LQN21] and Alonso and Casalino [AC19] acknowledge the explainability of the used models in EDM.

For the proposal, the RAPP tool aims to take on the preliminary works and combine functionalities from both communities: It is a data mining tool for educational data that includes fairness examinations and interventions to address responsibility when employing AI in educational institutes.

# 3 RAPP Tool

Making it possible to easily create various datasets from a single database with desired features and labels to train, save, and evaluate machine learning algorithms is the aim of the developed tool. For this, the GUI provides an intuitive way to load a particular SQLite database or a CSV file[5] and specify the initial settings for the machine learning pipeline. The demanded features and target labels can be derived by querying the database. Several settings are detected automatically such as the prediction type (classification, regression), the target variable (last column by default), and categorical features. The supported estimators for classification are *decision trees, random forest, support vector machine, naive bayes, and logistic regression* and for regression *linear regression, elastic net, bayesian ridge, decision tree regressor*, and *kernel ridge*. An *artificial neural network* with two hidden layers is also available for both of these task types. Experienced users can modify the configuration for their needs. Fig. 1 displays the user interface for the settings.

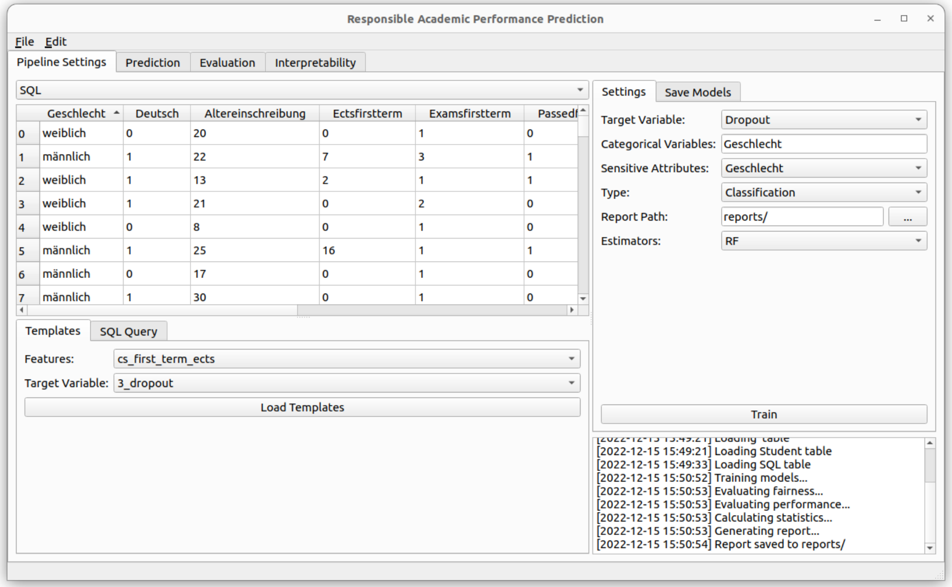
In the following, we will outline the two main uses and functionalities of the RAPP tool: APP and supporting the decision-making process whilst designing a responsible APP.

## 3.1 Academic Performance Prediction

### 3.1.1 Pipeline

At the front of the RAPP tool lies the ability to setup and train APP models over the implemented ML pipeline. The pipeline is outlined in Fig. 2. First, the pipeline's settings have to be specified. This includes the selection of a dataset to use for training as well as picking the ML algorithms to train.

