Personalized Allocation of Patients to Physiotherapists with Machine Learning Techniques

Cecilia Henningsson (BME19), Line Engdahl Höie (BME19)

I. Abstract

Osteoarthritis (OA) is a heterogeneous inflammatory joint disease that affects around 240-250 million people worldwide. The disease is a prominent cause of disability and a leading source of societal expenses in older adults. As of today, personalized OA treatment is considered essential and is currently being addressed in several osteoarthritis guidelines. To explore the possibilities of precision medicine, machine learning (ML) has been implemented to allocate patients to therapists in the psychotherapy field. It is therefore argued that allocation using ML can be of interest in OA physiotherapy as well. The aim of this study is to implement a supervised classification ML model to predict the best mapping between onboarding patient and physiotherapist. The OA patient and physiotherapist data collected between the years 2019-2022 was provided by the Swedish telehealth company Joint Academy (JA). The data, used for this study, consisted of 8 separate subsets. All subsets of data were processed according to general methods. Python was the programming language of choice, where algorithms, e.g. Random Forest Classifier (RF) and gradient booster XGBoost were explored and implemented with the software library *sci-kit learn*. Four different models were benchmarked against a baseline model. The baseline model received an F1-score of 71.31% on the test set, which corresponds to 20% of the final data set. The final XGBoosted model received an F1-score of 68.24% on the test set. The final model is not appropriate to implement in a care system in its current state. It can be further improved with better feature engineering, improved imputation techniques, and explore different target variables.

II. Introduction

ONE of the world's most prevalent disabling conditions is osteoarthritis [1]. The degenerative joint disease is causing an increasing global health problem that is held accountable for societal costs as well as individual complications [1], [2]. OA is the most common joint disease in the world as it affects around 240-250 million people worldwide [1], [3]. Since our society has an increasing obesity as well as an aging population, the prevalence of osteoarthritis is expected to increase rather than decrease [1]. In Sweden, osteoarthritis is one of the fastest-growing national diseases and by 2032, one in three over the age of 45 is expected to receive the diagnosis [4], [5]. Due to the heterogeneous nature of the disease and the eminent prevalence of comorbidities, personalized OA treatment is deemed as essential [1]. With the aim of treating and minimizing this dominant disease, many companies have explored the field of what can be done with physical therapy, and how it can be accessible for a majority of affected people [1].

Submitted June 13, 2022
Email address:{ce8234he-s@student.lu.se, li7581en-s@student.lu.se}
Supervisor: Emma Sjögren, Joint Academy

A. Osteoarthritis

Osteoarthritis can essentially be translated to "inflammation of bone and joint cartilage" and is a common degenerative joint disease in every possible joint, e.g. hips, hands, and knees [1], [6]. Osteoarthritis affects so-called synovial joints, which are of articular surfaces and an articular capsule surrounding a fluid-filled joint cavity [7]. The joint surfaces at the end of the bones are embedded with a layer of connective tissue, allowing the bones to rub against each other with minimal friction [8]. The joint cavity, called synovium, encloses synovial fluid, loose connective tissue, vascular units, as well as macrophage-like type A cells and fibroblast-like type B cells [9]. Type A cells phagocytize cellular debris in the synovium and release factors that stimulate B cells to produce the lubricating components of the synovial fluid [9].

Differentiated fibroblasts, called chondrocytes, are the cells responsible for the maintenance of healthy articular cartilage. Chondrocytes produce type II collagen and proteoglycans that generate elasticity and tensile strength [10]. This helps the underlying bone of weight-bearing joints to absorb weight and shock [9]. Normally, chondrocytes maintain an equilibrium between degradative catabolic activity and synthetic anabolic activity. However, with osteoarthritis, there is an increase in the expression of degradative enzymes, and a decrease in synthetic enzyme expression [11]. This results in weaker and less elastic articular cartilage. Degradation products fall into the synovial cavity, which causes macrophages and lymphocytes to release proinflammatory cytokines that cause further inflammation and loss of cartilage [11], [12]. Degradation of cartilage during a long period of time can cause the bones to rub against each other, resulting in bone eburnation which causes pain, stiffness, and immobility for the patient. [13].

Today, osteoarthritis is often not discovered or diagnosed until the patient has reached a moderate or severe stage of the disease. Thus the joint tissue is at risk of being permanently

and irreversibly damaged. This places high demands on treatment and prediction and opens up the discussion of what needs to be improved. Moreover, because of the fact that the clinical diagnostic techniques regarding OA practiced today do not meet the needs regarding preventing disease progression [14].

The first line of OA treatment involves non-pharmacological methods such as education, exercise, weight loss, and physiotherapy. Further treatment involves medication and surgery [15]. As of today, there are no disease-modifying pharmacological solutions to OA, leaving drug treatment being mainly used to reduce inflammation and pain [1], [16], [17]. Implementing surgery, the mortality-adjusted lifetime risk of total knee replacement at the age of 50 is 8,1% for men and 10,8% for women. The mortality-adjusted lifetime risk of total hip arthroplasty is 7,1% for men and 1,6% for women [1]. Joint replacement surgery is cost-effective if performed on patients with end-stage osteoarthritis [1]. However, as the yearly amounts of total joint replacements expand, the demand to reduce costs in arthroplasty is high. In the United States, this financial load on the health care system has led to the employment of bundled care programs to decrease post-acute care costs [2].

