MACHINE LEARNING - HOMEWORK 1
ENSEMBLE AND SVM MODELS FOR CLASSIFICATION TASKS

Cătălin-Alexandru Rîpanu 341C3
Automatic Control and Computers
National University of Science and Technology POLITEHNICA Bucharest
`catalin.ripanu@stud.acs.upb.ro`
Laboratory Assistant: Mihai Trascău

1 May 2024

Abstract

This project aims to explore the implementation of widely used machine learning ensemble algorithms (**Random Forest**, **Extra Trees**, **Support Vector Machine**, and **Gradient Boosted Trees**) employing a variety of techniques renowned for their simplicity and efficacy. The performance evaluation of these algorithms is multifaceted, considering crucial metrics including **Precision**, **Recall**, **Overall Accuracy**, and **F1 Score**. To facilitate implementation and execution, Python libraries **sklearn** and **xgboost** are leveraged, offering an extensive array of specialized functions tailored for Machine Learning tasks. This comprehensive approach seeks to provide insights into algorithmic behavior and performance across a range of evaluation criteria, fostering a deeper understanding of ensemble learning methodologies in practical contexts.

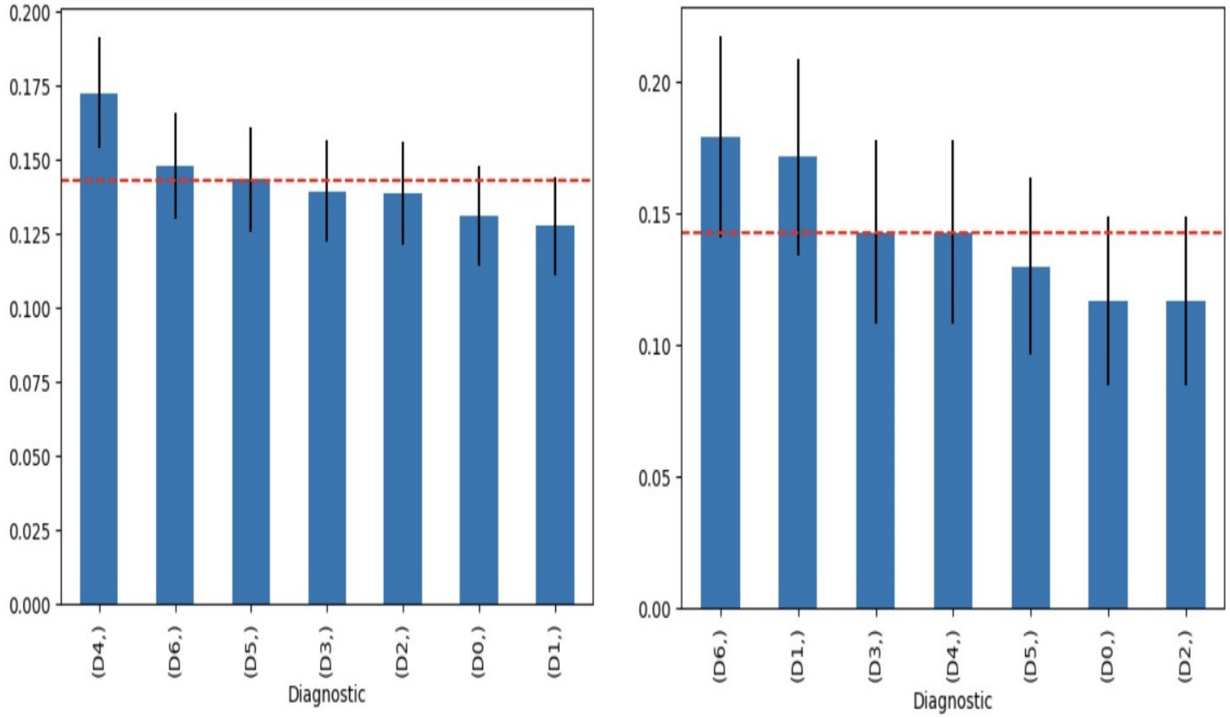
1 Data Analysis

The initial stage comprises thorough Data Analysis, commencing with an evaluation of class balance within both the training and test datasets. Following this assessment, detailed graphs and histograms are created to depict the distribution of attribute values, meticulously segmented into percentiles with a granularity of 10%. This graphical representation offers valuable insights into the underlying data structure. Finally, the significance of each attribute in predicting the target variable, labeled **"Diagnostic"**, is elucidated through individualized plots. These plots provide a nuanced understanding of attribute relevance, facilitating informed decision-making in subsequent modeling processes.

