Project Goals

In this study, we hoped to understand and predict the likelihood of complications for U.S. births. We define complications at birth to be a delivery that requires a Cesarean section, the presence of an infection, or birth weights that are significantly lower than the mean. Research has suggested that multiple maternal factors are responsible for influencing birth weight of newborns such as age, weight, and previous health history (Al-Shawwa & Abu-Naser, 2019). Similarly, birth weight can be predicted using other measurements from the mother, such as symphysio-fundal height/abdominal girth measurements which were found to be highly correlated to the birth weight of newborns (Ademowore et al, 1989). These studies motivated our research towards predicting birth weight from features of a pregnancy term.

The data explored comes from the US Births (2018) dataset on Kaggle, which is directly sourced from the Center for Disease Control (CDC) website. Birth complications is not a straightforward response variable, so we began our exploratory analysis using 'low birth weight' as a proxy for complications. We found low birth weight helpful to explore the data, but also found that other common and serious complications were oppositely affected by some of the predictors of low birth weight. So we decided to break apart complications into 3 major response variables: low birth weight, C-section delivery, and onset of infections during pregnancy.

One of the key indicators of good health outcomes during birth is being able to quickly diagnose issues that may occur and then promptly and effectively treat them before they escalate. By creating models that can accurately predict the risk of possible complications of a birth, hospital systems can more accurately forecast patients who are at a high risk of having a complication and resultantly prepare both the patients and providers. Shared decision-making is becoming more prevalent in medicine and with improved information disseminated to patients and healthcare teams symmetrically, patients and doctors can make data-driven decisions to improve outcomes.

Exploratory Analysis

The dataset contains 3,801,534 births and 55 different features that detail different aspects of the mother and newborn's hospital visit. In our dataset, we added the column 'target' based on the weight of the baby which indicates if the baby is underweight or not. We then performed exploratory analysis to see how the frequency of normal and underweight newborns are distributed using different predictors.

Mothers below the age of 24 and above the age of 35 had more underweight than normal weight babies, as shown in Figure 1.1. A more detailed view of each group of newborns can be seen in Figure 1.2. Comparing the weights of mothers at the time of delivery, newborns that were classified as underweight had mothers whose weights deviated more than those of normal weight infants. Figures 1.3 and 1.4 display the difference in weight and body mass index between the mothers of each group of newborns. A new mother was more likely to give birth to an underweight baby, but those with 1 or 2 previous births had a higher frequency of normal weight newborns. Mothers of more than 3 previous births experienced higher rates of underweight births than normal births. This trend is demonstrated in Figure 1.5. A new column was created to measure the number of months between the mother's last menstruation cycle and the birth of the child. This was calculated by finding the difference between the baby's month of birth and the mother's last recorded menses. We can see in Figure 1.6 that mothers who maintained a regular menstrual cycle past one month into the pregnancy were much more likely to give birth to an underweight baby.

Low Birthweight Analysis

Considering birth weight of a child to be the best indicator of the child's health, we began our modeling to determine that given a set of predictors, would the child be underweight at birth or not. This insight will help the parents to take precautionary measures to improve their chances of delivering a child with healthy weight.

The baseline accuracy for this model is 93.27%. We tested logistic regression on the cleaned dataset which gave an accuracy of 93.4% but a recall of only 6%. So, we ran the model again with upsampled data for the minority class and its recall went up to 70% However, the accuracy took a hit, down to 80%. We tested the other models without upsampling the data as accuracy was degrading with upsampled dataset due to more instances of the negative class present. Next, we ran a decision tree with criterion as entropy at each level which gave us an accuracy of 89 % and a recall of 64%. Next, we ran a bagging model which comparatively gave a higher accuracy of 93.3% along with the recall of 63.2%. The random forest model run on this data with n_estimators as 15 gave similar test accuracy as the bagging model, but with a lower recall of 62.6%. Finally, we ran the gradient boosting model, came up with n_estimators as 800 and max depth of 4 using hyperparameter tuning and got the best accuracy of **94%** along with the best recall of **66%**. The train and test accuracies and recalls can be found in Figure 2.1. The important features which came out from this model can be seen in Figure 2.2. With NMD being the most important factor and others such as pre-pregnancy weight, mothers age, prior births given by the mother follow.