To avoid invasive and costly surgical procedures, non-pharmacological methods are deemed as key treatment and are recommended by Osteoarthritis Research Society International (OARSI), as well as The European Society for Clinical and Economic Aspects of Osteoporosis, Osteoarthritis, and Musculoskeletal Diseases (ESCEO) [15]. Non-pharmacological methods have proven to be particularly important to larger weight-bearing joints such as knees [18], [19]. According to a 24-year follow-up on the impact of painful knee osteoarthritis on mortality, knee pain, with or without OA, increases mortality. Knee pain correlated mortality was even more prevalent among patients with higher BMI. However, by using non-pharmacological methods, such as prevention of comorbidities and weight loss, mortality risk was reduced [20]. Due to the heterogeneous nature of OA and the involvement of comorbidities, personalized treatment is deemed as fundamental [1], [21]. As of today, research on clinical predictors of response to various treatments is being addressed in many osteoarthritis guidelines [1].

B. Background

In 2020, a paper from the University of Sheffield was published, on how to match patients to therapists with ML techniques. The ”proof-of-concept” study aimed to develop a data-driven method to match patients being treated with psychotherapy to therapists. This was achieved through implementing a machine learning classification algorithm, specializing in pinpointing particular patient subgroups to specific therapists. The method included ML algorithms such as a Chi-Squared Automatic Interaction Detector (CHAID) algorithm and a Random Forest algorithm. The data was also cross-validated and evaluated using ods.

The study resulted in the identification of different subgroups of patients that were better suited for specific therapists. Therefore, the conclusion stated that machine learning can in fact help better the outcome of therapy treatment, by attaining strategic allocations [22].

C. Joint Academy

Joint Academy is a Swedish digital healthcare company with a focus on technology and science [23]. The company was founded in 2014 with the aim to reduce expensive high-risk treatments and to expand care access for clinically verified hip or knee OA patients [23], [24]. Today, JA operates a digital clinic for back and chronic joint pain by using telemedicine [23]. Treatment involves evidence-based non-surgical measures, such as education, personalized exercises, and guidance from a licensed PT [25]. As of now, the allocation of patient to physiotherapist is based on PT availability. However, Joint Academy wish to explore the possibilities of making the matching more personalized by basing it on different parameters [26].

In the study *Willingness for surgery and health-related quality of life after six months in a digital osteoarthritis self-management program* research showed that 47% of study participants no longer wanted surgery after 6 months of JA digital self-management program. It can therefore be argued that increased partition time in first-line self-management care could help delay or avoid total joint replacements. The unwillingness to undergo surgery that the study has shown, can also be associated with enhanced health-related quality of life correlated to the JA program [27]. Another study on the JA six months self-management program has shown that 69% of participants have reported an overall improvement in health, 85% have reported pain improvement, and 84% have reported better physical function. After six months, 42% of 228 participants also stated that they had stopped using pharmacological treatment, such as pain killers [28]. These numbers state a clear relevance in using digital and personalized physiotherapy which, according to a socio-economic study published by Lund University, has also proven to be cost-effective compared to the current face-to-face model of care [29]. The question, however, remains what the next step is to make OA physiotherapy more efficient.

D. Aims

Being the largest OA caregiver in Sweden [5], Joint Academy is, in accordance with international guidelines, in possession of large amounts of data that can be of interest in the AI and machine learning field. Working together with Joint Academy, the aim of this project is to research the possibility to optimize personalized first-line OA treatment by using machine learning, with the ambition of improving and individualizing the treatment further. By constructing a supervised classification model, the prospect of attaining individualized allocation of patients to physiotherapists will

be explored. Joint Academy has provided data sets containing anonymous patient and therapist data collected during 156 weeks between the years 2019-2022.

The manner of development of OA depends on different risk factors such as age, obesity and joint injuries [3]. By studying and classifying the relations and interactions between these factors, researchers hope to reduce the risk of disease development. This can be seen as a strong motive for the development and use of classification and prediction models that can analyze large sets of patient data [14].

As it was concluded in a previous study, allocating therapists and patients with ML techniques can be profitable in treatment results [22]. However, Joint Academy has not explored this in the field of OA. Therefore, our goal is to prove that it is possible for physiotherapy as well.

E. The Concepts of Machine Learning

Considering that this report revolves around machine learning, a quick introduction to the subject will be given below. The general task of ML is to recognize patterns in data, and to automatically improve through experience. It is a common type of artificial intelligence that is used for prediction problems, to achieve a more precise prediction without actually having to program an algorithm to do so. This is helpful when attempting to analyze large sets of data where the complexity of finding trends in the data increases exponentially [30].

When analyzing a data set, one must first ascertain whether or not the desired output is included in the data set. The desired output is defined as the *target variable* or *label*, and is what the model is meant to predict. The remaining columns are called features. Additionally, the data set is divided into a training and test set, which permits the algorithm to both train on data to attain a good prediction, as well as test and evaluate that prediction on unknown data separated from the training. To further understand the different problems ML can tackle, let us look at a large class of learning problems, supervised learning [30].

