Efficient Machine Learning in Diabetes Prediction and Risk Calculation

Cao, A.; Tummarakota, V.; Vempati, C.

Introduction:

Diabetes mellitus, or commonly known as diabetes, is a chronic disease where the pancreas does not produce sufficient amounts of insulin or the body cannot effectively use the insulin it produces. Diabetes is caused by many factors such as age, weight, lifestyle, and diet, and causes many other diseases and health conditions such as heart attacks, strokes, blindness, and even death. In 2014, 8.5% of adults aged 18 and older had diabetes. In 2019, 1.5 million people died from diabetes. In healthcare and medicine, AI and Machine Learning are used to solve problems that conventional medical techniques cannot be used to solve alone, and in recent years, have been on the rise to solve medical or healthcare-related problems. Due to the urgent nature of medicine and healthcare, especially when dealing with patients whose lives are in risk or high in priority or in situations of high risk and urgency, not only high performance is important, but computational resources and speed are very crucial in choosing which models will be used to solve healthcare and medical problems with Machine Learning. In this study, we use supervised Machine Learning models to not only predict the risk of diabetes, but also determine which factors significantly influence the risk and development of diabetes and how much influence these factors have in the development of diabetes. We will also investigate how specific factors, heart disease and gender, influence the risk and development of diabetes. Due to the urgency and delicacy of the situation, we will not only select the model with high performance, but also quick training time as well.

Current Approach and Methods:

The dataset we used for this study is the Diabetes prediction dataset from Kaggle which contains features such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, blood glucose level, and diabetes. Age, body mass index (BMI), HbA1c level, and blood glucose level are continuous numerical features while hypertension, heart disease, smoking history, and diabetes are discrete categorical features with diabetes as the output feature. The dataset originally contained 9500 patients with diabetes and 91,500 patients with no diabetes.

Due to the heavy class imbalance shown in this dataset, we used SMOTE and ADASYN classification balance methods to balance the number of entries in each class. Since we want to closely investigate the role of gender, which is a categorical feature, in the risk and development of diabetes, we used dummy encoding to create new features, male (indicating whether the patient is a male), female (indicating whether the patient is a female), and other (indicating whether the patient identifies as neither).

Since diabetes is a discrete categorical feature and is the output that we are looking for in this study, we will be using a supervised categorical machine learning approach. In this study, we implement Gaussian Naive Bayes, Logistic Regression, Random Forests, Decision Tree-based Bagging, AdaBoost, XGBoost, CatBoost, LightGBM, and Gradient Boosting. For hyperparameter tuning, we used LBFGS, Newton-Cholesky, and Newton-Cg solver methods for logistic regression while for the rest of the models, we used Bayesian Optimization. For the models that will be tuned with Bayesian Optimization, we'll go 15 iterations of 5 cross-validation rounds (Gradient Boosting, which is historically known for having longer training phases compared to its peers, will be the exception with 5 iterations of 5 cross-validation rounds). As a criteria of efficiency, we choose the model with not only a high performance according to ROC-AUC, but also fast training (ideally less than an hour) and response/prediction (ideally less than half of a second to a whole second) times as well.

Results:

With SMOTE:

a. Baseline

Model Name	**Training Time**	**Prediction Time**	Accuracy	**ROC-AUC**	Precision	Recall	F-Beta (Beta = 2)
Gaussian Naive Bayes	00:00:00.03331	00:00:00.00132	0.66787	0.91755	0.6003	0.99462	0.87913
Logistic Regression	00:00:01.64413	00:00:00.00125	0.88421	0.9615	0.88417	0.88291	0.88316
Random Forest	00:00:13.65429	00:00:00.00576	0.97607	0.99778	0.97355	0.97847	0.97748
Bagging	00:00:02.87225	00:00:00.00207	0.97251	0.99473	0.98034	0.96408	0.96729
AdaBoost	00:00:05.2386	00:00:00.01281	0.93716	0.98883	0.93332	0.94091	0.93938
XGBoost	00:00:01.52772	00:00:00:00056	0.9747	0.99716	0.99046	0.95837	0.96462
CatBoost	00:00:22.98753	00:00:00.00057	0.98115	0.99788	0.99402	0.96793	0.97304
LightGBM	00:00:01.2492	00:00:00.00088	0.97612	0.99734	0.9934	0.95837	0.96518
Gradient	00:00:20.	00:00:00.	0.95486	0.99373	0.96184	0.94684	0.9498

