Colin Michael

DSC 680-T302: Applied Data Science

Project 3 Milestone 3

May 25h, 2024

Predicting Diabetes Outcomes

Diabetes is a health condition that impacts how people turn food into energy. Diabetes prevents people from either making enough insulin to properly function or inhibits how insulin functions. This leads to blood sugar staying in people's bloodstream too long. Excessively high blood sugar causes health complications like heart failure, kidney failures, and vision loss (cdc.gov). For this project, I will be assuming the role of a data scientist employed by the American Center for Disease Control. My objective will be to utilized existing diabetes data sets to create predictive models aimed at helping to treat at-risk individuals for diabetes.

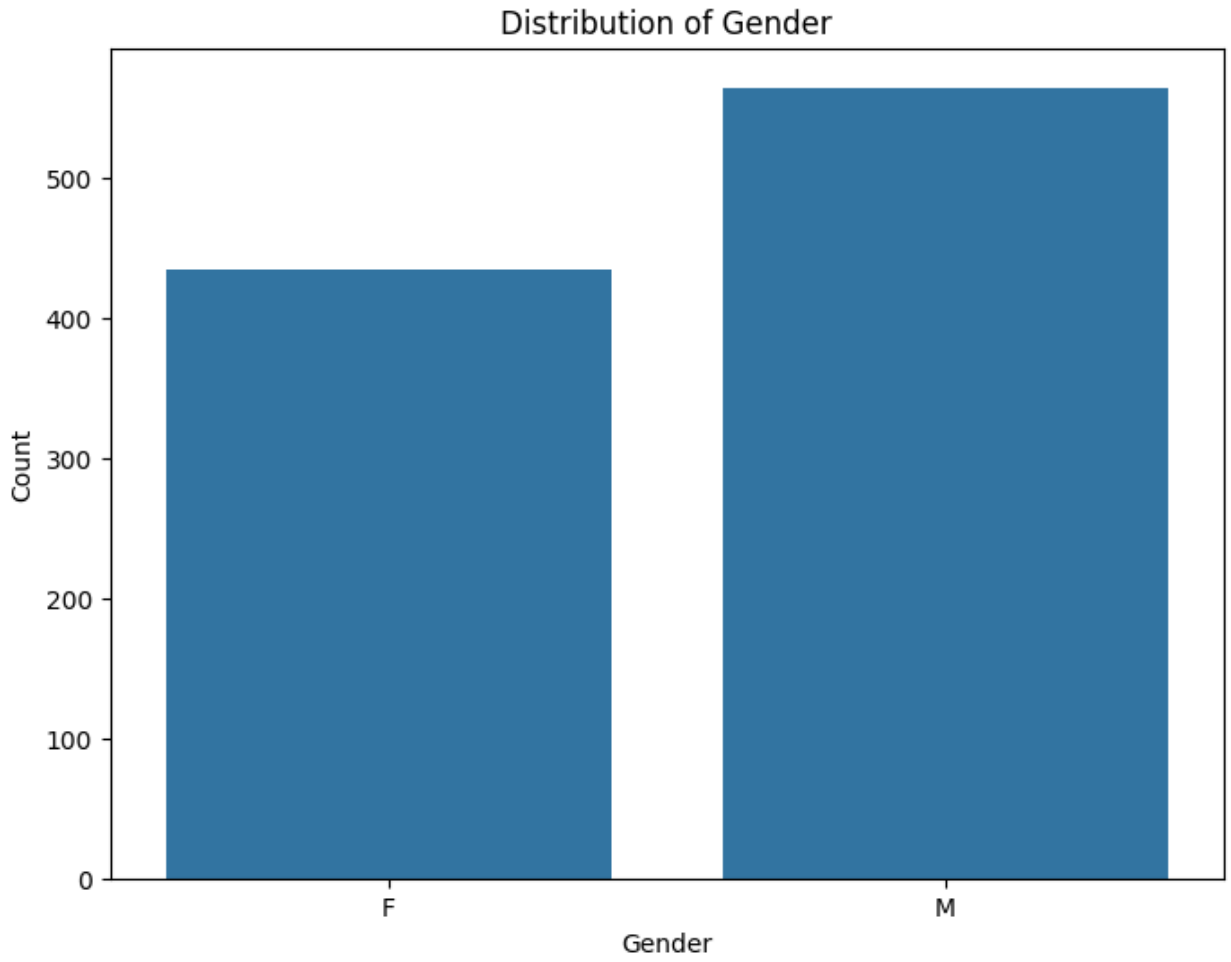
I decided to focus my final research project on diabetes outcomes because diabetes has impacted my family. I have had close family members struggle with diabetes, and I want to use data to better understand the issue. Diabetes is a terrible condition that impacts millions of people, and finding ways to better understand diabetes through data would be a great benefit to society.