Supervised learning is executed on data sets that contain both input and desired output. The goal of supervised learning is to train algorithms and classify or ”predict” labeled data correctly. This subcategory of ML is used for classification, which can be described as the process of recognizing and grouping target vectors into subgroups, often named classes. Supervised learning’s opposite, unsupervised learning, is a class of learning with unlabelled data [31]. This paper will not cover the unsupervised learning and is therefore left for the reader to explore themselves.

In this report, the data will consist of both inputs and desired target vector, and therefore supervised learning is being implemented. The learning problem is of a classification type, hence supervised learning is preferable.

F. Machine Learning methods used - technical description

Feature Engineering: In machine learning, a feature is any computable input, e.g. age, sex, car model, monthly income etc, that can be employed in a predictive model [32]. The purpose of feature engineering is to increase the performance of the ML algorithm by selecting and modifying the most significant variables from raw data, as well as make it compatible with the chosen ML algorithm [32], [33].

Model Evaluation: Model evaluation is a measurement of the performance of the ML model when new data, similar to the training data, is applied [34]. To measure the performance of the models, four common metrics to use are accuracy (1), recall (2), precision (3), and F1-score (4). These metrics are calculated through equations based on the ratio between true positives, true negatives, false positives, and false negatives, corresponding to mispredictions of the model. The four ratios can be evaluated in its simplicity by adding them to a matrix, a so called *confusion matrix*. In the columns, the negative and positive predictions (0 vs 1) are displayed, whilst the rows display correct and incorrect predictions. Therefore, a higher diagonal value of the confusion matrix corresponds to a better predictive model [35]. For a visual representation of each row and column, please regard table 1. Recall, precision, accuracy, F1-score as well as a confusion matrix were examined for each model to determine the highest performing one.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{4}$$

		Prediction outcome	
		0	**1**
actual value	**0**	True Negative	False Positive
	1	False Negative	True Positive

Overfitting and Underfitting: Two main sources for insufficient predictive performance in ML are overfitting and underfitting of the data. Overfitting and underfitting occurs when the machine learning model is incapable to generalize well from training data to new input data from the problem domain [36]. An underfitted model cannot follow

the underlying relationship of the data and will likely execute a considerable amount of mispredictions. Underfitting occurs when the data is insufficient, which can be prevented by using more data or by implementing feature selection [37]. Overfitting occur when the model begin to learn from noise and erroneous data entries due to being trained on too much data. These wrongfully learned model concepts do not apply on new data from the problem domain, which causes the model to perform poorly [36], [37].

Multicollinearity: Multicollinearity occurs when multiple columns contain similar information, such as weight and BMI, and therefore obtain a strong positive correlation. At the same time, multicollinearity can occur due to a negative correlation caused by opposite information, such as male and female gender. Multicollinearity can cause the model to memorize the training data, rather than learning the relationships in the data. This results in very high performance on the training data, but lesser results on new data [38].

Baseline model: A baseline is a basic, non-complex model used as a reference in ML projects. The aim is to find more developed solutions to achieve a higher score than the baseline [39].

XGBoost: XGBoost is a gradient boosting algorithm used for supervised learning, e.g. classification and regression [40], [41]. Gradient boosting has the principle of an ensemble, sequentially learning from data and improving prediction performance. XGBoost tries to minimize the residual error when adding more trees by combining models that are marginally better than random guessing, so-called weak learners. [41].

Resampling: Resampling is a commonly used technique to adopt whilst dealing with imbalanced data sets. There are two common resampling techniques in ML called oversampling and undersampling [35]. Oversampling is performed by enlarging the quantity of data points in the minority class to equally distribute the values in the training set. Undersampling instead entails reducing the number of data from the majority class. Deleting training data is not preferred. However, if the data set is very large, containing thousands or millions of data points, undersampling can be an appropriate approach [35].

III. Method

The approach of allocating patients with physiotherapists can be sectioned into two different procedures, according to the framework of ML: data preparation and machine learning. Data preparation often includes gathering, identifying, organizing, and cleaning data. A majority of the time was spent on data wrangling and feature engineering. The model training, on the other hand, entailed an extensive iteration process. This process included improving model performance by hyperparameter tuning and gradient boosting, but also by going back to the data set and preprocessing it moreover. The method section of this project is therefore, instead of being chronological, divided into different subgroups, perhaps implemented several times, during various parts of the project.

A. Data preparation

1) Feature engineering: Initially, eight different data frames containing anonymous patient and therapist data were uploaded on the web-based computational environment Jupyter Notebook. The intention of the data preparation was to merge all data frames into one large set. To prepare for merge, different computer programming methods were applied on the rows and columns of the data frames. This process is called feature engineering and was performed using the programming language Python 3.

To sort out redundant data, columns containing only a single value or over 70% missing or undefined values were removed from each data frame. Likewise, identical rows were erased from each data frame.

A proper merge requires only a single line per merge variable, in this case patient and/or therapist ID, to ensure a qualitative merge with a reasonable amount of rows. Hence, the following step was to ensure each data frame only consisted of one row for each patient or therapist ID. To extract information from the data sets, different mathematical techniques were applied to the data, e.g. one-hot encoding, mean, and summation.

2) Data merge: To construct input data for the ML algorithm, the data merge had to be done with high precision. The goal of the final data set was to have the data consist of one unique row per patient, which included both patient and therapist-related data. To achieve this, the different data sets were merged on two common variables, patient ID and therapist ID. Due to some data frames containing both therapist ID and patient ID, all eight data sets were able to be merged into one successfully.

