**Random Forest Classifier Model**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Original (imbalanced) dataset | | | Under Sampled Dataset | | |
| Metrics | 10 Features | 20 Features | 37 features | 10 Features | 20 features | 37 features |
| Accuracy | 0.91 | 0.90 | 0.90 | 0.71 | 0.73 | 0.73 |
| Precision | 0.43 | 0.38 | 0.31 | 0.72 | 0.72 | 0.72 |
| Recall | 0.08 | 0.12 | 0.12 | 0.68 | 0.77 | 0.77 |
| F1-score | 0.13 | 0.18 | 0.17 | 0.70 | 0.74 | 0.74 |
| AUC | 0.77 | 0.78 | 0.79 | 0.77 | 0.80 | 0.80 |

**Results overview:**

**Analysis of Original (Imbalanced) dataset without under sampling:**

1. **High accuracy but low recall:**

* Across all datasets , accuracy is relatively high (around 0.90-0.91). However, recall is very low, particularly in the 10-feature dataset (0.08). This means the model is not identifying true positive cases (people with a heart disease) effectively.
* Given the context , might need to prioritize having higher recall over high accuracy , as missing cases of heart diseases can have serious consequences.

1. **Decline in Precision and F1-Score as Features Increase:**

* Precision and F1-score decrease slightly as more features are added. For instance, precision goes from 0.43 in the 10-feature dataset to 0.31 in the 37-feature dataset.
* This hints at the possibility that adding more features may be introducing noise rather than useful information, potentially because the new features may not be strongly correlated to the target variable (heart disease).
* Model could also be overfitting .

1. **AUC score:**

A line graph with different colors

Description automatically generated

* AUC values are in the same range across the datasets (ranging from 0.77 to 0.79), with a slight increase as more features are added.
* The curves are well above the diagonal baseline, which represents random guessing (AUC = 0.5), indicating that the model performs better than random chance.
* Given the ROC curve's shape and the low recall scores reported earlier, it appears the model has difficulty detecting true positive cases (heart disease cases) with high sensitivity. This limitation is consistent with the low recall and F1-scores observed in the earlier metrics.

**Conclusion**

The current models provide high accuracy but struggle with low precision, recall and overall F1-score, which is a significant drawback for heart disease prediction. Improving recall should be a priority, possibly through rebalancing techniques and feature selection .

**Analysis of Under sampled dataset :**

1. **Recall Significantly Improved**:

* In the original imbalanced datasets, recall was very low (0.08 to 0.12), indicating the model was missing most cases of heart disease. With under sampling, recall increased to 0.68, 0.77, and 0.77 across the three feature sets, making the model much more effective at identifying true positives.

**2. Precision and F1-score:**

* Precision increased substantially, from as low as 0.31 in the original datasets to 0.72 across all under sampled datasets. The models are now more accurate in the positive predictions they make, while also capturing more true positives.
* F1-scores also improved greatly (0.13-0.18 in the original datasets versus 0.70-0.74 in the under sampled ones), indicating that the models now have a more balanced precision and recall.

**3.Accuracy :**

* Accuracy has dropped in the under sampled datasets (from ~0.90 to ~0.71-0.73), which is expected when using under sampling. This decrease in accuracy is likely due to the balanced nature of the dataset, where fewer negative cases may lead to an increase in false positives.

**4.AUC:**

A line graph with different colors

Description automatically generated

* AUC has improved slightly (from ~0.77-0.79 to ~0.77-0.80).
* There is an increase for the 20 and 37 feature sets compared to the original dataset, suggesting an improvement in model performance due to under sampling. The 10-feature AUC remains the same (0.77), indicating that under sampling mainly benefits models with a larger feature set.

1. **Feature count impact:**

* With under sampling, the 20 and 37-feature models both achieved high recall and F1-scores.
* There is no improvement in the metric values between the 20-feature dataset and the 37-feature dataset . Thus, adding more features beyond the initial 20-feature dataset may be redundant due to low correlation to the target variable.

