**Machine Learning Project Report**

**Enhancing Heart Disease Prediction by Identifying Key Features:**

**A Comparative Analysis of Machine Learning Models**

**Team Members:**

Hafila Morais Max Morais

Bharghav Chava

Chandrika Reddy Kamjula

Saiteja Bajoji

**Dataset:**

<https://archive.ics.uci.edu/dataset/45/heart+disease>

**Abstract:**

Heart disease is a leading cause of mortality globally, and predictive modeling can be a crucial tool in its early detection and management. This study focuses on a dataset comprising 14 features to tackle the problem through binary classification. A range of machine learning algorithms has been applied, including Logistic Regression, KNN(K-nearest neighbors) SVM (Support vector machines), and DA(Discriminant Analysis), Neural Nets alongside ensemble techniques such as Random Forest, CatBoost, and other Bagging and Boosting methods. Notably, the analysis of feature importance from the Random Forest and CatBoost models facilitated the selection of the top five most significant features. Subsequent analysis using these key features led to the development of models that demonstrated improved recall ratios and accuracies of up to 85%. This approach highlights the potential of using a reduced set of critical features for effective heart disease prediction.

**Introduction:**

In the United States, heart disease is the leading cause of death for both men and women, across all ethnicities. Heart disease cost the United States about **$239.9 billion** from 2018 to 2019 [1].

It is on the rise worldwide, concurrent with the increase in obesity rates and unhealthy food consumption habits. In 2019, the condition, which includes heart disease and stroke, was responsible for an astounding one-third of all deaths worldwide [2].

In addition, research indicates that among the long-term consequences of COVID-19 (dubbed long covid) is a higher likelihood of heart disease developing in otherwise healthy individuals and worsening of the condition in those already afflicted [3].

Gaining a better understanding of the risk factors surrounding heart disease, by way of data analysis, is paramount. Quantifying the level of risk for the major risk factors will doubtless prove valuable in diagnosing and treating heart disease.

This project involves building a classification model that analyzes various heart disease related metrics to classify, with reasonable accuracy, whether the person has heart disease or not. This is done in the hopes of gaining a better understanding of the correlation between the risk factors and heart disease diagnosis.

Classification models in this vein have several important real-world applications. Hospitals can make use of such a model (provided the accuracy is high enough) to assess the likelihood of a patient suffering a heart attack or being afflicted with heart disease. Based on this information they can assess which patients need immediate medical attention, drawing potentially life-saving attention to those at risk while also reducing unneeded hospitalizations for those who are not vulnerable. In the wake of the severe shortages of hospital beds and resources (like oxygen tanks) seen during the covid pandemic [4], serious research has been focused on reducing unnecessary hospitalizations [5]. A classification model like this would prove useful in such endeavors.

Businesses can also benefit from such classification models by refining them for ease of use and packaging them for users, such as by incorporating them in a fitness tracker or a health app. Other potential use-cases include, for example, AI health bots that take a person’s medical history and generate a personalized risk index for various diseases using such classification models. Using this information as an advisory, the person can schedule a visit to a doctor if needed.

In specific, this paper compares an omnibus of Machine Learning models, including some popular and versatile techniques like ensemble models, Support Vector Machines and Neural Networks, with the goal of finding the best quantity as well as the quality of the models that were implemented. These have been implemented in the search for better, more accurate, more well-balanced models for predicting heart diesease, with good success.

The models implemented in this paper have all also been both oversampled and undersampled, to treat the imbalance observed in the classes of the target variable. Oversampling was performed using a technique called SMOTE, and undersampling was performed using the random undersampler that comes built-in with the imblearn python package. A closer look will be taken at these methods in one of the following sections.

The significant advantage of this study over similar works lies in its comprehensive comparative analysis across a broad spectrum of machine learning models. A total of 13 different models have been implemented, including sophisticated techniques such as XGBoost, Support Vector Machines, Discriminant Analysis, and Quadratic LDA. This diverse array enables a robust evaluation of which models are most effective for predicting heart disease under varying conditions.

Furthermore, a strategic reduction in the number of features through feature importance analysis not only clarifies key trends within the data but also enhances model interpretability. This focused approach allows the model to identify the most impactful factors contributing to heart disease, potentially reducing the number of diagnostic tests required. This can be particularly beneficial in healthcare settings, where reducing the number of tests not only cuts costs but also alleviates the patient's burden, without compromising the accuracy of heart disease predictions. This paper's methodologies and findings offer valuable insights into efficient and cost-effective approaches for early heart disease detection.