I will be using a csv dataset from Kaggle, which was originally from the laboratory of Medical City Hospital of Iraq. The data includes 14 column: No. of Patient, Sugar Level Blood, Age, Gender, Creatinine ratio(Cr), Body Mass Index (BMI), Urea, Cholesterol (Chol), Fasting lipid profile, including total, LDL, VLDL, Triglycerides(TG) and HDL Cholesterol , HBA1C, Class (the patient's diabetes disease class may be Diabetic, Non-Diabetic, or Predict-Diabetic).

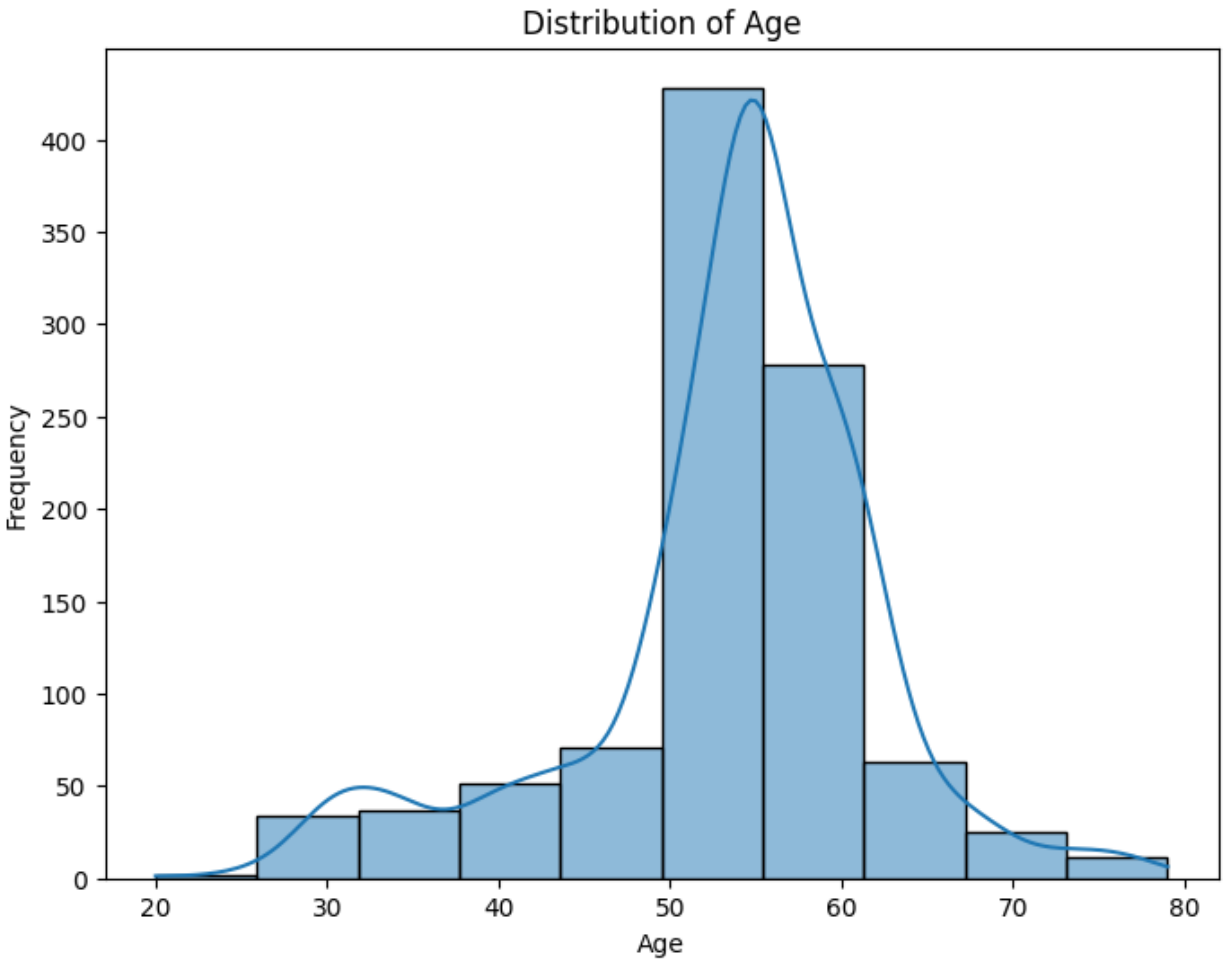
My chief aim is to use this data set to predict future outcomes on diabetes cases. I am planning on running a decision tree predictive model to my dataset. I believe this will be a good fit because my dataset includes a column 'Class', which identifies if a patient has diabetes. Also, I will run a linear regression with 'Class' as my dependent variable.

