Diabetes Detection Using Machine Learning with Feature Engineering

Awais Naeem Neil Roy Nicholas Strohmeyer Ahmet Gunhan Aydin

Abstract

In this project, we compare the performances of a range of industry-leading machine learning models on the BRFSS Diabetes Health Indicator datasets from Kaggle to gain insights on possible risk factors of diabetes and to provide benefit in helping identify diabetes in its early stages. We perform pre-processing on the data and then train models using Logistic Regression, SVM, MLP, Random Forest, XGBoost, Random Forest Top 5, and XGBoost Top 5, using an extensive grid search to optimize key hyperparameters for each model. We evaluate and analyze model performance using weighted accuracy, recall, precision, F1 score, along with AUC ROC scores for both balanced and imbalanced datasets. We found that XGBoost performs the best, with accuracy, recall, and precision of 76%, 73%, and 80%, respectively, for the balanced dataset and 85%, 81%, and 81% for the imbalanced. XGBoost Top 5 had 9% less weighted precision, but otherwise performed identically in other metrics, using only the top 5 input features.

1. Introduction and Background

Diabetes Mellitus refers to a class of chronic diseases that are characterized by excessive blood sugar. This excess causes chronic issues and can lead to failure of vital organs such as the eyes, nerves, heart, blood vessels and kidneys (American Diabetes Association, 2010). Furthermore, diabetes greatly increases the risk of stroke, heart disease and developing various forms of cancer, thus posing a significant health risk to all individuals with the disease (cdc, 2023).

It was estimated that 422 million people worldwide had diabetes in 2014. This number rose from just 108 million in 1980 (WHO). In the past 30 years, there has been a rise in type 2 diabetes in particular. In the United States, about 11.3% of the population has diabetes, this figure is 14.7% for adults (over 18). Notably, about 8.5 million of these adults were not aware they had diabetes or did not report it (cdc, 2023).

This last fact, combined with the fact that preventive measures can have a significant impact on delaying the early onset of type 2 diabetes in adulthood, underscores the importance of developing predictive systems that are able to detect diabetes in the early stages and diagnose prediabetic individuals. Some studies have suggested predicting the early onset of diabetes using machine learning techniques, and one group of authors was able to achieve over 90% prediction accuracy by training/evaluating the Random Forest classifier on a dataset containing 14 risk factors associated with the diabetes (Zou et al., 2018). Another study includes a fairly good survey of machine learning methods that have been tried by the community to predict the early onset of diabetes (Sisodia & Sisodia, 2018). Given the success that other studies have found on predicting diabetes, we aim to deploy a variety of ML models on this task to compare their performances. In this project we aim to:

- train various leading ML classification models on the a health indicators dataset with multiple risk-factors to compare the ability in an attempt to develop a model that can classify an individual as non-diabetic, prediabetic or diabetic, comparing the relative performances of the models against each other.
- analyze the risk factors to find which ones play a significant role in predicting diabetes using exploratory data analysis.

The remaining content is organized as followed: in Section 2, we describe the datasets we used for our models. In section 3, we do the exploratory data analysis to find patterns and discover insights from the datasets. In section 4, we discuss the pre-processing steps taken to prepare the data for use in model training. In section 5, we describe our models and how we trained them on the datasets. In section 6, we report our results and discuss our findings. Finally, in section 7, we summarize the work we did and our main learnings.

2. Data Description

We will be using the Diabetes Health Indicators data set (from a Kaggle competition) which originates from the 2015 Behavioral Risk Factor Surveillance System, a health-related survey collected by the CDC on annual basis from over 400,000 Americans since 1984 (kag). Originally, this data had a subset of 441,455 responses with a total of 330 features but then it was reduced to 22 features representing a mixture of medical indicators and demographics. From the Kaggle competition, we decided to work with a couple of extracted data sets from the original data set.

2.1. Multi-Class Imbalanced Data

This dataset contains around 253,680 survey responses with 21 input features. The target variable has three classes namely non-diabetic (0), prediabetic (1) and diabetic(2). There exists a class imbalance in the dataset such that non-diabetic, prediabetic and diabetic classes have 213704, 4632 and 35345 samples, respectively.

2.2. Binary Class Balanced Data

The balanced dataset contains around 70,692 survey responses with 21 input features. The responses are divided evenly between two classes i.e., non-diabetic (0) and diabetic (1).

We take the above two datasets to be a reliable representative sample of US adults although not the most recent and may not tansfer to other countries/regions of the world

2.3. Input Features

There are a total of 21 feature variables which constitute the input training data for our project. The feature space is a mix of nominal, ordinal and numeric features. Out of the 21 risk indicators, 14 are binary categorical features, 6 are ordinal ranked and only 1 is a continuous variable. Using this feature space, we will train our machine learning models such that they are able to predict whether a person is non-diabetic, pre-diabetic or diabetic. A full description can be found in 5 in the appendix.

3. Exploratory Data Analysis

To identify the general patterns in the data and the effect different features have on the output, we performed exploratory data analysis to uncover interesting insights given below:

1. Individuals who have trouble walking are 3x likely (Figure 1) and those with high blood pressure are 2x likely to have diabetes (Figure 2)

2. People having diabetes report an overall lower experience of general health (Figure 3) and a higher income level contributes to more risk of diabetes (Figure 4)

3. An individual with a higher BMI is more likely to have diabetes (Figure 5)

4. As someone ages, they become progressively more likely to have diabetes (Table 1, Figure 6)

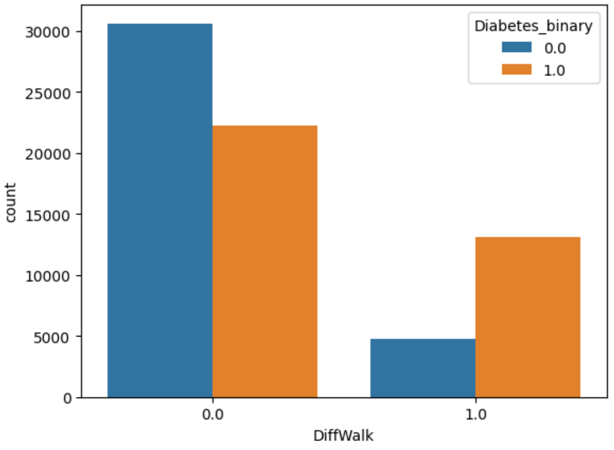

Figure 1. Walking Difficulty experienced by Non-Diabetic Person (0) vs Diabetic Person (1)