**Research Design:**

In this project, the Cleveland Coronary Artery Disease dataset from the University of California UCI (Irvine) was used. The data was collected by Dr. Robert Detrano, M.D., Ph.D of the Cleveland Clinic Foundation.

Among the 76 features that the complete dataset comprises, a subset of 14 features were identified to be relevant for the task at hand and were thus used for this project. The risk factors that were identified as the most significant and hence were chosen as the features in the model are listed below:

*Describing the dependent variables:*

1. Age:Age of the patient in years. (Integer)
2. Sex: Sex of the patient. (Categorical)

* 1: Male
* 0: Female

1. Chest pain type: Indicates the nature of chest pain. (Categorical)

* Value 1: typical angina
* Value 2: atypical angina
* Value 3: non-anginal pain
* Value 4: asymptomatic

Typical Angina (Value 1): Chest pain related to heart disease. Symptomised by pain or tightness in the chest. Triggers include stress and strenuous activity. Relieved by rest or nitroglycerin.

Atypical Angina (Value 2): Chest pain that does not fit the usual criteria for angina. It typically varies in intensity and location. It does not follow the usual patterns of trigger and relief by physical activity and rest, respectively.

Non-Anginal Pain (Value 3): Chest pain that is unrelated to the heart. Instead it usually involves the muscles, bones, or the respiratory system. Characterized by a sharp pain. Triggers include movements and breathing.

Asymptomatic (Value 4): No symptoms. Indicates that the individual has an underlying heart condition while lacking chest pain.

1. Trestbps: Resting blood pressure (in mm Hg), taken upon admission to the hospital (Integer)

Refers to the resting blood pressure (in mm Hg), taken upon admission to the hospital. A crucial indicator in assessing cardiovascular health.

1. Chol: Serum cholesterol (in mg/dl) (Integer)

Refers to the serum cholesterol level (in mg/dl). Less than 200 mg/dL is desirable, between 200-239 mg/dL is borderline high, 240 mg/dL and above is high.

1. Fbs: Whether fasting blood sugar > 120 mg/dl (Categorical)

* 1: Yes
* 0: No

Indicates whether a person's fasting blood glucose level is higher than 120 milligrams per deciliter (mg/dL).

1. Restecg: Resting electrocardiographic results (Categorical)

* Value 0: normal
* Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
* Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

Refers to the findings from a resting electrocardiogram (ECG or EKG), which records the electrical activity of the heart while the person is at rest.

1. Thalach: Maximum heart rate achieved (Integer)

Refers to the maximum heart rate achieved during stress testing (in bpm). An important indicator when assessing the heart's function and capacity under physical stress.

1. Exang: Exercise induced angina (Categorical)

* 1: Yes
* 0: No

Refers to exercise-induced angina, used to indicate whether chest pain occurs as a result of physical exertion.

1. Oldpeak: ST depression induced by exercise relative to rest (Integer)

Refers to the measurement of ST depression observed on an ECG during exercise compared to rest.

A higher value suggests a greater degree of myocardial stress and ischemia, which may indicate more severe coronary artery disease. A value of zero or close to zero typically indicates no evident ischemia under the test conditions.

1. Slope: The slope of the peak exercise ST segment (Categorical)

* Value 1: upsloping
* Value 2: flat
* Value 3: downsloping

Refers to the shape of the ST segment on an ECG during the peak of exercise in a stress test.

Value 1: Upsloping - Considered normal

Value 2: Flat - This can be a sign of ischemia

Value 3: Downsloping - Often considered the most worrisome and is associated with ischemia and the presence of coronary artery disease.

1. Ca: Number of major vessels (0-3) colored by flouroscopy (Integer)

Refers to the number of coronary arteries that can be seen with a deposit of plaque during a fluoroscopy procedure, typically assessed during a coronary angiogram. This metric counts the number of major coronary vessels with visible blockages. The values range from 0 to 3, with each number indicating the presence of blockages in that many coronary arteries.

1. Thal: Thallium stress test (Categorical)
   * Value 3: normal
   * Value 6: fixed defect
   * Value 7: reversable defect

Normal: No detectable abnormalities. Thallium is distributed evenly in all areas of the heart muscle.

Fixed defect: A certain area does not absorb thallium both at rest and during exercise, suggesting scar tissue from a previous heart attack.

Reversible defect (transient defect): An area of the heart takes up thallium at rest but not during exercise, indicating ischemia during physical activity.

*Describing the target variable:*

Num: Diagnosis of heart disease (angiographic disease status) (Integer)

* + Value 0: < 50% diameter narrowing
  + Value 1: > 50% diameter narrowing

The target variable refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0). This project follows their lead; thus the target variable has been converted into a binary variable. (0 = no ; 1,2,3,4 = yes).

