Binary Classification of Diabetes

I. Background:

We leveraged Kaggle to identify a dataset with extensive patient data from Electronic Health Records (EHRs), containing medical records related to one's risk of getting diabetes.

Our dataset contains 100,000 entries of medical and demographic data of patients, indicating whether an individual has diabetes with a 1-label and 0-labeled otherwise. The data contains various features like gender, age, hypertension, heart disease, smoking history, BMI (Body Mass Index), HbA1c (Hemoglobin A1c) levels, and blood glucose levels – where hypertension and heart disease are indicated present with a 1-label and 0-labeled otherwise. All features, except for gender described with 'Female/Male', are numerical.

These features are commonly used medical evaluations and measurements that support a diabetes diagnosis in patients. High BMI, HbA1c and blood glucose, for example, are closely linked to a higher risk of diabetes (Figure 1). HbA1c levels, the measure of an individual's average blood sugar level over the past 2-3 months, are considered particularly important and insightful as an indicator of diabetes. If the patient has an HbA1c over 6.5%, this is considered a strong indicator of diabetes. However, it is important to recognize that HbA1c levels also vary by age, as the value naturally increases as an individual ages. Based on studies, healthy individuals in age groups 20-39 are expected to have HbA1c levels less than 6%, those between 40-59 should have an HbA1c below 6.1% and ages 60 or older are expected to have an HbA1c 6.5%. Interestingly, although the average age of patients in the dataset is 41, a significant 36.8% of individuals who have diabetes have an HbA1c level of 6.5% or higher.

Similarly, it should be noted that gender and age may play a role in influencing other features as well. Men are generally known to have higher prevalence of hypertension than women, especially in early adulthood. This can be related to differences in hormone production between men and women, and overall the lower estrogen levels of men. Similarly, postmenopausal women have a higher likelihood for developing hypertension for the same reason. Blood glucose levels are also variable between genders and age ranges. Men are generally shown to have higher fasting plasma glucose than women and children.

Additionally, BMI is another metric used to suggest risk of diabetes, with higher levels indicating higher risk. BMI ranging from 18.5-24.9 is considered normal, 25-29.9 is considered overweight, and anything above 30 is considered obese. Although this feature is linked closely to patients being high-risk, only about 11% of individuals diagnosed with diabetes have a BMI greater than 25.

With all of these metrics, and respective biological relevance in mind, we aimed to build our predictor.

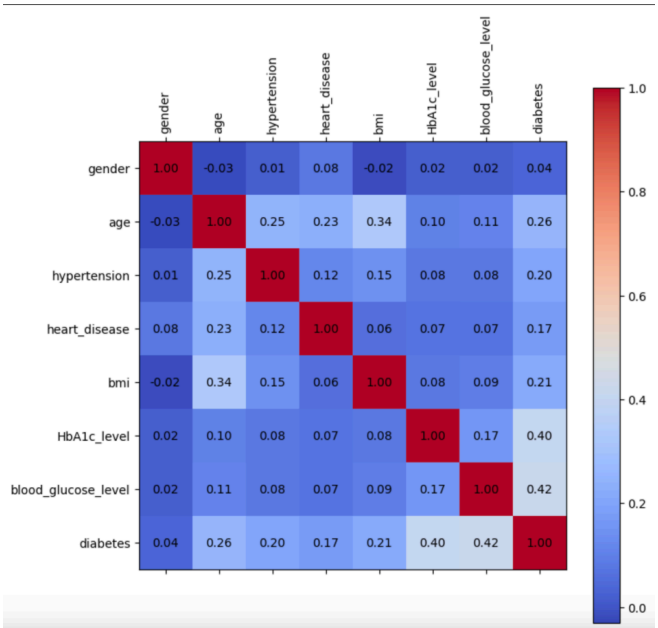

Figure 1: Pearson correlation plot of all the features.

