PREDICTING DIABETES AMONG WOMEN USING MACHINE LEARNING MODELS

Department of Data Analytics and Information Systems,

McCoy College of Business, Texas State University.

Vennela Innamuri, Sukanya Ramasani

Ankith Reddy Ragipindi, Bhuvana Chandra Kolipaka

ABSTRACT:

Diabetes is a chronic illness that can affect at any age due to various factors including pregnancy. This paper focuses on data on women and their bio markers. The models used to study predictions and correlation between variables in this paper are Ordinary Least Square, Logistics Regression, Naïve Bayes, KNN, Support vector machine, Discriminant analysis, Decision tree, Neural networks, and ensemble models like bagging, boosting and Random Forest. Performed outlier observations, feature engineering to capture more underlying data and normalization to balance the dataset. Performed Oversampling and Undersampling then produced classification report to evaluate each model. This paper aims to identify key features for predicting diabetes, analyze correlation between the biomarkers in the dataset and outcome variable, and recommend the best model.

INTRODUCTION

Diabetes is one of the most pervasive diseases in the world. It's a chronic illness caused by the lack of inherent ability to regulate sugar in blood. Insulin, produced by the pancreas, is the component responsible for controlling sugar levels. The prevalence of this chronic illness in the modern era despite the advancements in the medical field can be attributed to many factors including auto immune diseases, hormonal imbalance, lifestyle choices and use of medication. These factors are contributing to the onset of diabetes regardless of age. As diabetes don't have a cure, prevention and early detection is important. To achieve this, machine learning models are being used. This paper focuses on understanding diabetes among women recognizing the impact of hormonal fluctuations women experience regularly, particularly during pregnancy. These effects last longer even post pregnancy, sometimes lasting throughout their lifetime.

LITERATURE REVIEW

Sisodia and Sisodia (2018) explored the prediction of diabetes using machine learning algorithms. In this study, the models used were Decision tree, SVM and Naïve Bayes on the PIMA Indian Diabetes database. Naive bayes exhibited the highest accuracy of 76.30%. This research highlighted the significance of early detection of diabetes and potential of machine learning in diagnosing in medical field. This opens up new avenues for further automation and exploration of alternative algorithms.

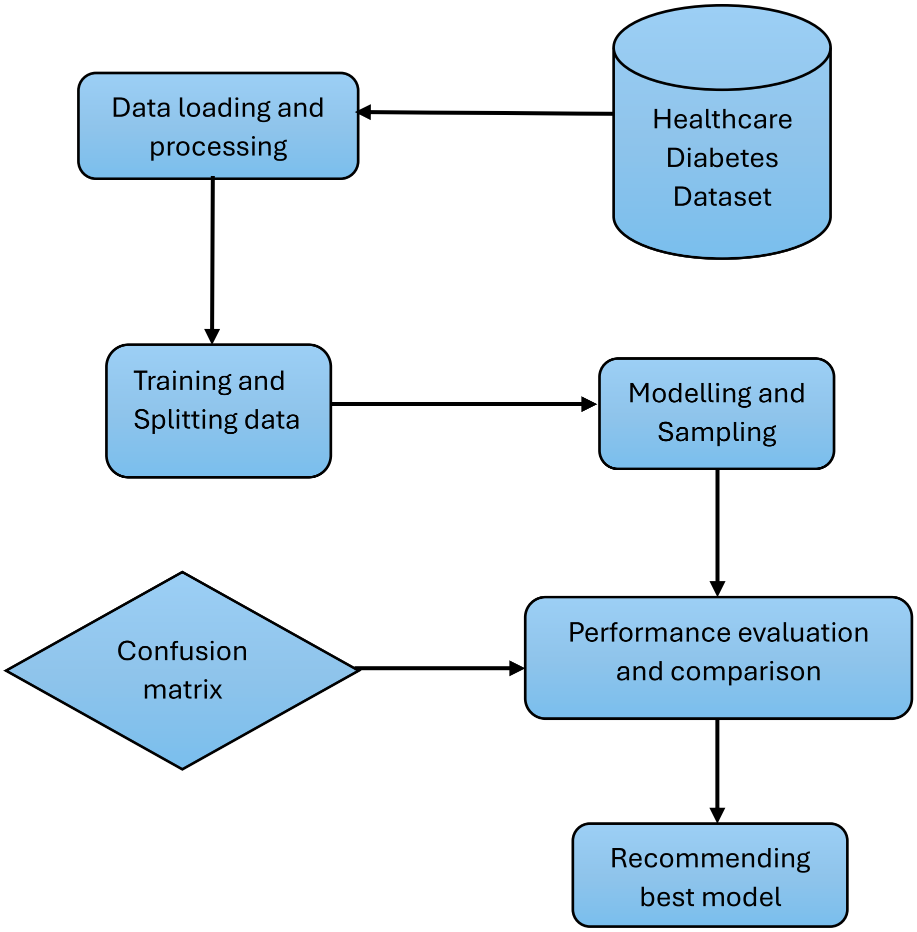
Sarwar et al. (2018) investigated the application of 6 machine learning models, SVM, KNN, LR, DT, RF and NM on the PIMA Indian dataset for predicting diabetes. The results indicated that SVM and KNN achieved highest accuracy of 77% suggesting their suitability for predicting analytics in healthcare. However, the study does acknowledge the limitations related to dataset and missing values. A larger dataset with more attributes and no NAN value will be proved higher accuracy.