In this section, a closer look will be taken at the various methods and models that constitute this paper. To start, a flowchart that describes the operation flow of the model will be shown. After which will be discussed some of the more novel methods that were used during model implementation. Then the models themselves will be looked at, with a brief description about the working of each model.

*Flowchart:*

A diagram of a data processing process

Description automatically generated

*Methods used:*

SMOTE:

SMOTE is short for Synthetic Minority Oversampling Technique. SMOTE balances class distribution by creating synthetic samples for the minority class through interpolation of existing instances,

Random Under sampler:

Random Under sampler mitigates class imbalance by randomly removing instances from the majority class, aiming to achieve a more balanced distribution between classes.

*Models Implemented:*

1. Bagging Classifier:

Bagging is short for Bootstrap Aggregating. It is an ensemble model. It works by building multiple independent base models and then aggregating them all to get a final model. Each of the base models decision tree in our case is trained on a random subset of the training data with replacement.

1. AdaBoost Classifier:

AdaBoost is short for Adaptive Boosting. It is an ensemble model. It works by sequentially training a set of weak learners decision tree in our case on the weighted versions of the training data, adjusting the weights of incorrectly classified instances at the start of every iteration.

1. XGBoost Classifier:

XGBoost is short for Extreme Gradient Boosting. It is an ensemble model, and a powerful one at that. It works by sequentially building a series of decision trees, with each tree learning from the errors of its predecessors. XGBoost employs gradient descent optimization to minimize a loss function, thereby enhancing speed and accuracy.

1. CatBoost Classifier:

CatBoost is short for Categorical Boosting. It is a gradient boosting algorithm designed to handle categorical features efficiently. It can automatically handle categorical data without the need for manual preprocessing. CatBoost incorporates ordered boosting, which improves the training process by considering the natural order of categorical features. It also handles missing data well and is a fast and competent model.

1. Random Forest Classifier:

Random Forest Classifier is an ensemble learning method that constructs a multitude of decision trees during training. Each tree is trained on a random subset of the training data and features, and the final prediction is made by aggregating the predictions of all individual trees (voting for classification, averaging for regression).

1. Decision Tree Classifier:

Decision Tree Classifier is a supervised learning algorithm used for classification tasks. It partitions the feature space into a set of disjoint regions based on feature values, with each region corresponding to a class label. Decision trees are interpretable, easy to visualize, and capable of handling both numerical and categorical data.

1. Logistic Regression:

Logistic Regression is a statistical model used for binary classification tasks. Logistic Regression estimates the coefficients of the input features to fit the logistic curve to the training data. It is widely used due to its simplicity, interpretability, and efficiency, especially when the relationship between features and the target variable is approximately linear.

1. KNN (K-Nearest Neighbors):

KNN is a simple and intuitive classification algorithm that makes predictions based on the majority class of its K nearest neighbors in the feature space. It does not require training; instead, it stores all training instances and calculates distances to determine the neighbors. KNN's decision boundary is non-linear and can adapt to complex patterns in the data. However, its performance heavily depends on the choice of K and the distance metric, and it can be computationally expensive for large datasets.

1. Naïve Bayes Classifier:

Naïve Bayes Classifier is a probabilistic model based on Bayes' theorem and the assumption of feature independence. Naïve Bayes calculates the probability of each class given the input features and selects the class with the highest probability. It is efficient, scalable, and robust to irrelevant features, although its assumption of feature independence may not hold in all cases.

1. Neural Networks:

Neural Networks are a class of deep learning models inspired by the structure and function of the human brain. They consist of interconnected layers of neurons, each performing simple operations and learning representations of the input data.

1. Support Vector Machine:

Support Vector Machine is a powerful supervised learning algorithm used for classification. SVM finds the optimal hyperplane that separates instances of different classes by maximizing the margin, the distance between the hyperplane and the nearest instances (support vectors).

1. Discriminant Analysis:

Discriminant Analysis is a statistical technique used for classification tasks. It models the probability distributions of the input features for each class and uses Bayes' theorem to compute the posterior probability of each class given the input features.

1. Quadratic Discriminant Analysis:

Quadratic Discriminant Analysis (QDA) is a variant of Linear Discriminant Analysis (LDA) that relaxes the assumption of equal covariance matrices across classes. Unlike LDA, which assumes a common covariance matrix for all classes, QDA estimates a separate covariance matrix for each class. This flexibility allows QDA to capture more complex relationships.