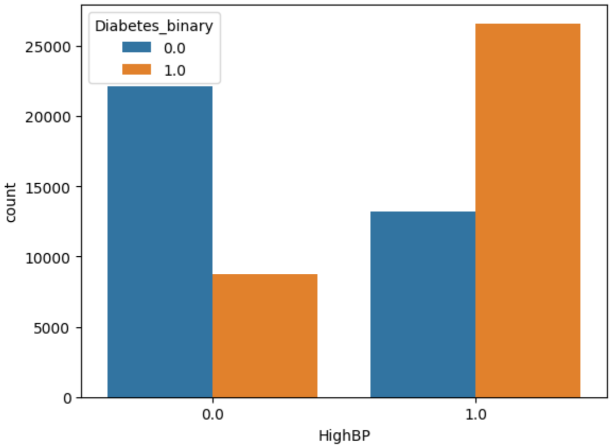

Figure 2. High Blood Pressure experienced by Non-Diabetic Person (0) vs Diabetic Person (1)

4. Data pre-processing

Raw data may exhibit noise, inconsistency and incompleteness, so the following data pre-processing steps were carried out on the complete data set.

4.1. Data Cleaning

Data cleaning involved the removal of duplicate, invalid and null values. Moreover, missing sample values were imputed or dropped across the features space. Finally, the data types of all the features were converted from string to int/floating point values.

4.2. Stratified Splitting into Train and Test datasets

To separate out the data for training and testing purposes, a stratified train-test split was performed to designate 80%

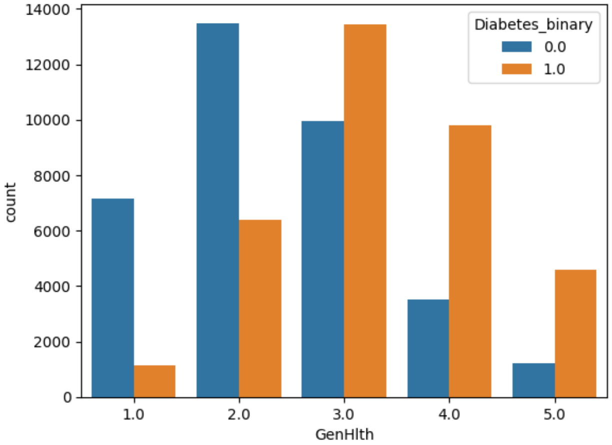

Figure 3. General Health Experience by Non-Diabetic Person (0) vs Diabetic Person (1)

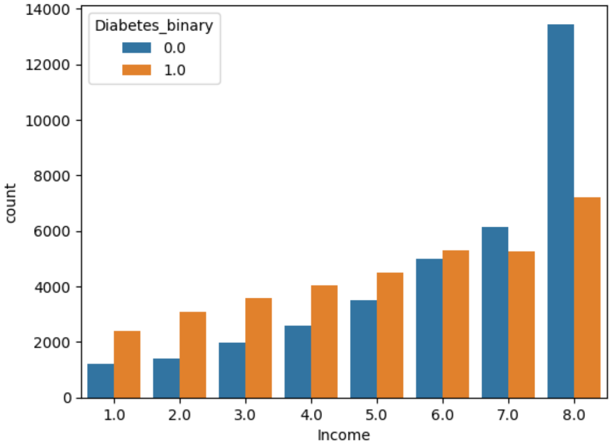

Figure 4. Income Levels of Non-Diabetic Person (0) vs Diabetic Person (1)

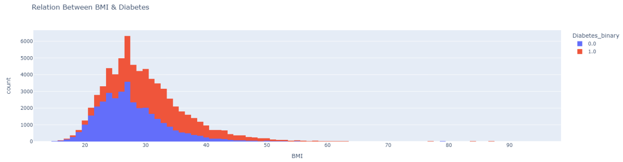

Figure 5. Body Mass Index for Non-Diabetic Person (Blue) vs Diabetic Person (Red)

data for training purposes and 20% data to evaluate model's performance. Since one of our data set possesses class imbalance, we imposed stratification split to preserve the distribution of target classes in both the training and test data set.

Table 1. Age Bucket

Bucket	Age (years)
1	18-23
2	24-29
3	34-39
...	...
12	75-80
13	80+

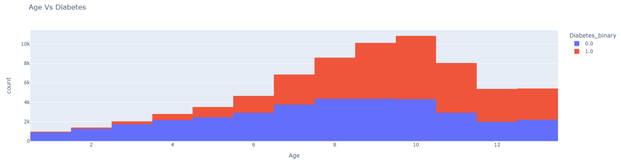

Figure 6. Age bucket for Non-Diabetic Person (Blue) vs Diabetic Person (Red)

4.3. Feature Encoding/Scaling

Each data set feature was bucketed into continuous, ordinal or binary categorical group. Then, we applied standard scalar normalization to continuous features, min-max scaling to ordinal features and one-hot encoding to binary categorical variables.

5. Learning/Modelling

To predict the diabetes onset using various risk factors, we have carefully chosen five distinct machine learning algorithms, each with its unique strengths and capabilities. These methods are Logistic regression, Support Vector Machine, Multi-layer perceptron, Random Forest, and lastly XGBoost. Each of these algorithms offers a distinct approach to diabetes detection, and their combination ensures a comprehensive and well-rounded exploration of the problem domain, increasing the likelihood of achieving accurate and robust results.

5.1. Training Procedure

To ensure the accuracy and robustness of our models, we adopt the Stratified K-Fold Cross Validation technique, which is particularly valuable when working with imbalanced datasets. By partitioning the data into stratified subsets, we can train and test our models in a way that ensures representation of all classes in each fold, preventing any class from being underrepresented during evaluation. In the quest for optimal model performance, we delve into hyperparameter tuning through a Grid Search process. This step is vital for finding the most suitable hyperparameters for each algorithm, maximizing their predictive capabilities.

To address the challenge posed by imbalanced data, we employ the F1-score as our metric and conduct a thorough grid search to optimize its performance.

5.2. Logistic Regression

Logistic Regression is a widely-used classification algorithm that operates by finding a decision boundary within a linear space. It achieves this by applying a transformation using the sigmoid function. The sigmoid function, also known as the logistic function, maps the linear combination of input features to values between 0 and 1, representing probabilities. The central goal of Logistic Regression is to minimize the cross-entropy loss during training. This loss function quantifies the dissimilarity between predicted probabilities and actual labels, and the optimization process adjusts model parameters to improve discriminatory capabilities.