**Overall summary and conclusion:**

The **accuracy** for the original dataset was relatively high due to the class imbalance, as the model favoured predicting the majority class (No Heart Disease). However, high accuracy can be misleading in an imbalanced dataset, as it may mask the model's inability to detect the minority class (heart disease).

The **accuracy** of the model decreased after under sampling. This is expected since under sampling reduces the number of majority class samples, making the model's accuracy more representative of its performance on both classes. The decrease in accuracy here is not necessarily negative, as it indicates a shift towards a more balanced detection of heart disease cases.

**Precision** is lower across the models (10, 20, and 37 features) on the original dataset compared to the under sampled dataset, with values of 0.43, 0.38, and 0.31 respectively. This lower precision reflects that the model was unable to reliably predict heart disease, with a high false positive rate due to class imbalance.

Additionally, the low **recall** values (0.08 to 0.12) from the original dataset confirm that many actual heart disease cases were missed, making this model unreliable for identifying the minority class (heart disease) .

Under sampling the dataset to address the class imbalance has been effective . It has a more balanced precision and F1-score and addressed the model's original limitation of low recall, making it far better at identifying cases of heart disease.

**Feature set selection:**

The 20-feature model consistently showed the best performance balance for both original and under sampled datasets. Increasing the features to 37 did not provide substantial improvements, suggesting that 20 features may be the best trade-off .

**Importance of Evaluation Metrics**:

* The **recall** metric, along with **F1-Score** and **AUC**, proved to be the most reliable indicators of performance for this application. **Recall** is critical because it indicates the model’s ability to detect true heart disease cases, which is the primary objective in a healthcare setting.
* **Accuracy** was misleading in the original dataset due to class imbalance; it was high but masked the models poor recall and precision for the heart disease class.
* The **AUC** for the under sampled datasets (0.77 to 0.80) indicates that the model has a moderate ability to distinguish between heart disease and no heart disease, and it performed similarly across feature sets.

**k-Nearest Neighbours (kNN) Classifier Model**

**Original Dataset - Results Overview:**

| **Dataset** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **AUC** |
| --- | --- | --- | --- | --- | --- |
| **10 Features** | 0.90 | 0.32 | 0.11 | 0.17 | 0.68 |
| **20 Features** | 0.90 | 0.36 | 0.13 | 0.19 | 0.70 |
| **37 Features** | 0.90 | 0.33 | 0.08 | 0.13 | 0.67 |

**Analysis of Metrics**

1. **Accuracy**: All datasets show a high accuracy of 90%, which might initially suggest strong classifier performance. However, accuracy can be misleading for imbalanced datasets, as it doesn’t account for the imbalance between classes (especially in binary classification with skewed class distributions).
2. **Precision and Recall**:
   * **Precision**: The precision values across the datasets are relatively low (0.32–0.36). This indicates a high rate of false positives (the classifier often misclassifies negative samples as positive)
   * **Recall**: The recall scores are also low (0.08–0.13), reflecting a high rate of false negatives where positive samples are classified as negative. This indicates the model struggles to identify true positives accurately.
3. **F1-Score**: The F1-scores are low across all datasets, especially with the 37-feature set. The F1-score, as the harmonic mean of precision and recall, is heavily influenced by the low recall values. This suggests that neither precision nor recall is strong, and the classifier’s balance between identifying true positives and avoiding false positives is suboptimal.

A graph of a line

Description automatically generated with medium confidence

1. **AUC (Area Under the ROC Curve)**:
   * The AUC values (0.67–0.70) show some variability but remain below 0.75, suggesting that the classifier has limited discriminative power for this problem. Ideally, higher AUC values (above 0.80) would indicate stronger performance. These AUC scores suggest that the classifier performs only slightly better than random chance.