[5] The CSV file is treated as a database.



|  | Geschlecht | Deutsch | Altereinschreibung | Ectsfirstterm | Examsfirstterm | Passedf |
|---|---|---|---|---|---|---|
| 0 | weiblich | 0 | 20 | 0 | 1 | 0 |
| 1 | männlich | 1 | 22 | 7 | 3 | 1 |
| 2 | weiblich | 1 | 13 | 2 | 1 | 1 |
| 3 | weiblich | 1 | 21 | 0 | 2 | 0 |
| 4 | weiblich | 0 | 8 | 0 | 1 | 0 |
| 5 | männlich | 1 | 25 | 16 | 1 | 1 |
| 6 | männlich | 0 | 17 | 0 | 1 | 0 |
| 7 | männlich | 1 | 30 | 0 | 1 | 0 |

Features: cs_first_term_ects
Target Variable: 3_dropout

Target Variable: Dropout
Categorical Variables: Geschlecht
Sensitive Attributes: Geschlecht
Type: Classification
Report Path: reports/
Estimators: RF

```
[2022-12-15 15:49:21] Loading  table
[2022-12-15 15:49:21] Loading Student table
[2022-12-15 15:49:33] Loading SQL table
[2022-12-15 15:50:52] Training models...
[2022-12-15 15:50:53] Evaluating fairness...
[2022-12-15 15:50:53] Evaluating performance...
[2022-12-15 15:50:53] Calculating statistics...
[2022-12-15 15:50:53] Generating report...
[2022-12-15 15:50:54] Report saved to reports/
```

Fig. 1: RAPP's Pipeline Settings Interface, 2022.

The data are queried over an SQLite database. While advanced users can enter custom queries on the database for feature engineering and feature selection, predefined feature and label sets were added for the given academic database to comfortably reuse and combine them in any desired pairing. The user can select, for instance, features such as credit points, grades, or number of passed exams, and target labels such as final GPA, achieved credits until semester $x$, or study duration. To ease working with different sets of features and labels we implemented an SQL templating engine which produces the final query based on the user's selections for a feature and a label set. This avoids combinatoric explosion which would arise if each feature-label pair's SQL query had to be implemented manually. The queried database then acts as a dataset for the machine learning pipeline.

Once the dataset is obtained, the features go through the pre-processing step of one-hot encoding any categorical features. After this, the data is split into training (80 %) and test (20 %) data.

Each of the user's selected models are trained on the training data. We also evaluate the performance over the training data via 5-fold cross-validation to capture how robust the models behave during training. The training concludes in an evaluation over various performance metrics as well as fairness metrics. Fairness is also audited directly over the dataset as well. The evaluation results are saved into a detailed PDF report file containing information over the demographics of the dataset as well as the performance and fairness results of each trained estimator.

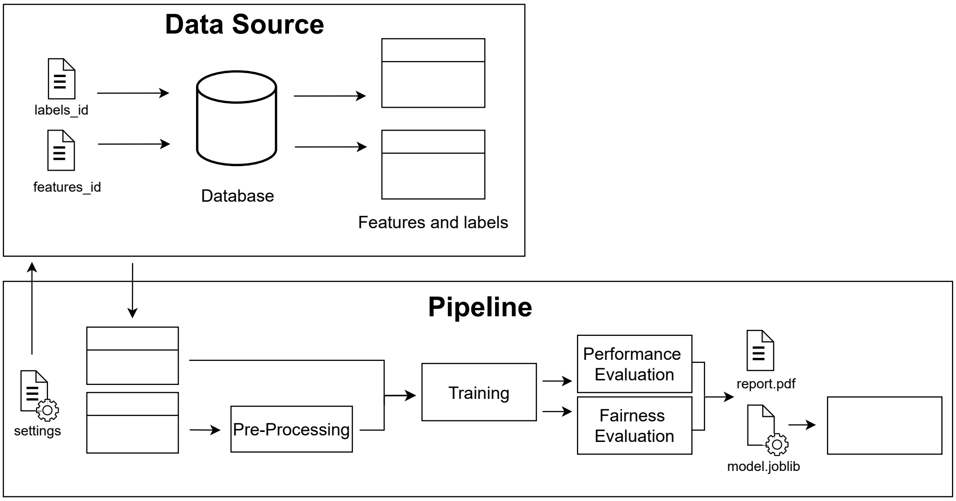


Fig. 2: RAPP's Machine Learning Pipeline, 2022.