5.3. Support Vector Machine

Support Vector Machine (SVM) is a powerful algorithm that addresses classification and regression tasks. It optimally determines a decision boundary in either the feature space or a transformed kernel space. SVM achieves this by solving a constrained optimization problem, aiming to maximize the average distance between data points of different classes and the decision boundary.

5.4. Multi-Layer Perceptron

The Multi-Layer Perceptron (MLP) is a powerful neural network architecture recognized as a ”universal function approximator.” It excels in transforming input features into specified outputs by leveraging a series of interconnected layers. These layers consist of linear transformations followed by nonlinear activation functions, allowing the network to capture complex patterns and relationships within the data. The key strength of MLP lies in its capacity to adapt and learn complex mappings from input to output. During training, the network minimizes a specified loss function by adjusting the weights and biases of its linear layers. This optimization process is achieved through the backpropagation algorithm, which computes the gradients of the loss with respect to the network's parameters and updates them accordingly. The hyperparameters for our MLP are the hidden layer size, activation function and the alpha value. These hyperparameters are chosen by grid search.

5.5. Random Forest Classification

Random Forest is a robust ensemble learning method that employs a collection of decision trees to enhance predictive accuracy. Each tree is trained on a different subset of the data, with random feature selections, promoting diversity. The final decision is made by aggregating individual tree outputs through averaging (for regression) or voting (for classification). This approach effectively captures complex patterns and is resilient to outliers, making Random Forest widely applicable in machine learning tasks. This method also provides feature importance analysis, crucial for understanding the relevance of various features in diabetes detection.

5.6. XGBoost

XGBoost (Chen & Guestrin, 2016), an extreme gradient boosting algorithm renowned for its exceptional predictive performance, stands out for its efficiency in handling large datasets. Leveraging a gradient boosting framework with regularization, XGBoost constructs a series of decision trees to iteratively correct errors made by preceding trees. The algorithm might incline toward the majority class for imbalanced dataset, compromising its ability to effectively discern patterns in the minority class. To address this, the hyperparameters of XGBoost, including tree depth, number of trees, learning rate, and the subsample ratio of the training instances (to prevent overfitting), underwent meticulous tuning through grid and randomized search methodologies.

5.7. Managing Imbalanced Data

A fundamental challenge for us was extreme imbalance in the multi-class data set. Thus, we experimented with several popular techniques for handling this issue: Synthetic Minority Oversampling Technique or ”SMOTE” (Chawla et al., 2002) and another, similar technique described in Khalila et al. (Khalilia et al., 2011).

The fundamental idea behind both methods is to balance the data during training. Models are fit on under-sampled subsets of the majority class (Khalilia et al., 2011) and by ”over-sampling” (**?**) sets of the minority class. In the former approach, a group of models are collected by training each model on a different balanced set. The final decision is obtained by a majority vote across the entire group. We present the results obtained by replicating (Khalilia et al., 2011) in the next section.

6. Results

The final evaluation of each model was performed using the same data set, pre-processing steps, and train-test split using the same random seed of 42. Each model was tuned using grid search for the best parameters. These steps allow us to make head-to-head comparisons among models ensuring the observed performance gaps are not due to variation in the inputs or sub-optimal parameter selection.

We find that all models perform nearly the same, with XGBoost having a slight edge across our chosen evaluation metrics. We believe the similarity in performance is due to the nature of the data itself, specifically the imbalance between classes and the fact that most models are able to find dominant trends that emerge even in our exploratory data analysis. Therefore, most models can achieve a good baseline accuracy. However, slight improvements upon this prove to be difficult even when using more modern, advanced models such as XGBoost.

6.1. Evaluation Metrics

To evaluate and compare the models we use the following metrics:

Accuracy – is the count of total correct predictions over the total records in a data set. A perfect system would predict every label correctly.

$$\frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

Recall - Recall is the proportion of true positives found by the classifier. Specifically, it is the number of predicted positives out of total positives

$$\frac{TP}{TP + FN} \tag{2}$$

Precision - is the number of total true positives predicted out of total positive predictions. Therefore a high precision tells us the classifier is making meaningful positive predictions.

$$\frac{TP}{TP + FP}$$

F1 Score - is the harmonic mean of precision and recall, which nicely balances the trade-off between the two in a single metric.

$$\frac{2PR}{P + R} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

AUC - The area under the receiver operator curve. AUC can be interpreted as the probability that the model predicts positive when given a positive example. An ideal model should approach 1 whereas the baseline is 0.5, which would be the equivalent of random guessing based on class proportions.

In order to obtain classifiers which optimally balance the precision-recall trade-off, we used the F1 score during grid search hyper-parameter tuning and weighted F1 score in the multi-class case. We also provide the confusion matrices of our binary classifiers Figure **??**.

6.2. Model Performance

In Tables 2 and 3, we present the above metrics for all classifiers across both data sets.

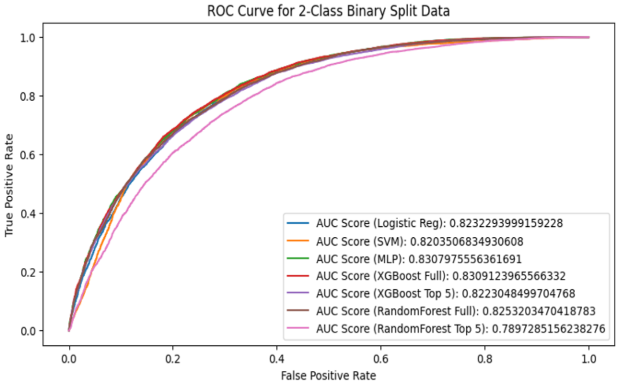

Figure 7. ROC curves are nearly identical across all models with slight difference in the Random Forest trained on only the top 5 features

6.3. Comparison to Related Experiment

The BRFSS dataset is one of the largest publicly available surveys regarding diabetes. Therefore, we now compare our results to a similar experiment done in 2019 (Xie et al., 2019) in which the authors also used a BRFSS dataset. They report achieving up to 82.4% accuracy on the binary classification task. Our best result is 76%. However, this is from training on a balanced sampled variant of the overall data set. In (Xie et al., 2019), the results are obtained from the imbalanced data set using SMOTE.