II. Predictive Task:

Our predictive task is to identify whether or not someone will have diabetes, based on a number of their health metrics. Our original dataset was already quite clean, however, we opted to one hot encode the feature regarding smoking history and change the gender features from a string to a binary value. Gender features were translated from string to binary with having 'Male' entries labeled with a 1, whereas 'Female' entries were labeled with a 0. However, after preliminary analysis, we actually chose to omit the smoking history data because it did not show a significant correlation with diabetes. Additionally, we found that one hot encoding is not very good for tree-based ensembles, which ended up being one of the models that best fit our dataset and task. Out of the remaining features, hypertension and heart disease had the least correlation with diabetes and were also highly imbalanced. Omitting these values in our dataset improved metrics in all of the models except the tree classifiers. Our tree classifier models performed the best because it uses a decision tree structure to classify the data into categories. Since the

hypertension and heart disease features are classified with binary values, it is built to fit the methodology of this model. Therefore, omitting these features reduces the accuracy. In summary, as this is the case, when we were working with the Random Forest Classifier, we ultimately chose to keep all features except for smoking history due to the lower correlation.

To continue, in our exploratory analysis it is noticeable that there is a high correlation between HbA1c levels and diabetes with a score of 0.622 on a scale from 0 to 1. In trying to identify the best model for our dataset, we first took into consideration the information we know about how diabetes is currently diagnosed. HbA1c (hemoglobin A1C) levels, a parameter in our dataset, is critical information that, as a reminder, captures a patients average blood glucose level over the past 2-3 months. It is measured from a simple blood test, and is a key way to determine if a patient tests positive for diabetes. A high HbA1c test result (>6.5%) can warrant a diabetes diagnosis all on its own, while oftentimes other metrics are also considered. We wanted to utilize this information, highlighting the importance of A1c values. One of the baselines we created includes a threshold model. If the HbA1c level was above a certain threshold, diabetes was predicted, if it was below that level, then it was predicted to not have diabetes. Another baseline was the logistic regression model with the HbA1c levels and blood glucose levels as the features. Both of these models provide a good baseline as they are simple and incorporate features that are highly correlated with diabetes.

In order to evaluate the performance of our model, we split the dataset into training and testing datasets. The training dataset consisted of the first 90,000 rows and the testing dataset consisted of the last 10,000 rows. There were some problems with the dataset, so to resolve these issues, we eventually decided to use some different techniques that will be discussed in the next section.

III. Model:

After doing exploratory analysis, it is noticeable that the data is very imbalanced with only 8,500 labels with diabetes and 86,500 labels without diabetes. There are also other features such as heart disease and hypertension that were imbalanced. Imbalanced datasets are a problem as the model will not be able to effectively learn patterns in the minority class. This also indicates that a well-performing model that is based on an imbalance dataset is not the most accurate when making predictions. Therefore, resolving these issues, and thereby increasing the reliability of the model, was a top priority.

In order to deal with this problem we tried two methods– SMOTE and Cost Sensitive Learning. SMOTE, which stands for Synthetic Over Sampling Technique, is a method to generate new, unique data points for the minority label. We chose this method instead of simply oversampling such that overfitting can be reduced. On the other hand, Cost Sensitive Learning patterns is a method of training the model such that there are different costs for misclassification errors. For example, if the minority label was predicted as a majority label, a false negative, it would have a higher cost than vice versa. Since there are more opportunities for the label to be a majority, if the model predicts the majority label, it is more likely to be correct; It is important to note that simply using the Cost Sensitive approach does not reflect on the accuracy of the model. To try this method, we used the balanced class weight parameter in ridge classification. This particular model did not work quite well compared to the others. Therefore, using SMOTE on our training set seemed to be a good way to deal with the imbalances as reflected through the performance on our testing set. There was also an option to use the balanced class weight parameter in our best model, which was useful to deal with the imbalances in the features.

When we first created our training and testing datasets, the data was unbalanced. Although we ran SMOTE on our training data, the testing data remained unbalanced as it was created at random. Due to this, the model was resulting in good error metrics for the baseline model. From this, we concluded that something was wrong because although HbA1c levels and diabetes have a high correlation, there are other factors that are involved in classification. To create a more fair metric of comparison, we created a testing dataset where 50% of the labels were diabetic and the remaining 50% were not diabetic. This testing set was a fairer representation because the baseline models did not do well– just as we had expected.

