ICR - Identifying Age-Related Conditions (7313 Project Report)

Yasamin Ghahremani
zub11@txstate.edu

Tanzina Akter Tani
tanzinatani@txstate.edu

1 Abstract

Although aging is a fundamental and undeniable feature of life, there is an increasing awareness that emphasizes its diverse nature. In this ICR Kaggle project, we attempt to understand and identify age-related conditions through the use of various Machine Learning (ML) models. For this purpose, we have pre-processed the relative tabular data, and applied a total of 10 individual and ensemble classification methods. Based on our analysis, the accuracy of the Stacking, Majority voting, Bagging with Catboost, Random forest and XGBoost classification models showcased the highest results when applied class weight for imbalanced data. When the SMOTE is applied for class balancing, the Bagging with Catboost gave the best accuracy.

2 Introduction

In recent years, while there have been advances in living conditions and scientific progress, the percentage of our lives where we spend in good health has essentially remained the same since 1990. This means that even though we're living longer, we're often spending the remainder of our lifespans with years marked by health challenges [15].

The aging process of humans is a multifaceted biological phenomenon that unravels over time, and is marked by a complex interplay of various factors. This time-dependent progression involves a gradual decline in both cognitive and physiological functions. This leads to a decreased capacity to respond effectively to stressors [8]. Based on reports from the World Health Organization, "Between 2000 and 2050, the proportion of the world's population over the age of 60 years will double from around 11 to 22 percent. The absolute number of people aged over 60 years is also expected to increase from 605 million to 2 billion over the same period" [9]. Not only thus but the landscape of health challenges has also shifted over time, moving from infectious diseases to chronic non-communicable ones, many of which are closely tied

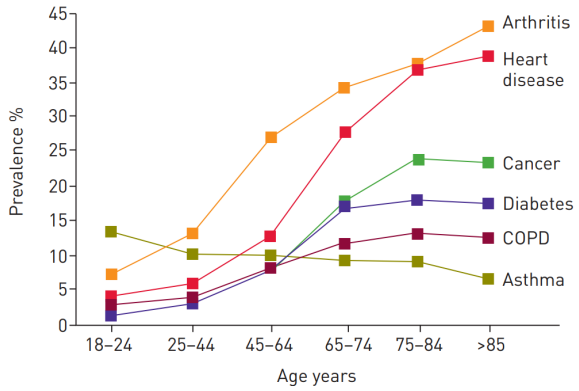

Figure 1: Growth and prevalence of some chronic conditions as a function of age [9]

to advancing age. Additionally, the global mental well-being of older adults has seen little improvement over the past three decades [13]. Therefore, analyzing the impact of these diverse experiences on humans' biological susceptibilities, and the pace at which they affect health outcomes on individuals has emerged as a key focus in the realm of biological aging research.

One notable aspect is that chronological age stands out as a significant risk factor for conditions like frailty, age-related morbidity, and mortality. Despite this general association, there exists considerable diversity in health outcomes among individuals sharing the same chronological age. This variability suggests that the aging process at an individual level is marked by heterogeneity, emphasizing the need for a nuanced understanding of the factors influencing aging and its impact on health [10].

Established in 2015, InVitro Cell Research (ICR) is a privately funded company whose work focuses on regenerative and preventive personalized medicine. Their mission is to explore interventions that slow and potentially reverse the biological aging process, while also preventing prevalent age-related diseases. Most recently, they hosted a coding competition entitled "Identify Age-Related Conditions" on Kaggle, which attempts to address this problem and develop an accurate model for identifying age-related conditions based on health records. This competition has been defined as a binary classification problem.

