applied sciences

MDPI

Article

Feature Selection with Small Data Sets: Identifying Feature Importance for Predictive Classification of Return-to-Work Date after Knee Arthroplasty

Harald H. Rietdijk [1,*], Daniël O. Strijbos [2,3], Patricia Conde-Cespedes [4], Talko B. Dijkhuis [1], Hilbrand K. E. Oldenhuis [1] and Maria Trocan [4]

[1] Lectoraat Digital Transformation, Hanze University of Applied Sciences, 9747AS Groningen, The Netherlands; t.b.dijkhuis@pl.hanze.nl (T.B.D.); h.k.e.oldenhuis@pl.hanze.nl (H.K.E.O.)
[2] Department of Public and Occupational Health, University of Amsterdam, Amsterdam UMC, 1105AZ Amsterdam, The Netherlands; d.strijbos@nijsmellinghe.nl
[3] Department of Health Innovations, Nij Smellinghe Hospital Drachten, 9202NN Drachten, The Netherlands
[4] Institut Supérieur d'Électronique de Paris (ISEP), 75006 Paris, France; patricia.conde-cespedes@isep.fr (P.C.-C.); maria.trocan@isep.fr (M.T.)
* Correspondence: h.h.rietdijk@pl.hanze.nl; Tel.: +31-6-51-91-6262

Abstract: In recent decades, the number of cases of knee arthroplasty among people of working age has increased. The integrated clinical pathway 'back at work after surgery' is an initiative to reduce the possible cost of sick leave. The evaluation of this pathway, like many clinical studies, faces the challenge of small data sets with a relatively high number of features. In this study, we investigate the possibility of identifying features that are important in determining the duration of rehabilitation, expressed in the return-to-work period, by using feature selection tools. Several models are used to classify the patient's data into two classes, and the results are evaluated based on the accuracy and the quality of the ordering of the features, for which we introduce a ranking score. A selection of estimators are used in an optimization step, reorganizing the feature ranking. The results show that for some models, the proposed optimization results in a better ordering of the features. The ordering of the features is evaluated visually and identified by the ranking score. Furthermore, for all models, higher accuracy, with a maximum of 91%, is achieved by applying the optimization process. The features that are identified as relevant for the duration of the return-to-work period are discussed and provide input for further research.

Keywords: machine learning; feature selection; knee arthroplasty

Citation: Rietdijk, H.H.; Strijbos, D.O.; Conde-Cespedes, P.; Dijkhuis, T.B.; Oldenhuis, H.K.E.; Trocan, M. Feature Selection with Small Data Sets: Identifying Feature Importance for Predictive Classification of Return-to-Work Date after Knee Arthroplasty. *Appl. Sci.* **2024**, *14*, 9389. https://doi.org/10.3390/app14209389

Academic Editor: Tomasz Kocejko

Received: 20 September 2024
Revised: 7 October 2024
Accepted: 10 October 2024
Published: 15 October 2024

1. Introduction

In recent decades, the number of cases of knee arthroplasty (KA) among people of working age has increased and will continue to increase in the future [1]. In countries such as the United States of America and the United Kingdom, for example, the majority of KA cases in 2035 will consist of people of working age. For this group of patients, the cost of treatment comes not only from medical care but also from the additional costs of sick leave. Despite positive results in improving knee function and pain relief for this group of patients, the return-to-work (RTW) period remains significant [2,3]. In the Netherlands, the percentages of people who have not returned to work after, respectively, 6 and 12 months are 33% and 13%. Between 2015 and 2017, this represented a total sick leave cost of 29.6 million euros for the Dutch workforce due to knee osteoarthritis [4].

These figures show that optimizing RTW following KA is of paramount importance. To achieve optimized RTW, closer collaboration between medical and occupational care should be considered, providing an integrated clinical pathway from preoperative care to RTW. In 2021, Strijbos et al. [5] showed the feasibility of such an integrated clinical pathway

based on two pillars, namely, timely integration of medical and occupational care and increased participation of patients in the process. The integrated clinical pathway, called back at work after surgery (BAAS), forms the basis of a multicenter study started in 2022 [6], with the aim of proving the effectiveness and economic value of this pathway. For the patient, such collaboration should lead to better adapted physical therapy, more suitable personal goals, better occupational support, and more personal advice from the hospital.

An important step in the evaluation of the clinical pathway is to determine the features that influence the length of the RTW period. This can be performed using feature selection (FS). FS is an essential process in data mining and machine learning (ML). In ML applications, it is used to optimize performance and accuracy by reducing the amount of data that must be processed and removing irrelevant, redundant, and noisy features [7,8]. Furthermore, FS can be used to improve the interpretation of the generated models. By removing irrelevant characteristics, the true factors that influence the outcome become more evident [9]. Another important factor in the accuracy of ML applications is the number of instances in the data set. We know from clinical studies that a small sample size can result in imprecise estimates or failure to reach statistical significance. Similarly, in machine learning, the optimal amount of training data is crucial to effective performance [10], as insufficient data can lead to under- or over-fitting and affect the accuracy of predictions.

Since FS often relies on the evaluation of the predictions created by the ML models, the same concerns about accuracy are valid for FS methods. In many cases, especially in clinical research, it can be complicated or even impossible to collect robust numbers of samples due to the limited available time, the limited number of available cases, or both. The data we use in our study highlight these challenges, with a sample size of 109 and more than 90 features.

Research Goals

Using data collected in the BAAS study, our study aims to achieve the following goals: (1) Determine if we can achieve a meaningful ranking of the features and useful performance of the ML models using FS tooling. (2) Establish whether it is possible to improve the feature ranking and performance of the ML models by applying an optimization procedure. (3) Discuss the results of feature selection and optimization to investigate whether meaningful conclusions can be drawn about the importance of features for the RTW period. The method used to achieve these goals is illustrated in Figure 1.

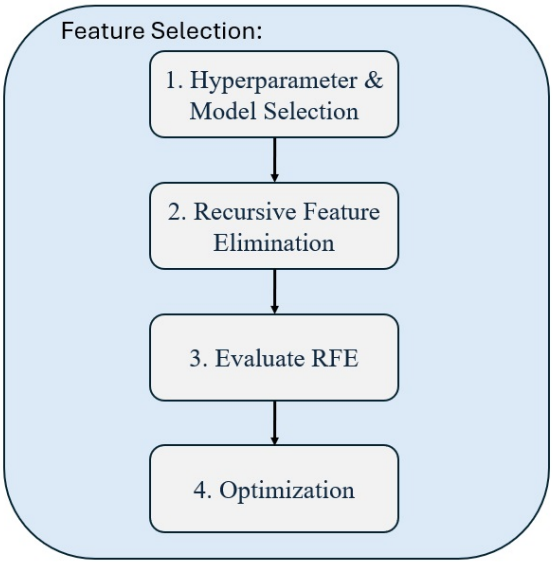

Figure 1. The method used to achieve the research goals.

The rest of the paper is organized as follows; In Section 2, we discuss the information that can be found in the data set, test the quality of the data, and describe some measures taken to prepare the data for further use. Section 2.3 describes the process that was followed to perform the feature selection. In Section 3, the results of the different steps described in Section 2.3 are presented. Finally, the conclusion and suggestions for further research are discussed in Section 4.

2. Materials and Methods

The BAAS study was carried out in two high-volume KA hospitals in the Netherlands. In the period between November 2021 and November 2023, 152 patients were recruited from the Elizabeth Tweesteden Ziekenhuis (ETZ) and Nij Smellinghe (NS), which perform approximately 700 and 350 KAs per year, respectively. In addition to the primary outcome given by the RTW, which is the first day back at work regardless of the number of hours worked, and the definitive RTW, which is the first day the patients have worked the number of hours as stated in their contract, large amounts of other data were collected for each patient. These data include, among others, demographic data such as age, length, sex, weight, clinical data such as the type and side of the operation, number of days in the hospital, number of days of sick leave before the operation, the results of several questionnaires and the ability and fitness test, but also mobility data collected with the use of a PAM 2.0 accelerometer. The Physical Activity Monitor (PAM) accelerometer is a wearable device that measures acceleration in both the horizontal and vertical directions and is used to monitor daily physical activity [11]. It also provides the option to give feedback through the Artis (Peercode B.V.) application.

Out of the total set of collected data, 44 different features were selected to be used in our study. The inclusion criteria for this selection were that the data have to be unambiguous numerical values, or it should be possible to represent the different values in ordinal encoding. Furthermore, data must be collected preoperatively or in the first six weeks after the operation. Finally, the information represented in the data cannot be derived from other selected features; for example, the *retirement date* was not used, as it is strongly related to the age of the patient. In the following section, a description of the selected features will be given. A complete list of all features and their type and value ranges can be found in Appendix A.