Models We Tried	Strengths	Weaknesses
Naive Bayes: Probabilistic classification method based on the Bayes theorem with the assumption that the features are linearly independent	• Simple implementation and computationally efficient • Training and Prediction tasks are fast due to feature independence so it is ideal for real-world applications	• Can struggle with class imbalances or overall imbalance data • Can overfit smaller datasets.
Random Forest Classifier: An ensemble learning method that combines predictions from multiple decision trees.	• Ensemble features give it a higher accuracy and robustness. • It is less prone to overfitting because of bagging and averaging	• If the dataset is imbalanced, the performance can be negatively impacted. • Hyperparameters can require meticulous hypertuning.
Logistic Regression: Instances are classified based on its probability and threshold.	• Easy to implement and can give interpretable results. • Well suited for binary outcomes like the binary label of presence of diabetes.	• Performs poorly with imbalanced data unless resampling tools are used. • This model can struggle with nonlinear relationships between features.

Figure 2: Strengths and weaknesses of some of the models we used.

IV. Literature Review:

The information in our dataset was sourced from Electronic Health Records (EHRs). EHRs are digitized records of patient data that contain large amounts of information, including medical history, diagnoses, treatment plans, and outcomes. EHRs are maintained as standard practice by healthcare providers and a very reliable source of information for a task like ours. The diabetes prediction parameters found in our dataset are tracked and measured by healthcare professionals, making it an optimal source to develop an accurate prediction model.

There is no record of this dataset being used for other studies, though it is a publicly available dataset. While we have no ability to track if predictive models have been used on our exact dataset, there are a plethora of other models and currently-implemented tools in healthcare to predict and diagnose diabetes.

In one particular study, an automatic diabetes prediction system has been developed using a private dataset of females in Bangladesh from a textile factory (Ref 1). The model used SMOTE and ADASYN approaches to manage the problem with class imbalance. The creators of the model also utilized a number of techniques including ML, Random Forest, SVM, Logistic Regression, and KNN to optimize the algorithm accuracy. A model with 81% accuracy was created in the XGBoost classifier with an ASASYN imbalance-mediation approach. The final project also yielded an Android application for prediction of diabetes. The author's workflow from start to finish was very similar to the approach we took for our model (Figure 2).

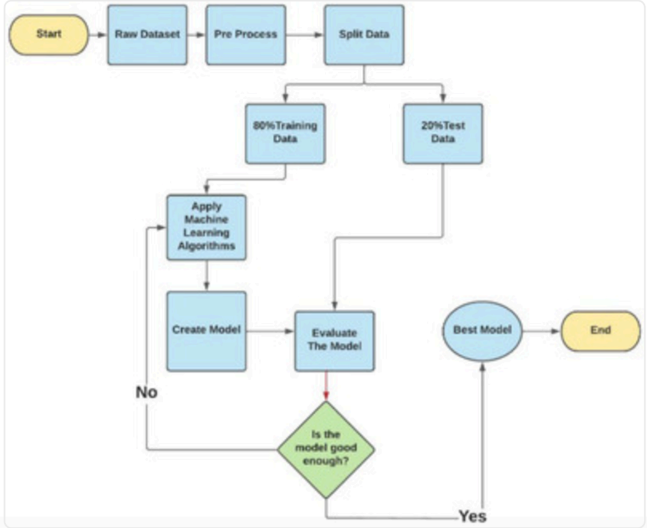

Figure 3: Schematic of workflow for model development, including creating test and training sets and model evaluation

When comparing our final model against the models implemented in this study, we found that our final accuracy of 88% was slightly higher than the study's (Figure 4). This may be related to our incredibly imbalanced dataset, with less than 9% of our data being of patients with diabetes. Although we did our best to balance this data using SMOTE and other metrics, we had a relatively small amount of data to train and test our predictor on diabetes patients.

Performance metrics of various classifiers with SMOTE technique in the merged dataset

Classifier	Precision	Recall	F1 Score	Accuracy	AUC
Logistic regression	0.78	0.77	0.77	77%	0.88
KNN	0.78	0.76	0.76	76%	0.85
Random forest	0.78	0.78	0.78	78%	0.87
Decision tree	0.75	0.73	0.73	73%	0.75
Bagging	**0.80**	**0.79**	**0.79**	**79%**	**0.87**
Adaboost	0.79	0.78	0.78	78%	0.85
XGboost	0.78	0.78	0.78	78%	0.84
Voting	0.79	0.79	0.79	79%	0.86
SVM	0.78	0.75	0.76	75%	0.87

Figure 4: Evaluation metrics to assess the different models from the literature study.