Before getting into the predictive data science methods, I want to do an exploratory analysis on my dataset. This is to get a high-level understanding of how the working data set is structured.

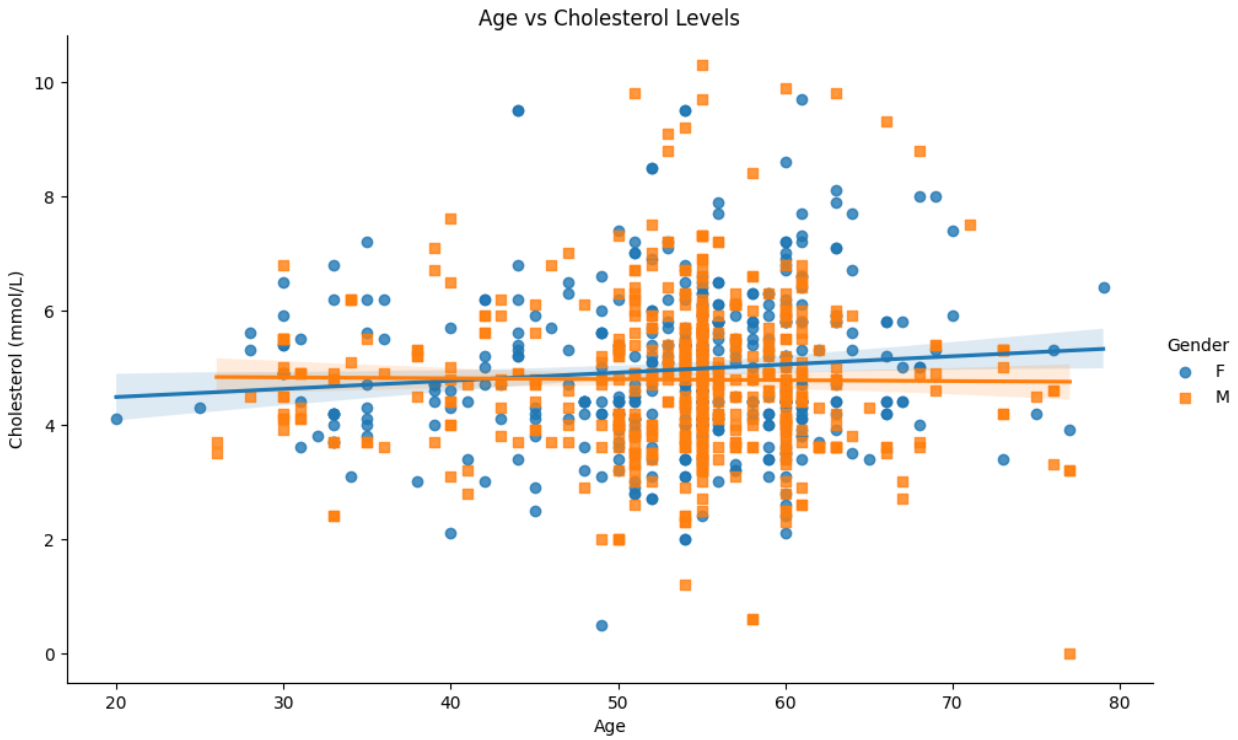
First, I created a bar chart to show the distribution of gender in the data. There are slightly more male participants than female:



Next, I created a histogram to show the distribution of age in the study:



Last, I created a scatterplot between age and cholesterol levels, divided by gender. There is a slightly positive correlation between age and cholesterol levels:



I decided to create a random forest model for my diabetes analysis. A random forest model is a machine learning method that uses classification and regression. Decision trees predict outcomes by splitting data based on features.