The analysis delves into the **Point-Biserial Correlation** and **Chi-Squared Statistic** to quantitatively determine the degree of correlation with the prediction target. This approach employs two distinct analyses, necessitated by the mixed nature of attributes within the dataset (some categorical, others purely numerical). By leveraging these statistical techniques, we gain a comprehensive understanding of the relationship between individual attributes and the prediction target, facilitating informed feature selection and model refinement.

The initial dataset underwent 5 random splits, ensuring unbalanced partitions each time. Subsequently, these divisions were allocated to the models of interest to assess their predictive capabilities relative to the information contained within the file. Through these tests, we aim to discern the sufficiency and relevance of the data concerning the target variable's prediction. This rigorous evaluation provides insights into the adequacy of available data and its utility in informing predictive modeling endeavors.

In general, when the p-value of attributes concerning another attribute is less than 0.05, it is commonly inferred that those attributes are dependent on each other. In this scenario, several attributes demonstrate correlation with the prediction target. Consequently, if these variables exhibit extreme values (outliers) and are not appropriately scaled, there is a risk that the model (whether it's an ensemble method or Support Vector Machine (SVM)) may fail to yield satisfactory results. This underscores the importance of preprocessing steps such as outlier detection and feature scaling to enhance the robustness and performance of the predictive model.



(a) Train Set

(b) Test Set

Figure 1: Class distribution in both datasets

A table for reporting the Point-Biserial value for all numerical attributes:

Feature	Point-Biserial Statistic Value	P-Value Statistic Value
Regular_fiber_diet	0.004	0.859
Sedentary_hours_daily	-0.014	0.524
Age	0.316	5.993e-46
Est_avg_calorie_intake	-0.038	0.093
Main_meals_daily	-0.126	2.802e-08
Height	0.113	5.472e-07
Water_daily	0.362	1.105e-60
Weight	0.592	2.952-182
Physical_activity_level	0.142	3.662e-10
Technology_time_use	-0.014	0.517

This table presents the Point-Biserial correlation values for each numerical attribute in the dataset. The Point-Biserial coefficient quantifies the correlation between a binary variable (the target variable) and a continuous variable (the numerical attribute). Positive values indicate a positive correlation, while negative values indicate a negative correlation. These values provide valuable insights into the relationship between each numerical attribute and the target variable, aiding in feature selection and model interpretation.

It's evident that, for instance, the attribute "Main_meals_daily" exhibits a correlation lower than 0.05, indicating its dependence on the target variable. Furthermore, it's the sole attribute negatively correlated with the target, suggesting that as a patient's food consumption decreases, the severity of the diagnosis increases (assuming D0 represents the least severe diagnosis and D6 the most severe).