2.1. Selected Features

The features can be categorized into five different groups. The first group contains personal information such as demographics, professional information, and specifics of the operation. The second group provides information on the impact of the KA on day-to-day life and professional activities. The third group contains the results of several fitness and ability tests. The fourth group contains the results of several questionnaires that had to be completed at set moments during the process, and the final group consists of the four measures of daily activity obtained with the Pam accelerometer.

In the personal information group, there are twelve features, namely *age*, *gender*, *height*, *weight*, the *kind of operation* that was performed, which is a total KA (TKA) or a unicompartimental KA (UKA), and the *length of stay* in the hospital and the *hospital* in which the operation was performed. Further personal information covered the patient's professional activities: how many *hours per week* were meant to be spent at work, is the patient the *breadwinner*, and is the employment salaried or is the patient an entrepreneur. In addition, patients were asked if they thought that the problems were *caused by their job* and if it would be possible to *continue working without an operation*. The answers to these questions were translated to a five-point ordinal scale. Most of these features appear in the iMTA Productivity Cost Questionnaire (IPCQ) [12].

The second set, which contains information on the clinical picture, has six features: the *number of sick days* taken before the operation, the *number of problematic days at work*, which is also part of the IPCQ, and the *amount of work* that could be done on these days, the

number of problematic non-working days, the *percentage of work* that could be done in general, and the *disability rating* that was assigned. In this context, problematic indicates that the normal level of productivity or activity could not be achieved, whereas the features in the fourth set focus on the personal perception of the problems.

Several fitness and ability tests, which form a predicted work ability score, make up the third set of eight features. These values include the de Morton Mobility Index (*DEMMI*) [13] measuring static and dynamic balance, the maximum distance covered in a six-minute walk test (*6MWT*) [14], and the results of two exercises that require patients to move from sitting to standing without using their hands. The *30CRT* [15] counts the number of times the patient can do this in 30 s, and the *5XCRT* [16] is the number of seconds it takes the patient to do this 5 times. The last ability test is the floor-to-waist lift test (*LFTT*) [17]. If, for some reason, the patient is not capable of performing this test, the therapist will fill in a replacement score based on the evaluation of the patient's condition. Where possible, these values were measured before and six weeks after the operation.

The set of results of several questionnaires filled out by patients had a total of fourteen features, divided into seven results before and seven results obtained six weeks after the operation. The Work, Osteoarthritis, and Joint Replacement Questionnaire (*WORQ*) [18] gives a measure of how patients perceive difficulty in performing work-related activities. The Central Sensitization Inventory (*CSI*) [19] scores pain-sensitization-related symptoms. And finally, the Knee Injury and Osteoarthritis Outcome Score(*KOOS*) [20] quantifies the KA-related experience. The results of this questionnaire were divided into five domains: symptoms, pain, activities, participation in sports, and quality of life.

The last set was collected using the accelerometer and contained four features. For these features, pre- and post-operation activity data were collected using the PAM 2.0 accelerometer. This wearable device records the movement of the patient and produces activity scores every 15 min. The value of this score is based on the amount and intensity of the movement performed. The intensity is calculated using the acceleration and direction of the action. Using these data, the *average-inactivity-moments* before the operation and in the six weeks following the operation were calculated. Each inactivity moment is a period of 30 min in which the activity does not rise above a certain threshold, which means that one hour without activity counts for two inactivity moments. The second value calculated using these data is the *average-number-of-steps-per-day* for the same periods.

2.2. Data Preprocessing

As mentioned in Section 1, 152 patients were recruited in the BAAS study. Of these 152 patients, the collected data from 42 of these patients had to be excluded from the data set because the operation did not take place, the patients experienced additional complications, such as COVID or infections, that affected the duration of the rehabilitation period, or due to other reasons, such as changing employment during the observation period, resulting in an initial data set of 110 rows.

Every row has 44 columns containing the features described in Section 2.1. Additionally, each row has a target variable, namely the definite return-to-work value, which is the number of weeks between the date of the operation and the full RTW. For classification, the rows were split into two classes based on whether or not this value was above or below the average RTW value.

In some cases, not all of the data were complete in the input files. In general, this was due to the fact that the default value was not filled in. In these cases, we assumed the default value was meant to be filled in and used this value. In the few remaining cases, the average value of the feature was imputed. For six patients, there were no accelerometer activity data for the period before or after the operation. This meant that the average number of inactivity moments and the number of steps before or after the operation were missing. Here, the data were imputed by calculating the average value of the missing feature for the seven patients with the value nearest to the average that was present, i.e., the corresponding average before or after.

In all steps of the process, the same split of 30%/70% was used for the test and training sets. Due to the limited number of samples in the data set, there is a risk of an uneven split between the training and the test sets. To verify that this did not occur, all features were passed through a Kruskal–Wallis test and a Mann–Whitney *U* test. Neither test indicated a significant difference between the test and the training set, indicating that the values of the different features for both sets were evenly spread over all possible values. The results of these tests per feature are given in Appendix A.

As a final step in the data preparation, a visual evaluation of the quality of the feature data and the distribution of the test and training data was performed. This evaluation consisted of generating sets of scatterplots of the training and test sets of the values of all features against the RTW values and the corresponding boxplots of the feature values, for example, Figures 2 and 3. Based on this evaluation, one outlier, with an RTW value of 61.7 was excluded from the data set, resulting in a final data set of 109 rows. This choice is confirmed by the fact that the Z value for this value in the Grubbs' test was 5.1, which is significantly higher than the critical value of 3.4 for this sample size.

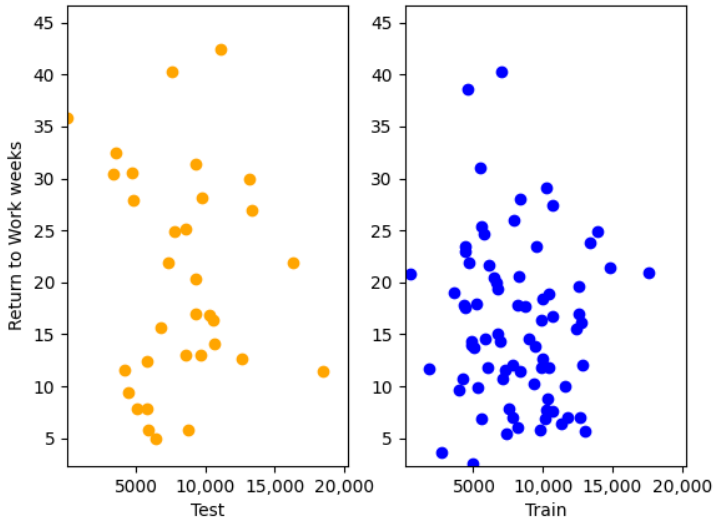

Figure 2. Scatterplot of the average number of steps post operation against weeks to definite return to work, for both the test set (**left**) and the training set (**right**).

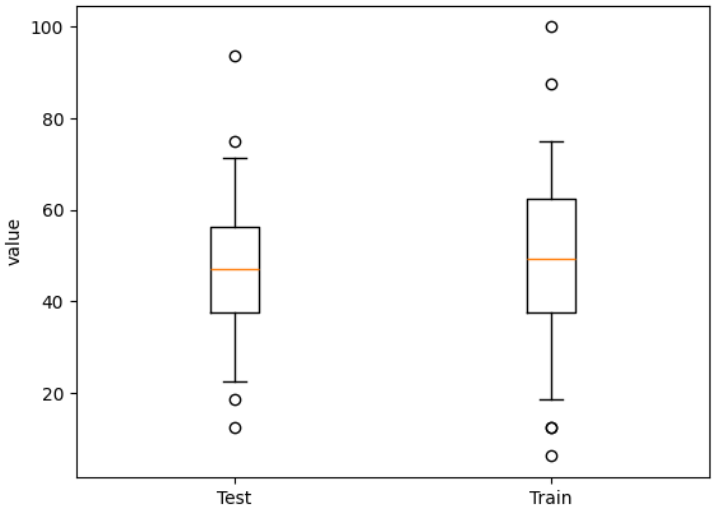

Figure 3. Boxplot of the values of KOOS questionnaire results for the quality of life domain at 6 weeks after the operation, for both the test set (**left**) and the training set (**right**).

2.3. Method

The objective of our study is to arrive at a meaningful prediction of the features of the given data set that are relevant to predict the definitive RTW. For this, it is important to carefully look at the characteristics of the data set. The challenge with the given data set lies in the fact that there is a limited number of samples compared to the number of features that are considered, 109 and 44, respectively. This ratio can easily lead to over-fitting of the models used [21], so special care should be taken when selecting the appropriate features and settings of the hyperparameters for the estimators used. To accommodate these considerations, the following steps were taken to arrive at a selection of the most important features.