While diabetes prediction is quite common data sources in the literature for machine learning models, predicting or diagnosing diabetes in the clinic is typically based on naive thresholding. If one's fasting glucose is above 126 mg/dL and their HbA1c is above 6.5%, the patient may be given an official diabetes diagnosis. So, while patient data is great for optimizing a machine learning model, practical, state-of-the-art implementations for prediction of whether a patient has diabetes, is not in high demand. As aforementioned, doctors and healthcare professionals

can leverage raw values from blood tests and other clinical evaluations to detect diabetes.

IV. Results and Conclusion:

In conclusion, our exploratory analysis and predictive task highlights the significant roles that different features play in diagnosing an individual with diabetes. Among the models that we tested, the Random Forest Classifier outperformed the alternatives, helping us achieve the best performance in terms of the error metrics. The model also was able to handle complex relationships between features in the dataset, which simpler models struggled to achieve. For example, one of the models that failed was the baseline threshold, taking into account only the HbA1c, with a threshold of above 6.5%. This model was too simplistic for our predictive task as we found that there were many other factors that played a critical role in classifying a patient with diabetes. On the other hand, another model that we tried was the Logistic Regression model. The problem with this model for our predictive task was not that it was too simplistic but rather how the Logistic Regression model works. Logistic Regression deals with discrete values, and applying this to our dataset presented issues as the relationship between features in our data was likely nonlinear. After trying numerous models for this predictive task, like Logistic Regression and the baseline threshold just to name two, ultimately, we narrowed down options to just two – Random Forest Classifier and Decision Tree – and found that the Random Forest model won, outperforming all other models. This is because we found that the Decision Tree tends to overfit the data. Ultimately, a result of capturing too much noise in the data. Alternatively, the Random Forest model avoided this issue by using its ensemble nature.

To continue refining our Random Forest Classifier model, all of the features except for a patient's smoking history were taken into consideration, eliminating the possibility that the model was too simplistic. By taking almost all of the features, the model then made the predictions both more accurate and more realistic to true diabetes testing methods.

In addition to taking in extra features, the success of our model was highly contributed by the effectiveness of previously mentioned data engineering techniques, like SMOTE. With this method, the imbalances in the data were addressed. Using these techniques, the results of the model significantly improved prediction fairness, reducing biases towards the majority class. By having the class weight be balanced as a parameter, the model was able to generalize well, avoiding cases of overfitting. In summary, the combination of data engineering, balancing data and the model's strengths was crucial in achieving strong predictive performance for diabetes diagnosis.

Model	Balanced Accuracy	BER	F1-Score	Recall
Threshold (Hb1Ac levels $> 0.65 = 1$)	0.505	0.2716	0.679	0.574
Logistic Regressor (Hb1AC and blood glucose levels)	0.4804	0.2716	0.6859428571	0.6002
Ridge Classifier	0.5047	0.2716	0.6789	0.5744
Logistic Regressor	0.7827	0.2414	0.7173	0.6126
Gaussian Distribution	0.7724	0.2447	0.7202	0.63
Decision Tree Classifier	0.8799	0.158	0.8124	0.684
Random Forest Classifier	0.8803	0.1573	0.8133	0.6854

Figure 5: Comparison of models and their error metrics.

References:

1. Brownlee, Jason. "SMOTE for Imbalanced Classification with Python." Machine Learning Mastery, 16 August 2020, https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/. Accessed 3 Dec. 2024.
2. Everlywell. "What Is Normal HbA1c by Age?" Everlywell Blog, https://www.everlywell.com/blog/hba1c/what-is-normal-hba1c-by-age/. Accessed 3 Dec. 2024.
3. MedlinePlus. "Hemoglobin A1C (HbA1c) Test: MedlinePlus Lab Test Information." Medlineplus.gov, 6 Sept. 2022, medlineplus.gov/lab-tests/hemoglobin-a1c-hba1c-test/.
4. Tasin, Isfafuzzaman, et al. "Diabetes Prediction Using Machine Learning and Explainable AI Techniques." Healthcare Technology Letters, U.S. National Library of Medicine, 14 Dec. 2022, pmc.ncbi.nlm.nih.gov/articles/PMC10107388/.