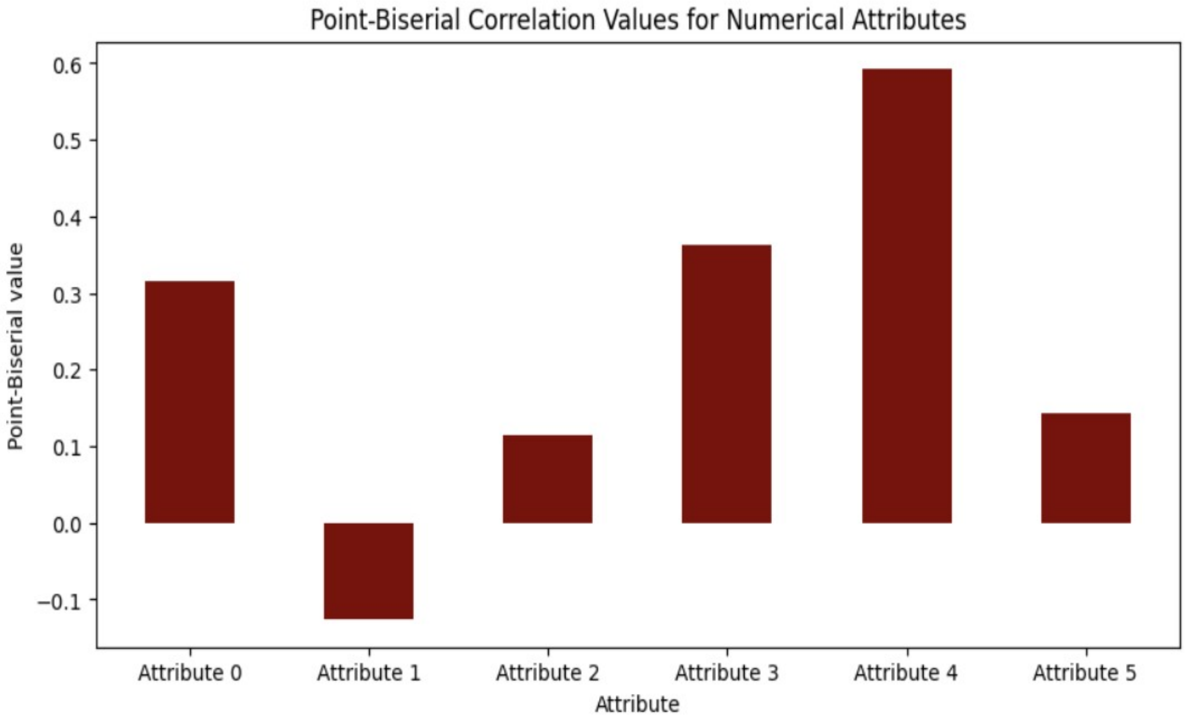


Figure 2: Feature correlations

Where:

- Attribute 0 -> Age
- Attribute 1 -> Main_meals_daily
- Attribute 2 -> Height
- Attribute 3 -> Water_daily
- Attribute 4 -> Weight
- Attribute 5 -> Physical_activity_level

A table reporting the Chi-Squared statistic values for all categorical attributes:

Gender	Chi-Squared Statistic Value	P-Value Statistic Value
Technology_time_use	594.505	7.853-118
Calorie_monitoring	111.023	3.542e-17
Smoker	30.310	0.006
Snacks	706.247	9.669e-131
Alcohol	313.742	4.536e-50
High_calorie_diet	208.210	1.149e-36
Diagnostic_in_family_history	565.486	1.165e-111
Transportation	275.744	3.519e-39

This table presents the Chi-Squared statistic values calculated for each categorical attribute in the dataset. The Chi-Squared test assesses the independence between categorical variables, making it particularly useful for understanding relationships between categorical attributes and the target variable. Higher Chi-Squared values indicate stronger associations between the categorical attribute and the target variable. Analyzing these values provides valuable insights into the relevance of each categorical attribute in predicting the target variable and aids in feature selection and model interpretation.

In this table, it's evident that all categorical attributes (whether ordinal or nominal) exhibit positive correlation with the target attribute "Diagnostic". This observation suggests a consistent pattern where the values of categorical attributes tend to align with the severity levels of the diagnosis. Such alignment underscores the potential predictive power of these categorical attributes in discerning the severity of diagnoses.