3 Dataset

3.1 Explanatory Data Analysis (EDA)

In order to tackle with the problem defined, we first explored the dataset, it's features. We started with the analysis and exploration of the characteristics of the train dataset. We found that the train dataset included 617 rows of data, along with 58 features

Feature	AB	AF	AM	AR	AX	AZ	BC	BD
mean	0.47	3502.01	38.96	10.12	5.54	10.56	8.05	5350.38
std	0.46	2300.32	69.72	10.51	2.55	4.35	65.16	3021.32
min	0.08	192.59	3.17	8.13	0.70	3.39	1.23	1693.62
max	6.16	28688.18	630.51	178.94	38.27	38.97	1463.69	53060.59

Table 1: Some of the differences observed in feature values, along with their Max, Min, Mean, and Standard deviation

representing health characteristics of the patients. The dataset also included an "id" column, individually defined for each patient, that due to the nature of the problem, was later removed. Among the columns were 56 anonymized health characteristics, most represented by two-letter names and numerical, while an "EJ" was also defined that included categorical values of "A" and "B". The final column showcases the binary label of each row of data, with the label of 0 representing not having any of the defined medical conditions, and 1 indicating the individual having the defined medical condition associated with aging.

We then started to statistically analyze the values in the train dataset and its columns, and found that there was a large difference in the numerical values of different features, leading us to follow what we had learned in class and later apply scaling to the dataset.

In the next stage, we attempted to study and explore the distribution of the features, with special attention to their class (target) values. This stage wasmainly executed using the Seaborn library in Python. As shown in the plots in Figure 2, we observed that the data for the classes follows a similar distribution, yet for the class of "0", the data experiences spikes that indicate higher values than those spikes for the other class.

Seeing this distribution, we decided to explore the balance of the train dataset, and to understand how it is. For this purpose, the pie chart showcased in Figure 3 distributions was extracted. Based on the chart, 82.5 percent of the train dataset has the label of 0, while only 17.5 percent have a label of 1.

In addition, to gain a better insight on the correlations among the variables, the following heatmap of correlations among the variables was extracted and shown in Figure 4. Among the most noted correlations were among "CL" and "DV", "FD" and "EH", and "BD" and "CS".

3.2 Submission Data

To test and compare the results, the hosts of the competition have provided a test dataset which has 5 rows of features without any class labels. Therefore, it is only

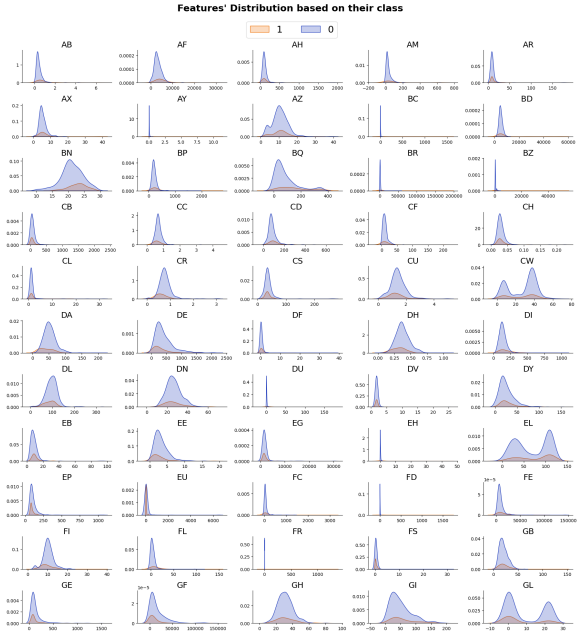

Figure 2: Features Distribution based on their class

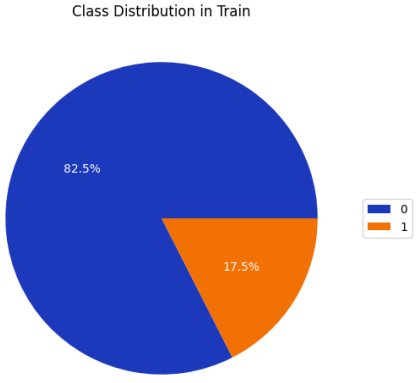

Figure 3: Class imbalance was observed in the train dataset

Class	Feature	Missing Count
0	BQ	60
0	CB	2
0	CC	2
0	DU	1
0	EL	54
0	FL	1
0	GL	1
1	CC	1
1	EL	6
1	FC	1
1	FS	1

Table 2: Missing Counts for Features in Each Class

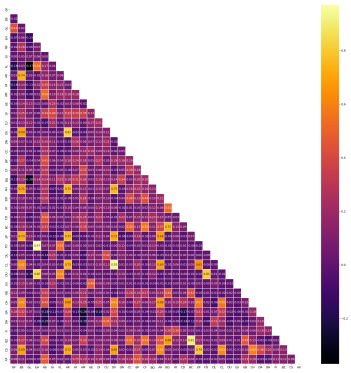

Figure 4: Heatmap of correlations among variables in the Train dataset

used for the submission to the Kaggle ICR competition. We will refer to this test set as 'submit-test' for clarity.