The first step is to determine which machine learning models should be used to perform the evaluation. All programming in this study was performed in Python 3.12, using the Scikit-Learn 1.4.0 tooling. This toolkit provides eleven algorithms for performing feature selection. In this study, we are interested in finding the set of features that gives optimal results for the prediction model and the corresponding ranking of the features. Because performance will be evaluated on the test set, it is necessary to be able to set the number of selected features. Taking these constraints and also the fact that we are working with a small data set into consideration reduces the choice to two options, namely Recursive Feature Elimination (RFE) and Select From Model (SFM), both using similar methods to evaluate which features to select. Of these algorithms, we will use RFE, because the implementation in Scikit-Learn provides good information about the feature ranking and can more easily be tuned to vary the number of selected features.

Since RFE is used, there is a restriction on the possible estimators that can be used. Suitable estimators must have the attribute $coef\_$ or $feature\_importances\_$ to provide information on the possible ranking of the features used. The $coef\_$ attribute gives the weight or coefficient matrix for models where this is possible, whereas the $feature\_importances\_$ attribute gives the relative importance of each feature.

In Scikit-Learn, a total of 54 regressors and 40 classifiers, generally called estimators, are available, grouped into several types, such as linear models, tree models, and ensemble methods. No assumptions were made about the performance of the different types of estimator models, so only attribute restrictions were taken into account in the selection process. Although the visual evaluation of the distribution of the different features compared to the RTW, carried out by inspecting the scatterplots, indicated little correlation, both the regressor and classifier estimators were considered.

Taking these restrictions into consideration, the following classifiers were used: From the discriminant analysis group, the LinearDiscriminantAnalysis; from the ensemble models, the AdaBoostClassifier, the RandomForestClassifier, the ExtraTreesClassifier, and the GradientBoostingClassifier; from the linear models, the LogisticRegression, and the SGDClassifier; from the support vector machines, the LinearSVC; and finally from the tree models, the DecisionTreeClassifier.

With these estimators, the following process was executed to arrive at a ranking of the different features in the data set.

1. Select the base estimator hyperparameter settings to be used in the process.
2. Use Recursive Feature Elimination to order the features based on their importance.
3. Evaluate the results of the RFE to determine if further optimizations are possible.
4. Perform this optimization on one or several selected estimators.

2.4. Determine the Hyperparameter Settings

The RFE selector from the Scikit-Learn toolkit takes a given estimator and recursively prunes the set of features to the required number of features until the requested number of features is selected. It does so by eliminating the least important features from the data set provided [22]. Part of the optimization process of estimator selection is to determine the specific settings of the estimator used, such as, for example, the maximum tree depth

for tree models. However, the RFE selector accepts only one specific estimator at a time, limiting the hyperparameter setting to one specific choice.

To determine a suitable set of hyperparameters for each estimator, the first step in the ranking process is to select the most optimal setting for each estimator using the GridSearchCV tooling from Scikit-Learn. The selected estimators are submitted to this tool with a selection of options for the most relevant parameters. The optimal settings of the hyperparameters are then determined by a cross-validated grid search over these options. In this step, the full set of features is used.

In all steps, for classification, the training data were split into two classes, where the first class contained samples where the RTW was less than average, and the second class contained those with a higher-than-average RTW.

2.5. Recursive Feature Elimination

The resulting best estimator is used in the Recursive Feature Elimination selector. The RFE selector accepts a given estimator and a data set with a given number of features. It then selects the n most important features, where n is a parameter to be supplied to the selector. The submitted features are ranked, where all selected features have rank 1 and the remaining features have a ranking higher than n.

In theory, the RFE process should rank the features in such a way that the explanatory variables are ranked first and the confounding variables last. This implies that at a certain number of selected features, a maximum for the performance metrics will be reached, after which these metrics will decrease.

Given that a data set was used with relatively few samples and many features, it is important to be able to check for over-fitting and evaluate the performance of the estimators on all subsets of the test data, where the subsets consist of all test samples but only the selected features. In this process, this was carried out by removing features in a step-by-step manner, running the RFE to select all but one of the supplied features. After each run, the resulting quality of the fitted model in the test set was stored, containing only the selected features. For the classifiers, the accuracy, F1, precision and recall scores are stored. These results can then be used to find the feature selection that gives the best performance for the given model. Over-fitting can be detected when both the results of the training set and the test set remain the same after adding more features. In general, in these cases, the training set has a 100% accuracy, whereas the test set scores lower, see, for example, Figure 4.

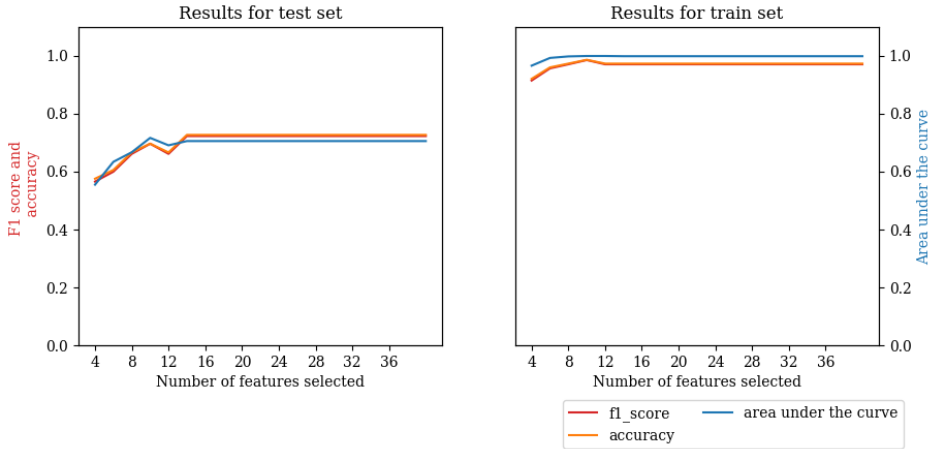

Figure 4. Results for the test (**left**) and training set (**right**) of the AdaBoost Classifier, clearly showing the effect of over-fitting after 14 features.

2.6. Evaluation of the RFE Results and Optimization

Because different models use different classification mechanisms, we should expect variation in the ranking of the features for the selected estimators. Therefore, to be able to

provide a balanced analysis of the features that have a predictive value on the RTW, it is important to compare the RFE results of multiple estimators.

Furthermore, since the data set used is limited, which could possibly lead to over-fitting, the ranking of the RFE selector should be verified. Ideally, the resulting order of the features after the RFE selection should be such that the features containing the most information on the RTW are ranked first and possible confounding features are ranked last. This means that when the number of selected features is plotted against the performance metrics of the resulting estimator, it should be shaped like a convex upward function. To eliminate the influence of over- or under-fitting, this evaluation will be performed using the test set.

One final aspect that should be taken into account is the fact that the estimators used for the RFE selection process were selected using the complete feature set. Reducing the number of features might have consequences for the optimal settings of the hyperparameters.

Based on these three observations, in the final step of the process, the proposed procedure to achieve optimization is to reorganize the features, starting with their ranking as calculated by the RFE selector, adjusted up or down depending on the observed change in accuracy of the estimator after adding this feature to the feature set. This process will be repeated several times, recalculating the accuracy for the estimator on the test set, for each subset of features, until the maximum performance and the corresponding selection of features are found. Using different settings for the hyperparameters for the selected estimator, this process will also be used to verify the correctness of the hyperparameter settings.

3. Results

The results can be divided into three parts. First of all, the data were used to select the optimal hyperparameter setting using the GridsearchCV tool. The resulting settings were used in the following steps of the research. These results are not presented here because they have little impact on the evaluation. The second set of results consists of the feature ranking that was obtained by applying the RFE selection using the selected estimators. The results of this process were used for the final part, the optimization of the RFE selection.

3.1. Classification

In Table 1, the results of the RFE selection process are presented. In the second column, the number of features for which the estimator achieves the highest accuracy is given. The next four columns give the accuracy, F1 score, precision, and recall on the test set, for the estimator using the selected number of features. The number in brackets in the accuracy column is the accuracy of the estimator that was generated through the grid search function, using the full feature set.

Table 1. Classifiers' accuracy, F1, precision and recall scores.

Classifier	Features	Accuracy [1]	F1 Score	Precision	Recall
Linear SVC	12	0.85 (0.76)	0.85	0.86	0.85
Logistic Regression	15	0.82 (0.73)	0.82	0.82	0.82
Linear Discriminant Analysis	28	0.79 (0.73)	0.79	0.79	0.79
Gradient Boosting Classifier	10	0.79 (0.70)	0.79	0.79	0.79
Ada Boost Classifier	14	0.73 (0.73)	0.72	0.73	0.72
Random Forest Classifier	6	0.73 (0.61)	0.72	0.75	0.72
SGD Classifier	6	0.73 (0.55)	0.73	0.73	0.73
Decision Tree Classifier	12	0.67 (0.70)	0.67	0.77	0.69

[1] The accuracy of the base estimator using the full feature set is given in brackets.