3) Data imputation: Following the completion of the data frame merge, data imputation was subsequently implemented to fill in missing values. Depending on the data type of a specific column, different strategies were used. Missing values in columns containing dates were replaced with the Python object *datetime*. Since the actual date remained unknown, the missing spot was filled with a deviating value that would differ significantly from other dates in the data frame. This was done with the aim of the algorithm to mark it as abnormal, and to disregard it in potential decision trees. Missing values in columns containing *boolean* objects were imputed with the median value, whilst data types *ints* and *floats* where imputed with mean values. Considering the classifier that was being used, and its inability to handle the data type string, all string columns required modification. All columns containing strings were one-hot encoded, whereas missing values in a string column therefore received the value of zero.

By choosing the target vector to be based on the ”rating” column, all users who did not have a rating after 12 weeks, the initial program length at JA, were discarded. Thereafter the column ”rating” was excluded from the data set to form a target vector, in the form of a NumPy array. Hence, two separate data sets were created, features and labels.

In supervised ML models, a well-defined target vector is required to be able to map the relationships between the target variable and the input data. This implies that binary output, such as 1 for a good match, and 0 for an insufficient match, must have a binary target vector. The rating column consisted of ratings between 1-5, with 5 being the highest score. The rating distribution of the target vector was heavily weighted to the rating 5, and consequently, it was decided that a rating of 5 was equal to a good match. Therefore all entities with 5 as a rating were labeled as class 1 and a rating of 1-4, including patients with missing values, was imputed with the value zero.

4) Standardization and data split: To feature scale, standardization was implemented. Non-numerical data types were separated from the train and test sets. To avoid standardization of patient and therapist ID:s, these columns were converted to strings during the standardization process, and then returned to their original type once the process was completed. Before standardization was implemented, the target vector and data were split into a train and test subset, whereas 80% of the data were assigned to training, leaving 20% for testing. The percentages were based on common practice and the size of the data set. Standardization was executed with the sklearn class *StandardScaler*.

B. Machine Learning

1) Baseline model: In this study, a baseline model was designed to randomly distribute the value of 1, indicating an adequate match, with a probability of around 70%. Non-matches, represented by 0, were distributed with a weight of the remaining 30%. This weighting was based on the target variable-ratio before undersampling.

2) Model fitting: ,
Random Forest: As previously mentioned, a supervised ML classification algorithm was a favorable approach since a predictive model based on prior patient and PT data was going to be created. Therefore the sklearn ensemble learning method RandomForestClassifier was implemented. The Random Forest Classifier was implemented with different setups of hyperparameters which resulted in three different models.

XGBoost: In addition to Random Forest, an alternative model was implemented within the framework of the algorithm XGBoost. The data used remained unaltered, and default parameters for the algorithm were used.

3) Data correlation: To avoid overfitting of the models, the correlation rate of the final data set was analyzed for multicollinearity. Columns with a correlation over 0.7 or under -0.8 were therefore removed from the final training data set.

4) Resampling: To prevent the model from always presenting one outcome, the data was resampled. The final data set, containing thousands of rows, was deemed sufficiently large to implement undersampling, and was performed using *imblearns* pre-built under-sampler *RandomUnderSampler*.

5) Hyperparameter tuning: After adjusting the correlation rate and undersampling to receive a more diverse and realistic model, hyperparameter tuning was conducted to enhance performance. Hyperparameter tuning involved choosing the most optimal set of hyperparameters to obtain the best performing model without overfitting. Hyperparameter optimization was performed in two different methods: by manual iteration and by plotting validation curves. The optimal value of each hyperparameter was combined in order to achieve the best possible score. All validation curves can be found in the appendix.

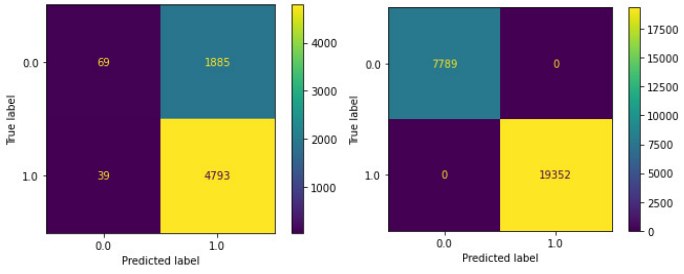
IV. Result

The total data set contained 33 927 rows and 118 columns. Hence, the data frame consisted of 33 927 unique patient IDs. After the data set was balanced according to the undersampling method, 22 364 rows and patient IDs remained.

Table I
DIFFERENT PERFORMANCE MEASUREMENTS FOR ALL FOUR DECISION TREE MODELS AND FOR THE BASELINE MODEL. ALL METRICS ARE BASED ON THE TEST DATA, EXCEPT ”F1-SCORE TRAIN SET”

	Model 1	Model 2	Model 3	Model 4	Baseline
Recall	0.9919	0.5906	0.6125	0.6016	0.7136
Precision	0.7177	0.7992	0.7504	0.7882	0.7125
F1-score	0.8328	0.6793	0.6744	0.6824	0.7131
F1-score train set	1.0	1.0	0.5828	0.6322	0.5859
Accuracy	0.7165	0.6029	0.5790	0.6824	0.5911

A. Model 1 - Random Forest



Test data Train data

Figure 1. Confusion matrix for test and train data prediction of model 1.