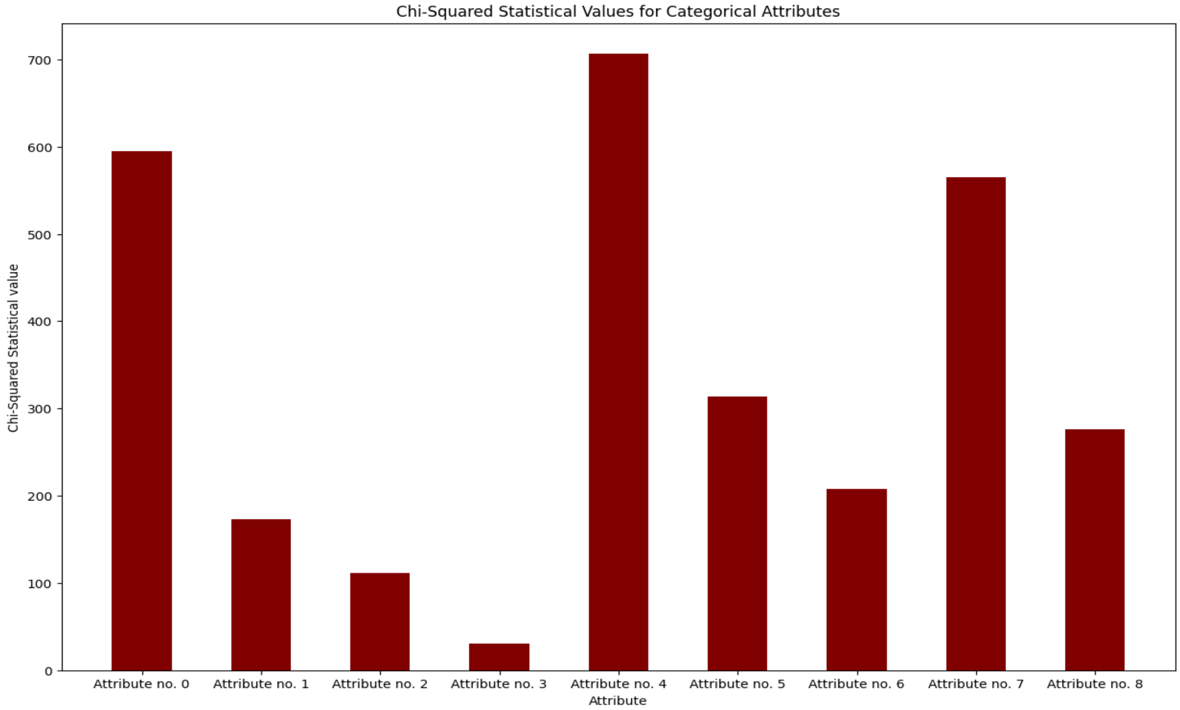


Figure 3: Feature correlations

Where:

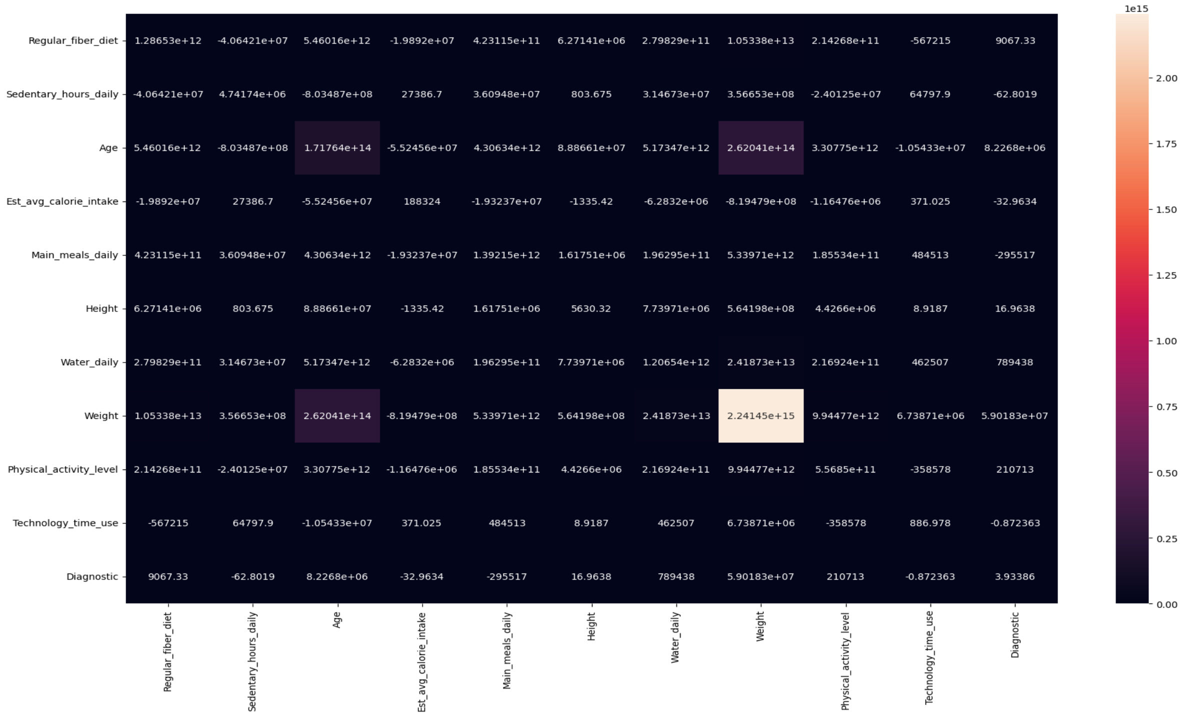
- Attribute no. 0 -> Gender
- Attribute no. 1 -> Technology_time_use
- Attribute no. 2 -> Calorie_monitoring
- Attribute no. 3 -> Smoker
- Attribute no. 4 -> Snacks
- Attribute no. 5 -> Alcohol
- Attribute no. 6 -> High_calorie_diet
- Attribute no. 7 -> Diagnostic_in_family_history
- Attribute no. 8 -> Transportation

Based on the histogram, it can be inferred that none of the categorical attributes exhibit negative correlation with the target variable.

