Multiclass Classification Approaches for Arthritis Classification

Mikolaj Hilgert
Master Data Science in Business and Entrepreneurship
Jheronimus Academy of Data Science
S-Hertogenbosch, The Netherlands
m.a.hilgert@tilburguniversity.edu

I. Introduction

Arthritis, a group of conditions characterised by inflammation and pain in the joints, affects millions of individuals and presents diverse subtypes, each requiring tailored treatment strategies [1]. The ability to identify the exact type of arthritis would greatly benefit improving patients lives and help in early detection and management. This project therefore aims to classify arthritis types in adults using data from the National Health and Nutrition Examination Survey (NHANES), which is an comprehensive dataset that captures socio-demographic, clinical, and diagnostic information [2].

The primary objective of this project is to develop a machine learning pipeline capable of accurately predicting the presence and types of arthritis in patients. Through a data-mining approach, the project aims not only to advance the understanding of arthritis subtype classification but also to demonstrate the practical applications of machine learning in the field of healthcare. This problem statement can be translated into a set of research questions provided below.

Main research question:

"To what extent can machine learning techniques accurately predict the type of arthritis in adults within the NHANES population?"

Sub-research questions:

1) What are the significant associations between participant characteristics and arthritis types in the NHANES dataset?
2) Among various classification models and evaluation metrics, which combination yields the best performance in predicting arthritis types?
3) Which features are most influential in predicting arthritis types?

If a machine learning model can be developed to a high enough standard, as is vital for healthcare applications it can help millions of people around the world receive arthritis diagnoses in a more timely manner. This report begins with a review of related works, then introduces the methodology used for this project. Following this, the experimental evaluation examines the decisions, insights, and results of the analysis. This is followed by a thorough discussion and final conclusion.

II. Related Work

As noted by [3], arthritis encompasses more than 100 conditions that affect the joints, bones, and surrounding tissues. The most common type is osteoarthritis, which occurs when the cartilage in joints begins to degenerate and break down, hence its alternative name, degenerative arthritis. Rheumatoid arthritis is a chronic inflammatory autoimmune disorder, where the immune system attacks the tissue surrounding joints responsible for fluid production, which aids in smooth movement. Psoriatic arthritis, on the other hand, is also a chronic autoimmune disease; in this case, the body attacks both the joints and the skin. It is linked to the psoriatic skin condition.

The study by [4] explored the risk factors that are associated with psoriatic arthritis (PsA) among individuals with psoriasis. The researchers used the same NHANES data utilised in this study, and applied algorithms to select key variables from 38 potential predictors. The results presented the key predictors for their Borutamodel included: age, fasting glucose, education level, thyroid disease, hypertension, and chronic bronchitis. The model achieved a area under the curve (AUC) score of 0.781, and Brier scores of 0.186 for the testing set, indicating a good fit. The study introduces a novel risk assessment model that effectively identifies and highlights key predictors for PsA, offering valuable insights into its risk factors.

[5] identifies several risk factors that are often associated with arthritis. These include age, with the risk increasing for older individuals, some sexes are more likely to develop arthritis. Family predisposition to arthritis can also elevate risk, with obesity also being a predictor as weight can add stress to joints. Smoking is also linked to higher risk of developing rheumatoid arthritis. Existing joint injuries can increase risk of osteoarthritis later in life, with people who served in the military being more likely to have arthritis. With veterans being 1.67 times more likely have arthritis compared to the overall U.S. adult population.

In a paper by [6] on Deep Learning-Based Classification of Inflammatory Arthritis, the researchers utilised three-dimensional (3D) joint shape patterns derived from high-resolution peripheral quantitative computed tomography (HR-pQCT) scans of hand joints. The authors trained a neural network on these shapes and aimed for it to differentiate

between rheumatoid arthritis (RA), PsA and healthy control samples (HC). The model was found to be accurate and achieved AUC scores of 0.82 for HC, 0.75 for RA, and 0.68 for PsA.