**Results and discussion:**

All these models were run on Google Colab Notebooks. The results and all the relevant tables have been provided in the Appendix section.

The graph below compares the ratios of sensitivity and specificity of the original and resampled variants of all 13 models.

**All models in Original Analysis**

Among these, it has been determined that the top 4 models are

**Best models in Original Analysis**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Models | Accuracy | Precision | | Recall | |
|  |  | 0 | 1 | 0 | 1 |
| KNN (5) | | | | | |
| Oversampled | 0.82 | 0.81 | 0.83 | 0.83 | 0.80 |
| Random Forest Classifier | | | | | |
| Undersampled | 0.78 | 0.76 | 0.81 | 0.83 | 0.73 |
| CatBoost Classifier | | | | | |
| Oversampled | 0.82 | 0.79 | 0.85 | 0.87 | 0.77 |
| XGBoost Classifier | | | | | |
| Undersampled | 0.78 | 0.79 | 0.77 | 0.77 | 0.80 |
|  |  |  |  |  |  |

**A graph of different types of sensitivity

Description automatically generated with medium confidence**

**Best models in Original Analysis**

The Best Model is KNN (5) Oversampled with a recall ratio of 83%-80% and a accuracy of 82%.

Post-analysis using feature importance is a valuable step in refining machine learning models, particularly in healthcare applications like predicting heart disease. This process can reveal whether certain features are adding noise rather than value, obscuring the underlying trends that are crucial for accurate predictions.

It has been observed that two of the best performing models were CatBoost and Random Forest. Thus, feature importance analysis can been performed with respect to these two models, and the identified features were then used to make a trimmed version of the dataset. This ranking helps in identifying the most influential features that are truly driving the predictions.

Trimming or removing features from a dataset streamlines the analysis by eliminating unnecessary test results. Feature importance and selection techniques are instrumental in identifying which tests are most predictive of heart disease, allowing healthcare providers to focus on those that offer the most diagnostic value.

This approach has several key benefits:

1. **Cost Efficiency**: By reducing the number of tests, healthcare providers can lower the overall cost of diagnosis, making it more accessible and affordable for patients.
2. **Time Savings**: Fewer tests mean quicker turnaround times from testing to diagnosis, which is crucial in conditions where early detection significantly impacts the treatment outcome.
3. **Patient Comfort**: Reducing the number of tests alleviates the physical and psychological burden on patients, who may otherwise undergo stressful and invasive procedures unnecessarily.
4. **Increased Accuracy**: With a more focused set of data derived from the most relevant tests, models can potentially yield higher accuracy, reducing the likelihood of misdiagnosis and ensuring that treatment is appropriately targeted.
5. **Resource Optimization**: Healthcare resources, including time and equipment, can be better allocated toward interventions that are proven to be effective, enhancing overall healthcare delivery.

Therefore, the strategic use of feature importance and selection not only improves predictive modeling but also contributes to more patient-centered, efficient, and cost-effective healthcare practices.

As this was done twice, once for each model, we end up with two such trimmed datasets. Post analysis has been performed with both of these trimmed datasets, thus resulting in two distinct post analyses. All relevant post analysis result tables have been provided in the Appendix section.

***Post analysis – 1 (Random Forest Feature Importance)***

In case of post analysis 1, the features that remain after feature reduction are –

* Chest pain type
* Thalach AKA Maximum heart rate achieved
* Thallium stress test
* Oldpeak AKA ST depression induced by exercise relative to rest
* Age

A patient would need to take just four kinds of test to get a heart disease classification, compared to roughly 11 before feature reduction. This is significantly more convenient to the patient, and would help reduce any unnecessary tests being prescribed to the patient.

A graph with red and white stripes

Description automatically generated

*Feature importance of Random Forest*

The post-analysis was also done to all of the 13 models and the best 4 models were identified to be as follows:

***Best models in post analysis -1***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Models | Accuracy | Precision | | Recall | |
|  |  | 0 | 1 | 0 | 1 |
| Bagging Classifier | | | | | |
| Oversampled | 0.72 | 0.70 | 0.74 | 0.77 | 0.67 |
| AdaBoost Classifier | | | | | |
| Undersampled | 0.73 | 0.73 | 0.73 | 0.73 | 0.73 |
| Decision Tree | | | | | |
| Undersampled | 0.68 | 0.68 | 0.69 | 0.70 | 0.67 |
| QDA | | | | | |
| Undersampled | 0.80 | 0.78 | 0.82 | 0.83 | 0.77 |

A graph of a graph showing different types of sensitivity

Description automatically generated with medium confidence

Best models in post analysis -1

Random forest key features importance did not provide us with better results than the original model itself. The features deemed important by Random Forest might not actually be the most predictive ones but are simply more influential in the way Random Forest constructs decision trees.