This observation suggests that there are no instances where the values of categorical attributes inversely relate to the severity levels of the diagnosis. Understanding this absence of negative correlation aids in refining our understanding of the relationships between categorical attributes and the target variable, contributing to more accurate predictive modeling and decision-making processes in medical diagnostics or similar domains.

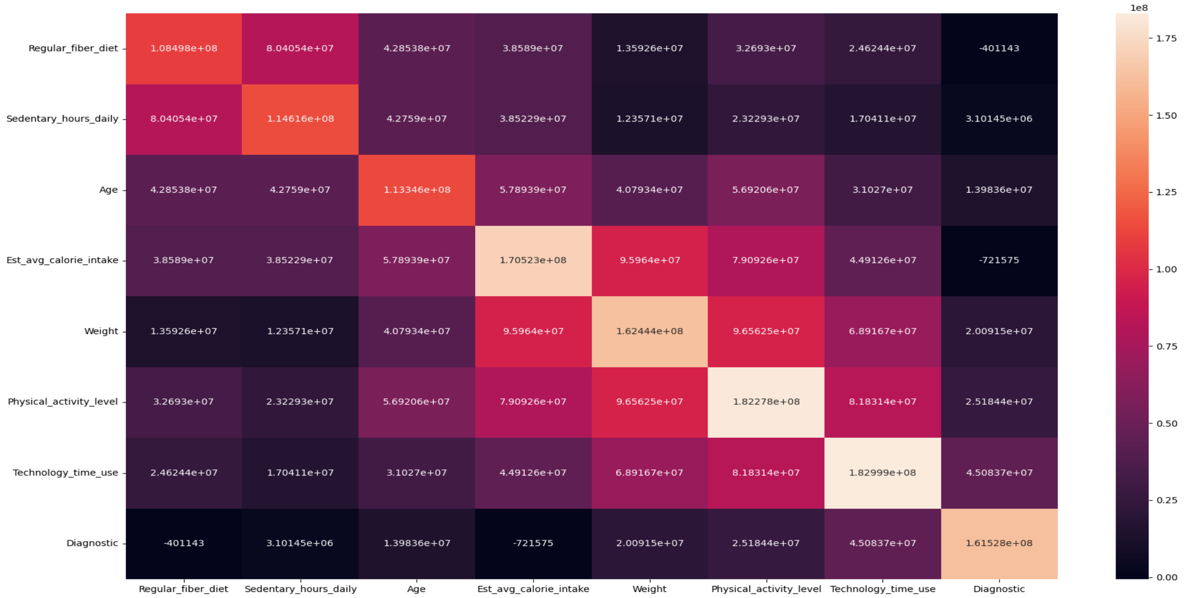
Additionally, the attribute "Weight" exhibits the highest variance compared to all other attribute variants. The highest covariance between attributes is observed between "Age" and "Weight". Moreover, all covariances with respect to the target attribute fall within the same domain, indicating no discrepancies in this regard.

This analysis highlights the variability and interrelationships among attributes, shedding light on their collective impact on predictive modeling outcomes.