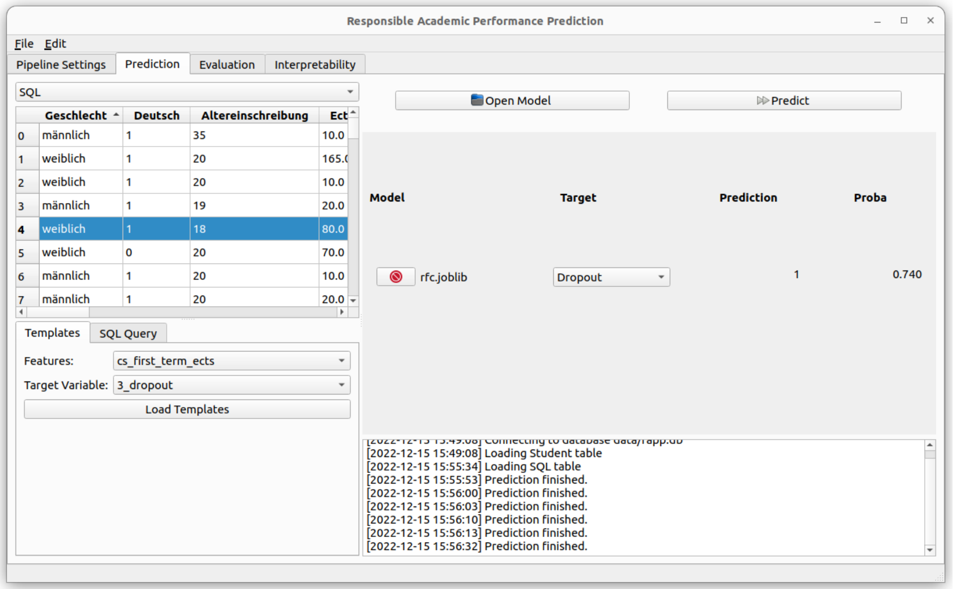
After the trained models are evaluated, the users can decide which models they want to save in order to use them later to predict on new data.

### 3.1.2 Prediction

To tackle the task of identifying at-risk students early, this tool includes a prediction interface as shown in Fig. 3. This interface enables the user to make predictions based on individual student's academic data. The user can then identify students who are more likely to benefit from the institution's support programs.

In order to predict the students' performances, new data from students as well as compatible models, i. e., models that have been trained with the same features, are required in the prediction interface. It is possible for the user to load various models trained for different target variables to predict several targets from the same features simultaneously. Once new data and selected models are loaded into the GUI, the features go through a pre-processing step and are then fed into the loaded models for the prediction. Fig. 3 shows an example of multiple targets being predicted with the data of a single student.

After the prediction has been run, the interface updates and displays the predictions of the models for each of the selected targets. It is also possible to load multiple models for one specific target to employ ensemble learning. In case of classification, we apply majority voting whereas in regression tasks the mean of the predicted values is used.



|  | Geschlecht | Deutsch | Altereinschreibung | Ect |
|---|---|---|---|---|
| 0 | männlich | 1 | 35 | 10.0 |
| 1 | weiblich | 1 | 20 | 165.( |
| 2 | weiblich | 1 | 20 | 10.0 |
| 3 | männlich | 1 | 19 | 20.0 |
| 4 | weiblich | 1 | 18 | 80.0 |
| 5 | weiblich | 0 | 20 | 70.0 |
| 6 | männlich | 1 | 20 | 10.0 |
| 7 | männlich | 1 | 20 | 20.0 |

Features: cs_first_term_ects

Target Variable: 3_dropout

| Model | Target | Prediction | Proba |
|---|---|---|---|
| rfc.joblib | Dropout | 1 | 0.740 |

```
[2022-12-15 15:49:08] Connecting to database data/rapp.db
[2022-12-15 15:49:08] Loading Student table
[2022-12-15 15:55:34] Loading SQL table
[2022-12-15 15:55:53] Prediction finished.
[2022-12-15 15:56:00] Prediction finished.
[2022-12-15 15:56:03] Prediction finished.
[2022-12-15 15:56:10] Prediction finished.
[2022-12-15 15:56:13] Prediction finished.
[2022-12-15 15:56:32] Prediction finished.
```