In our tests using the imbalanced *binary* data set, we also obtained similar accuracy scores of about 84-85% across models. Ultimately we chose to use the balanced binary data set for the final report. It can be seen from the results given in (Xie et al., 2019) that their sensitivity (recall) is quite low (38%) compared to our results using balanced data. Their specificity (ability to detect a negative case) is much higher(90%), but this is not surprising considering most examples are negative cases. Thus, we believe our observations and results are generally consistent with those in (Xie et al., 2019), if not slightly better. Furthermore, in (Xie et al., 2019), the authors report income and BMI being among the most predictive features, which is consistent with the results we discuss next.

6.4. Risk Factor Analysis

We collected feature importance metrics from XGBoost and Random Forest to rank order the top 5 most informative inputs to the model. These are listed on Table 4.

We retrained the XGBoost and Random Forest models using only these top 5 features, and the results are given on Tables 2, 3 and in Figure 11. We observed little to no performance degradation for XGBoost. The decrease is slightly more apparent in the Random Forest model (Figure **??**). This is possibly explained by the fact that Random Forest random-

Table 2. Binary 50/50 Data

Classifier	Accuracy	Precision	Recall	F1	AUC
Logistic Regression	0.75	0.74	0.76	0.75	0.82
MLP	0.75	0.75	0.74	0.75	0.83
SVM	0.75	0.72	0.80	0.76	0.82
Random Forest	0.75	0.73	0.79	0.76	0.83
Random Forest Top 5	0.72	0.71	0.76	0.73	0.79
XG Boost	0.76	0.73	0.80	0.76	0.83
XG Boost Top 5	0.75	0.72	0.79	0.76	0.82

Table 3. Multiclass Imbalanced Data

Classifier	Accuracy	Precision	F1
Logistic Regression	0.85	0.80	0.81
MLP	0.85	0.80	0.81
SVM	0.85	0.80	0.79
Random Forest	0.84	0.80	0.81
Random Forest Top 5	0.83	0.78	0.80
XG Boost	0.85	0.81	0.81
XG Boost Top 5	0.85	0.72	0.81
Undersampling RF	0.61	0.68	0.69

* F1 and precision refer to weighted calculations. Undersampling RF refers to results from testing methods given in (Khalilia et al., 2011)

Table 4. Top 5 Important Features

Random Forest	XGBoost
BMI	High BP
Age	High Chol
High BP	Gen Health
Gen Health	Age
Income	BMI

The top 5 features selected by both models. Compare to trends revealed in the exploratory data analysis

izes its choice of features to split decision trees on during the training stage. Thus, it's more likely to need a broader set of inputs to achieve optimal performance.

6.5. Managing Imbalanced Data

For binary classification, the results in Table 2 are from using a balanced data set. As mentioned previously, the models trained on the imbalanced set were giving higher accuracy (0.85), but this was mostly due to the fact that the models were exploiting the imbalance of the majority class.

To manage imbalance in the multi-class task, we tested the method described in (Khalilia et al., 2011). Pre-diabetic (class 1) was the smallest in size and therefore determined the size of the training sets. The other two classes were randomly and exhaustively under-sampled. The final model consisted of 46 sub-models due to the approximate 46:1 majority to minority class ratio. The 46 models predict on the entire test set and each prediction is counted as a vote. Counting up the votes, we then collected the evaluation metrics and present those results on Table 3.

The confusion matrices (Figure 9) provide more insight. Balancing encourages the model to focus more on pre-diabetic and diabetic examples when training. However, at inference time, this model produces more false positives (in both class 1 and 2) than models trained on the entire imbalanced set, leading to overall worse performance. We also see the overall probability of predicting class 1 increase when using this method (Figures 13 and 14). These results suggest the model is perhaps better calibrated using this technique (although this is not an explicit measure of calibration).

However, this raises questions about what is best for a given application. It could be the case that it is more costly to predict pre-diabetic or diabetic cases incorrectly than it is to predict non-diabetic cases incorrectly. Furthermore, a classifier which incorrectly predicts more pre-diabetic or diabetic cases out of the negative examples may be useful for flagging ambiguous cases for closer analysis. While not in the scope of our experiment, these considerations would be of interest in practical use cases.

6.6. Difficulty Predicting Prediabetic Cases

All models struggled the most with predicting class 1 (prediabetic). This makes sense for the models trained on imbalanced data where class 1 makes up less than 2% of all examples. However, this remains the case for the model adjusted for the data imbalance where it is still only able to achieve 0.365 recall of prediabetic cases 9. In 12 we hypothesize this could be due to a sort of masking effect in feature space. Notably, a model trained on the imbalanced data is always least confident predicting class 1 (Figure 13).

7. Conclusion

In this project, we apply 7 different ML models to the BRFSS diabetes datasets from Kaggle, and compare the models' performances in predicting whether or not someone has diabetes, given other their health indicators. We used stratified sampling and feature encoding/scaling to preprocess the data for both the three-class imbalanced and 2-class balanced datasets. Each model was optimized using a grid search to find the best hyperparameter options for the given model type using both datasets, and then used for inference. For XGBoost and Random Forest, we had a version of the model using all parameters along with a version using only the top 5 parameters, obtained using feature importance analysis.

Through our empirical studies, we found that all the models perform almost exactly the same using the same train and test set splits, getting similar accuracy, precision, recall, F1, and AUC ROC scores for both the 3-class imbalanced and the 2-class balanced datasets, respectively. XGBoost was the best performer (albeit slightly) with a accuracy of 76%, precision of 73%, and recall of 80% for the balanced dataset and accuracy of 85%, precision of 81%, and recall of 81% for the imbalanced dataset. For XGBoost, the top 5 model performs almost as well as the full model, indicating that most of the correlation is captures using a limited number of input features for this dataset. Random Forest top 5 was the lowest performing model, although not by much. Given all of the models performed similarly after hyperparameter optimization, we can conclude that it would be difficult to get more performance out of the models without enhancing our feature space. Given this is a Kaggle dataset, however, this is not really possible.

The code of the project can be found in Applied ML Course Project.

References

World health organization diabetes fact sheet. URL https://www.who.int/news-room/fact-sheets/detail/diabetes. [Accessed 04-12-2023].

Diabetes Health Indicators Dataset — kaggle.com. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset. [Accessed 04-12-2023].

What is diabetes?, Sep 2023. URL https://www.cdc.gov/diabetes/basics/diabetes.html. [Accessed 04-12-2023].

American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diabetes Care*, 33 Suppl 1 (Supplement_1):S62–9, January 2010.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Chen, T. and Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL http://doi.acm.org/10.1145/2939672.2939785.