The model included all default hyper parameters except for *n-estimators*, which was given the value 500. The test data

confusion matrix displayed an imbalanced distribution of class 1 versus class 0 predictions, where 1.6% of all predictions were of class 0. The train data confusion matrix contained no mispredictions.

B. Model 2 - Random Forest

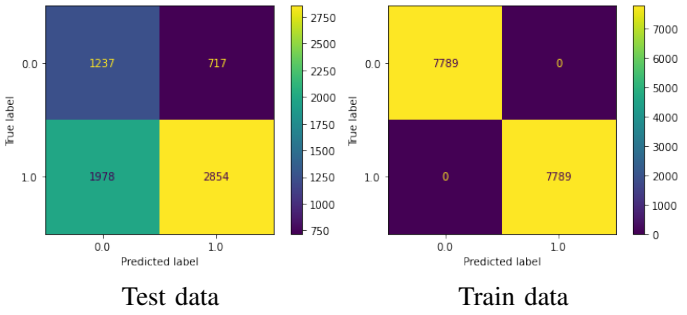


Figure 2. Confusion matrix for test and train data prediction of model 2

The parameters used in model 1 was implemented in model 2. The false positive and false negative rate in the test data set were more equalized after implementing undersampling. The models performance on the train data remained at zero mispredictions.

C. Model 3 - Random Forest

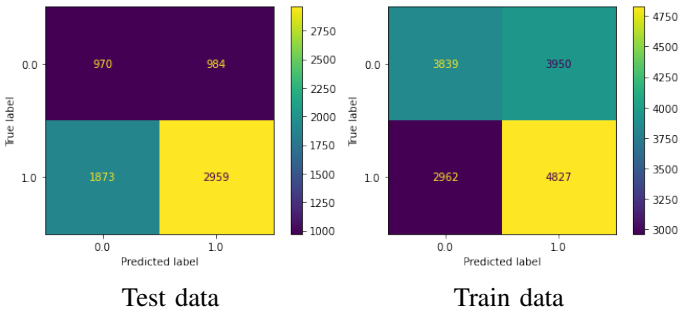


Figure 3. Confusion matrix for test and train data prediction of model 3

Model 3 was trained with parameters generated from hyper parameter tuning. The following parameters were used: n_estimators=700, max_depth=16, min_samples_leaf=3, max_features='auto', max_leaf_nodes=10, random_state= 1, max_samples=9. The confusion matrix displays misclassification on the train data set, which differs from former results with 100% correct classifications. The test set confusion matrix showed no significant change of performance in contrast to model 2.

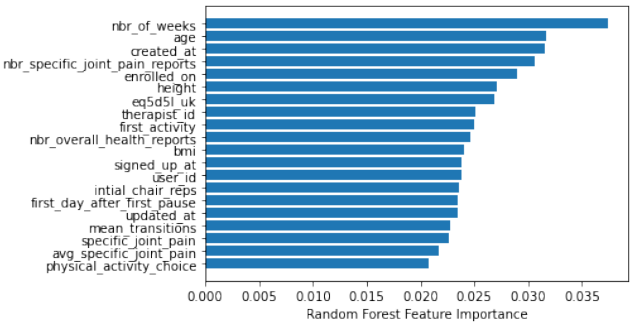


Figure 4. Feature Importance plot displaying the 20 most significant features in Random Forest Model 3

D. Model 4 - XGBoost

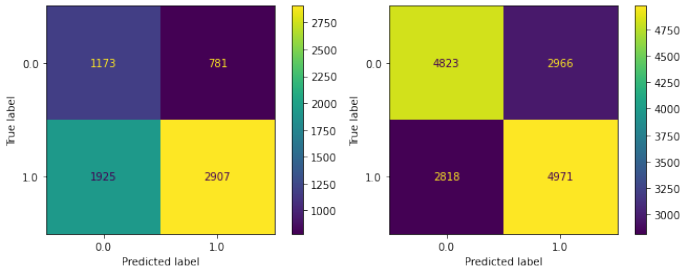


Figure 5. Confusion matrix for test and train data prediction of model 4

Model 4 included all default parameters for the XGBoost algorithm. The model showed slightly higher F1-scores for both the training and test data set.

V. Discussion

A. Baseline Model

To argue if the finished models are better than random guessing, it is common to compare model evaluation metrics to the given baseline model. In this project, the emphasis was on the F1-score, presenting a predictive performance by combining the metrics recall and precision. As previously mentioned, the baseline model was weighted according to the class ratio of the target vector. By using this 70/30 ratio, rather than a 50/50 distribution which would correspond to random guessing, the baseline model received an improved predictive performance and thus, a higher F1-score. F1-score of the baseline model was 71.31% on the test data (I).

B. Random Forest Models - Model 1

The Random Forest Classifier was chosen to be implemented on models 1-3. The operation of the RF algorithm was deemed appropriate due to its simplicity and powerful predictive performance on large data sets.