Tasin et al (2023) proposed an automatic diabetic prediction system which utilizes various machine learning models like XGBoost classifier on two datasets: PIMA Indian and female Bangladeshi patients. Through techniques like SMOTE and ADASYN techniques, the system achieved 81% accuracy and demonstrated versatility through domain adaptation. The XGBoost framework was deployed into a website and smart phone application enabling instant diabetic prediction.

Hasan et al (2020) did research on patients demonstrating that diabetes among adults (over 18 years old) has risen from 4.7% to 8.5% in 1980 to 2014 respectively and rapidly growing second and third countries. The author implemented the LDA, Naïve Bayes, Gaussian Process classification, SVM, Neural Network, Logistic Regression and Random Forest. All techniques were performed with extensive experiments on the outlier and missing values and performed with the maximum accuracy. Random forest is the best model with accuracy of 0.94% classified for the data when the trained model will be user-friendly interface.

Xue et al (2020) used supervised machine learning algorithms such as SVM, Naïve Bayes, LightGBM to predict diabetes. This study analyzed 520 patients and SVM showed the highest accuracy, highlighting the importance of early detection of diabetes. The study also underscores the evolving role of machine learning in revolutionizing diabetes risk prediction, SVM has come out as helpful tool for clinical practitioners in making informed decisions.

MATERIALS AND METHODS:



Dataset: The dataset in this paper is sourced from the National Institute of Diabetes and Digestive and Kidney Diseases. The dataset has 2768 records of individual women. The dataset in total has 9 variables: 8 predictors and one outcome variable indicating (0) non-diabetic or (1) diabetic status. The dataset consists of 1816 non- diabetic and 952 diabetic females. The predictors are various bio marks such as insulin, glucose, blood pressure, BMI, Age. The objective is to study the effect of each bio marker on the outcome variable.

Table 1: Description of attributes

S.no	Attributes	Description
1	Pregnancies	Number of times the individual pregnant throughout life
2	Glucose	Plasma glucose concentration in blood post 2 hours in an oral glucose tolerance test
3	Blood pressure	Diastolic blood pressure (mm hg)
4	Skin thickness	Triceps skin fold pressure (mm)
5	Insulin	2-Hour serum insulin (mu U/ml)
6	BMI	Body mass index
7	Diabetes Pedigree Function	Score on Likelihood of diabetes based on family history.
8	Age	Age of the individual
9	Outcome	Class 0 or 1