Boosting	72583	00091					

b. With Hyperparameter Tuning/Optimization

Model Name	Optimization Type	**Training Time**	**Prediction Time**	Accuracy	**ROC-AUC**	Precision	Recall	F-Beta (Beta = 2)
Gaussian Naive Bayes	Bayes	00:00:10.78046	00:00:00.00049	0.8377	0.91755	0.87771	0.78284	0.80014
Logistic Regression	Newton-Cg	00:00:31.36576	00:00:00.0009	0.88443	0.96151	0.88448	0.88302	0.88331
Logistic Regression	Newton-Cholesky	00:00:02.59841	00:00:00.00079	0.88443	0.96151	0.88448	0.88302	0.8831
Logistic Regression	LBFGS	00:00:13.08582	00:00:00.00094	0.88437	0.96151	0.88446	0.88291	0.88322
Random Forest	Bayes	01:09:52.75836	00:00:00.01361	0.9782	0.99799	0.9772	0.97902	0.97865
Bagging	Bayes	00:46:55.23581	00:00:00.05251	0.98102	0.99785	0.98906	0.97276	0.97598
AdaBoost	Bayes	00:26:51.89815	00:00:00.05193	0.9724	0.99653	0.99042	0.95376	0.96087
XGBoost	Bayes	00:01:36.69933	00:00:00.00084	0.97503	0.99667	0.99519	0.95442	0.9623
CatBoost	Bayes	00:43:14.84239	00:00:00.00106	0.98186	0.99787	0.99314	0.97023	0.97473
LightGBM	Bayes	00:11:22.44298	00:00:00.00127	0.98153	0.99799	0.9857	0.97704	0.97876
Gradient Boosting	Bayes	02:09:10.38001	00:00:00.00415	0.98224	0.99833	0.98358	0.98067	0.98125

<u>With ADASYN</u>

a. Baseline

Model Name	**Training Time**	**Prediction Time**	Accuracy	**ROC-AUC**	Precision	Recall	F-Beta (Beta = 2)
Gaussian	00:00:00.	00:00:00.	0.6731	0.87884	0.60383	0.99725	0.88228