III. Proposed Methodology and ML Techniques

The The CRoss Industry Standard Process for Data Mining (CRISP-DM) Model is a methodology for data mining, analytics, and data science projects [7]. It consists of six distinct phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment. As shown in Figure 1, transitions between phases are iterative, reflecting the methodology's flexibility and robustness. Given these characteristics, it has been selected as the methodology for this project.

Fig. 1: CRISP-DM Model

The individual steps of the model are elaborated on in following subsections.

A. Business Understanding

The goal of this project is to predict the type of arthritis a patient has based on their clinical and diagnostic data, enabling for quicker and more precise diagnoses for patients. The project's success will be measured by its F1 score, which balances precision and recall, this is because we are aiming to minimise false positives and false negatives. This metric is crucial to ensuring both accurate identification and minimising misclassification. A higher F1 score reflects a robust model that is a requirement for the medical industry, due to the high stakes and consequences associated with health.

B. Data Understanding

This step involves exploring the raw NHANES dataset which contains data on patient demographics, lab results, diet, examination results, used medications and patient questionnaires. Initial exploration is focused on identification of relevant features by accessing literature and using Random Forest for feature importance, as well as understanding the data distributions, and assessing data quality.

C. Data Preparation

In this phase, the dataset will be prepared for modelling by addressing missing values, outliers, and issues such as multicollinearity. Feature engineering techniques like feature scaling and log transforming will be applied. Further transformations, such as one hot encoding and the application of clustering and association rules mining to enhance the predictive power of the features, with the goal of improving the models performance.

D. Modelling

This project can be understood as an multi-class classification problem. As such, several models will be compared, including Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), artificial neural networks (ANN), and logistic regression. The dataset will be split into training and testing sets. Hyperparameter tuning using grid search will be conducted to optimise model performance for the target F1 score metric. A 5-fold cross-validation is also applied to reduce the risk of over-fitting. The focus will be on identifying the best-performing model based on the aforementioned F1 score and other complementary metrics.

E. Model Evaluation

The primary evaluation metric for this project is the F1 score, as explained in the Business Understanding section. Additionally, supporting metrics such as precision, recall, and accuracy are considered. By comparing models based on these metrics, the best-performing and, therefore, clinically applicable model will be assessed to determine if both the data science and clinical requirements have been met.

IV. Experimental Evaluation

A. Dataset Description

The provided NHANES dataset consists of six csv files. From these six available, this study makes use of five of them: namely `demographic.csv`, `labs.csv`, `diet.csv` and `examination.csv`. They are merged on the `SEQN` column, as it is a shared unique identifier amongst all of the datasets. The selected target variable for this project is, `MCQ195`, and is found within the `questionnaire.csv` file. The relevant values for this variable are: 0: No Arthritis, 1: Osteoarthritis or degenerative arthritis, 2: Rheumatoid arthritis, 3: Psoriatic arthritis. As well as; Other (4), Refused (7) or Don't Know (9).

After merging all of the datasets, the dataset consists of 5588 rows.

Inline with the literature review, the list below outlines all of the predictor variables that have been selected. Some features

have also been added based on a Random Forest feature importance analysis. Random Forest is an ensemble learning method, which is also used for feature selection, thanks to its inherent ability to rank features based on their importance [8].