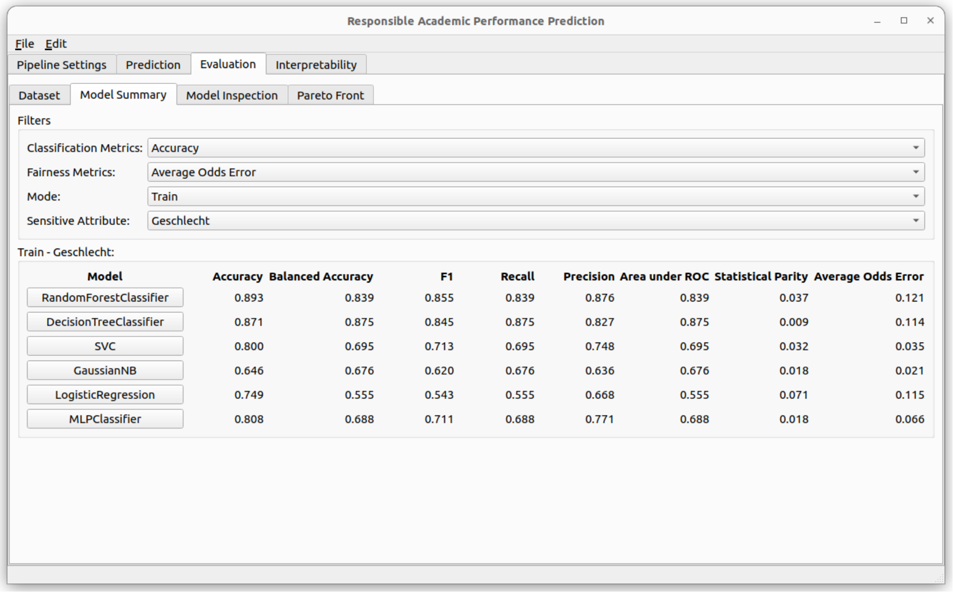
Fig. 3: RAPP’s Prediction Interface, 2022.

## 3.2 Decision Support System

The tool acts as a decision support system by providing the user statistical insights of the dataset as well as an extensive evaluation of the models’ performance and fairness. The models are automatically evaluated on the training and test data as they progress through the pipeline. The evaluation is displayed in the GUI, part of it is shown in Fig. 4, and is also generated as a LaTeX report, that is automatically compiled as a PDF file.

**Dataset.** The dataset tab contains a contingency table that displays the label $y \in \{0, 1\}$ and the sensitive attribute. This allows the user to comprehend the relationship between the sensitive attributes and the students’ performances.

**Performance Metrics.** As for stability reasons, the evaluation for the training data is always done with 5-fold cross-validation. The type of task that was selected beforehand determines the suitable metrics. Classification metrics included in the tool are *accuracy*, *balanced accuracy*, $F_1$, *recall*, *precision*, and *area under ROC*. As for regression metrics, the tool implements *mean absolute error*, *mean squared error*, *max error*, and $R^2$.



Filters

Classification Metrics: Accuracy

Fairness Metrics: Average Odds Error

Mode: Train

Sensitive Attribute: Geschlecht

Train - Geschlecht:

| Model | Accuracy | Balanced Accuracy | F1 | Recall | Precision | Area under ROC | Statistical Parity | Average Odds Error |
|---|---|---|---|---|---|---|---|---|
| RandomForestClassifier | 0.893 | 0.839 | 0.855 | 0.839 | 0.876 | 0.839 | 0.037 | 0.121 |
| DecisionTreeClassifier | 0.871 | 0.875 | 0.845 | 0.875 | 0.827 | 0.875 | 0.009 | 0.114 |
| SVC | 0.800 | 0.695 | 0.713 | 0.695 | 0.748 | 0.695 | 0.032 | 0.035 |
| GaussianNB | 0.646 | 0.676 | 0.620 | 0.676 | 0.636 | 0.676 | 0.018 | 0.021 |
| LogisticRegression | 0.749 | 0.555 | 0.543 | 0.555 | 0.668 | 0.555 | 0.071 | 0.115 |
| MLPClassifier | 0.808 | 0.688 | 0.711 | 0.688 | 0.771 | 0.688 | 0.018 | 0.066 |

Fig. 4: RAPP’s Decision Support System Interface, 2022.