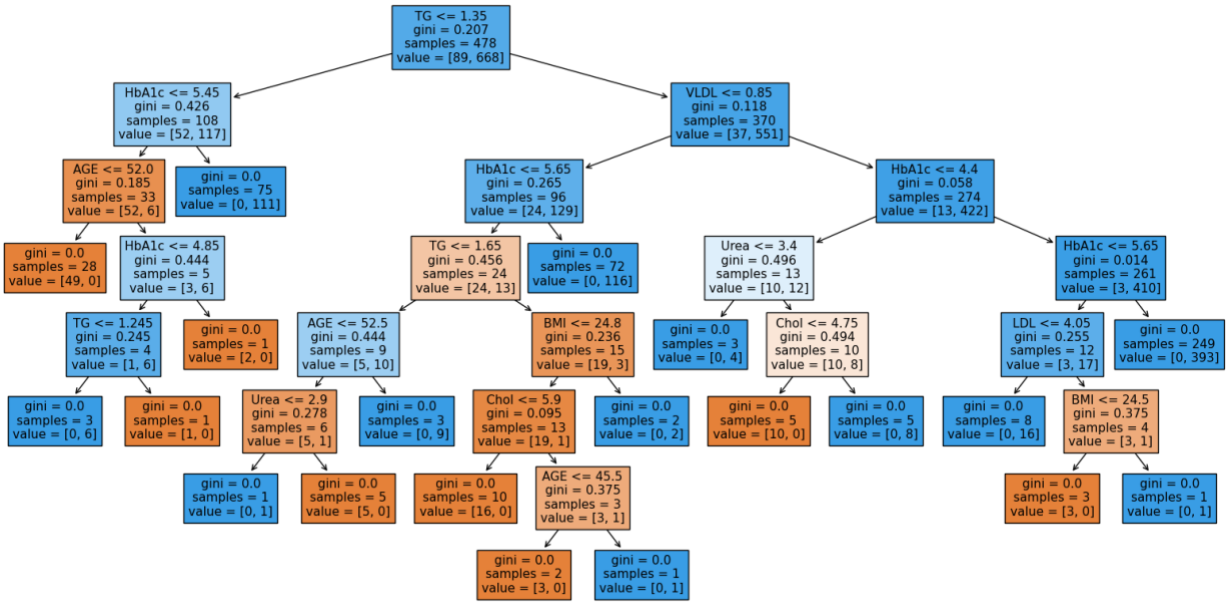
Predicting diabetes class, which I filtered to be ‘N’ or ‘Y’, is the dependent variable I am trying to measure. My independent variables are BMI, HbA1c, Cholesterol, Age, TG, VDL, LDL, Urea count, HDL, creatine, and gender. Here are the results matrix of the random forest model:

```
Random Forest Accuracy: 0.9842105263157894
Classification Report:
              precision    recall  f1-score   support

           N       0.90      0.95      0.93        20
           Y       0.99      0.99      0.99       170

    accuracy                           0.98       190
   macro avg       0.95      0.97      0.96       190
```

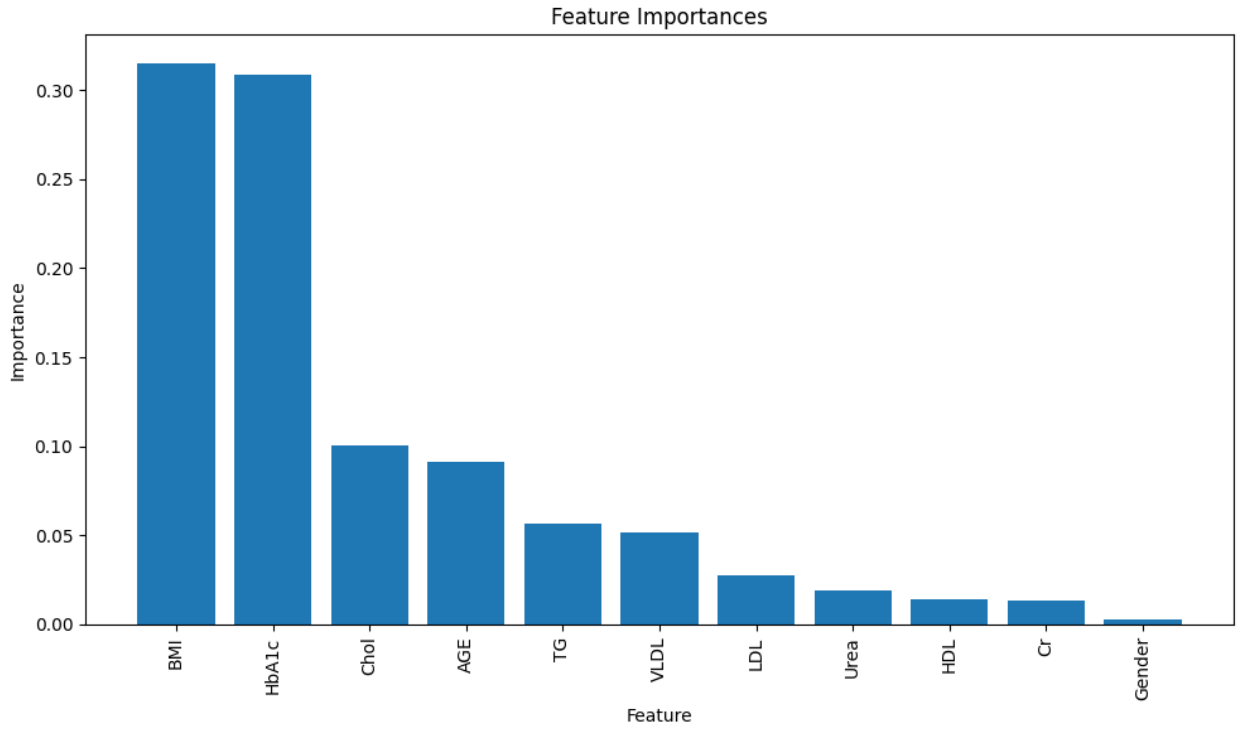
```
weighted avg       0.98      0.98      0.98       190
```