- `MCQ195` - Type of arthritis diagnosed.
- `RIAGENDR` - Gender of the participant.
- `RIDAGEYR` - Age at the time of screening (top-coded at 80 years for individuals aged 80 and above).
- `INDHHIN2` - Total household income (reported as a range in dollars).
- `HSD010` - General health status, according to interviee.
- `OSQ080` - Doctor-diagnosed bone fracture after age 20.
- `MCQ080` - Diagnosed as overweight by a doctor or health professional.
- `DLQ050` - Serious difficulty walking or climbing stairs.
- `DLQ060` - Difficulty dressing or bathing.
- `DMQMILIZ` - Served on active duty in the U.S. Armed Forces, Reserves, or National Guard (excluding training; includes activation for service or operations).
- `SMQ020` - Smoked at least 100 cigarettes in a lifetime.
- `BPQ020` - Diagnosed with high blood pressure (hypertension).
- `BPQ080` - Diagnosed with high blood cholesterol.
- `DIQ010` - Diagnosed with diabetes (excluding during pregnancy).
- `MCQ070` - Diagnosed with psoriasis.
- `MCQ082` - Diagnosed with celiac disease (also called sprue).
- `HUQ051` - Number of doctor visits in the past year (excluding hospital stays, ER visits, home visits, or phone calls).
- `HUQ071` - Hospitalized overnight in the past year (excluding ER stays).
- `WTMEC2YR` - Full sample 2-year MEC exam weight.
- `INDFMPIR` - Ratio of family income to poverty guidelines.
- `WTDRD1` - Dietary Day 1 sample weight.
- `LBXSAPSI` - Alkaline phosphatase (IU/L).
- `URDACT` - Albumin creatinine ratio (mg/g).
- `URXCRS` - Creatinine in urine (µmol/L).

B. Experimental Settings

Evaluation Criteria: As mentioned in the Business Understanding, the main metric is the F1 score, with other metrics used as auxiliary measures.

Data Preparation: As part of the data preparation process, null values were either removed or replaced. For the target feature, missing values were replaced with 0, indicating no arthritis. The numerical features were then analysed using a pair plot to check for skewness, which led to the application of log scaling or MinMax scaling as needed. A correlation heatmap was then generated to identify any (too) highly correlated features, but none were found in this case. Following this, the data was clustered to uncover potential insights, and association rule mining was employed to enhance predictive power.

Clustering and Association Rule Mining: As mentioned, in order to improve the predictive power of the models, clustering and association rule mining techniques were used.

Clustering: The K-Means clustering algorithm was used to identify natural groupings within the data. The elbow method was applied to determine the optimal number of clusters, which was found to be 5. The silhouette score was then calculated to evaluate the quality of the clusters. For visualisation, PCA was used to reduce the dimensionality of the data. Additionally, hierarchical clustering with the average linkage method was performed, and a dendrogram was plotted to visualize the hierarchical structure. This method however resulted in a lower silhouette score, as such the K-Means clusters were used for later feature engineering.

Association Rule Mining: This process involved using the FP-Growth algorithm to extract frequent itemsets within the dataset, with a minimum support threshold of 0.05 to exclude infrequent itemsets. Additionally, continuous variables were transformed into binary categories based on their median values: 1 if the value exceeded the median, and 0 otherwise. Using the identified frequent itemsets, association rules were then generated. These rules were further filtered to include only those with the target variable (arthritis) in their consequents. The confidence values of these filtered rules were subsequently added to the dataset as a new feature, and the top three rules were interpreted.

Training and Test Data: This project makes use of the NHANES dataset to build and evaluate the predictive model. Following the cleaning and preparation process, a 70-30 train/test split is applied to ensure the model has sufficient data for training and evaluation. However, as shown in Figure 2, the dataset exhibits significant skewness. To address this imbalance, we will implement data balancing techniques to counteract this. Though, the '3.0' arthritis class is tiny and can be considered too imbalanced.

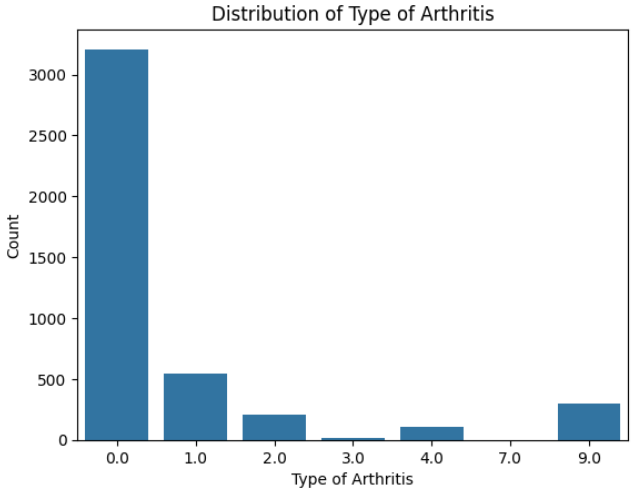

Fig. 2: Distribution of Arthritis Type in dataset

C. Results

This section will present the key findings relevant to addressing the research questions, which will be analysed further in the discussion section.