**Fairness Notions.** The fairness of the models’ predictions is assessed with regard to the sensitive attributes in order to detect potential discrimination. Similarly to the performance metrics, the notions are determined by the task type. Classification tasks implement *statistical parity*, *predictive equality*, and *equality of opportunity* [DL19, BHN19]. While statistical parity is one of the most commonly used fairness notions, recent work suggests a focus on *equalized odds* (requires predictive equality as well as equality of opportunity) as the go-to notion for APP systems [DD22]. Accordingly, the tool integrates *average odds error* [Be18] which quantifies equalized odds. For regression tasks we use the *individual fairness* and *group fairness* notion as introduced by Berk et al. [Be17].

To measure fairness criteria in classification, we use the absolute difference of the outcomes between two groups. Generally, a lower value describes less discrimination. Because group sizes greater than two (non-binary genders, multiple nationalities) might occur in the dataset, we use the maximum value of the absolute differences between all group pairs [Ž17]. This measures the maximal discrimination a classifier has achieved between two groups.

**Pareto Front: Performance and Fairness Trade-Off.** Due to the existence of a performance and fairness trade-off [BFT12], the trade-off can be visually examined in order to select the best trained models to use for predictions. The Pareto-efficient models, i. e., models that optimize both a particular performance metric and fairness measure, can then be identified. The fairness tab includes scatter points of the selected models in a

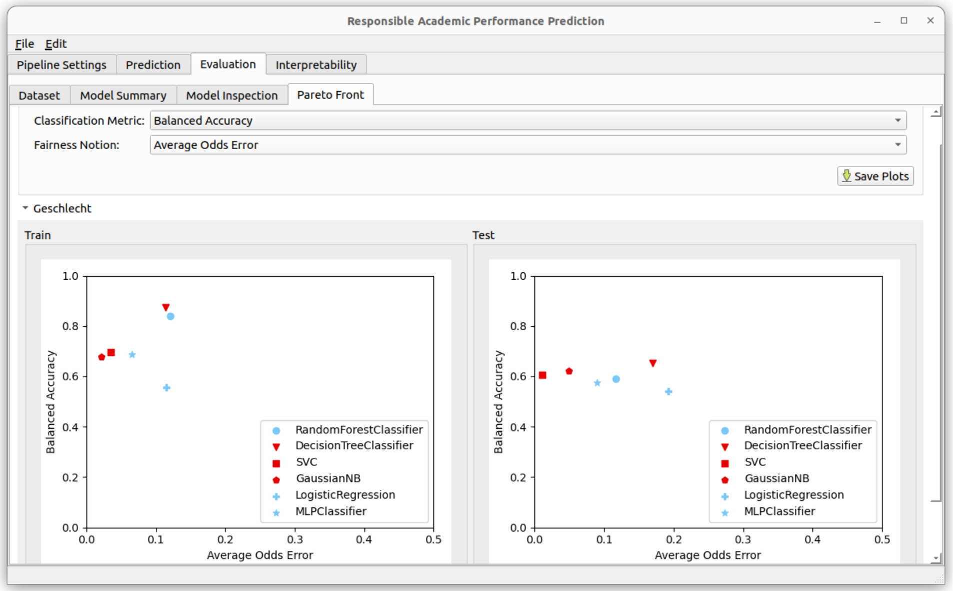

Classification Metric: Balanced Accuracy
Fairness Notion: Average Odds Error

Fig. 5: RAPP's Pareto Front Evaluation, 2022.

performance-fairness plot (see Fig. 5). The Pareto front, i. e., the set of all Pareto-efficient models [JS08], is shown in a different color to differentiate them from Pareto-dominated points. Pareto-efficient models are displayed in red whereas Pareto-dominated models are displayed in lightblue. This visualization limits the decision-making space for the user as only Pareto-efficient models are of interest. Because Pareto optimal solutions are first shown and the decision-maker selects her/his preferred model afterwards, this is a posteriori method in decision-making.