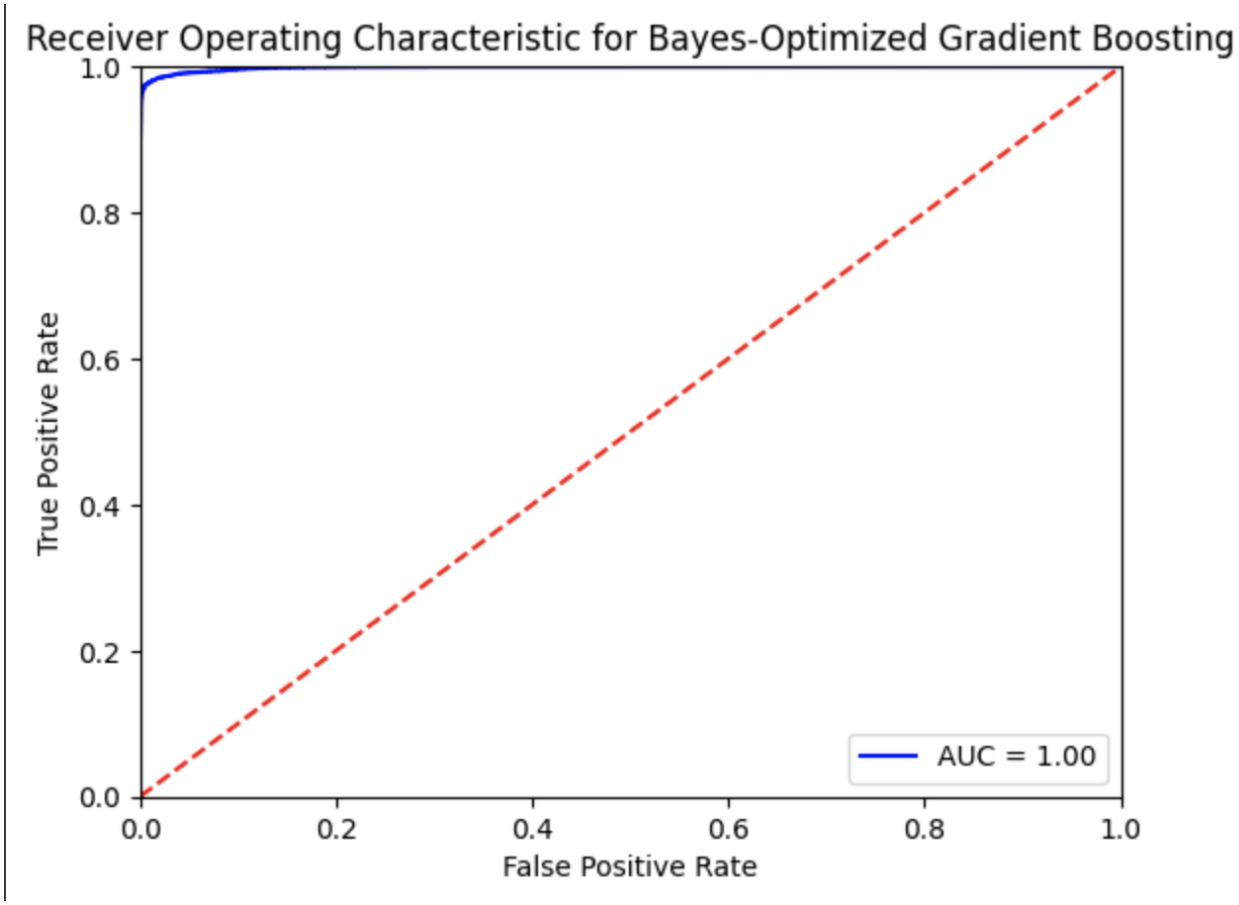
Naive Bayes	04858	00114					
Logistic Regression	00:00:01.52206	00:00:00.00082	0.83488	0.91094	0.81196	0.86949	0.85734
Random Forest	00:00:24.76945	00:00:00.00824	0.97584	0.99779	0.96539	0.98682	0.98246
Bagging	00:00:05.43833	00:00:00.00253	0.97393	0.99515	0.97679	0.97067	0.97189
AdaBoost	00:00:05.76573	00:00:00.01348	0.91676	0.97972	0.88888	0.95166	0.93841
XGBoost	00:00:01.13154	00:00:00.0007	0.96906	0.99657	0.98085	0.9565	0.96127
CatBoost	00:00:58.33225	00:00:00.00087	0.98202	0.998	0.99392	0.96979	0.97452
LightGBM	00:00:02.85959	00:00:00.00185	0.97409	0.99703	0.98757	0.96001	0.9654
Gradient Boosting	00:00:38.46797	00:00:00.00071	0.94015	0.99065	0.92361	0.95902	0.95173