Khalilia, M., Chakraborty, S., and Popescu, M. Predicting disease risks from highly imbalanced data using random forest. *BMC medical informatics and decision making*, 11:1–13, 2011.

Sisodia, D. and Sisodia, D. S. Prediction of diabetes using classification algorithms. *Procedia Computer Science*, 132:1578–1585, 2018. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2018.05.122. URL https://www.sciencedirect.com/science/article/pii/S1877050918308548. International Conference on Computational Intelligence and Data Science.

Xie, Z., Nikolayeva, O., Luo, J., and Li, D. Building risk prediction models for type 2 diabetes using machine learning techniques. *Preventing Chronic Disease*, 16, September 2019. ISSN 1545-1151. doi: 10.5888/pcd16.190109. URL http://dx.doi.org/10.5888/pcd16.190109.

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., and Tang, H. Predicting diabetes mellitus with machine learning techniques. *Frontiers in Genetics*, 9, 2018. ISSN 1664-8021. doi: 10.3389/fgene.2018.00515. URL https://www.frontiersin.org/articles/10.3389/fgene.2018.00515.

Appendix

Table 5. BRFSS 2015 Data Description

Feature Name	Description	Type
Diabetes	The respondent has been told they have diabetes (or are pre-diabetic)	Dependent
High BP	Respondent has been told by a medical professional they have high blood pressue	Binary
High Chol	Respondent has been told ... they have high cholesterol	Binary
Stroke	Respondent has been told ... that they have endured a stroke	Binary
HeartDiseaseor Attack	Respondent reports having suffered from heart disease or attack	Binary
Chol Check	Respondent has had cholesterol checked in last 5 years	Binary
BMI	The individual's body mass index	Continuous
Smoking	Respondent has smoked at least 100 cigarettes in your life ?	Binary
Physical Activity	Respondent engaged in physical activity outside of work in past 30 days	Binary
Fruits	Respondent consumes fruit at least once per day	Binary
Vegetables	Respondent consumes vegetables at least once per day	Binary
Heavy Alcohol Consump.	More than 14 (male) or 7 (female) drinks per week	Binary
Gen Hlth	Self reported level of general health from	Ordinal
Ment Hlth	Days out of last 30 respondent felt their mental health was not good	Ordinal
PHYS Hlth	Days out of last 30 respondent felt their physical health was not good	Ordinal
Diff Walk	Does the respondent have serious difficulty walking	Binary
NoDocBcCost	Was there a time in past 12 months the respondent could not see doctor due to cost	Binary
AnyHealthCare	Does the respondent have any kind of healthcare coverage	Binary
Sex	Indicate male or female	binary
Age	14 levels of age category	binary
Edu	The highest level completed by the individual	Ordinal
Income	A level of income self-reported by respondent	Ordinal

Table 6. All feature descriptions and data processing steps in detail can be found at https://www.kaggle.com/code/alexteboul/diabetes-health-indicators-dataset-notebook

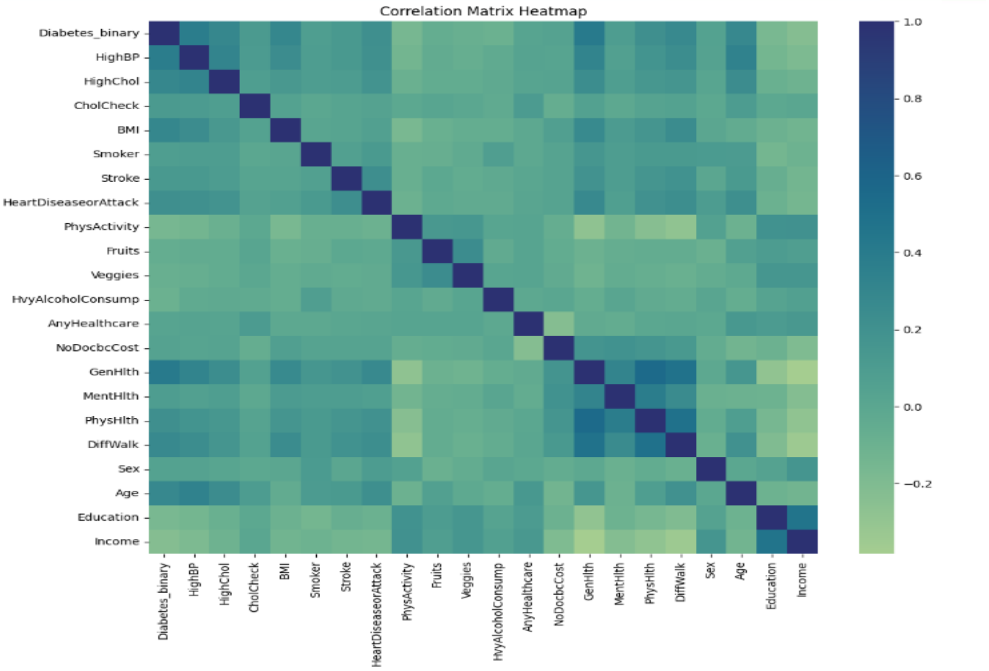

Figure 8. Visualization of correlations among all input features from the diabetes BRFSS data set