In classification we aim to maximize the performance metric whereas a maximization of the performance in regression corresponds to minimizing the error. For contextual conveniences, we maximize the negative error in regression to yield for the same plot.

## 4 Case Study

The RAPP tool is developed as part of a research project concerning itself with designing a socially responsible framework on how to approach APP in higher education. For this, we conducted a case study over data given to us by the Heinrich Heine University Düsseldorf. The case study was concerned mostly with probing of which prediction tasks show most-promising performances and to estimate possible algorithmic fairness problems. Hereby, the prediction tasks differed in their combination of input features as well as at-risk definition for prediction.

| | Prediction | | |
|---|---|---|---|
| Input | Dropout | MA Adm. | SDS |
| ECTP + Exam stats | 0.65 | 0.63 | 0.67 |
| Grades + Exam stats | 0.68 | 0.67 | 0.61 |
| Specific modules | 0.74 | 0.62 | 0.63 |

Tab. 1: Overview of exemplary training results over CS students in their first semester. Displaying the best performing balanced accuracy achieved by any trained model over combinations of selected feature sets and the prediction of student dropouts, master program admission (MA Adm.), and finishing in standard duration of study (SDS).

As we were interested in any combination of these predefined features and prediction goals, the RAPP tool was a great help in leveraging the combinatorial explosion problem into a manageable set of selectable templates, allowing us to quickly train and store models for each combination. Fig. 4 displays one such training result as reported within the tool, allowing comparison of the trained models over various performance and fairness measures. Tab. 1 shows exemplary results conducted with the RAPP tool over computer science (CS) students after their first semester.

## 5 Limitations and Future Work

Since the tool is still in development, new opportunities for future improvements present themselves constantly. These enhancements include changes to the tool's architecture, as well as making the prediction process more transparent to the user.

The tool as it currently is comes with SQL templates designed for our database in use. However, in order to allow other educational institutions to target at-risk students, the tool allows to write a different set of SQL templates and to load any SQLite database, making the tool essentially database agnostic. Still, this requires proficiency with writing SQL queries and modulating them into the templating engine, a skill that end users might not have. Here, the ease of use could be enhanced.

While allowing to inspect potentially exhibited discrimination by the trained models, it is not yet possible to train models with fairness-interventions in mind. In the future we would like to incorporate ways to train models with fairness-accounting measures such as pre-, in-, and postprocessing [DL19, Fr19, PS20]

## 6 Discussion

The tool helps in investigating which fairness constraints are met by any trained model and thus guides the user in their decision making of which model to employ, but by no means does the tool alone help achieving the overall goal.

Approaching RAPP includes to find a suitable definition for algorithmic fairness by involving both, the institute's stakeholders as well as the affected student body [KLM22]. While the notion of equalized odds seems to be a desirable fairness constraint [DD22], the student body appears to favor demographic parity [Ma20]. Further, the potential damage caused by misclassifications needs to be carefully considered. All these above points are not meant to be resolved by the RAPP tool but rather need to be part of the conceptualization when planning to employ such a system *before* actual employment of the system takes place. However, the RAPP tool helps to investigate whether potential concerns are dealt with appropriately by the trained models or not.

## 7 Conclusion

In this paper, we presented the RAPP tool for developing responsible academic performance prediction systems. The tool tackles two main tasks: designing, training, and analyzing different APP tasks, and acting as a decision support system for selecting the best suited models in a fairness-sensitive and socially responsible context.