The results show that, in general, the RFE selection process improves the accuracy of the estimator, with the exception of the Decision Tree classifier. This last result, combined with the observation that for both the AdaBoost and Random Forest classifier, two ensemble classifiers that take the Decision Tree estimator as a base estimator, show little improvement,

indicate that for tree models, the RFE selection does not necessarily order the features in the right order for the given data set. This is also illustrated in the graphs in Figure 5, showing the accuracy, F1, and area under the curve scores for the Decision Tree Classifier on the test and training set. The fluctuation in accuracy with each added feature shows that the order in which the features are added does not result in the expected convex upward function. The fact that the recall is much lower than the precision indicates that the underlying algorithm focuses on the retrieved elements, more than the relevant elements.

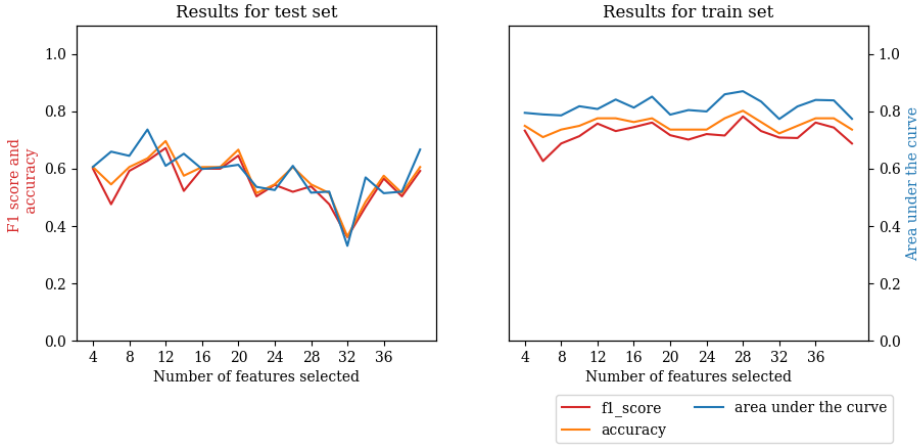

Figure 5. Results for the test (**left**) and training set (**right**) of the Decision Tree Classifier, showing the performance metrics, f1-score, accuracy, and area under the curve, plotted against the number of features used for the model.

As illustrated in the plot in Figure 6 and the improved accuracy results, RFE selection applied on the other models has better performance. Based on this assumption, which will be verified in the optimization process, we can conclude that the resulting feature ranking gives us information on the importance of the features with respect to the RTW. There are four classifiers that have an acceptable accuracy score for classification. In Table 2, the selected features and the ranking of these features are given for features that appear in more than one model. The results are ordered according to the number of appearances and the average ranking. The resulting scores between 79% and 85% are well above the results that were expected after visual evaluation of the scatterplots, which showed very little correlation.

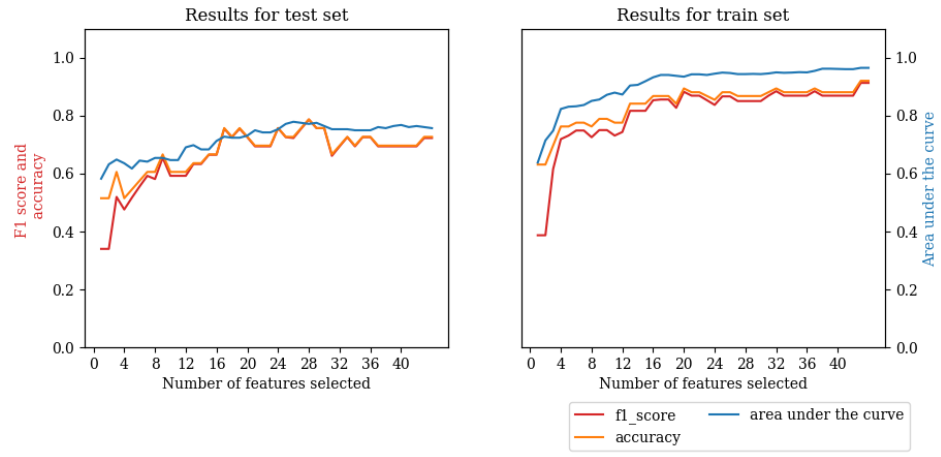

Figure 6. Results for the test (**left**) and training set (**right**) of the Linear Discriminant Analysis Classifier, showing the performance metrics, f1-score, accuracy, and area under the curve, plotted against the number of features used for the model.

Table 2. Selected features for Linear Support Vector Classification (LSVC), Gradient Boost Classification (GBC), Logistic Regressor Classification (LR), and Linear Discriminant Analysis Classification (LDA).

Feature (Average Ranking)	LSVC	GBC	LR	LDA
operation-type (1)	1	1	1	1
5XCRT-T0 (8)	6	3	9	14
prob-unpaid-days (8.5)	5	9	7	13
CSI-T1 (8.8)	9	2	12	12
days-in-hospital (2.7)	2		3	3
problems-caused-work (3.7)	3		4	4
nr-of-days-off-work (8.7)	7		11	8
average-inactivity-post (10.7)	12	10		10
CSI-T0 (13.3)	10		13	17
WORQ-T1 (13.7)	11		14	16
type-of-work (2)			2	2
gender (5)			5	5
work-capable-wo-oper (5)	4		6	
breadwinner (7)	8			6
nr-prob-days (14)			10	18
height (14.5)		6		23
hours-per-week (14.5)		5		24
30CRT-T1 (15)			15	15
preop-capa (18)			8	28

3.2. Optimization

The results presented in Section 3.1 indicate that the RFE selection performs worse when applied to a tree-based model. In Section 2.3, we also discussed the risk of over-fitting. The optimization process described in Section 2.6 aims to compensate for both of these problems. Figure 4 shows the result of the AdaBoost Classifier after the RFE selection and the GridSearchCV step, clearly indicating the occurrence of over-fitting. Because the AdaBoost Classifier is used with a Decision Tree base estimator, this model is the best choice to use in the full optimization process, in which both hyperparameters are varied and feature ranking is adjusted according to the evolution of the model performance when features are added.

As can be seen in Figure 6, and on the left-hand side of the figures in Appendix C, the accuracy results on the test set after the RFE selection process still show significant fluctuations before the optimum number of features is reached. This suggests that improvements can be achieved by applying the optimization process to these classifiers as well. For these models, we choose not to apply the hyperparameter part of the optimization process, since over-fitting appears to be less of an issue.

3.2.1. Full Optimization of the AdaBoost Classifier

In the AdaBoost Classifier optimization process, repeated RFE selection and reordering of the features is applied to a total of 45 different configurations. This number is created by taking nine possible Decision Tree estimators, used as the base estimator for the AdaBoost Classifier, where the *n_estimator* parameter can have any of the following values, 5, 10, 15, 20, or 25. The Decision Tree estimators used were created taking all possible combinations for the parameters *max_depth* and *min_samples_split* from the possible choices, 1, 2, or 3, and 0.1, 0.2, or 0.3.

Table 3 shows a selection of the results of the full optimization. A complete overview of the results can be found in Appendix B. As can be seen in this table, the improvement lies between 3% and 24%, with an average improvement of 14%. The maximum accuracy achieved is 91%, an improvement of 18% compared to the original performance of the AdaBoost Classifier. Figure 7 shows that even when over-fitting occurs during the RFE selection, the optimization generates a better fit. Figure 8, in which the results of the estimator

that has the highest accuracy are plotted, shows that, despite a significant improvement in accuracy, the resulting ordering still does not show the expected convex upward function which should be the result of ordering the features based on their importance related to RTW.

Table 3. Comparison of the results of Recursive Feature elimination and the optimization algorithm.

Classifier	RFE Result	Optimization Result	Improvement
ADA15.3.0.1	0.73	0.91	0.18
ADA20.2.0.1	0.64	0.88	0.24
ADA10.3.0.3	0.76	0.88	0.12
ADA10.2.0.1	0.67	0.88	0.21
ADA20.1.0.1	0.76	0.88	0.12
ADA20.2.0.2	0.64	0.88	0.24
ADA5.2.0.1	0.61	0.85	0.24
ADA10.1.0.3	0.61	0.85	0.24
ADA15.2.0.2	0.61	0.85	0.24
ADA20.3.0.1	0.79	0.82	0.03
ADA5.1.0.1	0.61	0.76	0.15

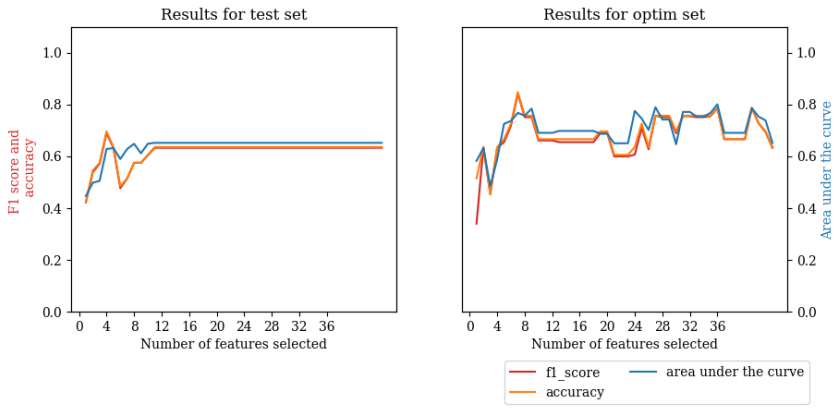

Figure 7. Results for the test set before (**left**) and after optimization (**right**) for the AdaBoost Classifier, showing that over-fitting is reduced during the optimization.