4 Methodology

For the purpose of the competition, we followed what we had learned throughout the Machine Learning course and designed the methodology showcased in Figure 5.

4.1 Data Preprocessing

This section will discuss how the dataset is preprocessed. In the train set, there are in total 131 missing values. The missing values correspond to feature columns and classes are shown in Table2. As previously mentioned, the train and submit test sets have only one categorical column – "EJ" with 'A' and 'B' values. It is converted to binary 0 and 1 respectively. The Id of patients are dropped in the preprocessing step for both provided sets. This column is only added on the result to the submission file on Kaggle.

To handle the missing values of the training set, two processes have been considered, mean and median. At first, all the blank spaces are filled with 'nan' and then based on the mean of each feature column, the blank are replaced with the mean value. Same procedures are adopted for median as well. The step has been done with the help of SimpleImputer library in Python. The train set is then split into X for features and y for labels and converted to a Numpy array. The train set is split into train and test with a ratio of 75:25, to enable us to evaluate the model's result with the test set. For data imbalance problems, two techniques are considered, one is class weight without balancing the dataset, another is the SMOTE Oversampled method with RandomOverSampler to oversample the minority class 1.

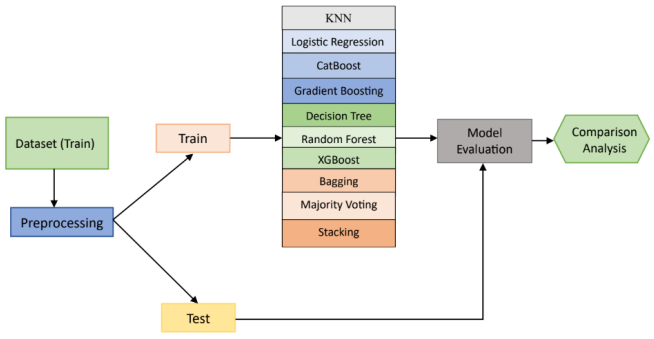

Figure 5: Research project methodology

4.2 Machine Learning models for Classification

K-Nearest Neighbor classifier:
The KNN method is probably among one of the most prominent methods for classification. For this classifier, the number of neighbors K must first be carefully selected, then the distance between the query instance and current instance from the data are calculated. Different distance metrics can be used, and among those include methods are some of the ones discussed during the course, such as Euclidean, Manhattan, Hamming. For this purpose, the distance and index of the instances are collected as sorted in ascending order and k entries are picked from it. Based on the picked entries, mode of the labels corresponding to K are returned for classification problems [6].

Logistic Regression:
Logistic regression is a machine learning algorithm that predicts the probability of the class belonging. It analyzes the relationship between independent and dependent variables. It works for the categorical or discrete value. The continuous output of linear regression is transformed to categorical output using Sigmoid function and gives the probability result in between 0 and 1 range [3].

Decision Trees:
Here, the most relevant features are selected through splitting the data based on the most informative criteria like Gini impurity or information gain. This process of splitting continues recursively until the stopping conditions are met. The leaf nodes of the resulting tree are the class labels, and the tree is traversed for the new data points to predict the class [14].

Random Forest:
The Random Forest algorithm is an ensemble learning technique based on Decision Trees. It entails building numerous decision trees during training and combining their predictions. Random forest incorporates randomness by taking into account a random subset of characteristics at each split and by training each tree on a random sample of the data, which prevents the overfitting and creates a more robust model [1].