First, related to the first (association rules) research question. Many association rules were identified, however when sorting by confidence and extracting the top 3, the following interpretation is made:

- Individuals with good general health, in Cluster 4, and who have not smoked 100 cigarettes in their lifetime are strongly associated with having no arthritis.
  Support: 6.3%, **Confidence:** 98.4%, **Lift:** 1.22.
  This indicates that 6.3% of the population satisfies this rule, with a 98.4% likelihood that the outcome (no arthritis) is true when the conditions are met.
- Not overweight individuals in Cluster 4 with good general health are highly likely to report no arthritis.
  Support: 8.2%, **Confidence:** 97.9%, **Lift:** 1.21.
  This means that 8.2% of the population is represented by this rule, and when the conditions are met, there is a 97.9% chance of no arthritis.
- Not overweight individuals in Cluster 4 with moderate doctor visits in the past year are closely associated with having no arthritis.
  Support: 5.5%, **Confidence:** 97.8%, **Lift:** 1.21.
  This rule covers 5.5% of the population, with a 97.8% probability of no arthritis under the specified conditions. The lift of 1.21 indicates a meaningful positive association.

Four machine learning models were compared in this study. Each model had its hyperparameters tuned. Random Over Sampling (ROS) was applied, as it performed better than SMOTE. The target metric for this project is F1 score, so it is compared to the results from all models in the image below. The Macro (arithmetic mean) F1 score is used, as the dataset is heavily imbalanced, and it is important to predict the different arthritis types accurately, weighted results can be misleading.

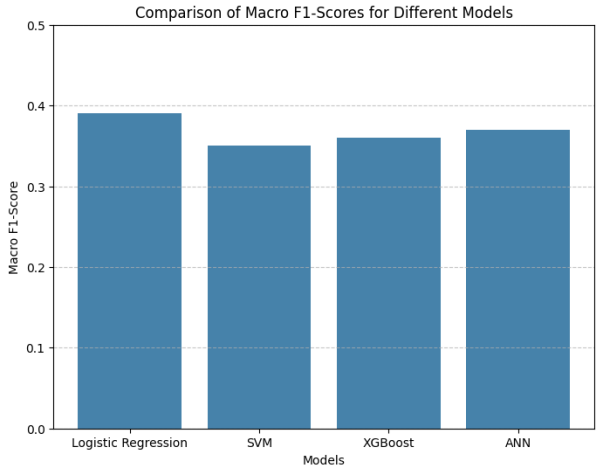

Fig. 3: Macro F1 Scores

From the results above, we can see that the Logistic Regression model has performed the best. It is clear that the general results of all the models are poor and around the 0.35 range. This is a clear indication that there is not enough predictive power in the predictor selected variables. Nonetheless, with the Logistic Regression being the best performer, the confusion matrix can give more information.

The confusion matrix is pictured in Figure 4. In which it is clear that the majority class (no arthritis) seems to perform quite well, but for the other classes this performance begins to worsen, with the worst being for the Psoriatic arthritis.