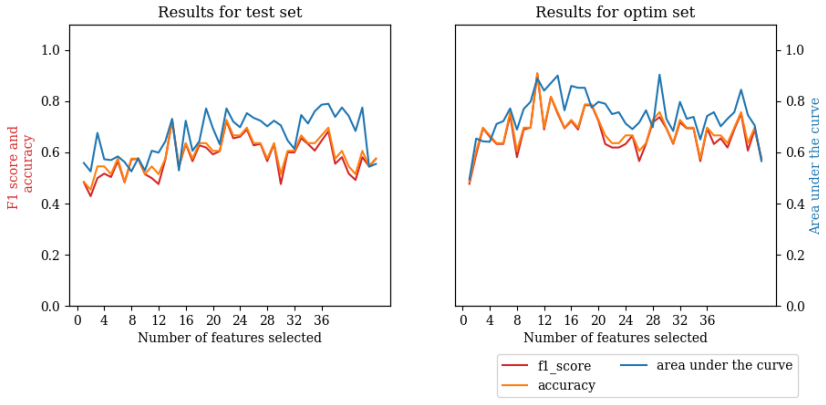

Figure 8. Results for the test set of best performing AdaBoost Classifier after RFE (**left**) and optimization (**right**).

3.2.2. Optimization of the Best Performing Classifiers

Table 4 shows the results of performing the reorganizing part of the optimization process on the four models that had the highest accuracy score. Looking at Figures 9 and A1–A3, we can see that, despite the fact that for all classifiers the maximum accuracy is

higher, only for the LR and LDA classifiers, the ordering of the features has improved by applying the optimization to the RFE results.

Table 4. Results for the optimization of the best-performing classifiers.

Classifier	RFE Result	Optimization Result	Improvement	Nr. of Features [1]
Linear SVC	0.85	0.88	0.03	37 (12)
Gradient Boosting Classifier	0.79	0.88	0.09	6 (10)
Logistic Regression	0.82	0.88	0.06	23 (15)
Linear Discriminant Analysis	0.79	0.85	0.06	16 (28)

[1] The original number of features for the optimal accuracy is given in brackets.

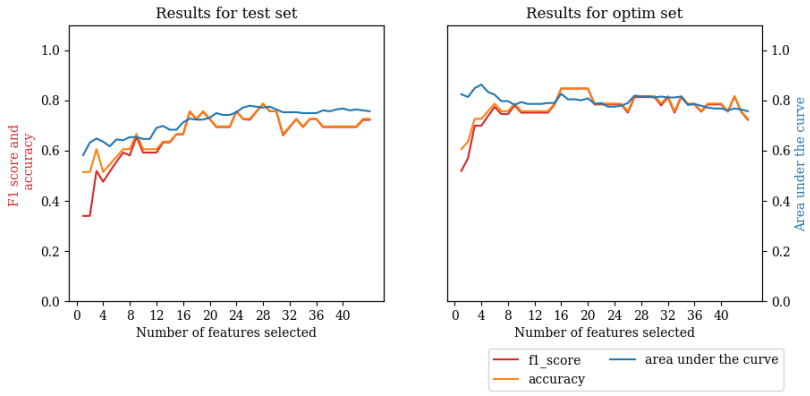

Figure 9. Results for the test set before (**left**) and after optimization (**right**) for the Linear Discriminant Analysis Classifier, showing a better ordering of features as a result of the optimization.

In the optimization process used here, we evaluate the improvement based on the maximum accuracy score obtained by the ordering, resulting ultimately in models that achieve higher scoring, but not necessarily better ordering. The evaluation of the resulting ordering is then performed based on visual interpretation of the plots such as Figure 9. To be able to focus more on the resulting ordering, we investigated the possibility of assigning a score to the outcome of a selection process. As a first step, we defined the following value:

$$S_{ordering} = A_m^2 \times (1 - \sum_{i=1}^{N} P_i \times D_i \times (a_{i-1} - a_i)), \tag{1}$$

where

$$A_m = \text{maximum accuracy of the model with this ordering,} \tag{2}$$

$$N = \text{number of features,} \tag{3}$$

$$P_i = \begin{cases} 0 \text{ when } a_{i-1} < a_i \text{ and i} < F_m \\ 0 \text{ when } a_{i-1} > a_i \text{ and i} > F_m \\ 1 \text{ otherwise} \end{cases} \tag{4}$$

$$F_m = \text{Number of features used in highest scoring model} \tag{5}$$

$$D_i = \frac{|i - F_m|}{N - F_m} \tag{6}$$

$$a_i = \text{accuracy score of the model using the first i features,} \tag{7}$$

$$a_0 = \frac{1}{\text{number of classes}}, \tag{8}$$

This score has a maximum of 1, which is obtained when the model has perfect accuracy and all confounding variables are ranked last. The preliminary results, given in Table 5,

indicate that the values of the $S_{ordering}$ score confirm the conclusions drawn by visual evaluation of the graphs. This suggests that this value can be a useful starting point for further research on optimizing feature rankings.

Table 5. Results for $S_{ordering}$.

Classifier	$S_{ordering}$ after RFE	$S_{ordering}$ after Optimization
Linear SVC	0.60	0.19
Logistic Regression	0.46	0.60
Linear Discriminant Analysis	0.37	0.57
Gradient Boosting Classifier	0.53	0.35
AdaBoost Classifier	0.53	n/a
Random Forest Classifier	0.31	n/a
Decision Tree Classifier	0.28	n/a

4. Discussion

The results presented in Section 3 show that using the RFE selector can give some useful insights in the feature importance for a data set with limited sample numbers such as the one used in this study. Comparing the results of the fitted estimator in the test set to that of the training set, but also simply the development of the accuracy results of the estimator with each added feature, showed that for tree-based models, the RFE selection process is not performed optimally (see Figure 5) when looking at the ordering of features. Linear classification models, on the other hand, display an accuracy profile that is more like the expected convex upward function.

Furthermore, although the proposed optimization process drastically improved the performance of the AdaBoost Classifier, the resulting graph showed that the order of the features was still not optimal. For the LDA, LR, LSVC and GBC, we saw that despite improved performance, the optimization only resulted in more meaningful ordering for the LDA and LR classifiers.

Taking these observations into consideration, a discussion of the feature importance should be based on the results obtained from the LSVC, GBC after the RFE selection, and from the LR and LDA estimators after the optimization. The resulting features are listed in Table 6 in the same order as in Table 2, with the original ranking of the LR and LDA models given in brackets.

Looking at the features that appear in these models, it is clear that the most important feature is the type of operation. Together with the fact that the length of stay in the hospital (*days-in-hospital*) ranks fifth, this shows that the impact of the operation is very determinant for the length of the RTW period. Logically, total KA has more impact than unicompartimental KA, and the length of stay in most cases depends on the recovery of the operation.

A second interesting point is the fact that two of the higher ranking features reflect personal experiences of the rehabilitation process, namely the *CSI-T1* and *WORQ-T1* features. This indicates the importance of the mental factor in the duration of the RTW period. However, it does not provide any information on whether the experience of the patients is the cause of the longer rehabilitation process, or if actual slow recovery is the cause of the patient's experience. This should be investigated by looking at individual results and, for example, the correlation with the values of *prob-unpaid-days*, *problems-caused-work* and *type-of-work*, which are features that could provide more information on the patient's pain experience in relation to work.

One feature that is of particular interest is the average inactivity post-operation. As can be seen from the positive coefficient, a higher number of inactivity moments in the first six weeks leads to a higher RTW value. This is a feature that could provide input for coaching during the rehabilitation process [23]. However, more research is needed to determine what causes the higher number; this could, for example, be caused by an overload of movements before the inactivity moments.

Table 6. Selected features for Linear Support Vector Classification (LSVC), Gradient Boost Classification (GBC),Optimized Logistic Regressor Classification(LR), and Optimized Linear Discriminant Analysis Classification (LDA). Original ranking after RFE for LR and LDA is given in brackets.