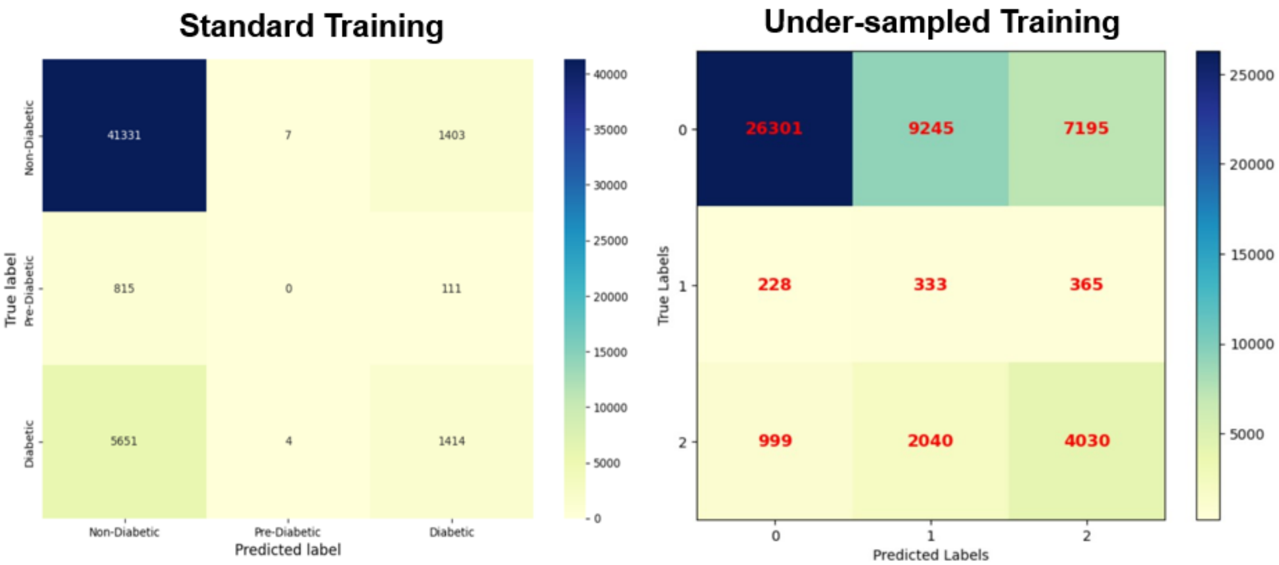

Figure 9. Left : The confusion matrix obtained from the Random Forest model trained on the entire imbalanced 012 data set. *Right :* The confusion matrix resulting from using the under-sampling and majority voting method for imbalanced data using random forests as underlying base models

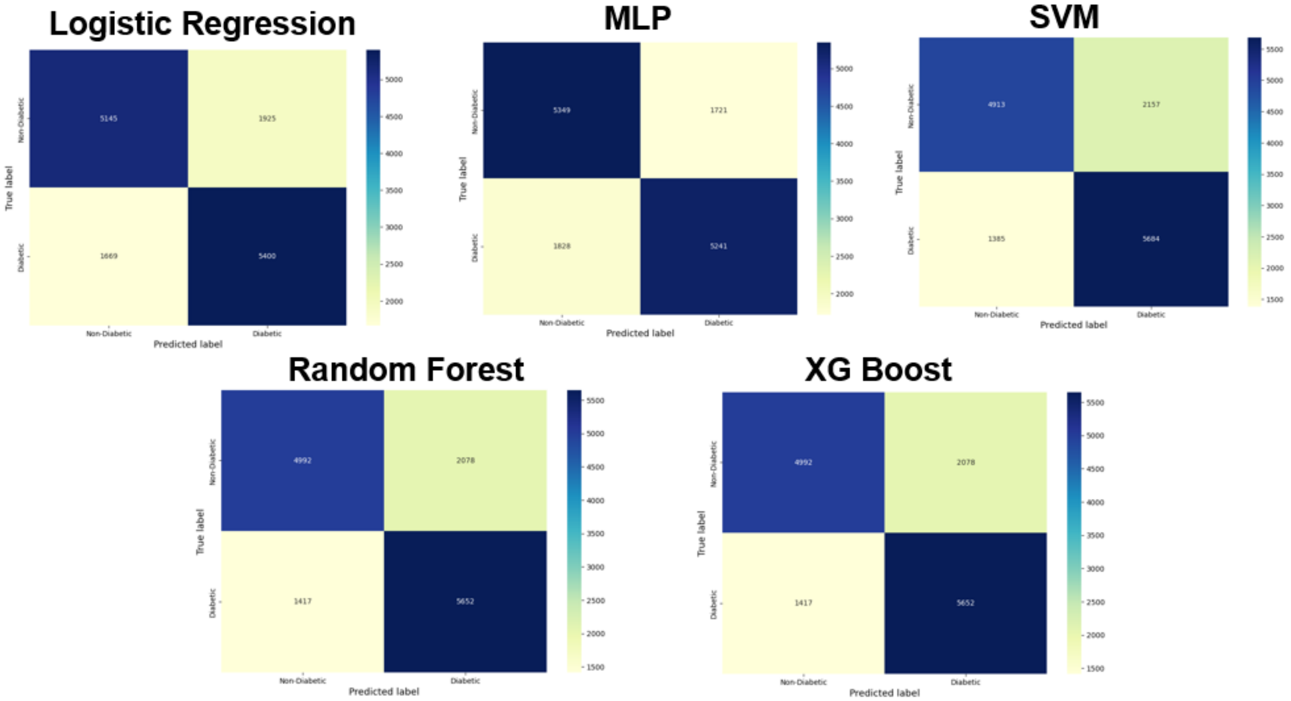

Figure 10. The confusion matrices from the 5 tested models in the binary classification task. A slight bias towards mis-classifying diabetic cases exists for the first 2 models, while the latter 3 show slight bias towards mis-classifying non-diabetic cases.