**Conclusions Based on Evaluation Metrics**

1. **Most Suitable Metrics**:
   * **Precision and Recall** are more reliable metrics in this case due to the likely class imbalance. Accuracy alone would be insufficient for evaluation, as it does not capture the classifier's struggles with true positive and false positive rates. The low recall values are particularly concerning, as they indicate many positive cases are being missed, which might be critical depending on the application.
   * **AUC** is also useful as it provides insight into the classifier’s ability to differentiate between classes. However, given the low AUC scores, the k-Nearest Neighbors (kNN) model may not be well-suited for this problem without further adjustments.
2. **Insights on Feature Selection**:
   * The 20-feature dataset provides slightly better precision, recall, F1-score, and AUC compared to the 10- and 37-feature datasets. This could suggest that the 20-feature set strikes a better balance between including relevant information and avoiding unnecessary noise.
   * The 37-feature set, which includes more attributes, shows a slight decline in recall and F1-score, possibly due to the inclusion of irrelevant or redundant features. This outcome reinforces the importance of careful feature selection and dimensionality reduction to improve classifier performance.

**Undersampled Dataset:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Metric** | **10 Features (Undersampled)** | |  | | --- | | **20 Features (Undersampled)** |  |  | | --- | |  | | | **37 Features (Undersampled)** | | --- |  |  | | --- | |  | |
| **Accuracy** | 0.69 | 0.72 | 0.69 |
| **Precision** | 0.68 | 0.71 | 0.69 |
| **Recall** | 0.71 | 0.74 | 0.70 |
| **F1-Score** | 0.70 | 0.72 | 0.69 |
| **AUC** | 0.73 | 0.77 | 0.74 |

**Now we compare and analyse each metric in the context of undersampling:**

1. **Accuracy**:
   * **Original Datasets**: Accuracy was consistent around 0.90 for all feature sets.
   * **Undersampled Datasets**: Accuracy dropped to around 0.69-0.72.
   * **Interpretation**: Accuracy has decreased, this is expected with undersampling, as it often reduces the dominance of the majority class and focuses the model on both classes more equally. This reduction may actually indicate a more balanced model rather than one biased towards the majority class.
2. **Precision**:
   * **Original Datasets**: Precision values were relatively low (0.32 to 0.36).
   * **Undersampled Datasets**: Precision has increased substantially to around 0.68-0.71.
   * **Interpretation**: The improvement in precision suggests that undersampling helped the model reduce false positives by training it to focus more on the minority class. This is a positive impact of undersampling.
3. **Recall**:
   * **Original Datasets**: Recall values were very low (0.08 to 0.13).
   * **Undersampled Datasets**: Recall values increased significantly to around 0.70-0.74.
   * **Interpretation**: The increase in recall indicates that the model is now better at identifying the minority class (heart disease cases). This suggests undersampling effectively helped the model recognize the minority class more reliably, likely due to a balanced dataset.
4. **F1-Score**:
   * **Original Datasets**: F1-Scores were low (0.13 to 0.19).
   * **Undersampled Datasets**: F1-Scores improved to around 0.69-0.72.
   * **Interpretation**: F1-score balances both precision and recall, and this increase reflects that the model has become better at managing both false positives and false negatives. The improved F1-score suggests that the model’s overall performance is more balanced with undersampling.
   * A graph of a number of data

     Description automatically generated
5. **AUC**:
   * **Original Datasets**: AUC values ranged from 0.67 to 0.70.
   * **Undersampled Datasets**: AUC values improved to around 0.73-0.77.
   * **Interpretation**: An increase in AUC indicates better discrimination between classes with the undersampled datasets. The undersampled model's ROC curves (as shown in the image) stay further from the diagonal random line, suggesting better classification ability.

**Impact of Undersampling**

* **Recall**: Undersampling significantly increased recall, indicating that the model has improved its ability to detect the minority class (heart disease cases). This shift is crucial, as higher recall generally means the model is less likely to miss true positives, an important factor in medical diagnoses.
* **Precision**: Precision also increased, which is a good sign. Often, undersampling can lead to a higher number of false positives, reducing precision, but our results show that precision was enhanced. This suggests that the model improved its ability to correctly predict positive cases even with a smaller data set.
* **F1-Score**: With higher F1-scores across the undersampled datasets, the model demonstrates a better balance between precision and recall, making it more suitable for applications where both true positives and false negatives are crucial.