Gradient Boosting:
Another ensemble learning technique that builds a predictive model using an ensemble of weak learners, typically decision trees. Contrary to Random Forest, Gradient Boosting builds trees sequentially, with each tree correcting the errors of the previous one. By gradually including weak learners into the group, the approach reduces a loss function [2].

XGBoost:
Extreme Gradient Boosting or XGBoost is a highly optimized implementation of gradient boosting. It operates through adding Decision Trees to an ensemble in a sequential manner, and each tree correcting the errors of the preceding ones. With capabili-

ties including handling of missing values, regularization, parallelization for efficiency, and custom objective functions, it improves classic gradient boosting. Through iteratively improving the model's predictions, the technique optimizes a defined loss function. Because of its efficiency and speed, XGBoost has gained a lot of popularity and is now frequently used for machine learning tasks, particularly in competitive and real-world settings [4].

Catboost:

CatBoost is a gradient boosting method designed to handle category features efficiently. It excels the analysis of real-world datasets by processing categorical variables automatically without manual encoding. Similar to Gradient Boosting, this method adds Decision Trees progressively to minimize a defined loss function. For faster training, it adds innovations such as ordered boosting and leverages advanced techniques such as oblivious trees [5].

Bagging:

Known as Bootstrap Aggregating, this method improves the performance of a base model by bootstrapping several instances on different subsets of the data. The final prediction is often derived through averaging in regression or majority voting in classification, which takes advantage of the variety introduced by different data subsets to increase generalization and reduce overfitting [11].

Majority Voting:

Majority voting is a type of ensemble approach in which each model predicts a class independently, and the final prediction is based on the class that obtains the most votes. It offers a simple and effective method for combining predictions from numerous models in a classification ensemble, harnessing their combined capabilities for more robust decision-making [12].

Stacked Generalization:

Stacking is an ensemble learning strategy that entails training multiple models and integrating their predictions to form a meta-model. Stacking, unlike bagging or boosting, has a two-level architecture. Base models are trained using training data at the first level, and their predictions become input characteristics for a higher-level model known as the meta-model or blender. To make the final forecast, the meta-model learns to blend the predictions of the basis models. Stacking takes advantage of the diversity of base models to capture different features of the data, which can result in higher predictive performance [7].

5 Results and Discussions

The two combinations of pre-processing are used for all the classifiers - 1. median for missing value, class weights to handle imbalance. 2. mean for missing value,

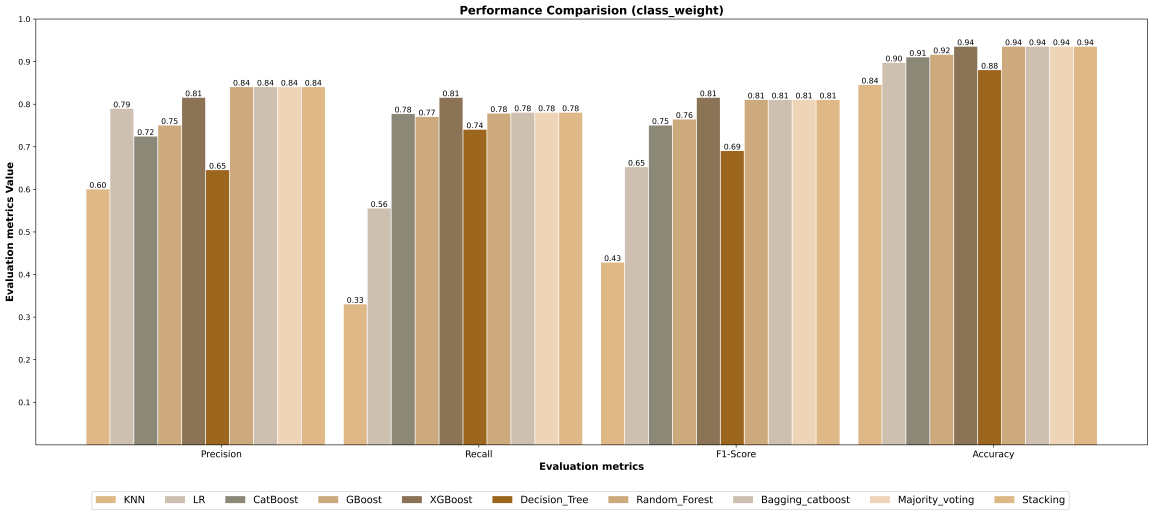

Figure 6: Performance metrics of results obtained through our analysis

SMOTE to handle imbalance. The results are measured with the scale of 0 to 1. The gridsearch method is applied to get the best combination of parameters of the classification models.