The best model in this post analysis 1 using random forest feature importance selection is Quadratic discriminant analysis under sample with a recall ratio of 87%-77% and an accuracy of 80%

***Post analysis – 2 (Cat Boost Feature Importance)***

In the case of post-analysis 2, the features that remain after feature reduction are –

* Chest pain type
* Thallium stress test
* Ca AKA Number of major vessels (0-3) colored by fluoroscopy
* Oldpeak AKA ST depression induced by exercise relative to rest
* Exercise induced angina

Similar to the previous scenario, in this case a patient would need to take just 5 tests, which is still a massive reduction than the initial scenario.

A graph of red and white bars

Description automatically generated

*Feature importance of Cat Boost*

The post-analysis was also done to all of the 13 models and the best 4 models were identified to be as follows:

**Best models in post analysis -2**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Models | Accuracy | Precision | | Recall | |
|  |  | 0 | 1 | 0 | 1 |
| CatBoost Classifier | | | | | |
| Original | 0.85 | 0.84 | 0.86 | 0.87 | 0.83 |
| Random Forest Classifier | | | | | |
| Original | 0.78 | 0.76 | 0.81 | 0.83 | 0.73 |
| Decision Tree | | | | | |
| Oversampled | 0.78 | 0.76 | 0.81 | 0.83 | 0.73 |
| Support Vector Machine | | | | | |
| Oversampled | 0.78 | 0.76 | 0.81 | 0.83 | 0.73 |

A graph of different sizes and colors

Description automatically generated

Best models in post analysis -2

Using just 5 features CatBoost Classifier was able to achieve the highest recall ratio of 87%-83% and an accuracy of 85% which is the highest among all the models in both the original analysis and both post analyses.

**Conclusion and Future work:**

In conclusion, our analysis reveals that a K-nearest neighbors (KNN) model with five neighbors outperforms other classification models in predicting heart disease. Utilizing feature importance scores from CatBoost and Random Forest models, five key features have been identified, that aid in significantly enhancing model performance.

As important as features are to a model, more features can also distract the model from identifying key issues and mask the model from identifying key trends in the model. By refining the dataset to include only these essential features, a more balanced recall ratio is achieved, alongside improved overall accuracies. The strategic use of CatBoost's feature importance was particularly instrumental in realizing these enhanced outcomes, demonstrating the value of targeted feature selection in predictive modeling of heart disease.

It's noteworthy that the CatBoost Classifier, using only five key features, achieved the highest recall ratio of 87-83% and an accuracy of 85%. This performance is the best among all models tested in both the original and subsequent analyses. Achieving both a high recall ratio and excellent accuracy demonstrates CatBoost’s capability to not only correctly identify a high proportion of actual positive cases (high recall) but also to maintain a high level of correct predictions across all cases (high accuracy).

The key feature captured by both random forest feature importance and Cat Boost feature importance is

* Chest pain type
* Thallium stress test
* Oldpeak AKA ST depression induced by exercise relative to rest

These features (tests) are crucial in the prediction of heart diseases. This can give the patient/doctors an idea of which tests to prescribe in case of suspected heart conditions. By identifying these key features—chest pain type, thallium stress test, and ST depression (Old peak) induced by exercise relative to rest—as crucial predictors, one can streamline the diagnostic process significantly.

Future studies could enhance the robustness of our findings by using the selected features on a larger dataset. Our initial model was trained on a relatively small sample of 303 entries from Cleveland Clinic, which may not represent the broader population and is susceptible to overfitting, leading to biased results. A larger and more diverse dataset would help validate our model and potentially uncover additional insights.

Moreover, refining the models through hyperparameter tuning could further improve their performance. Techniques such as GridSearchCV allow for systematic exploration of parameter combinations, optimizing the models for better accuracy, precision, and a more balanced recall ratio. These enhancements would contribute to developing a more reliable and effective tool for predicting heart disease.

Using SHAP (SHapley Additive exPlanations) values to assess local feature importance can also be an effective strategy to enhance model accuracy and interpretability. Unlike global feature importance, which provides an overview of the most influential features across the entire dataset, local feature importance provided by SHAP values offers detailed insights into how individual features affect each specific prediction.

**References:**

[1] Heart Disease Facts | CdC.gov (2023). <https://www.cdc.gov/heartdisease/facts.htm>.

[2] Roth, G.A. et al. (2020) 'Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019,' Journal of the American College of Cardiology, 76(25), pp. 2982–3021. <https://doi.org/10.1016/j.jacc.2020.11.010>.