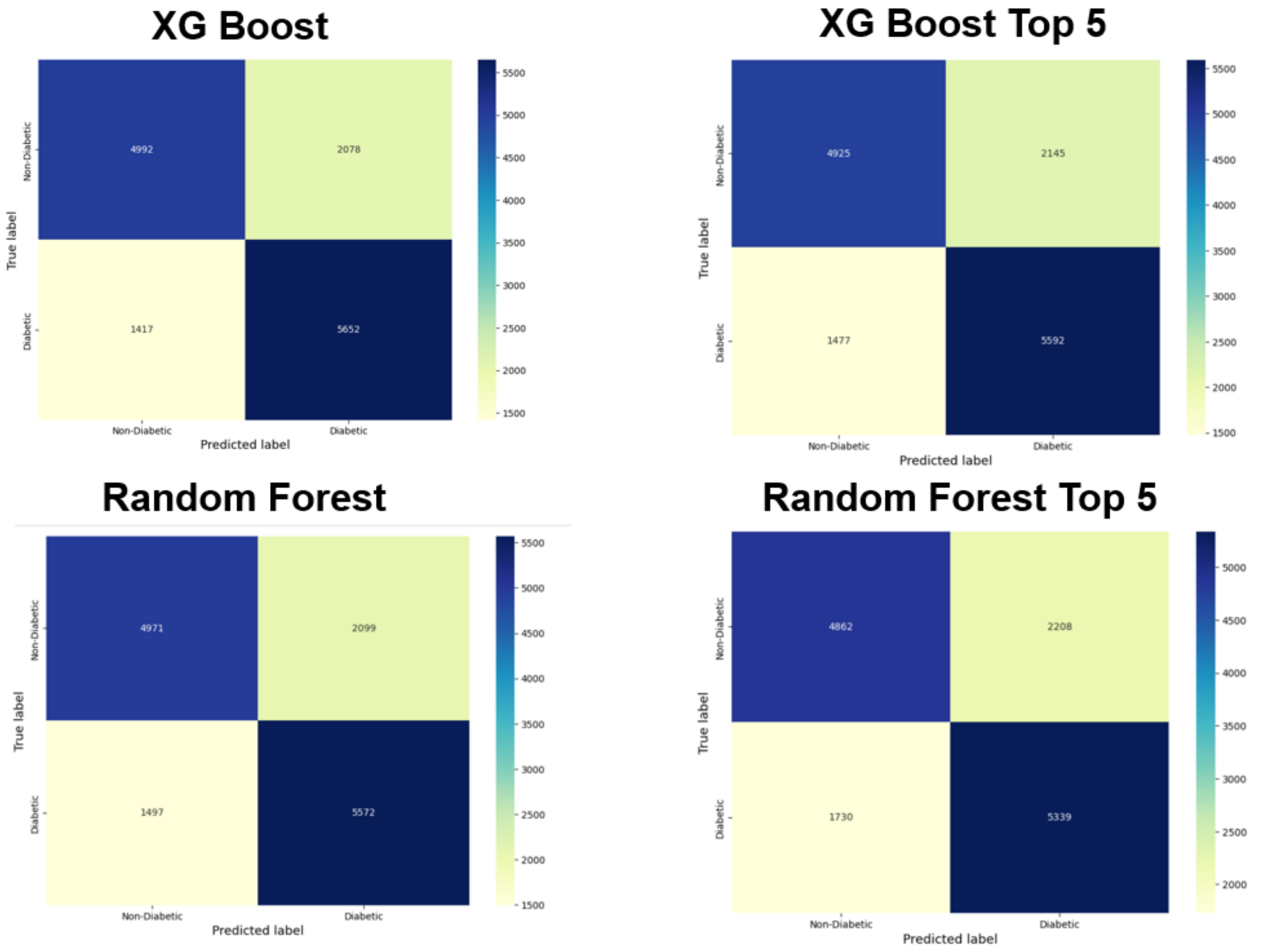

Figure 11. XGBoost, Random Forest and the corresponding ”top5” models compared side by side. The performance drop is slight but apparent

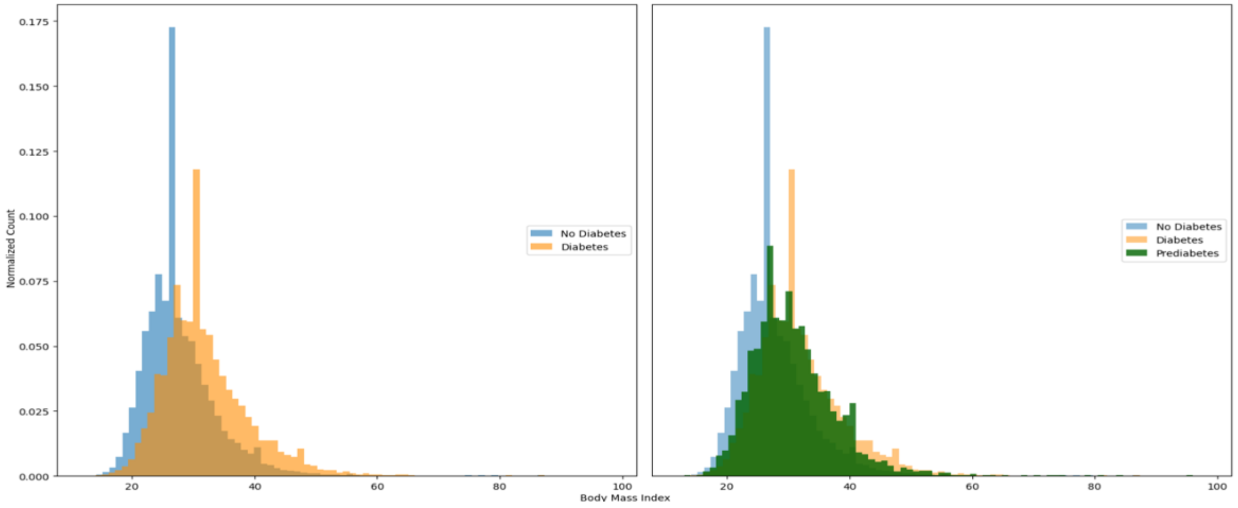

Figure 12. Body Mass Index, one of the strongest predictive features. *Left*: A clear trend seems to appear between non-diabetic and diabetic individuals, *Right* the pre-diabetic distribution appears as an average case of the other two while failing to clearly appear as its own distinct class. We suspect a similar phenomena across the full-dimensional feature space which, in addition to data imbalance, may help explain why this class is the hardest to predict

