Machine Learning - Assignment 2 - Type 2 Diabetes

1. Can survey questions asked from the CDC provide accurate predictions of whether an individual has diabetes? Do you recommend additional features in the data?
    a. The survey questions asked from the CC provide better prediction for the person who don't have diabetes, but not so good in predicting the person who do have diabetes. The main reason is the datasets are not a balanced dataset, the class distribution in non-diabetes occupies 86% of the whole datasets. So the first problem is the imbalanced dataset.
    b. In terms of feature of the data, the survey data offers a reasonable basis for predicting diabetes risk, however, using these survey-based features alone for diabetes prediction has some limitations, and additional data or features could enhance prediction accuracy, such as clinical data (Blood sugar levels, Family history of diabetes, Triglyceride and HDL cholesterol levels, Insulin resistance measures), Lifestyle and Socioeconomic Factors (Sleep patterns, Stress and cortisol levels, Occupation) and so on.
    c. Moreover, if we are looking at diabetes in general, and then we can take a look at gestational and type 1 diabetes. For gestational diabetes, we need to add a feature of "pregnancy", and for type 1 we would need to take a look at family members with type 1, given that type 1 is a genetic autoimmune disease, though type 1 is a mutation of several genes. Due to the mutations and cause of type 1, it seems unlikely to be part of the survey. But it was worth mentioning.

2. What risk factors are most predictive of diabetes risk?
    a. There are several risk factors which are directly harmful, while there certain things that are preventative, such as being physically active. Given the results, it seems a combination of avoiding risk factors, while also following the preventative measures is the best way to go. Have therefore chosen to split them up into two categories of directly harmful, and preventative.
    b. Directly harmful: HeartDiseaseorAttack, HighChol, HighBP, CholCheck,Stroke, DiffWalk, BMI, Age, Income
    c. Preventive: Veggie, GenHealth, PhysHlth, PhysActivity
    d. We are not taking in account the AnyHealthcare as that means a potential false positive, as if you do not have health care then you have no one to actually diagnose you with diabetes.
3. Can we use a subset of the risk factors to accurately predict whether an individual has diabetes?
    a. yes - We are not using the whole dataset to predict according to the feature selection process, we only select the high correlation features from the dataset for the prediction. The benefits to select the features instead of using the whole dataset will reduce the dimensionality, reduce the training time,

avoid the overfitting. Additionally, we also applied PCA to select the top 10 features to do the experiment whether the model performance can be improved, however it showed that the features we selected are very good for predicting if a person has diabetes or not, and using PCA to reduce features actually reduced our model performance.

4. Which machine learning models are best for classifying the disease? Compare models and explain why a model performed better based on the confusion matrix and minimizing false negatives.
    a. In this diabetes prediction, false negative is more important, so we should focus on `recall`(`sensitivity`) from confusion matrix to evaluate the model. we're primarily interested in how well the model identifies true positives (people with diabetes) while minimizing false negatives (those who are incorrectly classified as not having diabetes). Compare with recall of 3 models in the `Diabetes_binary=1 `, Decision Tree perform slightly better, which means better at identifying diabetic individuals (fewer false negatives).
    b. Considering this is a imbalanced dataset, precision-recall curves (PRCs) and the area under those curves may offer a better comparative visualization of model performance. So we plot Precision Recall curve to compare the model's ability to distinguish between the classes. Based on these PR-AUC results, Logistic Regression slightly outperforms (higher AUC) the other models in terms of precision and recall.

Group Collaboration:

We sat and worked together, discussing features and how to attack the assignment. Zoie sat up visualization while Emma sat up contingency table to look at the direct correlation between the binary diabetes column and other columns, in combination with the visualization graphs and contingency table it was a lot clearer to check which features were meaningful for our model(s). Using the logistic regression model with prediction also showed clear coefficient values for taking a deeper look at, such as a very high correlation between BP and diabetes. Zoie primarily coded, while Emma provided research, discussion and initial logistic regression model. Emma tried to provide some code for the AUC, but had some problems with “difference being too large” when just trying to add one cell with a few lines of code, therefore Zoie copied the code from discord and added it instead.

Additional reflection:

Analyzed the columns “Diabetes_binary” and “HeartDiseaseorAttack” which are both binary columns. I used crosstab to create a contingency table of those columns, the results are very interesting. If you look at the amount of people who do not have heart disease and have not had a heart attack, there is a gigantic amount who also does not have diabetes, and a small amount who does have diabetes (about 13.5% of people without heart disease have diabetes). However, if you look at the amount of people who have heart disease or have had a heart attack, the amount of people who also have diabetes is 49.1%.

HeartDiseaseorAttack	0.0	1.0
Diabetes_binary		
0.0	202319	16015
1.0	27468	7878

On the flipside if we look at a negative correlation, for example by taking a look at a classic example of Physical Activity in the column “PhysActivity”.
You see the amount of those who are not physically active, the amount of people with diabetes are 26.8%, it is clear that simply looking at one feature alone is not enough to prove whether or not being physically active prevents diabetes. However, looking at the physically active people, only 13.1% of them also have diabetes. It is clear that physical activity definitely helps.

PhysActivity	0.0	1.0
Diabetes_binary		
0.0	48701	169633
1.0	13059	22287