	Regular_fiber_diet	Sedentary_hours_daily	Age	Est_avg_calorie_intake	Main_meals_daily	Height	Water_daily	Weight	Physical_activity_level	Technology_time_use	Diagnostic
Regular_fiber_diet	1.28653e+12	-4.06421e+07	5.46016e+12	-1.9892e+07	4.23115e+11	6.27141e+06	2.79829e+11	1.05338e+13	2.14268e+11	-567215	9067.33
Sedentary_hours_daily	-4.06421e+07	4.74174e+06	-8.03487e+08	27386.7	3.60948e+07	803.675	3.14673e+07	3.56653e+08	-2.40125e+07	64797.9	-62.8019
Age	5.46016e+12	-8.03487e+08	1.71764e+14	-5.52456e+07	4.30634e+12	8.88661e+07	5.17347e+12	2.62041e+14	3.30775e+12	-1.05433e+07	8.2268e+06
Est_avg_calorie_intake	-1.9892e+07	27386.7	-5.52456e+07	188324	-1.93237e+07	-1335.42	-6.2832e+06	-8.19479e+08	-1.16476e+06	371.025	-32.9634
Main_meals_daily	4.23115e+11	3.60948e+07	4.30634e+12	-1.93237e+07	1.39215e+12	1.61751e+06	1.96295e+11	5.33971e+12	1.85534e+11	484513	-295517
Height	6.27141e+06	803.675	8.88661e+07	-1335.42	1.61751e+06	5630.32	7.73971e+06	5.64198e+08	4.4266e+06	8.9187	16.9638
Water_daily	2.79829e+11	3.14673e+07	5.17347e+12	-6.2832e+06	1.96295e+11	7.73971e+06	1.20654e+12	2.41873e+13	2.16924e+11	462507	789438
Weight	1.05338e+13	3.56653e+08	2.62041e+14	-8.19479e+08	5.33971e+12	5.64198e+08	2.41873e+13	2.24145e+15	9.94477e+12	6.73871e+06	5.90183e+07
Physical_activity_level	2.14268e+11	-2.40125e+07	3.30775e+12	-1.16476e+06	1.85534e+11	4.4266e+06	2.16924e+11	9.94477e+12	5.5685e+11	-358578	210713
Technology_time_use	-567215	64797.9	-1.05433e+07	371.025	484513	8.9187	462507	6.73871e+06	-358578	886.978	-0.872363
Diagnostic	9067.33	-62.8019	8.2268e+06	-32.9634	-295517	16.9638	789438	5.90183e+07	210713	-0.872363	3.93386

Figure 4: Feature covariances



	Regular_fiber_diet	Sedentary_hours_daily	Age	Est_avg_calorie_intake	Weight	Physical_activity_level	Technology_time_use	Diagnostic
Regular_fiber_diet	1.08498e+08	8.04054e+07	4.28538e+07	3.8589e+07	1.35926e+07	3.2693e+07	2.46244e+07	-401143
Sedentary_hours_daily	8.04054e+07	1.14616e+08	4.2759e+07	3.85229e+07	1.23571e+07	2.32293e+07	1.70411e+07	3.10145e+06
Age	4.28538e+07	4.2759e+07	1.13346e+08	5.78939e+07	4.07934e+07	5.69206e+07	3.1027e+07	1.39836e+07
Est_avg_calorie_intake	3.8589e+07	3.85229e+07	5.78939e+07	1.70523e+08	9.5964e+07	7.90926e+07	4.49126e+07	-721575
Weight	1.35926e+07	1.23571e+07	4.07934e+07	9.5964e+07	1.62444e+08	9.65625e+07	6.89167e+07	2.00915e+07
Physical_activity_level	3.2693e+07	2.32293e+07	5.69206e+07	7.90926e+07	9.65625e+07	1.82278e+08	8.18314e+07	2.51844e+07
Technology_time_use	2.46244e+07	1.70411e+07	3.1027e+07	4.49126e+07	6.89167e+07	8.18314e+07	1.82999e+08	4.50837e+07
Diagnostic	-401143	3.10145e+06	1.39836e+07	-721575	2.00915e+07	2.51844e+07	4.50837e+07	1.61528e+08

Figure 5: Selected Features covariances

The **SelectPercentile** method with a score function of **mutual_info_classif** and a **percentile** value of 35 is employed for feature selection.

This method selects the top features based on their mutual information with the target variable in a classification task. Mutual information measures the amount of information gained about one variable through the other variable. By specifying a percentile value of 35, the algorithm retains the top 35% of features with the highest mutual information scores, effectively reducing the dimensionality of the feature space while preserving the most informative attributes for classification.

This feature selection technique helps improve model performance by focusing on the most relevant features and mitigating the effects of noise or irrelevant attributes in the dataset.

The removed features are **Main_meals_daily**, **Height**, **Water_daily** and **Calorie_monitoring** which were likely among the 65% with lower mutual information scores. This implies that, according to the mutual information criterion, these features provided less predictive power regarding the target variable compared to the retained features.