Figure 1: Distribution of all the variables in the dataset

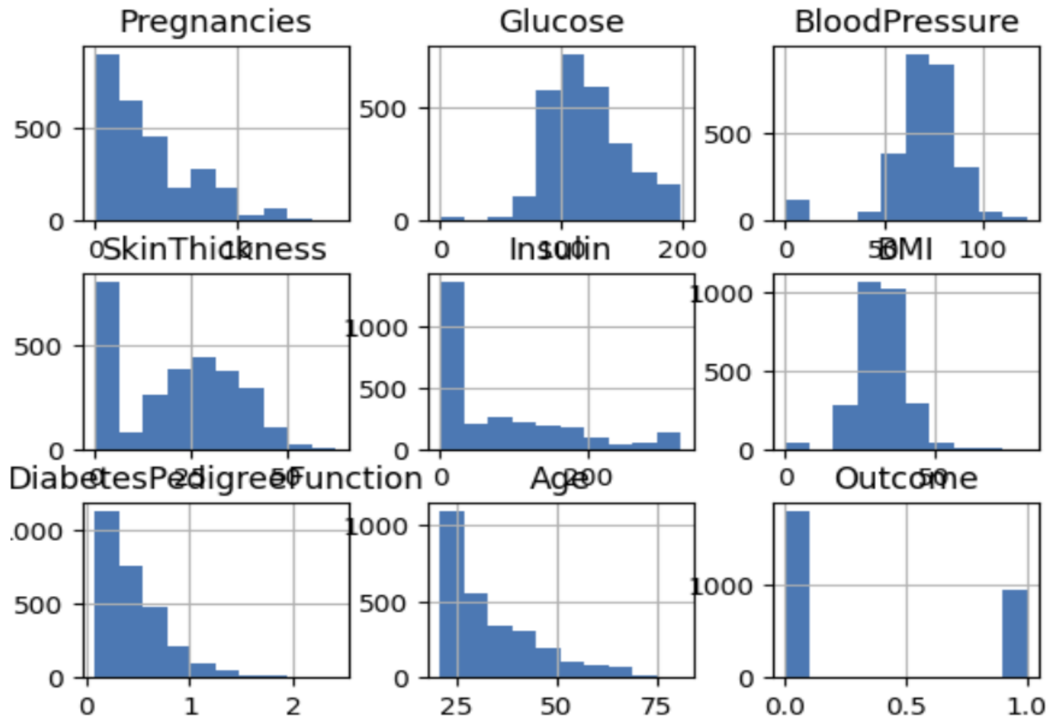


Figure 2: Heat map displaying correlations between each variable without feature engineering.