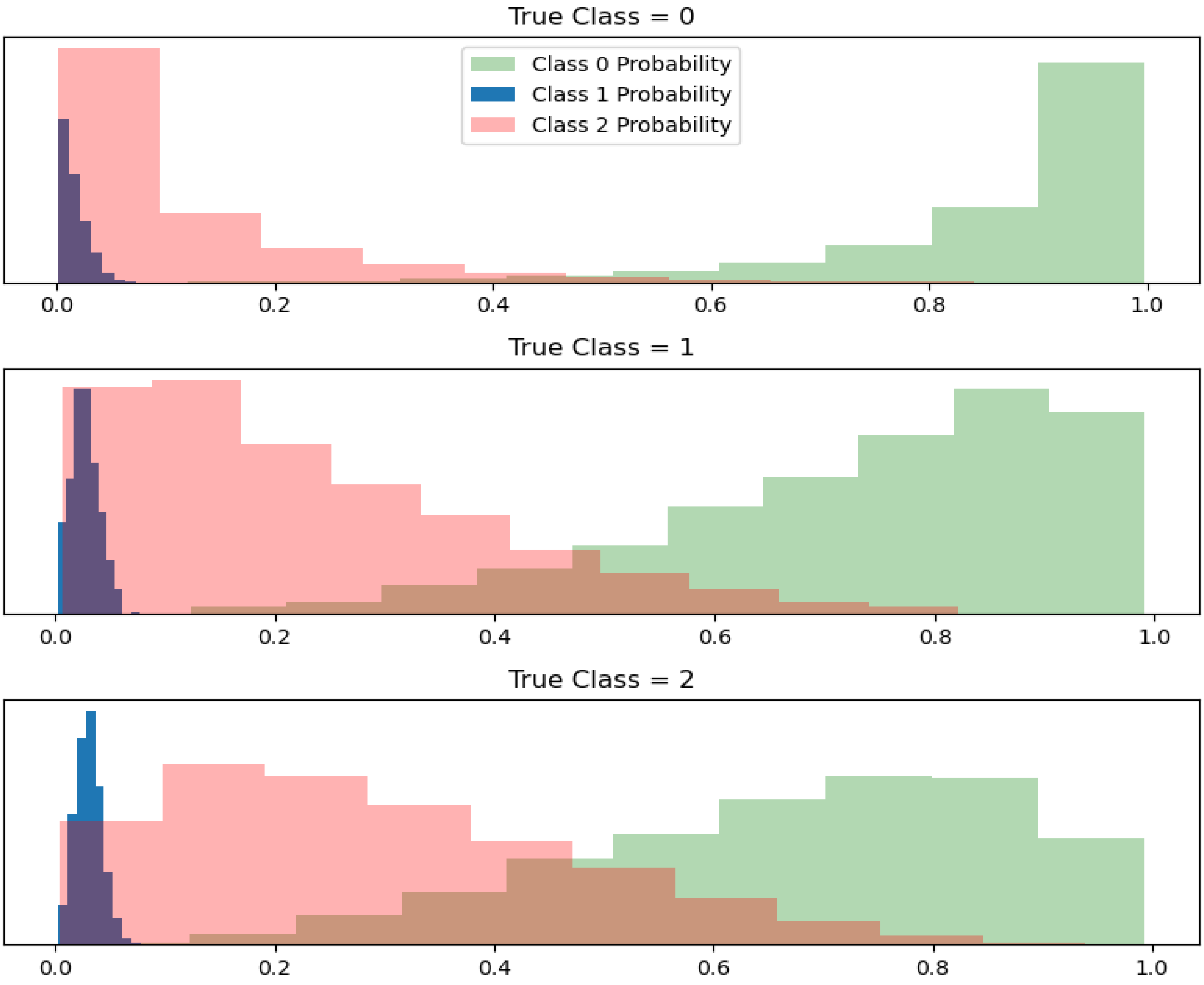

Figure 13. Probabilities of each class predicted by our random forest classifier when trained on the entire imbalanced dataset at once. We can see that class 1 is never considered to be likely. The model appears to be most confident in predicting class 0

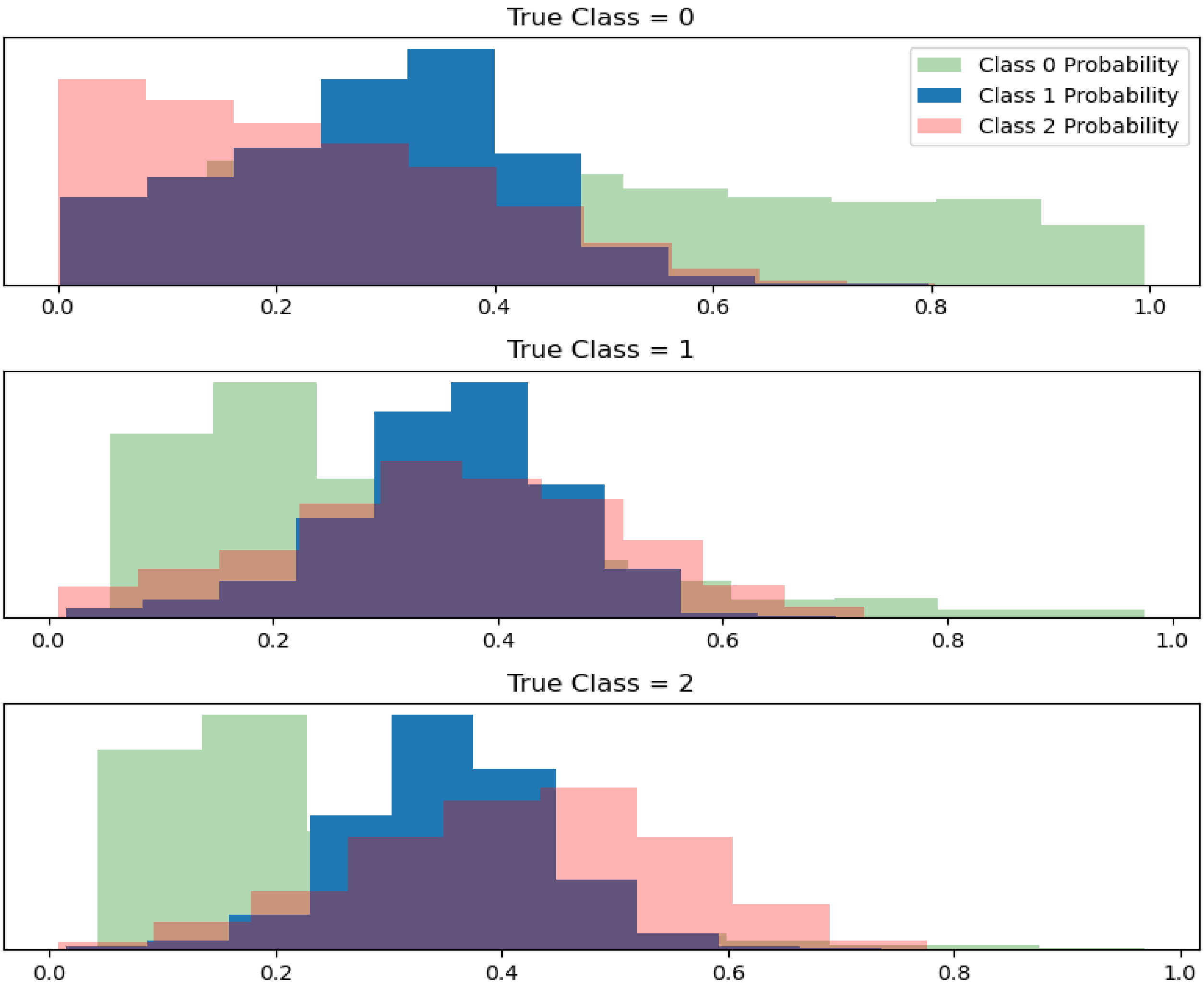

Figure 14. Probabilities of each class predicted by our random forest classifier trained using the undersampling method. Although class 1 is more likely to be predicted now, we can see the model may predict any of the 3 classes with relatively equal probability when the true label is class 1. Clearer separation exists when the true label is 0 or 2, and class 1 is typically seen as an intermediate case.