1. Median and class weights pre-process: From Figure 6, we can see the last four classifiers random forest, bagging (catboost), majority voting and stacking gave the same precision (0.84), recall (0.78), f1-score (0.81) and accuracy (0.94). For XGBoost, the f1-score and accuracy are the same as the last four classifiers but the precision is 0.81 and recall is 0.84. After the above results, the Catboost and Gboost gave the best result by giving 0.75 and 0.76 f1-score respectively. The KNN model gave the worst result for each metric. For logistic regression (LR), even though it gave a good precision score (0.79), the recall (0.56) and f1-score (0.65) are low which means the model can not handle the under-sampled class 1. After KNN and LR, the decision tree gave the worst result.

The balanced logarithmic loss is used as the submission evaluation metric for the ICR kaggle competition. The submission result using the submit-test dataset gave us the best private score of 0.459 and public score of 0.23.

2. Mean and SMOTE pre-process: For this step, the dataset were scaled using standard scaling. The bagging with catboost gave the best precision (0.89), recall (0.77), f1-score (0.83) and accuracy (0.94). Except KNN, LR and tree, all other classifiers have almost the same results. The decision tree gave the 0.68 f1-score and 0.87 accuracy which is better than KNN but worse than all other methods. The logistic regression gave almost the same results for f1-score and accuracy. However,

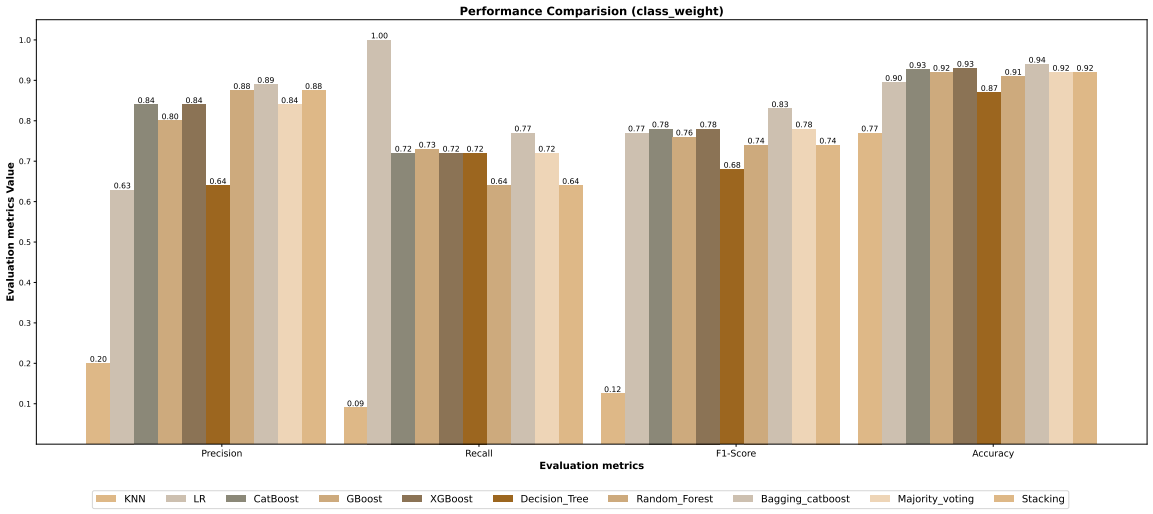

Figure 7: Performance metrics of results obtained through our SMOTE analysis

the recall is showing 1 but the precision is low (0.63). It implies that the model detects practically all positive instances but also generates a significant number of false positive predictions. The model's high recall value shows that it is sensitive to positive cases but not precise in its predictions.