**ROC and AUC Analysis**

The ROC curves (as shown in the image) for the undersampled datasets show a clear improvement, with AUC values now ranging from 0.73 to 0.77. This higher AUC indicates that the undersampling technique allowed the kNN classifier to achieve better separation between the classes, effectively reducing overlap between false positives and true positives.

**Logistic Regression Classifier Model**

**Original Dataset - Results Overview:**

| **Dataset** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **AUC** | **Confusion Matrix** | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **TN** | **FP** | **FN** | **TP** |
| **10 Features** | 0.91 | 0.52 | 0.09 | 0.15 | 0.80 | 54440 | 454 | 4959 | 491 |
| **20 Features** | 0.91 | 0.54 | 0.10 | 0.17 | 0.83 | 54413 | 481 | 4895 | 554 |
| **37 Features** | 0.91 | 0.53 | 0.11 | 0.18 | 0.84 | 54387 | 507 | 4868 | 582 |

**Metrics Analysis:**

1. **Accuracy:** While accuracy is a straightforward metric that indicates the overall proportion of correct predictions, it can be misleading in imbalanced datasets. In our case, the accuracy remained consistently high (around 0.91) across the three encoded datasets. However, this high accuracy was somewhat deceptive, as it masked poor performance in identifying heart disease cases. The model could predict the majority class (no heart disease) correctly, leading to inflated accuracy while failing to capture the minority class effectively.
2. **Precision and Recall:**

**Precision**: With values between 0.52 and 0.54, the model shows a moderate ability to reduce false positives but struggles to accurately identify true positives. This suggests that, while the model may correctly classify negative instances, it has a higher chance of misclassifying actual heart disease cases as non-heart disease.

**Recall**: Low recall values (ranging from 0.09 to 0.11) indicate that a substantial number of positive cases are missed (high FN), a significant concern in a medical context where undiagnosed cases could have severe consequences. Recall should be improved to make the model more sensitive to true heart disease cases.

1. **F1-Score:** The low F1-scores across all datasets (0.15–0.18) reflect the poor balance between precision and recall. This low score points to a significant limitation in the model’s predictive capability for the minority class. Improvements in either precision or recall are essential to increase the F1-score and make the model more reliable.
2. **AUC (Area Under the ROC Curve):** With AUC values increasing slightly with the feature count (0.80 to 0.84), there is a slight gain in the model's ability to differentiate between classes. However, these values are below 0.85, suggesting that the model has limited discriminatory power and may require further enhancements, such as more complex feature engineering or resampling, to effectively separate positive and negative classes.

A line graph with different colored lines

Description automatically generated

1. Confusion Matrix: The confusion matrix for each dataset indicates high values for true negatives (TN) but significantly lower true positives (TP). False negatives (FN) are particularly high, underscoring the model’s struggle to identify heart disease cases. Each matrix illustrates the impact of class imbalance, where a focus on maximizing TN and minimizing FP has led to an overlooked positive class.

A blue square with white text

Description automatically generated A blue square with white text

Description automatically generated A blue square with white text

Description automatically generated

**Conclusions Based on Evaluation Metrics**

1. **Most Suitable Metrics:**

Precision and Recall are essential in this case, with recall being particularly important given the model’s high FN rate. A high recall would ensure fewer missed heart disease cases, which is critical in a medical application. Accuracy is less meaningful due to the class imbalance, while AUC provides a broader view of model discrimination, though still below desired levels.

1. **Insights on Feature Selection:**

The 37-feature dataset performed marginally better, suggesting that additional features contributed useful information without overfitting. However, the marginal improvement also indicates that model improvements are likely needed beyond just feature count; resampling or more advanced techniques could further enhance the model’s sensitivity.