The confusion matrix corresponding to the training data displayed that the model predicted the labels 100% correctly (see figure 1). Thereby the matrix showed 0 mispredictions and the F1 score was equal to one (1, I). These results were

not realistic and gave a strong indication that the model might be overfitted.

When examining the evaluation metrics of the test data, it was observed that the F1-score had dropped around 20% in contrast to the F1-score on the train set (I). Because the predictive performance dropped significantly on the test set, it was confirmed that model 1 was overfitted. The F1-score on the test data of model 1 outperforms the benchmark F1-score for the test data, but due to overfitting model 1 was disregarded.

The evaluation metrics of model 1 (I), would normally correspond to a well-trained and high-performing model. However, by examining the true negative and false positive rate of the confusion matrix (1), it was observed that the model predominantly predicted class 1. The ratio 69:1885 was alarming since all of these predictions were supposed to indicate a non-match of class 0. The disproportionate outcome is favorable for predicting class 1 correct, but not the contrary. It is unlikely that this is a realistic outcome of PT matches. Hence the model has faulty logic, which implies that error lies within the final data set. This resulted in a defectively trained algorithm that presented a good score on the current train set but would perform poorly on new and more balanced test data.

C. Model 2

Due to the prediction error in model 1, it was argued that it was reasonable to evaluate whether or not a balanced data set would change the outcome. The target vector was imbalanced with a 70/30 ratio and was consequently undersampled to a 50/50 ratio. By balancing the target vector to an even distribution before being fitted, the false positive rate significantly decreased on the test data. This can be seen in figure 2, where the predictions of the test data are much more balanced. F1-score on the test data was lower than the benchmark test F1-score (I).

The first attempt at correcting the overfitting error in model 1 consisted of checking the data set for multicollinearity amongst columns. This however did not change the training data results (see figure 2).

D. Model 3

The third model was generated after using hyperparameter tuning on model 2, and its confusion matrix can be seen in figure 3. The effects of altering the hyperparameters were studied by plotting cross-validation curves (A). The scores on the training set showed somewhat characteristic curves, with corresponding limit values as to where they started to overfit. By choosing parameter values before the graphs diverged, overfitting could be prevented. The F1 training score on model 3 dropped significantly from 1.0 to 0.6322 (I).

The hyperparameter tuning on the test set was performed by using validation curves as well as altering the parameters manually. However, this resulted in little to no change in F1-score, which was unexpected and did not seem reasonable. This indicated that there might be an error in the cross-validation code or the final data set. A lower F1-score on the training set combined with low performance on the test set is a sign of an underfitted model. This might be the result of undersampling or by the removal of correlation columns.

The possibility of features not being needed should be further examined by looking at feature selection and eliminating features of low importance. The 20 most important features can be shown in Figure 4, where *nbr_of_weeks* and *age* were the two most significant features. Further research must be done to evaluate every possible feature but was not included in this study.

E. Model 4 - XGBoost model

In a last attempt to achieve better predictive performance, XGBoost was implemented on the data to reduce the residual when adding more trees to the forest. The XGBoost algorithm attained a slightly better result, increasing F1-score, accuracy, and precision on the test data set (see table I). This final model was deemed most suited for the project thesis. The models confusion matrix can be seen in figure 5.

However, the F1 test score of Model 4 did not surpass the baseline model's measure. One might argue that this result signifies that a weighted randomized model is the preferable one, but there might be some different perspectives worth reflecting over. An explanation for the surpassing predictive performance in the baseline model is the imbalanced test set. The principle of resampling is to only modify the training data, leaving the test data to an untouched ratio. Since the test set had a similar ratio as the baseline model weighting, F1-score increased compared to the F1 training score. The benchmark F1-score on the training set is approximately 5% less than model 4 training score, indicating that the benchmark model could perform worse on a differently distributed test set (I).

F. The Next Step

The final model did not have the ideal predictive performance, which questions whether allocation of patients to physiotherapists using ML techniques can be achieved with these models and data structure. When a model is underfitted, more information could be extracted with improved feature engineering, different imputation techniques, and more time and knowledge. Perhaps, another target vector than "rating" would be more effective. Since improvement in OA can be measured in increased mobility and decreased pain, these variables might be a better indicator as to whether the pathological state of the patient has improved.

With that being said, there are certain difficulties that come with choosing a target vector or with implementing these kinds of prediction problems, due to what is called the human factor. When handling human beings, a good predictive

model has to comprehend complex cognitive processes such as human decision making, etc. This will always reflect in the data and is inescapable when implementing ML techniques. No human can be predicted with 100% accuracy due to the fact that it is, in fact, a human being.

G. Ethics and Sustainable development

When discussing AI and machine learning one can not turn a blind eye to the ethical challenges that *the subject* faces, especially when the algorithms are used in healthcare. This field places higher demands on AI than others, since healthcare processes private and sensitive information. Also, the laws of ethics can often limit the innovation possibilities that AI otherwise provides.

A machine learning algorithm is always at risk of becoming biased. Since the project aim is to recommend a certain employee from a clinic, it is essential that bias does not occur. If the algorithm becomes biased, one physiotherapist can end up never being recommended, or another one getting non-varied work.

Algorithms like this can streamline back-office assignments that will improve the work quality and efficiency at medical facilities, which not only improves patient experiences but also reduces waste and saves resources. This is a development that promotes sustainable solutions that will benefit both patients and companies in the long run, by providing better care whilst not adding additional effort.