The KNN classifier gave the worst result even though the accuracy is 0.77. It shows us that accuracy is not a good metric for model evaluation. The precision, recall and f1-score is very low. The reason is even though the dataset is balanced with SMOTE, it can generate unrealistic synthetic or noise samples for class 1. Therefore, the KNN model could not be classified correctly for class 1. The effect on SMOTE is also shown in the final submission of the submit-test for ICR competition. It gave us a balanced log loss of 0.72 for both private and public scores.

6 Conclusion

This ICR Kaggle project explores the multifaceted nature of aging by employing various Machine Learning (ML) models to analyze and identify age-related conditions. The study involved the preprocessing of tabular data and the application and ensemble of different classification methods. The analysis revealed that Stacking, Majority Voting, Bagging with Catboost, Random Forest, and XGBoost classification models demonstrated the highest accuracy when implemented with the median imputation of the missing values and class weight adjustment for imbalanced data. Notably, when the Synthetic Minority Over-sampling Technique (SMOTE) was applied for class balancing, the bagging with Catboost demonstrated the highest accuracy. These findings contribute to the growing awareness and need for study of the diverse aspects of ag-

ing and highlight the efficacy of specific ML approaches in discerning age-related conditions.

References

[1] Ahmad Taher Azar, Hanaa Ismail Elshazly, Aboul Ella Hassanien, and Abeer Mohamed Elkorany. A random forest classifier for lymph diseases. *Computer methods and programs in biomedicine*, 113(2):465–473, 2014.

[2] Candice Bentéjac, Anna Csörgő, and Gonzalo Martínez-Muñoz. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54:1937–1967, 2021.

[3] Ekaba Bisong and Ekaba Bisong. Logistic regression. *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, pages 243–250, 2019.

[4] Najmeddine Dhieb, Hakim Ghazzai, Hichem Besbes, and Yehia Massoud. Extreme gradient boosting machine learning algorithm for safe auto insurance operations. In *2019 IEEE international conference on vehicular electronics and safety (ICVES)*, pages 1–5. IEEE, 2019.

[5] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*, 2018.

[6] Evelyn Fix and Joseph Lawson Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, 57(3):238–247, 1989.

[7] Hyunjin Kwon, Jinhyeok Park, and Youngho Lee. Stacking ensemble technique for classifying breast cancer. *Healthcare informatics research*, 25(4):283–288, 2019.

[8] David J Lowsky, S Jay Olshansky, Jay Bhattacharya, and Dana P Goldman. Heterogeneity in healthy aging. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, 69(6):640–649, 2014.

[9] William MacNee, Roberto A Rabinovich, and Gourab Choudhury. Ageing and the border between health and disease. *European Respiratory Journal*, 44(5):1332–1352, 2014.

[10] Aung Zaw Zaw Phyo, Rosanne Freak-Poli, Heather Craig, Danijela Gasevic, Nigel P Stocks, David A Gonzalez-Chica, and Joanne Ryan. Quality of life and mortality in the general population: a systematic review and meta-analysis. *BMC public health*, 20(1):1–20, 2020.

[11] J Ross Quinlan et al. Bagging, boosting, and c4. 5. In *Aaai/Iaai, vol. 1*, pages 725–730, 1996.

[12] Artittayapron Rojarath, Wararat Songpan, and Chakrit Pong-inwong. Improved ensemble learning for classification techniques based on majority voting. In *2016 7th IEEE international conference on software engineering and service science (ICSESS)*, pages 107–110. IEEE, 2016.

[13] Andrew J Scott, Martin Ellison, and David A Sinclair. The economic value of targeting aging. *Nature Aging*, 1(7):616–623, 2021.

[14] Yan-Yan Song and LU Ying. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2):130, 2015.

[15] Zhuoting Zhu, Danli Shi, Peng Guankai, Zachary Tan, Xianwen Shang, Wenyi Hu, Huan Liao, Xueli Zhang, Yu Huang, Honghua Yu, et al. Retinal age gap as a predictive biomarker for mortality risk. *British Journal of Ophthalmology*, 107(4):547–554, 2023.