**Undersampled Dataset - Results Overview:**

| **Dataset** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **AUC** | **Confusion Matrix** | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **TN** | **FP** | **FN** | **TP** |
| **10 Features**  **(Undersampled)** | 0.71 | 0.74 | 0.67 | 0.70 | 0.79 | 4127 | 1311 | 1821 | 3646 |
| **20 Features**  **(Undersampled)** | 0.76 | 0.75 | 0.78 | 0.76 | 0.83 | 3999 | 1439 | 1190 | 4277 |
| **37 Features**  **(Undersampled)** | 0.76 | 0.75 | 0.79 | 0.77 | 0.84 | 3983 | 1455 | 1131 | 4336 |

**Metrics Analysis:**

1. **Accuracy:** Accuracy decreased to around 0.71–0.76 due to undersampling, as expected. This drop reflects a more balanced model that focuses on both classes rather than prioritizing the majority class. Although lower, the accuracy here may represent a more honest assessment of model performance.
2. **Precision and Recall:** Precision improved substantially to 0.74–0.75, indicating a lower rate of false positives and suggesting the model is better at identifying true positives. With recall rising to 0.67–0.79, the model now successfully captures more heart disease cases. This improvement in recall is critical for health-related applications, as it reduces the number of missed diagnoses, contributing to more reliable predictions.
3. **F1-Score:** F1-scores also saw significant improvements, now around 0.70–0.77, indicating that the model’s balance between precision and recall is more optimal with undersampling. This improvement demonstrates the positive impact of undersampling on balancing class prediction.
4. **AUC:** The AUC values improved to around 0.79–0.84, with the best results for the 37-feature dataset. This increase confirms that the model’s discriminative power was enhanced by undersampling, allowing it to more accurately differentiate between positive and negative classes.

A line graph with different colored lines

Description automatically generated

1. **Confusion Matrix:** The confusion matrices reveal a notable improvement in TP values, with a simultaneous reduction in FN. This change indicates that the model is now more balanced, correctly identifying both positive and negative classes with greater accuracy. This balanced confusion matrix outcome confirms that undersampling mitigated the imbalance effects, yielding a more robust model for real-world applications.

A blue squares with white text

Description automatically generated A blue squares with white text

Description automatically generated A blue squares with white text

Description automatically generated

**Impact of Undersampling**

Recall: The model’s recall improvement suggests it is now better suited to detect true positive cases of heart disease, reducing the risk of missed cases.

Precision: Increased precision indicates fewer false positives, suggesting that undersampling helped the model refine its focus on true cases.

F1-Score: Higher F1-scores across all undersampled datasets signify an enhanced balance, making the model more applicable in contexts where both true positives and true negatives are important.

**Decision Tree Classifier**

* **Under sampling significantly improves TP rate (Recall):** Notice how the TP rate jumps drastically for the under sampled datasets. This means the model is much better at correctly identifying individuals with heart disease after under sampling. For example, with "10 features", the TP rate goes from 0.0704 to 0.6563. This is a crucial improvement, especially in a medical context where identifying true positives is vital.
* **Under sampling comes at the cost of accuracy and precision:** While under sampling helps with recall, it reduces accuracy and precision. This is because the model becomes more sensitive to the minority class (heart disease), leading to more false positives.
* **More features don't always mean better performance:** In the original datasets, using all features resulted in the highest TP rate (0.2461) but the lowest accuracy (0.8540). This suggests that some features might be noisy or irrelevant, hindering the model's performance.
* **F1-score provides a balanced view:** The F1-score, which balances precision and recall, is highest for the under sampled datasets. This highlights the importance of considering both metrics, especially when dealing with imbalanced datasets.
* **ROC AUC is generally low:** The ROC AUC values are relatively low across all datasets, indicating that the model's overall ability to discriminate between classes is not very strong.

**Analysis by Dataset:**

* **10 features:** This dataset shows the most dramatic improvement with under sampling, with a significant increase in TP rate and F1-score.
* **20 features:** Similar to "10 features," under sampling improves TP rate and F1-score, but the improvement is slightly less pronounced.
* **All features:** Under sampling still helps with recall, but the accuracy and precision drop more significantly compared to the other datasets.

**Conclusions:**

* **Under sampling is crucial for improving the model's ability to detect heart disease.** However, it's important to be aware of the trade-off with accuracy and precision.
* **Feature selection plays a role in model performance.** Using all features doesn't necessarily lead to the best results.