Feature (Average Ranking)	LSVC	GBC	LR	LDA
operation-type (1)	1	1	1 (1)	1 (1)
5XCRT-T0 (7)	6	3	8 (9)	11 (14)
CSI-T1 (8.8)	9	2	12 (12)	12 (12)
prob-unpaid-days (9.8)	5	9	10 (7)	15 (13)
days-in-hospital (2.7)	2		3 (3)	3 (3)
problems-caused-work (4)	3		5 (4)	4 (4)
nr-of-days-off-work (9.7)	7		14 (11)	8 (8)
average-inactivity-post (10.7)	12	10		10 (10)
WORQ-T1 (15)	11		18 (14)	16 (16)
type-of-work (2)			2 (2)	2 (2)
gender (4.5)			4 (5)	5 (5)
work-capable-wo-oper (5.5)	4		7 (6)	
breadwinner (7.5)	8			7 (6)
CSI-T0 (10.5)	10		11 (13)	(17)
hours-per-week (12)		5	19	(24)
30CRT-T1 (13.5)			13 (15)	14 (15)
30CRT-T0 (17)			21	13 (11)
KOOS-A-T0(22)			22	
KOOS-Q-T1(23)			23	

The final feature on the list of highest ranking features to be discussed is the *5XCRT-T0* feature. This value reflects the patient's preoperative ability to perform five sit-and-stand movements without using their hands. As such, it probably has a high correlation with the type of operation, which would explain the high ranking. This also needs to be investigated further.

When we compare the resulting ranking and the maximum accuracy of the optimization process with the original ranking after the RFE step, we notice that the order achieved in the optimization process improves the order of the rankings. This can be seen both visually and by comparing the $S_{ordering}$ value. However, since most of the high ranking features in the RFE process remain high ranking features after optimization, but in a slightly different order, we can conclude that some features may seem to be confounding but in combination with other features, are explanatory features. The improved maximum accuracy is achieved in the reordering of the low-ranking features. For the LR classifier, the two characteristics that were pushed forward and gave the final contribution to the highest score are given in Table 6. These two features underline the second point of this discussion.

5. Conclusions

In this paper, we present the following results:

- Dropping features using the RFE process improves the performance metrics.
- Tree-based models have poor results with small data sets and provide little information on feature importance.
- Optimization through reordering of features improves accuracy for most classifiers.
- The ordering of the features for the LR and LDA classifiers improves after optimization, also highlighting possible related features.
- Comparison of the results of the optimized LR and LDA models and the LSVC and GBC models after RFE provides meaningful information on feature importance and useful insights for further investigation.

In summary, we can say that although the correlation graphs indicate that there is not enough information in the given data set to produce accurate regression models, using feature selection methods and comparing the results of several estimators can provide relevant insight in feature ranking for a data set with limited samples. Furthermore, the given optimization process almost always results in higher accuracy, which is useful when models are used for further predictive classification. However, when the focus lies on obtaining an understanding of which features influence the outcome of the classification, we see that feature selection, with or without the optimization, should be used with care.

Introducing a scoring value, such as the proposed $S_{ordering}$, could prove to be very helpful in these cases.

Author Contributions: Conceptualization, H.H.R., P.C.-C. and M.T.; methodology, H.H.R. and P.C.-C.; software, H.H.R.; validation, D.O.S., P.C.-C., T.B.D., H.K.E.O. and M.T.; formal analysis, H.H.R., T.B.D. and P.C.-C.; investigation, H.H.R.; resources, D.O.S.; data curation, H.H.R. and D.O.S.; writing—original draft preparation, H.H.R.; writing—review and editing, D.O.S., P.C, T.B.D., H.K.E.O. and M.T.; visualization, H.H.R.; supervision, P.C.-C., H.K.E.O. and M.T.; project administration, H.K.E.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The BAAS study has been approved by the Medical Ethical Committee of the Amsterdam UMC, location AMC (reference ID: W21_454 # 21.504) and the Medical Ethical Committees of the local hospitals Nij Smellinghe (NS, reference ID: 19888) and Elizabeth Tweesteden Ziekenhuis (ETZ, reference ID: L1429.2021). The study is in line with the Medical Treatment Agreement Act (in Dutch: Wet Geneeskundige Behandelovereenkomst, WGBO), the Data Protection Act (in Dutch: Wet bescherming persoonsgegevens, WBP, per 28 May 2018 the EU General Data Protection Regulation) and the Code of Conduct for Responsible Use (FEDERA).

Informed Consent Statement: Informed consent has been obtained from all subjects and/or their legal guardian(s) before participating.

Data Availability Statement: The data sets used and/or analyzed during the current study are available from the corresponding author on reasonable request. The data involve patient information for which informed consent has been obtained. This consent was given under specific conditions to protect patient privacy. Despite the data being coded and our best efforts to maintain anonymity, there remains a small possibility that the data could be traced back to individual patients. Therefore, we cannot fully guarantee anonymity if the data were to be published publicly. Also, this dataset is part of a larger ongoing study, with additional publications planned in the future. Consequently, it is not feasible to make the data publicly accessible at this time. However, the data can be shared upon request for purposes such as internal quality control, provided that appropriate privacy and data security measures are in place.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

TKA	Total knee arthroplasty
UKA	Unicompartimental knee arthroplasty
RTW	Return to work
BAAS	Back at work after surgery
ML	Machine learning
FS	Feature selection
PAM	Physical Activity <onitor
IPCQ	iMTA Productivity Cost Questionnaire
DEMMI	de Morton Mobility Index
WORQ	Work, Osteoarthritis and Joint Replacement Questionnaire
CSI	Central Sensitization Inventory
KOOS	Knee Injury and Osteoarthritis Outcome Score
RFE	Recursive Feature Elimination
SFM	Select From Model
LSVC	Linear Support Vector Classification
GBC	Gradient Boost Classification
LR	Logistic Regressor Classification
LDA	Linear Discriminant Analysis Classification
ADA	AdaBoost Classification

Appendix A. List of Features

Table A1. Demographics and personal features.

Feature	Description	Mean	Range	Standard Deviation	Coefficient of Variation	Kruskal–Wallis *p*-Value
age	The age of the patient	58.49	47–65	4.2	0.07	0.7
height	The height of the patient	176.43	157–203	8.5	0.05	0.53
weight	The weight of the patient	90.22	64–122	12.96	0.14	0.31
gender	The Gender of the patient	0.55	0–1	0.5	0.91	0.44
hospital	The hospital in which the KA was performed	0.27	0–1	0.44	1.63	0.4
operation-type	The type of operation, Total KA or Unicompartemental KA	0.77	0–1	0.42	0.55	0.48
days-in-hospital	Length of stay in the hospital	2.28	1–7	0.8	0.35	0.56
hours-per-week	Number of hours per week on the patients contract	31.3	6–75	12.15	0.39	0.09
type-of-work	Is the nature of the job self employed or salaried	0.8	0–1	0.4	0.5	0.49
breadwinner	Is the patient the main provider of income	0.55	0–1	0.5	0.91	0.44
work-capable-wo-oper	Is it possible to continue working, in the opinion of the patient, without the operation	2.15	1–4	1.28	0.6	0.14
problems-caused-work	Are the knee problems caused by working activities, according to the patient	1.71	0–4	1.21	0.71	0.07

Table A2. Job-related features.

Feature	Description	Mean	Range	Standard Deviation	Coefficient of Variation	Kruskal–Wallis *p*-Value
prob-unpaid-days	Number of problematic days when the patient was not working	4.72	0–28	8.57	1.82	0.87
nr-prob-days	Number of problematic days when the patient was working	9.51	0–40	10.29	1.08	0.84
am-work-prob-days	Amount of work done on problematic working days	54.94	0–100	41.63	0.76	0.82
preop-capa	Amount of work that could be done before the operation as a percentage of the normal work capacity	6.22	1–10	2.11	0.34	0.36
Work-capacity-preop	Assigned percentage of disability leave	68.99	0–100	45.36	0.66	0.83
nr-of-days-off-work	Number of sick days taken before the operation	2.72	0–32	7.2	2.65	0.1

Table A3. Fitness test features.

Feature	Description	Mean	Range	Standard Deviation	Coefficient of Variation	Kruskal–Wallis *p*-Value
30CRT-T0	Number of sit-and-stand movements, without using hands, the patient can perform in 30 s. Before the operation	13.73	0–31	5.02	0.37	0.43
5XCRT-T0	Amount of time it takes the patient to perform 5 sit-and-stand movements, without using hands. Before the operation	11.58	0–37.44	4.95	0.43	0.68
6MWT-T0	Distance the patient can cover walking for 6 min. Before the operation	484.5	0–700	102.7	0.21	0.63
LFTT-T0	Results of the floor-to-waist lift test. Before operation	22	0–71	11.47	0.52	0.6
CD10-T0	Estimated replacement value, assessed by the therapist, for the lift test, if the patient could not perform this test. Before operation	7.72	0–10	2.4	0.31	0.16
30CRT-T1	Number of sit-and-stand movements, without using hands, the patient can perform in 30 s. 6 Weeks after the operation	10.57	0–20	4.72	0.45	0.39
5XCRT-T1	Amount of time it takes the patient to perform 5 sit-and-stand movements, without using hands. 6 Weeks after the operation	13.08	0–60	9.64	0.74	0.05
6MWT-T1	Distance the patient can cover walking for 6 min. 6 Weeks after the operation	414.19	0–658	144.61	0.35	0.07

Table A4. Questionnaire features pre-operation.