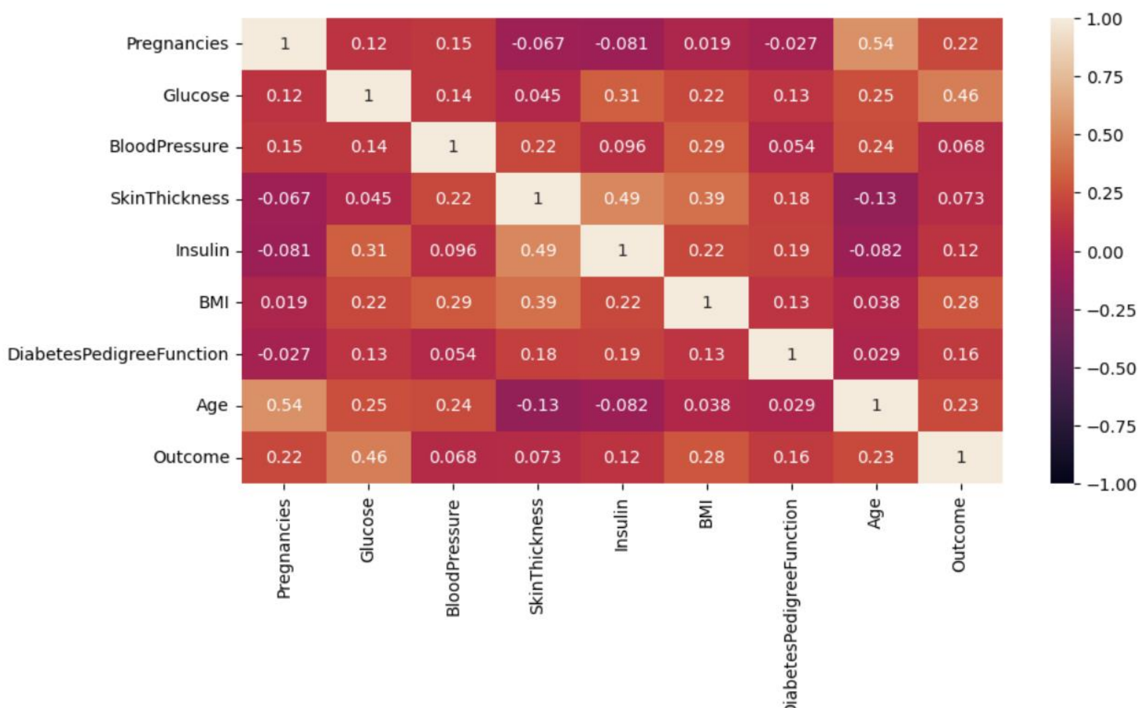
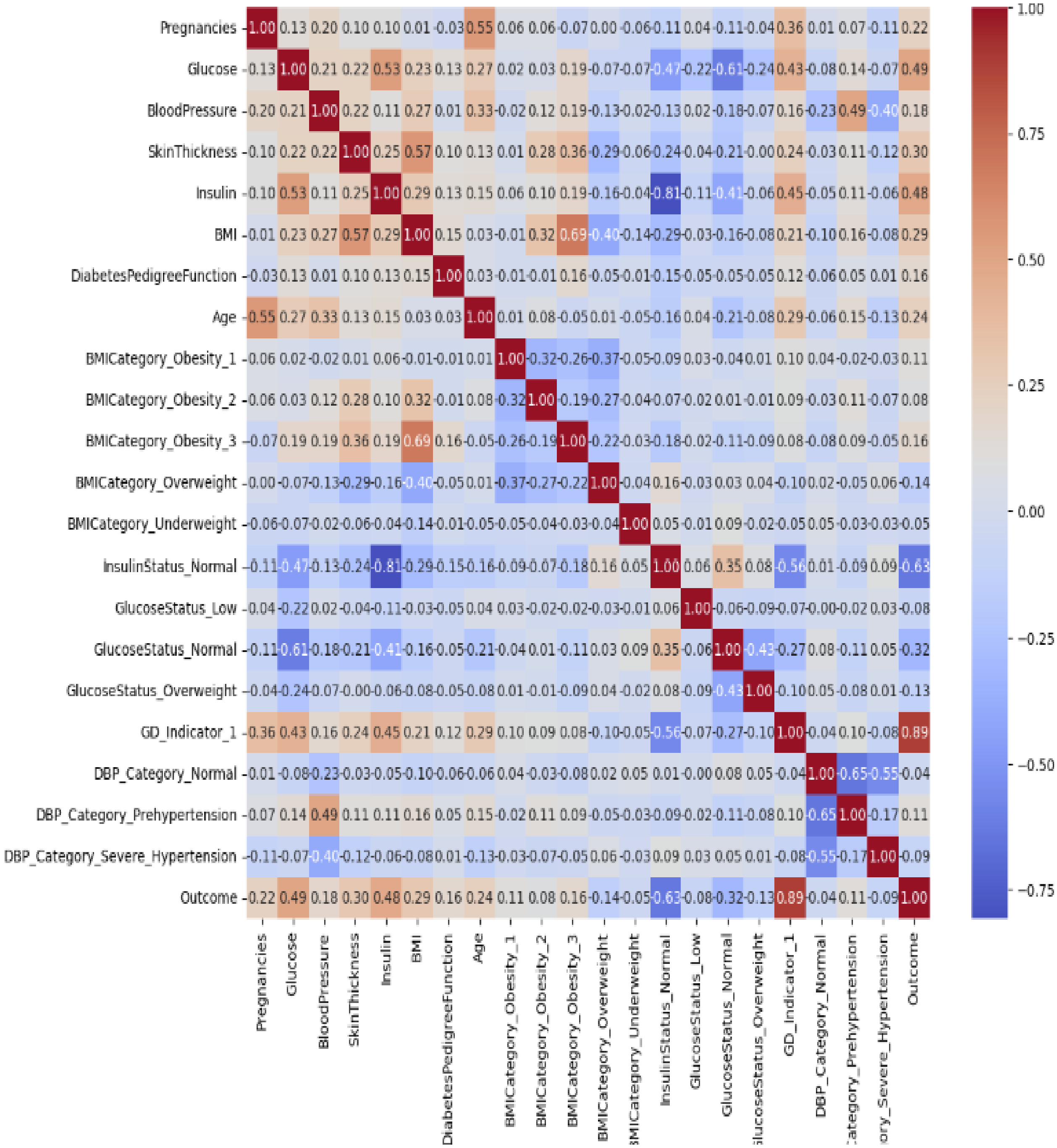


Figure 3: Heat map displaying correlations with feature engineering.