VI. Conclusion

The aim of this thesis was to optimize allocation of patients to physiotherapists by using ML techniques. From the five implemented and tuned machine learning models, it can be concluded that this is feasible. However, due to underfitting, the prediction performance is far from optimized. With that being said, the underfitted model indicates that there is more useful information available in the data set. With better feature engineering, improved imputation techniques, and perhaps a more suitable target vector, the model has a chance of improving significantly. Lastly, interfering with the model optimization is the difficulties that come with prediction problems involving human beings. In conclusion, there is more potential to the ML models, and the extent to how good this model can get is yet to be fully explored .

VII. Acknowledgements

Our special thanks to our supervisor, Joint Academy data scientist Emma Sjögren, for always taking her time to help and continuously pushing us to do our best. Emma has provided us with countless papers and blog posts, large sets of python code, useful lessons as well as entertaining anecdotes. This project would not have been possible without her support. We would like to express our gratitude to Joint Academy for this interesting and motivating project assignment. We would also like to thank Joint Academy for the access to the extensive patient and physiotherapist meta data. Finally, we would like to thank the Department of Biomedical Engineering and our course coordinators Josefin Starkhammar and Tomas Jansson, for this instructive ”klinnovation” course project.

The overall project included a *Fast AI* course, workshops, research, analysis, coding, supervisor meetings, and report writing. Both group members participated in the AI course and attended workshops and meetings. Cecilia Henningsson did the major part of the coding and programming whilst Line Engdahl Höie contributed with pseudocode and coding logic. Line EH was also mainly responsible for keeping a journal during the project.

With respect to report writing, Introduction and Background, were written by both Line EH and Cecilia H. Line EH made the initial draft of the Abstract and Method sections, whilst Cecilia H was responsible for drafting the Results, Conclusions, and Discussion sections, before the text and overall content were completed as a joint effort.

References

[1] Bierma-Zeinstra S Hunter DJ. Osteoarthritis. *the Lancet*, 393(10182):15, 2019.

[2] David A Crawford, Adolph V Lombardi Jr, Keith R Berend, James I Huddleston III, Christopher L Peters, Alexander DeHaan, Erin K Zimmerman, and Paul J Duwelius. Early outcomes of primary total hip arthroplasty with use of a smartphone-based care platform: a prospective randomized controlled trial. *The Bone & Joint Journal*, 103(7 Supple B):91–97, 2021.

[3] Jeffrey N Katz, Kaetlyn R Arant, and Richard F Loeser. Diagnosis and treatment of hip and knee osteoarthritis: a review. *Jama*, 325(6):568–578, 2021.

[4] Aleksandra Turkiewicz, Ingemar F Petersson, Jonas Björk, Gillian Hawker, Leif E Dahlberg, L Stefan Lohmander, and Martin Englund. Current and future impact of osteoarthritis on health care: a population-based study with projections to year 2032. *Osteoarthritis and cartilage*, 22(11):1826–1832, 2014.

[5] Joint Academy. Joint Academy, Årsrapport 2020. https://www.jointacademy.com/se/sv/arsrapport-2020/.

[6] K Hrishikesh Yadav and Sindhu Shashidharan. Effectiveness of retrowalking in osteoarthritis of knee–a review article. *International Journal of Advanced Research*, 4(2):215–220, 2016.

[7] Valentin LL Popov, Aleksandr MM Poliakov, and Vladimir II Pakhaliuk. Synovial joints. tribology, regeneration, regenerative rehabilitation and arthroplasty. *Lubricants*, 9(2):15, 2021.

[8] Ross E. Petty and James T. Cassidy. Chapter 2 - structure and function. In James T. Cassidy, Ross E. Petty, Ronald M. Laxer, and Carol B. Lindsley, editors, *Textbook of Pediatric Rheumatology (Fifth Edition)*, pages 9–18. W.B. Saunders, Philadelphia, fifth edition edition, 2005.

[9] René van Weeren. 11 - joint physiology: Responses to exercise and training. In Kenneth W. Hinchcliff, Andris J. Kaneps, and Raymond J. Geor, editors, *Equine Sports Medicine and Surgery (Second Edition)*, pages 213–222. W.B. Saunders, second edition edition, 2014.

[10] Tiziana Franceschetti and Anne M. Delany. Chapter 25 - mirnas in bone repair. In Chandan K. Sen, editor, *MicroRNA in Regenerative Medicine*, pages 653–683. Academic Press, Oxford, 2015.

[11] Shari M Ling and Joan M Bathon. Osteoarthritis: pathophysiology. *Johns Hopkins Arthritis Center*, 2012.

[12] Sunhee Jang, Kijun Lee, and Ji Hyeon Ju. Recent updates of diagnosis, pathophysiology, and treatment on osteoarthritis of the knee. *International Journal of Molecular Sciences*, 22(5):2619, 2021.

[13] R Lagier. Bone eburnation in rheumatic diseases: a guiding trace in today's radiological diagnosis and in paleopathology. *Clinical rheumatology*, 25(2):127–131, 2006.

[14] Afshin Jamshidi, Jean-Pierre Pelletier, and Johanne Martel-Pelletier. Machine-learning-based patient-specific prediction models for knee osteoarthritis. *Nature Reviews Rheumatology*, 15(1):49–60, 2019.