For the setup of APP tasks, the concurrent design and direct comparison of different tasks, i. e., different input features and target labels, was a main objective as the definition of academic performances differ depending on the viewpoints of users, the student body, or the application context. In order to assist the user which model or models are socially responsible when being employed to target interventions at at-risk students, extensive performance and fairness metrics are included. The metrics are viewable in the GUI itself but are also automatically exported to a PDF file. Assessing fairness metrics and highlighting the Pareto front of classical performance metrics and achieved fairness parities guides the user in the decision-making process of finding the most suitable model for their desired task.

Overall, the tool provides an interface to non–machine learning engineers to train, evaluate, and employ models in the APP domain by providing a simplified ML pipeline configuration and highlighting crucial trade-offs of the model accuracy vs. fairness, rendering responsible APP systems a step more accessible and approachable to everyone.

## Acknowledgments

This work was supported by the Federal Ministry of Education and Research (BMBF) under Grand No. 16DHB4020.

## Bibliography

[AC19] Alonso, José M; Casalino, Gabriella: Explainable artificial intelligence for human-centric data analysis in virtual learning environments. In: International workshop on higher education learning methodologies and technologies online. Springer, pp. 125–138, 2019.

[AL19] Ahn, Yongsu; Lin, Yu-Ru: FairSight: Visual analytics for fairness in decision making. IEEE transactions on visualization and computer graphics, 26(1):1086–1095, 2019.

[Ba16] Badr, Ghada; Algobail, Afnan; Almutairi, Hanadi; Almutery, Manal: Predicting students' performance in university courses: a case study and tool in KSU mathematics department. Procedia Computer Science, 82:80–89, 2016.

[Be17] Berk, Richard; Heidari, Hoda; Jabbari, Shahin; Joseph, Matthew; Kearns, Michael; Morgenstern, Jamie; Neel, Seth; Roth, Aaron: A convex framework for fair regression. arXiv preprint arXiv:1706.02409, 2017.

[Be18] Bellamy, Rachel K. E.; Dey, Kuntal; Hind, Michael; Hoffman, Samuel C.; Houde, Stephanie; Kannan, Kalapriya; Lohia, Pranay; Martino, Jacquelyn; Mehta, Sameep; Mojsilovic, Aleksandra; Nagar, Seema; Ramamurthy, Karthikeyan Natesan; Richards, John T.; Saha, Diptikalyan; Sattigeri, Prasanna; Singh, Moninder; Varshney, Kush R.; Zhang, Yunfeng: AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. CoRR, abs/1810.01943, 2018.

[BFT12] Bertsimas, Dimitris; Farias, Vivek F; Trichakis, Nikolaos: On the efficiency-fairness trade-off. Management Science, 58(12):2234–2250, 2012.

[BHN19] Barocas, Solon; Hardt, Moritz; Narayanan, Arvind: Fairness and Machine Learning. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

[Bi20] Bird, Sarah; Dudík, Miro; Edgar, Richard; Horn, Brandon; Lutz, Roman; Milan, Vanessa; Sameki, Mehrnoosh; Wallach, Hanna; Walker, Kathleen: Fairlearn: A toolkit for assessing and improving fairness in AI. Microsoft, Tech. Rep. MSR-TR-2020-32, 2020.

[DD22] Dunkelau, Jannik; Duong, Manh Khoi: Towards Equalised Odds as Fairness Metric in Academic Performance Prediction. In: 2nd Workshop on Fairness, Accountability, and Transparency in Educational Data. July 2022.

[De13] Demšar, Janez; Curk, Tomaž; Erjavec, Aleš; Črt Gorup; Hočevar, Tomaž; Milutinovič, Mitar; Možina, Martin; Polajnar, Matija; Toplak, Marko; Starič, Anže; Štajdohar, Miha; Umek, Lan; Žagar, Lan; Žbontar, Jure; Žitnik, Marinka; Zupan, Blaž: Orange: Data Mining Toolbox in Python. Journal of Machine Learning Research, 14:2349–2353, 2013.