Heat map provides visual representation of correlation between variables with varying shades of color or easier understanding. Values closer to '+1' indicate strong positive relationship and '-1' indicates negative relationship while 0 means no correlation. Insulin is negatively correlated to InsulinStatus_Normal (-0.81), implying increase in insulin levels, the likelihood of decrease in InsulinStatus_Normal. Insulin and glucose have a strong negative correlation (0.63), suggesting that lower insulin levels are associated with higher chances of being diabetic.

Age and pregnancies have a moderate positive correlation implying the increase in age has likelihood in increase in pregnancies. 'SkinThickness' seems to have moderately strong positive relation with BMI. Diabetic pedigree function which is based on family history doesn't seem to have much correlation with other bio markers which could mean that the influence of these bio markers may not have anything to with genetically inherited diseases.

Data preparation and processing: The dataset was loaded into the Google Colab environment using pandas library. Imputation based on median was performed to filter out any 0's and replace with Nan. Outlier observation analysis was done on insulin as it is the main bio marker influencing diabetes. Interquartile formula (IQR):

IQR = Q3 - Q1
lower = Q1 - 1.5 * IQR
upper = Q3 + 1.5 * IQR, here, Q is quarter.

Figure4: Outlier observation on Insulin

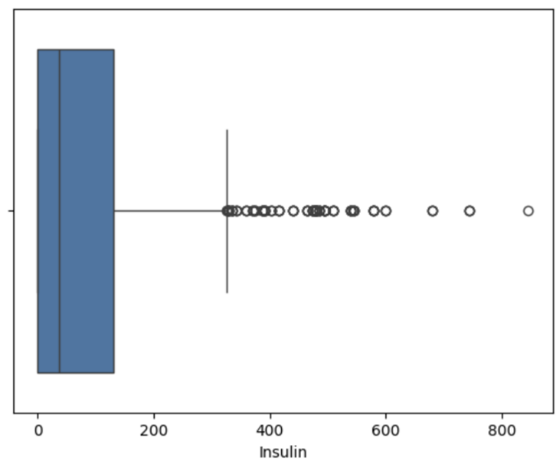
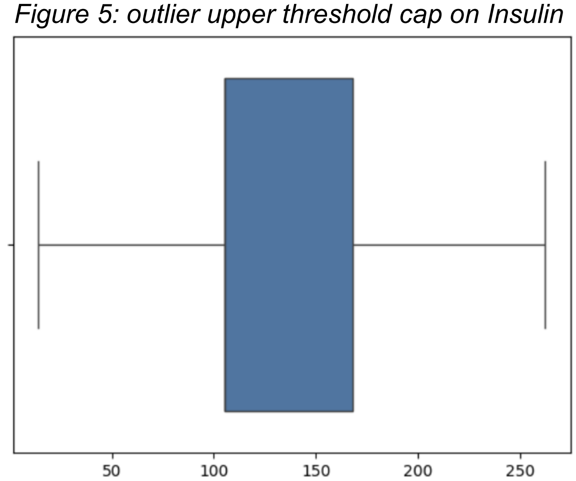


Figure 5: outlier upper threshold cap on Insulin



Feature Selection and Engineering: To capture nuanced relationships and underlying trends, new features are engineered with existing data by combing them or creating new functions. This paper categorized BMI into 5 different classes to understand how different BMI categories relate to the risk diabetes. The categories are Underweight, Normal, overweight, Obesity class 1, Obesity class 2, Obesity class 3.

Insulin has two categories, Normal and Abnormal. Based on the given threshold. With this, relating insulin with other variables is easier. Similarly, Glucose has been categorized into 4 classes: Low, Normal, Overweight, High. Next, created a binary variable for gestational diabetes with pregnancies and outcome variables. Then used blood pressure to categorize 4 new classes: Normal, prehypertension, Hypertension, severe hypertension. This categorization allows better interpretation and understanding of the relationship between variables and risk of diabetes.

One Hot Encoding and label encoding: Creating new features created new variables with categorical data types. By creating dummy variables with 'pd.get.dummies()' these categorical variables are converted into numerical data type as most machine learning models require numerical input. Next, create Dataframe with the new categorical variables is split from the original predictor variables to perform normalization and the concatenated. The further machine learning modeling is done on this new dataset.

Training and Testing: The new dataset is split into 80:20 ratio where 80% of the data is used for training and 20% of the data is validated. Oversampling is done with SMOTE() package and Undersampling is done with RandomUnderSampler package.