2 Model Training

Below are the relevant statistics for the two previously mentioned models, incorporating their respective configurations (scaling, depths, etc.):

Here are the measurements for the manual implementations:

2.1 RandomForest using sklearn

Parameters	Precision Mean	Recall Mean	F1 Mean	Precision Variance	Recall Variance	F1 Variance	Accuracy
criterion: entropy depth: 14 samples: 0.7 estimators: 300	0.74	0.68	0.67	**0.02**	**0.07**	**0.03**	0.70
criterion: entropy depth: 14 samples: 0.8 estimators: 300	0.76	0.71	0.69	**0.02**	0.06	0.02	0.72
criterion: entropy depth: 14 samples: 0.8 estimators: 200	**0.79**	**0.74**	0.72	0.01	0.06	0.02	**0.74**
criterion: entropy depth: 14 samples: 0.8 estimators: 200	0.78	0.73	0.72	0.01	0.05	0.01	**0.74**
criterion: entropy depth: 14 samples: 0.7 estimators: 300	**0.79**	**0.74**	**0.73**	**0.02**	0.04	0.01	**0.74**

Where:

- Accuracy Mean: 0.72
- Accuracy Variance: 0.0003

2.2 ExtraTrees using sklearn

Parameters	Precision Mean	Recall Mean	F1 Mean	Precision Variance	Recall Variance	F1 Variance	Accuracy
criterion: entropy depth: 14 samples: 0.7 estimators: 200	0.73	0.67	0.65	**0.02**	**0.08**	**0.04**	0.69
criterion: entropy depth: 14 samples: 0.7 estimators: 200	0.76	0.70	0.69	**0.02**	0.07	0.03	0.72
criterion: log_loss depth: 14 samples: 0.6 estimators: 300	0.76	0.72	0.71	0.01	0.06	0.02	0.73
criterion: log_losss depth: 14 samples: 0.7 estimators: 300	0.76	0.72	0.71	**0.02**	0.05	0.01	0.73
criterion: entropy depth: 14 samples: 0.8 estimators: 300	**0.79**	**0.74**	**0.74**	**0.02**	0.04	0.01	**0.75**

Where:

- Accuracy Mean: 0.72
- Accuracy Variance: 0.0004

2.3 Support Vector Machines using sklearn

Parameters	Precision Mean	Recall Mean	F1 Mean	Precision Variance	Recall Variance	F1 Variance	Accuracy
C: 5 gamma: scale kernel: rbf	0.81	0.8	0.81	**0.01**	0.01	**0.01**	0.81
C: 5 gamma: scale kernel: rbf	0.81	0.81	0.81	**0.01**	**0.02**	**0.01**	0.82
C: 10 gamma: scale kernel: rbf	0.84	0.83	0.83	**0.01**	**0.01**	0.01	0.83
C: 5 gamma: scale kernel: rbf	0.84	0.83	0.83	**0.01**	0.01	0.008	0.84
C: 5 gamma: scale kernel: rbf	**0.86**	**0.85**	**0.85**	**0.01**	0.005	0.005	**0.85**

Where:

- Accuracy Mean: 0.83
- Accuracy Variance: 0.0002

2.4 GradientBoosted Trees using xgboost

Parameters	Precision Mean	Recall Mean	F1 Mean	Precision Variance	Recall Variance	F1 Variance	Accuracy
learning_rate: 1e-07 depth: 14 estimators: 200	0.73	0.71	0.71	**0.01**	**0.08**	**0.03**	0.73
learning_rate: 1e-07 depth: 11 estimators: 300	0.73	0.72	0.71	**0.01**	0.05	0.02	0.73
learning_rate: 1e-05 depth: 13 estimators: 100	**0.76**	**0.74**	**0.73**	**0.01**	0.04	0.02	**0.74**
learning_rate: 1e-07 depth: 14 estimators: 100	0.74	0.72	0.71	**0.01**	0.05	0.02	**0.74**
learning_rate: 1e-07 depth: 14 estimators: 300	0.74	0.72	0.72	**0.01**	0.03	0.02	0.72

Where:

- Accuracy Mean: 0.73
- Accuracy Variance: 0.00007

These statistics provide a detailed overview of each model's configuration and performance metrics. By including information on feature scaling, selection methods, and evaluation scores, stakeholders gain deeper insights into the models' behavior and effectiveness for the given task.

In conclusion, it's evident that the choice of hyperparameters significantly influences the predictive power of all the models analyzed above.

This observation underscores the importance of careful tuning and optimization of hyperparameters to achieve optimal performance in predictive modeling tasks. By selecting appropriate hyperparameters, such as the number of estimators, maximum depth, kernel type, and regularization parameter, among others, practitioners can enhance the predictive capabilities of machine learning models and improve their effectiveness in real-world applications. Additionally, thorough experimentation and validation processes are essential for identifying the most suitable hyperparameter configurations for specific datasets and tasks, ultimately leading to more accurate and reliable predictions.