[3] Xie, Y. et al. (2022) 'Long-term cardiovascular outcomes of COVID-19,' Nature Medicine, 28(3), pp. 583–590. <https://doi.org/10.1038/s41591-022-01689-3>.

[4] The Associated Press (2021) 'Many hospitals with no beds left are forced to send COVID patients to cities far away,' NPR, 19 August. <https://www.npr.org/2021/08/19/1029378744/hospital-beds-shortage-covid-coronavirus-states>.

[5] Social Care Institute for Excellence (SCIE) (2023) Hospital discharge and pre-venting unnecessary hospital admissions (COVID-19) - SCIE. <https://www.scie.org.uk/providing-care/commissioning/hospital-discharge-admissions/>.

**Appendix :**

*Original Analysis:*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Models | Accuracy | Precision | | Recall | |
|  |  | 0 | 1 | 0 | 1 |
| Bagging Classifier | | | | | |
| Original | 0.75 | 0.70 | 0.83 | 0.87 | 0.63 |
| Oversampled | 0.78 | 0.76 | 0.81 | 0.83 | 0.73 |
| Undersampled | 0.77 | 0.75 | 0.79 | 0.80 | 0.73 |
| CatBoost Classifier | | | | | |
| Original | 0.77 | 0.75 | 0.79 | 0.80 | 0.73 |
| Oversampled | 0.82 | 0.79 | 0.85 | 0.87 | 0.77 |
| Undersampled | 0.77 | 0.75 | 0.79 | 0.80 | 0.73 |
| AdaBoost Classifier | | | | | |
| Original | 0.72 | 0.76 | 0.69 | 0.63 | 0.80 |
| Oversampled | 0.65 | 0.67 | 0.64 | 0.60 | 0.70 |
| Undersampled | 0.72 | 0.74 | 0.70 | 0.67 | 0.77 |
| Decision Tree Classifier | | | | | |
| Original | 0.73 | 0.77 | 0.71 | 0.67 | 0.80 |
| Oversampled | 0.70 | 0.70 | 0.70 | 0.70 | 0.70 |
| Undersampled | 0.67 | 0.67 | 0.67 | 0.67 | 0.67 |
| Random Forest Classifier | | | | | |
| Original | 0.77 | 0.74 | 0.81 | 0.83 | 0.70 |
| Oversampled | 0.78 | 0.76 | 0.81 | 0.83 | 0.73 |
| Undersampled | 0.78 | 0.76 | 0.81 | 0.83 | 0.73 |
| Logistic Regression | | | | | |
| Original | 0.83 | 0.78 | 0.92 | 0.93 | 0.73 |
| Oversampled | 0.83 | 0.78 | 0.92 | 0.93 | 0.73 |
| Undersampled | 0.83 | 0.78 | 0.92 | 0.93 | 0.73 |
| KNN | | | | | |
| Original | 0.78 | 0.74 | 0.84 | 0.87 | 0.70 |
| Oversampled | 0.82 | 0.81 | 0.83 | 0.83 | 0.80 |
| Undersampled | 0.80 | 0.76 | 0.85 | 0.87 | 0.73 |
| Naïve Bayes | | | | | |
| Original | 0.80 | 0.75 | 0.88 | 0.90 | 0.70 |
| Oversampled | 0.82 | 0.77 | 0.88 | 0.90 | 0.73 |
| Undersampled | 0.82 | 0.77 | 0.88 | 0.90 | 0.73 |
| Neural Networks | | | | | |
| Original | 0.80 | 0.78 | 0.82 | 0.83 | 0.77 |
| Oversampled | 0.83 | 0.78 | 0.92 | 0.93 | 0.73 |
| Undersampled | 0.82 | 0.79 | 0.85 | 0.87 | 0.77 |
| XGBoost | | | | | |
| Original | 0.73 | 0.75 | 0.72 | 0.70 | 0.77 |
| Oversampled | 0.73 | 0.72 | 0.75 | 0.77 | 0.70 |
| Undersampled | 0.78 | 0.79 | 0.77 | 0.77 | 0.80 |
| SVM | | | | | |
| Original | 0.83 | 0.78 | 0.92 | 0.93 | 0.73 |
| Oversampled | 0.83 | 0.78 | 0.92 | 0.93 | 0.73 |
| Undersampled | 0.83 | 0.78 | 0.92 | 0.93 | 0.73 |
| Discriminant Analysis | | | | | |
| Original | 0.83 | 0.78 | 0.92 | 0.93 | 0.73 |
| Oversampled | 0.83 | 0.78 | 0.92 | 0.93 | 0.73 |
| Undersampled | 0.83 | 0.78 | 0.92 | 0.93 | 0.73 |
| Quadratic Discriminant Analysis | | | | | |
| Original | 0.83 | 0.79 | 0.88 | 0.90 | 0.77 |
| Oversampled | 0.87 | 0.82 | 0.92 | 0.93 | 0.80 |
| Undersampled | 0.83 | 0.79 | 0.88 | 0.90 | 0.77 |