MODELS AND TECHNIQUES:

Ordinary least squares: Based on the OLS regression results it can be observed that R-squared is 0.843 i.e, the independent variables can explain variance of dependent variables by 84.3%. Therefore, the model is considered a good fit. Notably, BMI categories, Age, GlucoseStatus_Low have p-values more than 0.05 proving to be statistically insignificant to diabetes.

It can also be observed that Pregnancies, Insulin, BMI Categories, InsulinStatus_Normal, GlucoseStatus categories are negative coefficients indicating a negative relationship with the outcome variable, where increase in the independent variable leads to decrease in independent variable. The larger the coefficient the stronger the effect of the relationship.

Figure 5: OLS regression results with feature engineering

OLS Regression Results

Dep. Variable:	Outcome	**R-squared:**	0.843
Model:	OLS	**Adj. R-squared:**	0.842
Method:	Least Squares	**F-statistic:**	702.7
Date:	Mon, 06 May 2024	**Prob (F-statistic):**	0.00
Time:	11:38:23	**Log-Likelihood:**	697.91
No. Observations:	2760	**AIC:**	-1352.
Df Residuals:	2738	**BIC:**	-1222.
Df Model:	21		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-0.0935	0.074	-1.271	0.204	-0.238	0.051
Pregnancies	-0.0136	0.001	-9.914	0.000	-0.016	-0.011
Glucose	0.0010	0.000	3.619	0.000	0.000	0.002
BloodPressure	0.0011	0.000	2.588	0.010	0.000	0.002
SkinThickness	0.0018	0.001	3.466	0.001	0.001	0.003
Insulin	-0.0010	0.000	-8.805	0.000	-0.001	-0.001
BMI	0.0058	0.002	3.827	0.000	0.003	0.009
DiabetesPedigreeFunction	0.0336	0.011	2.956	0.003	0.011	0.056
Age	0.0004	0.000	0.919	0.358	-0.000	0.001
BMICategory_Obesity_1	-0.0082	0.019	-0.428	0.669	-0.046	0.029
BMICategory_Obesity_2	-0.0494	0.026	-1.923	0.055	-0.100	0.001
BMICategory_Obesity_3	-0.0501	0.037	-1.371	0.171	-0.122	0.022
BMICategory_Overweight	-0.0174	0.014	-1.202	0.230	-0.046	0.011
BMICategory_Underweight	0.0229	0.052	0.438	0.661	-0.079	0.125
InsulinStatus_Normal	-0.2149	0.014	-15.296	0.000	-0.242	-0.187
GlucoseStatus_Low	-0.0468	0.041	-1.150	0.250	-0.127	0.033
GlucoseStatus_Normal	-0.0444	0.020	-2.245	0.025	-0.083	-0.006
GlucoseStatus_Overweight	-0.0331	0.014	-2.409	0.016	-0.060	-0.006
GD_Indicator_1	0.8217	0.011	77.176	0.000	0.801	0.843
DBP_Category_Normal	0.1016	0.024	4.256	0.000	0.055	0.148
DBP_Category_Prehypertension	0.0929	0.023	3.993	0.000	0.047	0.139
DBP_Category_Severe_Hypertension	0.1011	0.028	3.670	0.000	0.047	0.155

Omnibus:	1583.562	**Durbin-Watson:**	1.983
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	12217.093
Skew:	2.693	**Prob(JB):**	0.00
Kurtosis:	11.788	**Cond. No.**	4.74e+03

As the outcome variable for diabetes has binary class 0 and 1, we chose binary classification models to predict outcome. In this paper, supervised learning is used as the models are trained and tested on the samplings of existing data. The techniques used are imputing, scaling, sampling, hyperparameter tuning using GridSearchCV. The models used are Logistics Regression, Decision Tree (fully grown & pruned), Naïve Bayes, K-Nearest Neighbor, Support Vector Machine, Discriminant Analysis, Neural Network, and Ensemble models bagging, boosting and Random Forest were also used. These models use sklearn package and produced classification reports for original, oversampling, Undersampling for each model.

RESULT AND DISCUSSION

In this study, all these models were applied on 'Healthcare-Diabetes' dataset. The data was split into 80% training data and 20% testing data. The main parameters compared and evaluated are recall, precision and accuracy.