The model has an overall accuracy of 98.4%, which is very good. 90% precision of N means that when the model predicts N, it is correct 90% of the time. The model has a 99% precision on Y, which is excellent.

The full random forest tree can be visualized here:



Feature importance, which shows how each independent variable impacts the dependent variable, is a critical part of random trees. Here is a bar chart of the features in this model:



The data shows that BMI is the greatest indicator of diabetes. A close second is HbA1c. Cholesterol and age are third and fourth, but there is a significant drop in their feature importances. This model suggests that BMI and HbA1c are the most important factors to focus on when looking at diabetes outcomes.

My dataset only includes 1,000 rows of data, which in the scope of a data science project is not very big. This limited dataset may hinder the reliability of my analysis. Also, the patient data is all from Iraq. A stronger dataset would include patient data from different parts of the world, to make my results more diverse.

I want to make sure my research project is done in an ethical fashion. I want to avoid reinforces negative stereotypes in my project, and instead focus on helping people. I do not want my research to be used to create prejudice against factors like Body Mass Index and Age.

I believe that my approach to applying data science towards diabetes research could be greatly expanded upon. I would like to gather a larger dataset in the future. Along with more

participants, I would like to gather data from people across the world. This would help control for environmental factors. It would be interesting to bring in other factors, such as income and education levels. I believe these data points would have strong correlations with diabetes outcomes.

10 Questions:

1. **Why did I choose to study diabetes?** I have family members impacted by the condition and have seen how it negatively impacts people.
2. **What made me choose a random forest model?** I think the data set having defined diabetes outcomes made it a great choice because it was easy to see which rows of data did and did not lead to diabetes.
3. **What surprised me most about my findings?** How relatively small the impact of cholesterol on diabetes outcomes was.
4. **What other data science methods would be applicable for my data set?** I would be interested in doing a k-nearest neighbor analysis on the data set.
5. **What other medical conditions would be beneficial to study?** I would like to do a similar research project on dementia.
6. **How could I improve my random forest model?** Adding data points like education level and income.
7. **What confused you the most about the project?** How to interpret the results of the forest
8. **How would you rate the quality of data?** The data set had a wide variety of variables that provided for interesting insights.

9. **What methods did you need to take to clean the data?** Some of the class records had trailing spaces that I had to clean up, for example, I had data records that were populated as ‘N ‘, and I had to convert them to ‘N’
10. **What other data sets would help this project?** I would like to bring in data sets that show how diabetes impacts various aspects of people’s lives.

Appendix:

Diabetes is an extremely complex health condition that has been studied by many different organizations. Preventative care is one of the best ways to get ahead of diabetes diagnoses. The Mayo Clinic suggests five tips for preventing diabetes (MayoClinic.org):

1. Lose extra weight
2. Be more physically active
3. Eat healthy plant foods
4. Eat healthy fats
5. Skip fad diets and make healthier choices.

These tips all support the findings from my data science project that Body Mass Index is the greatest predictor of diabetes outcomes. Losing weight, being active, and eating a healthy diet are all ways to control one’s Body Mass Index. It can be difficult, but the Mayo clinic, along with my data science project, support the argument that lowering one’s Body Mass Index will help prevent diabetes.

Sources:

Centers for Disease Control and Prevention. (n.d.). *Diabetes*. Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/index.html

Find open datasets and Machine Learning Projects. Kaggle. (n.d.). https://www.kaggle.com/datasets

Mayo Foundation for Medical Education and Research. (2023, March 24). *Diabetes prevention: 5 tips for taking control*. Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/in-depth/diabetes-prevention/art-20047639