Feature	Description	Mean	Range	Standard Deviation	Coefficient of Variation	Kruskal–Wallis *p*-Value
CSI-T0	Central Sensitization Inventory, measures the somatic and emotional symptoms common with a total maximum score of 100, before operation	25.37	0–56	9.92	0.39	0.16
WORQ-T0	Work, Osteoarthritis or joint-Replacement Questionnaire, to assess physical difficulty experienced in work, before operation	46.63	13.46–76.92	14.06	0.3	0.95
KOOS-A-T0	Knee Injury and Osteoarthritis Outcome Score, to evaluate symptoms and limitations (activities), before operation	53.94	13.24–89.71	16.15	0.3	0.6
KOOS-P-T0	Knee Injury and Osteoarthritis Outcome Score, to evaluate symptoms and limitations (pain), before operation	48.08	8.33–86.11	17.16	0.36	0.7
KOOS-Q-T0	Knee Injury and Osteoarthritis Outcome Score, to evaluate symptoms and limitations (quality of life), before operation	39.93	0–100	27.05	0.68	0.56
KOOS-S-T0	Knee Injury and Osteoarthritis Outcome Score, to evaluate symptoms and limitations (symptoms), before operation	50.46	7.14–89.29	16.81	0.33	0.76
KOOS-SP-T0	Knee Injury and Osteoarthritis Outcome Score, to evaluate symptoms and limitations (sport and recreation), before operation	34.48	0–100	37.08	1.08	0.37

Table A5. Questionnaire features post-operation.

Feature	Description	Mean	Range	Standard Deviation	Coefficient of Variation	Kruskal–Wallis *p*-Value
CSI-T1	Central Sensitization Inventory, measures the somatic and emotional symptoms common with a total maximum score of 100, 6 weeks after operation	24.07	0–61	10.24	0.43	0.36
WORQ-T1	Work, Osteoarthritis or joint-Replacement Questionnaire, to assess physical difficulty experienced in work, 6 week after operation	30.06	6–50	10.46	0.35	0.74
KOOS-A-T1	Knee Injury and Osteoarthritis Outcome Score, to evaluate symptoms and limitations (activities), 6 weeks after operation	60.67	4.41–100	23.75	0.39	0.63
KOOS-P-T1	Knee Injury and Osteoarthritis Outcome Score, to evaluate symptoms and limitations (pain), 6 weeks after operation	57.39	0–97.22	21.04	0.37	0.49
KOOS-Q-T1	Knee Injury and Osteoarthritis Outcome Score, to evaluate symptoms and limitations (quality of life), 6 weeks after operation	48.62	6.25–100	17.68	0.36	0.51
KOOS-S-T1	Knee Injury and Osteoarthritis Outcome Score, to evaluate symptoms and limitations (symptoms), 6 weeks after operation	57.44	21.43–85.71	15.16	0.26	0.68
KOOS-SP-T1	Knee Injury and Osteoarthritis Outcome Score, to evaluate symptoms and limitations (sport and recreation), 6 weeks after operation	45.05	0–100	33.4	0.74	0.37

Table A6. Features collected using the PAM accelerometer.

Feature	Description	Mean	Range	Standard Deviation	Coefficient of Variation	Kruskal–Wallis *p*-Value
average-inactivity-pre	Average number of 30 min periods of inactivity before operation	5.04	1.2–19	2.57	0.51	0.31
average-inactivity-post	Average number of 30 min periods of inactivity in the 6 weeks after operation	5	1–11.89	2.12	0.42	0.34
average-steps-pre	Average number of steps per day before operation	9703.74	63–22761.6	4056.78	0.42	0.61
average-steps-post	Average number of steps per day in the 6 weeks after operation	8239.77	125–18438.6	3429.61	0.42	0.97

Appendix B. Optimization Results for AdaBoost Classifier

Table A7. Comparison of the results of Recursive Feature Elimination and the optimization algorithm.

Classifier	RFE Result	Optimization Result	Improvement	Nr. of Features
ADA15.3.0.1	0.73	0.91	0.18	11
ADA20.2.0.1	0.64	0.88	0.24	20
ADA20.1.0.3	0.76	0.88	0.12	12
ADA20.1.0.2	0.76	0.88	0.12	12
ADA10.3.0.3	0.76	0.88	0.12	40
ADA10.2.0.1	0.67	0.88	0.21	4
ADA20.1.0.1	0.76	0.88	0.12	20
ADA20.2.0.2	0.64	0.88	0.24	21
ADA25.2.0.2	0.70	0.85	0.15	20
ADA25.2.0.3	0.70	0.85	0.15	34
ADA25.3.0.1	0.70	0.85	0.15	21
ADA5.2.0.1	0.61	0.85	0.24	10
ADA25.3.0.2	0.79	0.85	0.06	13
ADA5.3.0.1	0.73	0.85	0.12	13
ADA10.1.0.3	0.61	0.85	0.24	6
ADA10.2.0.3	0.67	0.85	0.18	13
ADA10.3.0.1	0.76	0.85	0.09	40
ADA10.3.0.2	0.79	0.85	0.06	25
ADA15.1.0.1	0.70	0.85	0.15	7
ADA15.1.0.2	0.70	0.85	0.15	7
ADA15.1.0.3	0.70	0.85	0.15	7
ADA15.2.0.1	0.70	0.85	0.15	16
ADA25.3.0.3	0.79	0.85	0.06	13
ADA15.3.0.2	0.76	0.85	0.09	40
ADA15.3.0.3	0.73	0.85	0.12	38
ADA15.2.0.2	0.61	0.85	0.24	37
ADA25.2.0.1	0.70	0.85	0.15	21
ADA20.3.0.3	0.73	0.85	0.12	23
ADA25.1.0.1	0.73	0.85	0.12	14
ADA25.1.0.2	0.73	0.85	0.12	14
ADA25.1.0.3	0.73	0.85	0.12	5
ADA5.2.0.3	0.67	0.82	0.15	11
ADA5.2.0.2	0.70	0.82	0.12	9
ADA15.2.0.3	0.67	0.82	0.1	23
ADA10.1.0.1	0.61	0.82	0.21	4
ADA5.3.0.3	0.73	0.82	0.09	4
ADA10.1.0.2	0.61	0.82	0.21	4
ADA5.3.0.2	0.70	0.82	0.12	2
ADA20.2.0.3	0.64	0.82	0.18	13
ADA20.3.0.1	0.79	0.82	0.03	29
ADA10.2.0.2	0.70	0.82	0.12	24
ADA20.3.0.2	0.76	0.82	0.06	23
ADA5.1.0.2	0.61	0.76	0.15	15
ADA5.1.0.3	0.61	0.76	0.15	15
ADA5.1.0.1	0.61	0.76	0.15	15

Appendix C. Optimization Results for LogisticRegression Classifier, GradientBoosting Classifier and LinearSVC

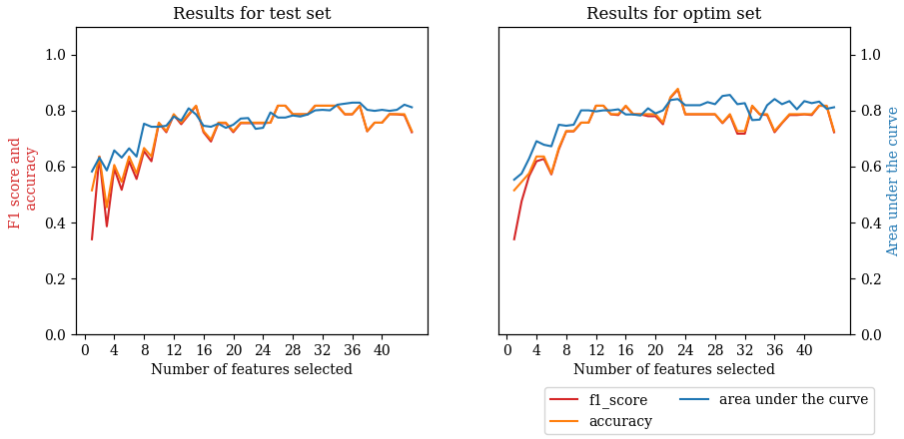

Figure A1. Results for the test set before (**left**) and after (**right**) optimization for the Logistic Regression Classifier. The reduced fluctuation at low number of features indicate a better ordering after optimization.