Comparison Analysis: We performed models with and without feature engineering. The least square regression model showed R square as 0.309 which means the independent variables can explain variance in dependent variable by 30%, which means worst fit. With feature engineering, OLS showed 84% indicating good fit. That is, independent variables can explain variance in dependent variable by 84%. We performed models with and without feature engineering. The least square regression model showed R-square as 0.309 which means the independent variables can explain 30% of the variance in dependent variable, suggesting worst fit. However, after implementing feature engineering, OLS showed R-square value of 84.3This indicates that the independent variables can explain 84.3% variance in dependent variable, suggesting good fit. This difference in R square value proves that the model's ability to explain variance in dependent variable has substantially improved. This indicates that the engineered features are more informative and better aligned with the target variable. This leads to a more accurate relationship between predictors and outcome variable.

Figure 6: OLS without feature engineering

OLS Regression Results

Dep. Variable:	Outcome	**R-squared:**	0.309
Model:	OLS	**Adj. R-squared:**	0.307
Method:	Least Squares	**F-statistic:**	154.0
Date:	Mon, 06 May 2024	**Prob (F-statistic):**	1.26e-214
Time:	22:46:14	**Log-Likelihood:**	-1350.5
No. Observations:	2759	**AIC:**	2719.
Df Residuals:	2750	**BIC:**	2772.
Df Model:	8		
Covariance Type:	nonrobust		

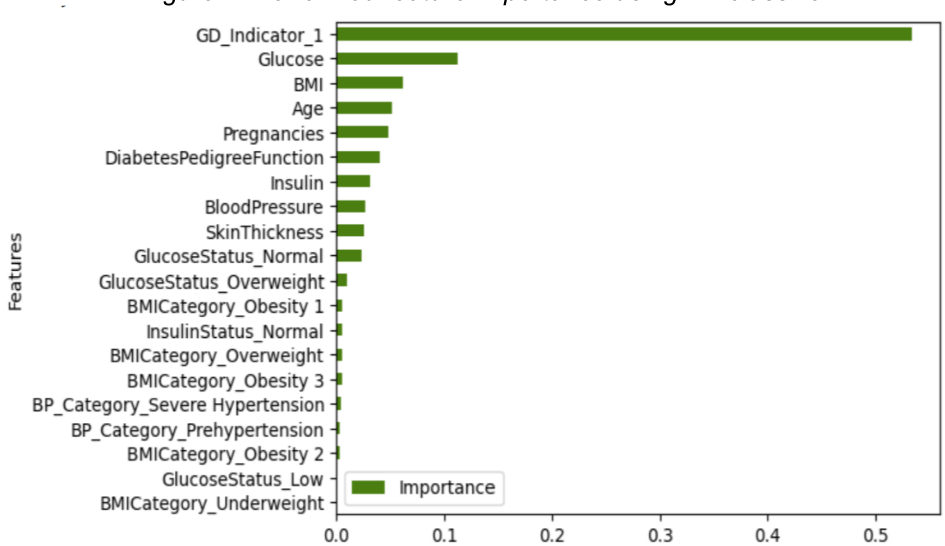
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-1.0026	0.055	-18.142	0.000	-1.111	-0.894
Pregnancies	0.0246	0.003	8.093	0.000	0.019	0.031
Glucose	0.0064	0.000	22.038	0.000	0.006	0.007
BloodPressure	-0.0005	0.001	-0.760	0.447	-0.002	0.001
SkinThickness	0.0008	0.001	0.769	0.442	-0.001	0.003
Insulin	-9.873e-05	0.000	-0.715	0.475	-0.000	0.000
BMI	0.0113	0.001	8.562	0.000	0.009	0.014
DiabetesPedigreeFunction	0.1228	0.024	5.217	0.000	0.077	0.169
Age	0.0019	0.001	2.392	0.017	0.000	0.003

Omnibus:	98.426	**Durbin-Watson:**	1.993
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	72.573
Skew:	0.299	**Prob(JB):**	1.74e-16
Kurtosis:	2.478	**Cond. No.**	1.55e+03

Table 2: Importance and ranking of features using random forest classifier.