*Post Analysis 1 (Random Forest top 5 features ):*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Models | Accuracy | Precision | | Recall | |
|  |  | 0 | 1 | 0 | 1 |
| Bagging Classifier | | | | | |
| Original | 0.75 | 0.71 | 0.80 | 0.83 | 0.67 |
| Oversampled | 0.72 | 0.70 | 0.74 | 0.77 | 0.67 |
| Undersampled | 0.70 | 0.67 | 0.75 | 0.80 | 0.60 |
| CatBoost Classifier | | | | | |
| Original | 0.73 | 0.68 | 0.85 | 0.90 | 0.57 |
| Oversampled | 0.77 | 0.71 | 0.86 | 0.90 | 0.63 |
| Undersampled | 0.73 | 0.68 | 0.85 | 0.90 | 0.57 |
| AdaBoost Classifier | | | | | |
| Original | 0.63 | 0.63 | 0.63 | 0.63 | 0.63 |
| Oversampled | 0.58 | 0.59 | 0.58 | 0.57 | 0.60 |
| Undersampled | 0.73 | 0.73 | 0.73 | 0.73 | 0.73 |
| Decision Tree Classifier | | | | | |
| Original | 0.67 | 0.67 | 0.67 | 0.67 | 0.67 |
| Oversampled | 0.63 | 0.63 | 0.63 | 0.63 | 0.63 |
| Undersampled | 0.68 | 0.68 | 0.69 | 0.70 | 0.67 |
| Random Forest Classifier | | | | | |
| Original | 0.72 | 0.68 | 0.78 | 0.83 | 0.60 |
| Oversampled | 0.70 | 0.68 | 0.73 | 0.77 | 0.63 |
| Undersampled | 0.72 | 0.67 | 0.81 | 0.87 | 0.57 |
| Logistic Regression | | | | | |
| Original | 0.80 | 0.75 | 0.88 | 0.90 | 0.70 |
| Oversampled | 0.80 | 0.75 | 0.88 | 0.90 | 0.70 |
| Undersampled | 0.80 | 0.75 | 0.88 | 0.90 | 0.70 |
| KNN | | | | | |
| Original | 0.70 | 0.66 | 0.77 | 0.83 | 0.57 |
| Oversampled | 0.62 | 0.61 | 0.63 | 0.67 | 0.57 |
| Undersampled | 0.72 | 0.70 | 0.74 | 0.77 | 0.67 |
| Naïve Bayes | | | | | |
| Original | 0.82 | 0.77 | 0.88 | 0.90 | 0.73 |
| Oversampled | 0.82 | 0.77 | 0.88 | 0.90 | 0.73 |
| Undersampled | 0.82 | 0.77 | 0.88 | 0.90 | 0.73 |
| Neural Networks | | | | | |
| Original | 0.73 | 0.68 | 0.82 | 0.87 | 0.60 |
| Oversampled | 0.78 | 0.76 | 0.81 | 0.83 | 0.73 |
| Undersampled | 0.80 | 0.76 | 0.85 | 0.87 | 0.73 |
| XGBoost | | | | | |
| Original | 0.72 | 0.69 | 0.76 | 0.80 | 0.63 |
| Oversampled | 0.70 | 0.67 | 0.75 | 0.80 | 0.60 |
| Undersampled | 0.72 | 0.70 | 0.74 | 0.77 | 0.67 |
| SVM | | | | | |
| Original | 0.80 | 0.74 | 0.91 | 0.93 | 0.67 |
| Oversampled | 0.80 | 0.74 | 0.91 | 0.93 | 0.67 |
| Undersampled | 0.80 | 0.74 | 0.91 | 0.93 | 0.67 |
| Discriminant Analysis | | | | | |
| Original | 0.82 | 0.76 | 0.91 | 0.93 | 0.70 |
| Oversampled | 0.78 | 0.74 | 0.84 | 0.87 | 0.70 |
| Undersampled | 0.80 | 0.75 | 0.88 | 0.90 | 0.70 |
| Quadratic Discriminant Analysis | | | | | |
| Original | 0.78 | 0.76 | 0.81 | 0.83 | 0.73 |
| Oversampled | 0.80 | 0.78 | 0.82 | 0.83 | 0.77 |
| Undersampled | 0.80 | 0.78 | 0.82 | 0.83 | 0.77 |