b. With Hyperparameter Tuning/Optimization

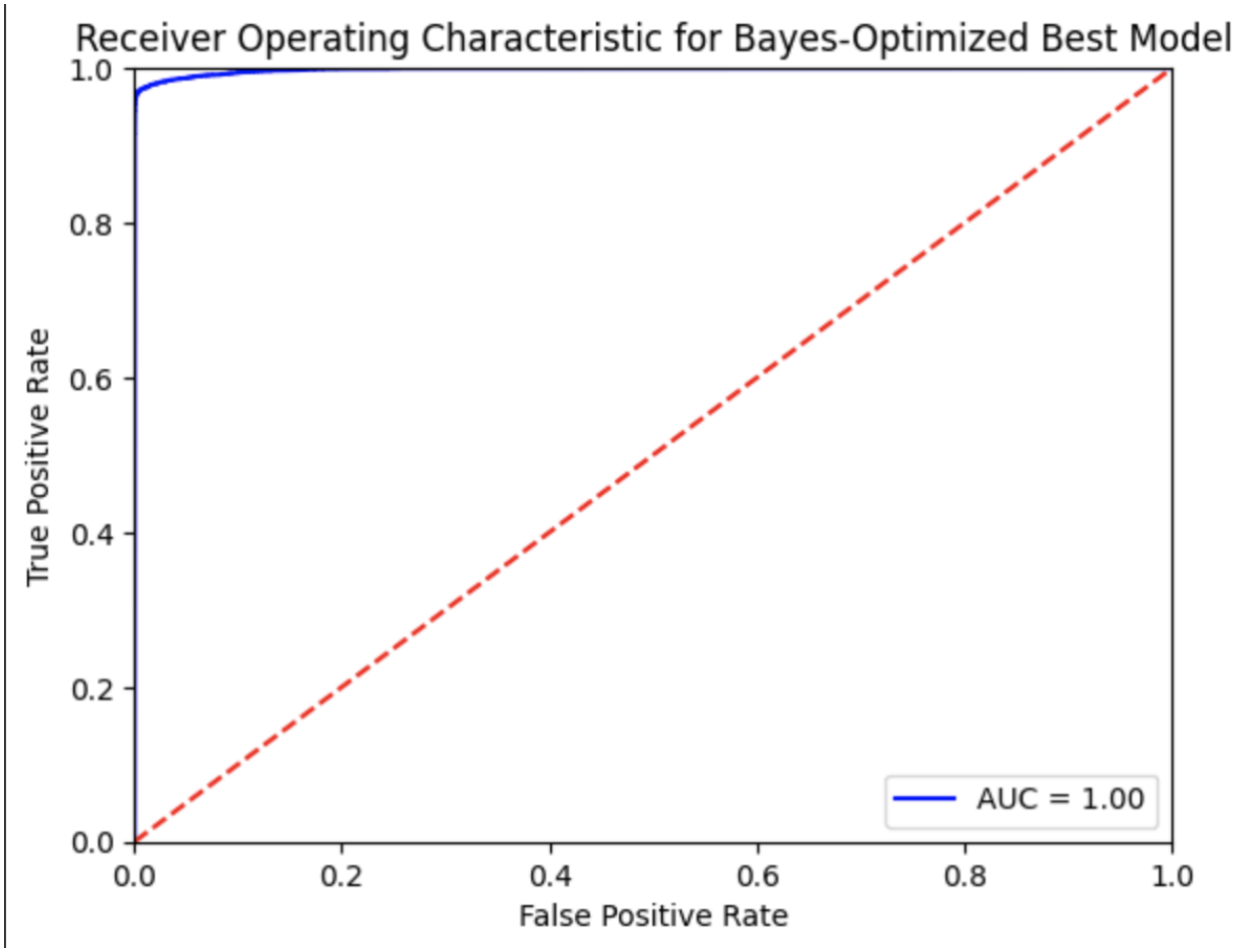
Model Name	Optimization Type	**Training Time**	**Prediction Time**	Accuracy	**ROC-AUC**	Precision	Recall	F-Beta (Beta = 2)
Gaussian Naive Bayes	Bayes	00:00:16.87278	00:00:00.00061	0.79531	0.87872	0.81779	0.75733	0.7687
Logistic Regression	Newton-Cg	00:00:52.53083	00:00:00.0005	0.8345	0.91087	0.81061	0.87081	0.85807
Logistic Regression	Newton-Cholesky	00:00:03.96393	00:00:00.00111	0.8345	0.91087	0.81061	0.87081	0.85807
Logistic Regression	LBFGS	00:00:24.30636	00:00:00.00047	0.83439	0.91087	0.81051	0.8707	0.85796
Random Forest	Bayes	01:37:47.73777	00:00:00.0178	0.97699	0.99797	0.9692	0.98506	0.98185
Bagging	Bayes	01:36:41	00:00:00	0.98289	0.99828	0.98931	0.97616	0.97876