[DL19] Dunkelau, Jannik; Leuschel, Michael: Fairness-Aware Machine Learning: An Extensive Overview. Working paper, available at `https://www3.hhu.de/stups/downloads/pdf/fairness-survey.pdf`, October 2019.

[Fr19] Friedler, Sorelle A.; Scheidegger, Carlos; Venkatasubramanian, Suresh; Choudhary, Sonam; Hamilton, Evan P.; Roth, Derek: A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. ACM, jan 2019.

[Gr20] Grönberg, Niku; Knutas, Antti; Hynninen, Timo; Hujala, Maija: An online tool for analyzing written student feedback. In: Koli Calling'20: Proceedings of the 20th Koli Calling International Conference on Computing Education Research. pp. 1–2, 2020.

[Ha09] Hall, Mark; Frank, Eibe; Holmes, Geoffrey; Pfahringer, Bernhard; Reutemann, Peter; Witten, Ian H: The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1):10–18, 2009.

[HK16] Hofmann, Markus; Klinkenberg, Ralf: RapidMiner: Data mining use cases and business analytics applications. CRC Press, 2016.

[HR20] Hu, Qian; Rangwala, Huzefa: Towards Fair Educational Data Mining: A Case Study on Detecting At-Risk Students. International Educational Data Mining Society, 2020.

[JS08] Jin, Yaochu; Sendhoff, Bernhard: Pareto-Based Multiobjective Machine Learning: An Overview and Case Studies. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 38(3):397–415, 2008.

[KL20] Kizilcec, René F.; Lee, Hansol: Algorithmic Fairness in Education. arXiv, 2020.

[KLM22] Keller, Birte; Lünich, Marco; Marcinkowski, Frank: How Is Socially Responsible Academic Performance Prediction Possible? In: Strategy, Policy, Practice, and Governance for AI in Higher Education Institutions, pp. 126–155. IGI Global, may 2022.

[LMP16] Livieris, Ioannis; Mikropoulos, Tassos; Pintelas, Panagiotis: A decision support system for predicting students' performance. Themes in Science and Technology Education, 9(1):43–57, 2016.

[LMZ19] Loukina, Anastassia; Madnani, Nitin; Zechner, Klaus: The many dimensions of algorithmic fairness in educational applications. In: Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications. Association for Computational Linguistics, pp. 1–10, August 2019.

[LQN21] Le Quy, Tai; Ntoutsi, Eirini: Towards fair, explainable and actionable clustering for learning analytics. In: EDM. 2021.

[Ma20] Marcinkowski, Frank; Kieslich, Kimon; Starke, Christopher; Lünich, Marco: Implications of AI (un-)fairness in higher education admissions. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. ACM, jan 2020.

[PS20] Pessach, Dana; Shmueli, Erez: Algorithmic Fairness. volume abs/2001.09784, 2020.

[RD22] Ruf, Boris; Detyniecki, Marcin: A Tool Bundle for AI Fairness in Practice. In: CHI Conference on Human Factors in Computing Systems Extended Abstracts. pp. 1–3, 2022.

[Sa18] Saleiro, Pedro; Kuester, Benedict; Stevens, Abby; Anisfeld, Ari; Hinkson, Loren; London, Jesse; Ghani, Rayid: Aequitas: A Bias and Fairness Audit Toolkit. arXiv preprint arXiv:1811.05577, 2018.

[SA20] Sandee, Jan Jaap; Aivaloglou, Efthimia: Gitcanary: A tool for analyzing student contributions in group programming assignments. In: Koli Calling'20: Proceedings of the 20th Koli Calling International Conference on Computing Education Research. pp. 1–2, 2020.

[Ž17] Žliobaitė, Indrė: Measuring discrimination in algorithmic decision making. Data Mining and Knowledge Discovery, 31:1060–1089, 2017.