*Post Analysis 2 (CatBoost top 5 features):*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Models | Accuracy | Precision | | Recall | |
|  |  | 0 | 1 | 0 | 1 |
| Bagging Classifier | | | | | |
| Original | 0.75 | 0.73 | 0.78 | 0.80 | 0.70 |
| Oversampled | 0.80 | 0.76 | 0.85 | 0.87 | 0.73 |
| Undersampled | 0.80 | 0.76 | 0.85 | 0.87 | 0.73 |
| CatBoost Classifier | | | | | |
| Original | 0.85 | 0.84 | 0.86 | 0.87 | 0.83 |
| aefOversampled | 0.85 | 0.84 | 0.86 | 0.87 | 0.83 |
| Undersampled | 0.85 | 0.84 | 0.86 | 0.87 | 0.83 |
| AdaBoost Classifier | | | | | |
| Original | 0.77 | 0.74 | 0.81 | 0.83 | 0.70 |
| Oversampled | 0.78 | 0.74 | 0.84 | 0.87 | 0.70 |
| Undersampled | 0.73 | 0.71 | 0.77 | 0.80 | 0.67 |
| Decision Tree Classifier | | | | | |
| Original | 0.78 | 0.76 | 0.81 | 0.83 | 0.73 |
| Oversampled | 0.78 | 0.76 | 0.81 | 0.83 | 0.73 |
| Undersampled | 0.77 | 0.75 | 0.79 | 0.80 | 0.73 |
| Random Forest Classifier | | | | | |
| Original | 0.83 | 0.79 | 0.88 | 0.90 | 0.77 |
| Oversampled | 0.83 | 0.79 | 0.88 | 0.90 | 0.77 |
| Undersampled | 0.83 | 0.79 | 0.88 | 0.90 | 0.77 |
| Logistic Regression | | | | | |
| Original | 0.78 | 0.74 | 0.84 | 0.87 | 0.70 |
| Oversampled | 0.78 | 0.76 | 0.81 | 0.83 | 0.73 |
| Undersampled | 0.78 | 0.74 | 0.84 | 0.87 | 0.70 |
| KNN | | | | | |
| Original | 0.78 | 0.73 | 0.87 | 0.90 | 0.67 |
| Oversampled | 0.78 | 0.73 | 0.87 | 0.90 | 0.67 |
| Undersampled | 0.78 | 0.73 | 0.87 | 0.90 | 0.67 |
| Naïve Bayes | | | | | |
| Original | 0.82 | 0.77 | 0.88 | 0.90 | 0.73 |
| Oversampled | 0.82 | 0.77 | 0.88 | 0.90 | 0.73 |
| Undersampled | 0.80 | 0.75 | 0.88 | 0.90 | 0.70 |
| Neural Networks | | | | | |
| Original | 0.80 | 0.76 | 0.85 | 0.87 | 0.73 |
| Oversampled | 0.80 | 0.76 | 0.85 | 0.87 | 0.73 |
| Undersampled | 0.82 | 0.77 | 0.88 | 0.90 | 0.73 |
| XGBoost | | | | | |
| Original | 0.78 | 0.73 | 0.87 | 0.90 | 0.67 |
| Oversampled | 0.78 | 0.73 | 0.87 | 0.90 | 0.67 |
| Undersampled | 0.77 | 0.72 | 0.83 | 0.87 | 0.67 |
| SVM | | | | | |
| Original | 0.75 | 0.71 | 0.80 | 0.83 | 0.67 |
| Oversampled | 0.78 | 0.76 | 0.81 | 0.83 | 0.73 |
| Undersampled | 0.78 | 0.74 | 0.84 | 0.87 | 0.70 |
| Discriminant Analysis | | | | | |
| Original | 0.82 | 0.77 | 0.88 | 0.90 | 0.73 |
| Oversampled | 0.83 | 0.79 | 0.88 | 0.90 | 0.77 |
| Undersampled | 0.83 | 0.79 | 0.88 | 0.90 | 0.77 |
| Quadratic Discriminant Analysis | | | | | |
| Original | 0.77 | 0.72 | 0.83 | 0.87 | 0.67 |
| Oversampled | 0.77 | 0.74 | 0.81 | 0.83 | 0.70 |
| Undersampled | 0.77 | 0.74 | 0.81 | 0.83 | 0.70 |