		.75314	.06175					
AdaBoost	Bayes	00:36:07 .87613	00:00:00 .03483	0.96595	0.99573	0.97964	0.95133	0.95686
XGBoost	Bayes	00:02:10 .28595	00:00:00 .00074	0.96048	0.99509	0.9667	0.95342	0.95605
CatBoost	Bayes	00:57:11 .72478	00:00:00 .00088	0.98235	0.99808	0.99028	0.97407	0.97727
LightGBM	Bayes	00:16:57 .74728	00:00:00 .00432	0.98092	0.9982	0.98311	0.97847	0.97939
Gradient Boosting	Bayes	03:05:38 .29679	00:00:00 .00179	0.98306	0.99852	0.98213	0.98385	0.98351

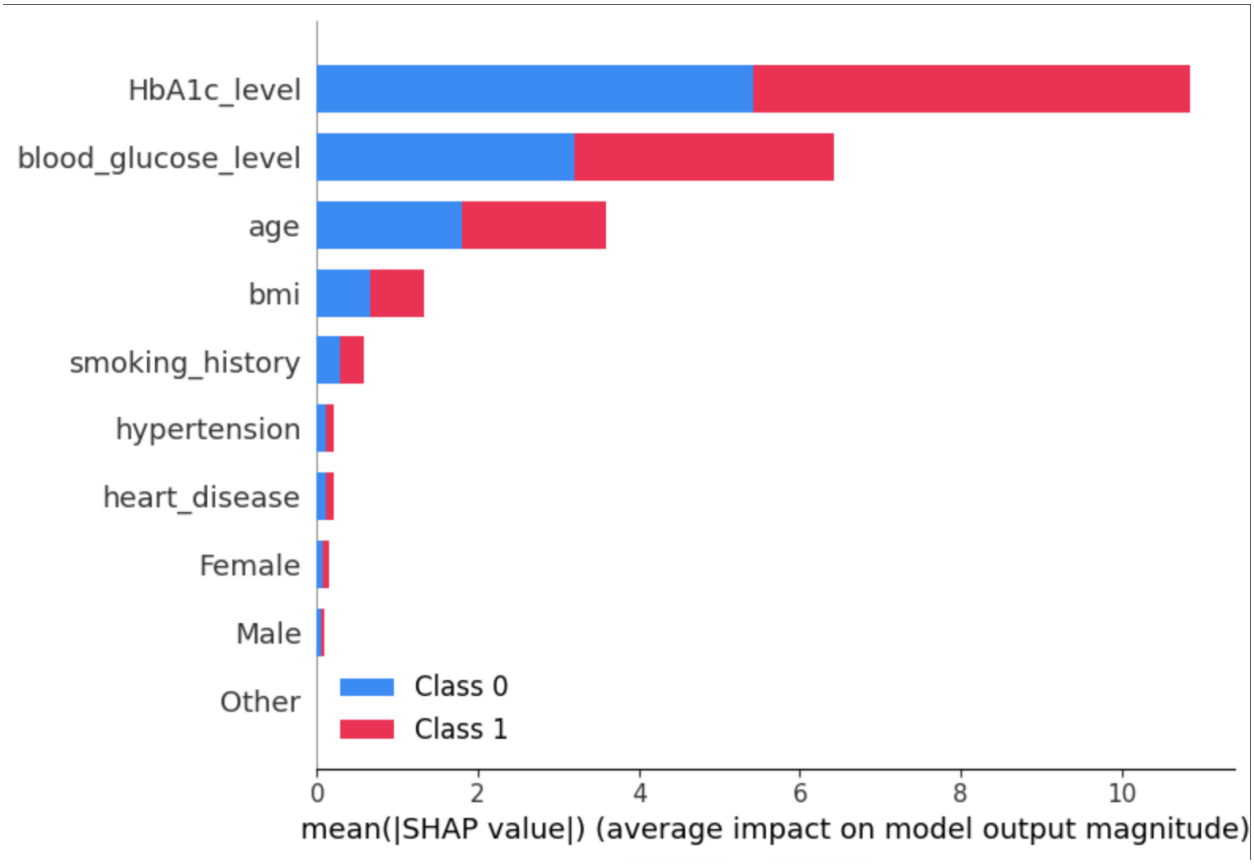
ROC-AUC of Gradient Boosting



ROC-AUC of LightGBM:



SHAP Interpretation Results based on LightGBM:



While Gradient Boosting model, trained with both SMOTE- and ADASYN-balanced datasets, is the best performing model in terms of ROC-AUC, it has the slowest training times when trained with both SMOTE- and ADASYN-balanced datasets. The fastest model trained is the Gaussian Naive Bayes trained with the ADASYN-balanced dataset, but it is among the worst-performing models in terms of ROC-AUC. All the models used in the study have response/prediction times of less than half of second. The best performing model in terms of efficiency (high ROC-AUC score and fast training time) is LightGBM trained with the ADASYN-balanced dataset (third-best scoring model and under an hour of training time).

According to the LightGBM model trained with the ADASYN-balanced dataset, the 5 most significant features are Hb1Ac level, blood glucose level, age, body mass index (BMI), and smoking history. While heart disease played a bit of a role in the development of diabetes, compared to the aforementioned features, it did not play as or that significant of a role in the risk and development of diabetes and in fact one of the least influential features in the risk and development of diabetes. The same can be said for gender. While being male or female played a bit of a role in the development of diabetes, and according to the interpretation results, females have a slightly higher chance of developing diabetes compared to their male counterparts, like heart disease, it did not play as or that significant of a role in the risk and development of diabetes and in fact the least influential features in the risk and development of diabetes.

Identifying as another gender other than male and female does not or rarely influences the risk and development of diabetes.

Future Directions and Conclusion:

Based on the performance metrics of our model, we believe that our model can effectively both detect patients who have actually developed diabetes properly and avoid misdiagnosing those without diabetes as diabetic while taking in consideration of the possible costs between these two situations. With the relatively fast training time while also high performance as well, we believe that our model can be applied to high-risk situations involving diabetes while also maintaining its high accuracy and performance.

In the future, we want to investigate how diabetes affects the development and risk of other diseases such as heart disease. We also want to investigate how diabetes can be developed in patients who are currently diagnosed as non-diabetic in the next year or so. For possible improvement of model performance, we may only consider the top significant features as input for our model. Also, as a possible suggestion of improvement of model performance and considering the relatively large size of the dataset, we might also implement Deep Neural Networks for our dataset. Also to put our results in real-life applications, we plan to develop a user-friendly web app in the sometime future.

References:

- Loke, A. "Diabetes." *World Health Organization*, World Health Organization, 5 Apr. 2023, www.who.int/news-room/fact-sheets/detail/diabetes.
- Tasin I, Nabil TU, Islam S, Khan R. Diabetes prediction using machine learning and explainable AI techniques. Healthc Technol Lett. 2022 Dec 14;10(1-2):1-10. doi: 10.1049/htl2.12039. PMID: 37077883; PMCID: PMC10107388.
- Anant Ram and Honey Vishwakarma 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1116 012135
- Fregoso-Aparicio, L., Noguez, J., Montesinos, L. et al. Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. Diabetol Metab Syndr 13, 148 (2021). https://doi.org/10.1186/s13098-021-00767-9
- Zou Q, Qu K, Luo Y, Yin D, Ju Y and Tang H (2018) Predicting Diabetes Mellitus With Machine Learning Techniques. Front. Genet. 9:515. doi: 10.3389/fgene.2018.00515
- Qin Y, Wu J, Xiao W, Wang K, Huang A, Liu B, Yu J, Li C, Yu F, Ren Z. Machine Learning Models for Data-Driven Prediction of Diabetes by Lifestyle Type. International Journal of Environmental Research and Public Health. 2022; 19(22):15027. https://doi.org/10.3390/ijerph192215027
- Aishwarya Mujumdar, V Vaidehi, Diabetes Prediction using Machine Learning Algorithms, Procedia Computer Science, Volume 165, 2019, Pages 292-299, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.01.047.
- K. Hasan et al.: Diabetes Prediction Using Ensembling of Different ML Classifiers