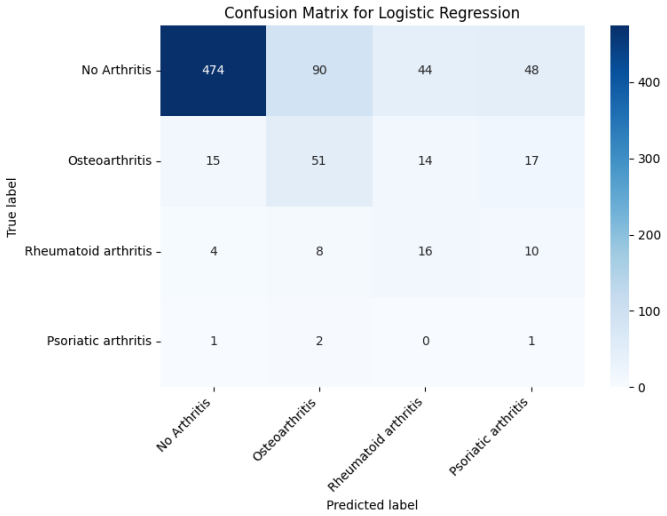

Fig. 4: Best Model Confusion Matrix

From the confusion matrix, relevant metrics can be calculated and have been assembled in Table 1.

TABLE I: Classification Report for Logistic Regression

Class	Precision	Recall	F1-Score	Support
No Arthritis	0.96	0.72	0.82	656
Osteoarthritis	0.34	0.53	0.41	97
Rheumatoid arthritis	0.22	0.42	0.29	38
Psoriatic arthritis	0.01	0.25	0.03	4
Accuracy			0.68	795
Macro Avg	0.38	0.48	0.39	795
Weighted Avg	0.84	0.68	0.74	795

The results show that the model performs well for the majority ”No Arthritis” class but struggles significantly with the minority classes, particularly ”Psoriatic arthritis,” likely due to the small sample size. While the overall metrics are boosted by the dominant class.

Next, although logistic regression assigns coefficients to features, it does not inherently clarify which terms most strongly drive the outcome. Instead these can be extracted from the XGBoost classifier, as features are ranked by their occurrence in the used trees. Since the performance delta between the two models is not so high, value can be gained from what features were influential in the classification:

TABLE II: Top 5 Feature Importances for XGBoost

Feature	Importance
doctor_visits_last_12_months_7	0.0591
dietary_day_one_sample_weight_above_median	0.0572
cluster_1	0.0400
age_above_median	0.0377
annual_household_income_2.0	0.0365

This suggests that the general health and situation of the person are strong indicators of the arthritis status of the person.

V. Discussion

The study explored the potential of machine learning models to classify arthritis subtypes within the NHANES dataset, yielding several insights but also underlined the general poor performance of the models. The Logistic Regression was identified as the best-performing model, but its overall effectiveness remained poor, with a macro F1 score of 0.39. When looking deeper at the results, it revealed that one of the classes (Psoriatic Arthritis) performed extremely poorly, with an F1 score of just 0.03. This delta when compared to the 0.82 score of the majority class indicates that the predictive variables were insufficiently representative of the distinct characteristics of the different arthritis types.

The dataset itself has posed a significant challenge in this project. While it is very comprehensive with many variables, the data distribution of the target variable is heavily imbalanced, with the "No Arthritis" class overwhelmingly dominating the dataset. Even with data balancing techniques like Random Over Sampling, the results did not improve significantly. This suggests that the issue seems to extend beyond simple class imbalance. However again, in the case of Psoriatic Arthritis, there were only 12 values of that class in the cleaned dataset, so naturally, the balancing techniques were very ineffective as there was not enough variation in the data.

The clustering and association rule mining provided some insights. For instance, individuals in good general health, who had not smoked, and were not overweight showed a strong association with the absence of arthritis. These patterns primarily confirmed findings documented in the literature review rather than offering new, unseen predictive power specific to arthritis subtypes. However, as noted by [9], distinguishing between types of arthritis can be challenging due to symptom overlap, which often leads to misdiagnosis. This challenge is supported by the evidence but may also be influenced by the aforementioned severe class imbalance. The variables used in this study are prevalent across different types of arthritis, making it difficult to differentiate between them based solely on clinical presentation. Additional data, such as MRI scans, as utilised in one of the studies outlined in the Related Works section, may be required for more accurate classification.