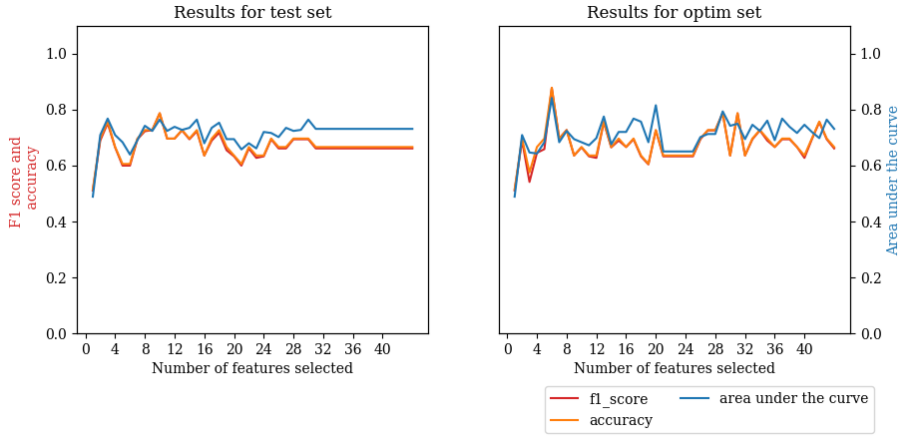

Figure A2. Results for the test set before (**left**) and after (**right**) optimization for the GradientBoosting Classifier, showing that although over-fitting after 32 features seems resolved, the ordering does not appear to be improved.

Figure A3. Results for the test set before (**left**) and after (**right**) optimization for the LinearSVC, showing that some important features are moved upward in the ordering, through the higher performance metric at low numbers of features, but not all, since the maximum appears at 36 features.

References

1. Price, A.J.; Alvand, A.; Troelsen, A.; Katz, J.N.; Hooper, G.; Gray, A.; Carr, A.; Beard, D. Hip and knee replacement 2 Knee replacement. *Lancet* **2018**, *392*, 1672–1682. [CrossRef] [PubMed]
2. Culliford, D.; Maskell, J.; Judge, A.; Cooper, C.; Prieto-Alhambra, D.; Arden, N.K. Future projections of total hip and knee arthroplasty in the UK: Results from the UK Clinical Practice Research Datalink. *Osteoarthr. Cartil.* **2015**, *23*, 594–600. [CrossRef] [PubMed]
3. Kurtz, S.; Ong, K.; Lau, E.; Mowat, F.; Halpern, M. Projections of Primary and Revision Hip and Knee Arthroplasty in the United States from 2005 to 2030. *J. Bone Jt. Surg.* **2007**, *89*, 780–785. [CrossRef]
4. Hardenberg, M.; Speklé, E.M.; Coenen, P.; Brus, I.M.; Kuijer, P.P.F. The economic burden of knee and hip osteoarthritis: Absenteeism and costs in the Dutch workforce. *BMC Musculoskelet. Disord.* **2022**, *23*, 364. [CrossRef] [PubMed]
5. Strijbos, D.O.; van der Sluis, G.; Boymans, T.A.; de Groot, S.; Klomp, S.; Kooijman, C.M.; Reneman, M.F.; Kuijer, P.P.F. Implementation of back at work after surgery (BAAS): A feasibility study of an integrated pathway for improved return to work after knee arthroplasty. *Musculoskelet. Care* **2022**, *20*, 950–959. [CrossRef] [PubMed]
6. Strijbos, D.O.; van der Sluis, G.; van Houtert, W.F.; Straat, A.C.; van Zaanen, Y.; de Groot, S.; Klomp, S.; Krijnen, W.P.; Kooijman, C.M.; van den Brand, I.; et al. Protocol for a multicenter study on effectiveness and economics of the Back At work After Surgery (BAAS): A clinical pathway for knee arthroplasty. *BMC Musculoskelet. Disord.* **2023**, *24*, 199. [CrossRef]
7. Bouchlaghem, Y.; Akhiat, Y.; Amjad, S. Feature Selection: A Review and Comparative Study. In *Proceedings of the E3S Web of Conferences*; EDP Sciences: Les Ulis, France, **2022**; Volume 351, p. 01046. [CrossRef]
8. Pudjihartono, N.; Fadason, T.; Kempa-Liehr, A.W.; O'Sullivan, J.M. A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Front. Bioinform.* **2022**, *2*, 927312. [CrossRef] [PubMed]
9. Lötsch, J.; Ultsch, A. Enhancing Explainable Machine Learning by Reconsidering Initially Unselected Items in Feature Selection for Classification. *BioMedInformatics* **2022**, *2*, 701–714. [CrossRef]
10. Rajput, D.; Wang, W.J.; Chen, C.C. Evaluation of a decided sample size in machine learning applications. *BMC Bioinform.* **2023**, *24*, 48. [CrossRef]
11. Slootmaker, S.M.; Chin A Paw, M.J.; Schuit, A.J.; Van Mechelen, W.; Koppes, L.L. Concurrent validity of the PAM accelerometer relative to the MTI Actigraph using oxygen consumption as a reference. *Scand. J. Med. Sci. Sport.* **2009**, *19*, 36–43. [CrossRef]
12. Bouwmans, C.; Krol, M.; Severens, H.; Koopmanschap, M.; Brouwer, W.; Roijen, L.H.V. The iMTA Productivity Cost Questionnaire: A Standardized Instrument for Measuring and Valuing Health-Related Productivity Losses. *Value Health* **2015**, *18*, 753–758. [CrossRef] [PubMed]
13. de Morton, N.A.; Davidson, M.; Keating, J.L. The de Morton Mobility Index (DEMMI): An essential health index for an ageing world. *Health Qual. Life Outcomes* **2008**, *6*, 1–15. [CrossRef] [PubMed]
14. Guyatt, G.H.; Sullivan, M.J.; Thompson, P.J.; Fallen, E.L.; Pugsley, S.O.; Taylor, D.W.; Berman, L.B. The 6-minute walk: A new measure of exercise capacity in patients with chronic heart failure. *Can. Med. Assoc. J.* **1985**, *132*, 919. [PubMed]
15. Özkeskin, M.; Özden, F.; Ar, E.; Yüceyar, N. The reliability and validity of the 30-second chair stand test and modified four square step test in persons with multiple sclerosis. *Physiother. Theory Pract.* **2022**, *39*, 2189–2195. [CrossRef] [PubMed]
16. Teo, T.W.L.; Mong, Y.; Ng, S.S.M. The repetitive Five-Times-Sit-To-Stand test: Its reliability in older adults. *Int. J. Ther. Rehabil.* **2013**, *20*, 122–130. [CrossRef]
17. Kuijer, P.P.; Gouttebarge, V.; Brouwer, S.; Reneman, M.F.; Frings-Dresen, M.H. Are performance-based measures predictive of work participation in patients with musculoskeletal disorders? A systematic review. *Int. Arch. Occup. Environ. Health* **2012**, *85*, 109–123. [CrossRef] [PubMed]
18. Kievit, A.J.; Kuijer, P.P.F.; Kievit, R.A.; Sierevelt, I.N.; Blankevoort, L.; Frings-Dresen, M.H. A reliable, valid and responsive questionnaire to score the impact of knee complaints on work following total knee arthroplasty: The WORQ. *J. Arthroplast.* **2014**, *29*, 1169–1175.e2. [CrossRef] [PubMed]
19. Kregel, J.; Vuijk, P.J.; Descheemaeker, F.; Keizer, D.; van der Noord, R.; Nijs, J.; Cagnie, B.; Meeus, M.; van Wilgen, P. The Dutch Central Sensitization Inventory (CSI): Factor Analysis, Discriminative Power, and Test-Retest Reliability. *Clin. J. Pain* **2016**, *32*, 624–630. [CrossRef] [PubMed]
20. Collins, N.J.; Misra, D.; Felson, D.T.; Crossley, K.M.; Roos, E.M. Measures of knee function: International Knee Documentation Committee (IKDC) Subjective Knee Evaluation Form, Knee Injury and Osteoarthritis Outcome Score (KOOS), Knee Injury and Osteoarthritis Outcome Score Physical Function Short Form (KOOS-PS), Knee Outcome Survey Activities of Daily Living Scale (KOS-ADL), Lysholm Knee Scoring Scale, Oxford Knee Score (OKS), Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), Activity Rating Scale (ARS), and Tegner Activity Score (TAS). *Arthritis Care Res.* **2011**, *63*, S208–S228. [CrossRef]
21. Park, Y.; Ho, J.C. Tackling Overfitting in Boosting for Noisy Healthcare Data. *IEEE Trans. Knowl. Data Eng.* **2021**, *33*, 2995–3006. [CrossRef]

22. Scikit-Learn. RFE—Scikit-Learn 1.5.1 Documentation. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html (accessed on 23 August 2024).
23. Van Buren, A.; Kwan, A.; Rietdijk, H.H.; Dijkhuis, T.B.; Conde-Cespedes, P.; Oldenhuis, H.; Trocan, M. A Clustering Approach for Personalized Coaching Applications. In *Advances in Computational Collective Intelligence. ICCCI 2024*; Communications in Computer and Information Science; Nguyen, N.-T., Franczyk, B., Ludwig, A., Treur, J., Vossen, G., Kozierkiewicz, A., Eds.; Springer, Cham, Switzerland, **2024**; Volume 2166, pp. 351–363. [CrossRef]