Importance Ranking	Features	Importance
1	Glucose	0.068136
2	BloodPressure	0.020638
3	SkinThickness	0.061488
4	Insulin	0.185742
5	BMI	0.042162
6	DiabetesPedigreeFunction	0.024733
7	Age	0.036132
8	BMICategory_Obesity_1	0.003246
9	BMICategory_Obesity_2	0.003949
10	BMICategory_Obesity_3	0.003595
11	BMICategory_Overweight	0.002028
12	BMICategory_Underweight	0.000001
13	BloodPressure	0.020638
14	GlucoseStatus_Low	0.000329
15	GlucoseStatus_Normal	0.020140
16	GlucoseStatus_Overweight	0.005391
17	GD_Indicator_1	0.364198
18	DBP_Category_Normal	0.003115
19	DBP_Category_Prehypertension	0.002615
20	DBP_Category_Severe_Hypertension	0.001965

Figure 7: Performed feature importance using RF classifier.



Model performance: From the above tabular columns, the performance metrics of various machine learning models across different data sampling techniques, focusing on binary classification outcomes ('0' and '1' classes). Each model is assessed based on accuracy, precision, recall, and potentially F1-scores, with these metrics serving as key indicators of their classification effectiveness.

Table 3: Accuracy, Precision, Recall of each model.

Models	Accuracy	Precision		Recall	
		0	1	0	1
Navie Bayes					
Original	88	99	73	83	99
OverSample	86	99	70	80	99
UnderSample	86	99	70	80	99
Decision Tree fully grown					
Original	100	100	100	100	99
OverSample	95	94	100	100	86
UnderSample	98	99	96	98	98
Decision Tree - pruned					
Original	98	99	96	98	99
OverSample	97	99	94	97	99
UnderSample	98	99	96	98	98
Logistic Regression					
Original	95	96	94	97	92
OverSample	97	98	95	98	95
UnderSample	98	99	97	98	98
KNN k=3					
Original	97	99	93	97	97
OverSample	98	100	94	97	100
UnderSample	94	99	85	92	99
Bagging					
Original	100	100	100	100	99
OverSample	100	100	100	100	99
UnderSample	100	100	100	100	99
Boosting;					
Original	100	100	100	100	99
OverSample	99	100	98	99	100
UnderSample	100	100	100	100	99
random forest					
Original	100	100	100	100	99
OverSample	100	100	100	100	99
UnderSample	100	100	99	100	99

Neural Network					
Original	68	68	0	100	0
OverSample	69	69	69	99	5
UnderSample	97	100	93	97	99
SVM					
Original	97	97	97	99	93
OverSample	96	98	92	96	96
UnderSample	96	98	91	96	96
Discriminant Analysis					
Original	95	94	99	100	86
OverSample	95	94	99	100	86
UnderSample	95	94	99	100	86

CONCLUSION

Predictive analysis of chronic diseases like diabetes has come a long way and changed the way medical practitioners and researchers gain insights from the data and make informed decisions. In this paper binary class models were used on 2769 records. It can be observed that logistic regression had the highest accuracy for predicting diabetes. For models like neural network the values had large differences and weren't balanced. While ensemble models have 100% indicating overfitting. Comparing accuracy, precision, recall values of all models, SVM, Logistic Regression, Decision Tree – Pruned can be recommended as best models. Their values are high and balanced as well.

LIMITATIONS AND FUTURE WORKS

To further improve prediction, larger data with zero missing values, more variables like fasting insulin, fasting glucose, temporal data will be needed. Relationships between predictor variables can be investigated more to draw out new features and capture underlying trends. Dataset can be tuned better to avoid overfitting. Confusion matrix can be done to count false positives, True positives, True negatives, and False negatives so that errors can be captured and rectified. Training and testing with larger dataset and more significant variables will provide more valuable insights for better accurate prediction.

REFERENCE

Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2022). Diabetes prediction using machine learning and explainable AI techniques. Healthcare technology letters, 10(1-2), 1–10. https://doi.org/10.1049/htl2.12039

Sisodia, D. and Sisodia, D.S., 2018. Prediction of diabetes using classification algorithms. Procedia computer science, 132, pp.1578-1585.

Sarwar, M.A., Kamal, N., Hamid, W. and Shah, M.A., 2018, September. Prediction of diabetes using machine learning algorithms in healthcare. In 2018 24th international conference on automation and computing (ICAC) (pp. 1-6). IEEE.

Hasan, M.K., Alam, M.A., Das, D., Hossain, E. and Hasan, M., 2020. Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8, pp.76516-76531.

Xue, J., Min, F. and Ma, F., 2020, November. Research on diabetes prediction method based on machine learning. In *Journal of Physics: Conference Series* (Vol. 1684, No. 1, p. 012062). IOP Publishing.