When examining the feature importances identified by the XGBoost algorithm, the results mostly align with prior research but fail to provide strong predictive differentiation for arthritis subtypes. This suggests that the current features selected from the dataset lack the information needed to effectively distinguish between the types of arthritis. Interestingly, the created cluster variable emerged as the third most important feature for the model. This highlights that applying clustering during the data exploration phase can yield results that, while not immediately apparent, may become significant or relevant.

In general, alternative approaches could be considered - such as with combination of data more related to subtypes of arthritides or using some more advanced algorithms could prove beneficial.

VI. Conclusion

Overall, the performance exhibited by the machine learning models in classification of the arthritis subtypes proved limited. While the performance was better than random chance, this is not enough in the medical field, where the consequences of wrong diagnosis can be extremely high and damaging to patients wellbeing. Nonetheless, while the Logistic Regression had performed the best among the tested approaches, future work could benefit from incorporation of other models, such as deep learning. Moreover, the inherent class imbalance has proved to be a large drawback in this implementation. Incorporation of more specific features relevant to arthritis subtypes could also enhance the model's effectiveness. Furthermore, inclusion of more domain specific insight, for example from knowledge experts, such as doctors could give the much needed domain insight into making better decisions that would boost the model performance.

[10]

References

[1] E. A. Fallon, M. A. Boring, A. L. Foster, E. W. Stowe, T. D. Lites, E. L. Odom, and P. Seth, "Prevalence of diagnosed arthritis — united states, 2019–2021," *MMWR Morbidity and Mortality Weekly Report*, vol. 72, pp. 1101–1107, 10 2023. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10578950/

[2] Centers, "National health and nutrition examination survey," Kaggle.com, 2014. [Online]. Available: https://www.kaggle.com/datasets/cdc/national-health-and-nutrition-examination-survey

[3] E. Sweeney, "3 main types of arthritis—rheumatoid, psoriatic and osteoarthritis," JNJ.com, 10 2023. [Online]. Available: https://www.jnj.com/health-and-wellness/arthritis-3-main-types-rheumatoid-psoriatic-osteoarthritis

[4] J. Zhan, F. Chen, Y. Li, and C. Huang, "Risk prediction model for psoriatic arthritis: Nhanes data and multi-algorithm approach," *Clinical Rheumatology*, 11 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/39585569/

[5] CDC, "Arthritis risk factors," Arthritis, 06 2024. [Online]. Available: https://www.cdc.gov/arthritis/risk-factors/index.html

[6] L. Folle, D. Simon, K. Tascilar, G. Krönke, A.-M. Liphardt, A. Maier, G. Schett, and A. Kleyer, "Deep learning-based classification of inflammatory arthritis by identification of joint shape patterns—how neural networks can tell us where to "deep dive" clinically," *Frontiers in Medicine*, vol. 9, 03 2022. [Online]. Available: https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2022.850552/full

[7] N. Hotz, "What is crisp dm? - data science pm," Data Science PM, 09 2018. [Online]. Available: https://www.datascience-pm.com/crisp-dm-2/

[8] R.-C. Chen, C. Dewi, S.-W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," *Journal Of Big Data*, vol. 7, 07 2020. [Online]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00327-4

[9] W. Saalfeld, A. M. Mixon, J. Zelie, and E. J. Lydon, "Differentiating psoriatic arthritis from osteoarthritis and rheumatoid arthritis: A narrative review and guide for advanced practice providers," *Rheumatology and Therapy*, vol. 8, pp. 1493–1517, 09 2021. [Online]. Available: https://link.springer.com/article/10.1007/s40744-021-00365-1

[10] "Chatgpt used for grammar in the report," ChatGPT, 2024. [Online]. Available: https://chatgpt.com/share/6756124b-d174-800d-93a6-d2dc2846e74a