[15] Nigel K Arden, Thomas A Perry, Raveendhara R Bannuru, Olivier Bruyère, Cyrus Cooper, Ida K Haugen, Marc C Hochberg, Timothy E McAlindon, Ali Mobasheri, and Jean-Yves Reginster. Non-surgical management of knee osteoarthritis: comparison of esceo and oarsi 2019 guidelines. *Nature Reviews Rheumatology*, 17(1):59–66, 2021.

[16] Kåre B Hagen, Geir Smedslund, Nina Østerås, and Gro Jamtvedt. Quality of community-based osteoarthritis care: a systematic review and meta-analysis. *Arthritis care & research*, 68(10):1443–1452, 2016.

[17] Sameer Akram Gohir, Paul Greenhaff, Abhishek Abhishek, and Ana M Valdes. Evaluating the efficacy of internet-based exercise programme aimed at treating knee osteoarthritis (ibeat-oa) in the community: a study protocol for a randomised controlled trial. *BMJ open*, 9(10):e030564, 2019.

[18] Gro Jamtvedt, Kristin Thuve Dahm, Anne Christie, Rikke H Moe, Espen Haavardsholm, Inger Holm, and Kåre B Hagen. Physical therapy interventions for patients with osteoarthritis of the knee: an overview of systematic reviews. *Physical therapy*, 88(1):123–136, 2008.

[19] Keith Sinusas. Osteoarthritis: diagnosis and treatment. *American family physician*, 85(1):49–56, 2012.

[20] RJ Cleveland, C Alvarez, TA Schwartz, E Losina, JB Renner, JM Jordan, and LF Callahan. The impact of painful knee osteoarthritis on mortality: a community-based cohort study with over 24 years of follow-up. *Osteoarthritis and cartilage*, 27(4):593–602, 2019.

[21] Erlangga Yusuf. Pharmacologic and non-pharmacologic treatment of osteoarthritis. *Current Treatment Options in Rheumatology*, 2(2):111–125, 2016.

[22] Jaime Delgadillo, Julian Rubel, and Michael Barkham. Towards personalized allocation of patients to therapists. *Journal of Consulting and Clinical Psychology*, 88(9):799, 2020.

[23] Joint Academy. About us: Joint Academy. https://www.jointacademy.com/us/en/about/.

[24] Håkan Nero, Jakob Dahlberg, Leif E Dahlberg, et al. A 6-week web-based osteoarthritis treatment program: observational quasi-experimental study. *Journal of medical Internet research*, 19(12):e9255, 2017.

[25] Anna Cronström, Leif E Dahlberg, Håkan Nero, and Catharina Sjödahl Hammarlund. "i was considering surgery because i believed that was how it was treated": a qualitative study on willingness for joint surgery after completion of a digital management program for osteoarthritis. *Osteoarthritis and cartilage*, 27(7):1026–1032, 2019.

[26] Emma Sjögren. personal communication.

[27] H Nero, S Lohmander, and LE Dahlberg. Willingness for surgery and health-related quality of life after six months in a digital osteoarthritis self-management program. *Osteoarthritis and Cartilage*, 28:S31, 2020.

[28] H Nero, S Lohmander, and LE Dahlberg. Improved patient outcomes by a first-line osteoarthritis self-management program delivered digitally. *Osteoarthritis and Cartilage*, 28:S164–S165, 2020.

[29] B Ekman, H Nero, LS Lohmander, and LE Dahlberg. Costing analysis of a digital first-line treatment platform for patients with knee and hip osteoarthritis in sweden. *Plos one*, 15(8):e0236342, 2020.

[30] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[31] Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová. Machine learning and the physical sciences. *Reviews of Modern Physics*, 91(4):045002, 2019.

[32] H Patel. What is Feature Engineering — Importance, Tools and Techniques for Machine Learning. https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10, August 2021.

[33] Brownlee J. Discover Feature Engineering, How to Engineer Features and How to Get Good at It. https://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/, September 2014.

[34] DataRobot. Model Fitting. https://www.datarobot.com/wiki/fitting/, sine anno.

[35] Alencar R. Resampling strategies for imbalanced datasets. https://www.kaggle.com/code/rafjaa/resampling-strategies-for-imbalanced-datasets/notebook, 2018.

[36] Brownlee J. Overfitting and Underfitting With Machine Learning Algorithms. https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/, March 2016.

[37] ML — Underfitting and Overfitting.

[38] Badr W. Why Feature Correlation Matters... A Lot! https://towardsdatascience.com/why-feature-correlation-matters-a-lot-847e8ba439c4, January 2019.

[39] Nair A. Baseline Models: Your Guide For Model Building. https://towardsdatascience.com/baseline-models-your-guide-for-model-building-1ec3aa244b8d, April 2022.

[40] dmlc XGBoost. About XGBoost. https://xgboost.ai/about, sine anno.

[41] Lutins E. Boosting in Machine Learning and the Implementation of XGBoost in Python. https://towardsdatascience.com/boosting-in-machine-learning-and-the-implementation-of-xgboost-in-python-fb5365e9f2a0, August 2017.

APPENDIX

A. Validation-curve plots for hyper parameters of Random Forest Classifier model 2.