C-Section Analysis

1 in 3 women in the US deliver via C-Section (up from 1 in 4 over the past 20 years), which is double the 10-15% target rates established by the WHO. Healthcare providers may recommend a C-section for medical reasons, and there are cases of women requesting to have an elective C-section for their first delivery. However, according to the American College of Obstetricians and Gynecologists, C-sections can introduce risks to babies and mothers and have a higher maternal mortality rate than natural births. With that said, planning for a vaginal birth and requiring an emergency C-section has a higher maternal mortality rate than a planned C-section. Predicting C-sections more accurately could reduce the amount of emergency procedures at scale and potentially improve patient outcomes.

When searching for significant risk factors for a C-section, random forests and logistic regressions both pointed to the mother's BMI before pregnancy and the mother's age during delivery. Women of higher ages and higher BMIs were at more risk of a C-section.

The baseline rate of C-section deliveries is about 32%, meaning that baseline accuracy is 68% (if predicting no C-sections). However, there is no operational benefit to predicting no true positives and planning for no C-sections. Running a logistic regression classification with 28 predictors was only able to improve precision to 62% and dropped recall down to 31% (Figure 3.1), meaning that more than half of the women that end up needing a C-section would be missed, which is worse than guessing.

Binary classification did not yield promising results. However, the probability outputs of the classification demonstrated a relationship with BMI. Using solely BMI as a univariate predictor of C-section delivery resulted in logistic regression probabilities that were directly correlated with the *average rate* of C-sections at each BMI (eg. at a BMI of 29.7, 30% of women underwent C-sections and the logistic regression probability output was also 30%).

Overall, these findings demonstrate a close to perfect linear relationship between levels of BMI and the average rate (or risk) of C-section (Figure 3.2). Running a linear regression on BMI and rate of C-section resulted in an R-Squared of 96.7%, which can be improved to an R-Squared of 99.9% when removing the extremes of BMI and modeling only the data corresponding to BMIs from 18.5-50 (Figures 3.3 and 3.4). With a coefficient of 0.0117 (standar error 0.0002), a woman's BMI is an *almost* direct approximation of her percent risk of undergoing a C-section.

Predicting if an individual patient will undergo a C-section is still unclear, but predicting across the aggregate of patient groups of normal, overweight, and obese women (Figure 3.5) within a healthcare sytem on average can be accurately forecasted by multiplying each pregnant patient's BMI by 117%.

Infection Analysis

According to a study by Cynthia J. Berg on pregnancy related mortalities in the US, it is estimated that 10% of all pregnancy deaths are due to the onset of an infection. Using logistic regression we have attempted to find a way of predicting whether or not an infection will occur during birth. The baseline prediction for this model is to assume all patients will not have an infection, which has an accuracy of 97.24%. Using logistic regression and a decision tree model we were able to achieve prediction accuracies detailed in the Figure 4.1. While we were unable to achieve a better accuracy than the baseline assumption, the logistic regression model can provide insights into which features are key indicators that a patient may have an infection during their birth.

Based on the logistic regression coefficients in Figure 4.2 a hospital can label a patient as high risk if they are positive for key indicators such as cigarette usage, needing a c-section, being unmarried, and needing to induce labor. Not only are these indicators significant based on the logistic regression model, but their impact on the possibility of getting an infection can be logically supported as well. C-sections and induction of labor are both surgical procedures that open your body up to the outside environment, thus increasing the possibility of infection. Marital status may seem out of place, but in a study done by Andrée-Anne Fafard St-Germain et al. on the infection rate among married and unmarried mothers it was confirmed that unmarried mothers are indeed at higher risk for infection. The main reason being they are exposed to a higher rate of sexually transmitted diseases.

Conclusions and Insights

Addressing the complications at birth related to low birth weight, infections and requiring C-section by exploring the dataset and coming up with models has led us to the conclusion that the following factors should be considered important by parents to avoid any undesirable complications at birth:

1. Low Birth weight: Mother’s age, Non menses duration during pregnancy, and pre pregnancy weight.
2. Infections: Cigarette usage and Marital status
3. C-section possibility: BMI and Mother’s age

Using these key indicators medical professionals can better identify patients that are at high-risk for complications and can adjust their care plan accordingly.

Figures

Figure 1.1

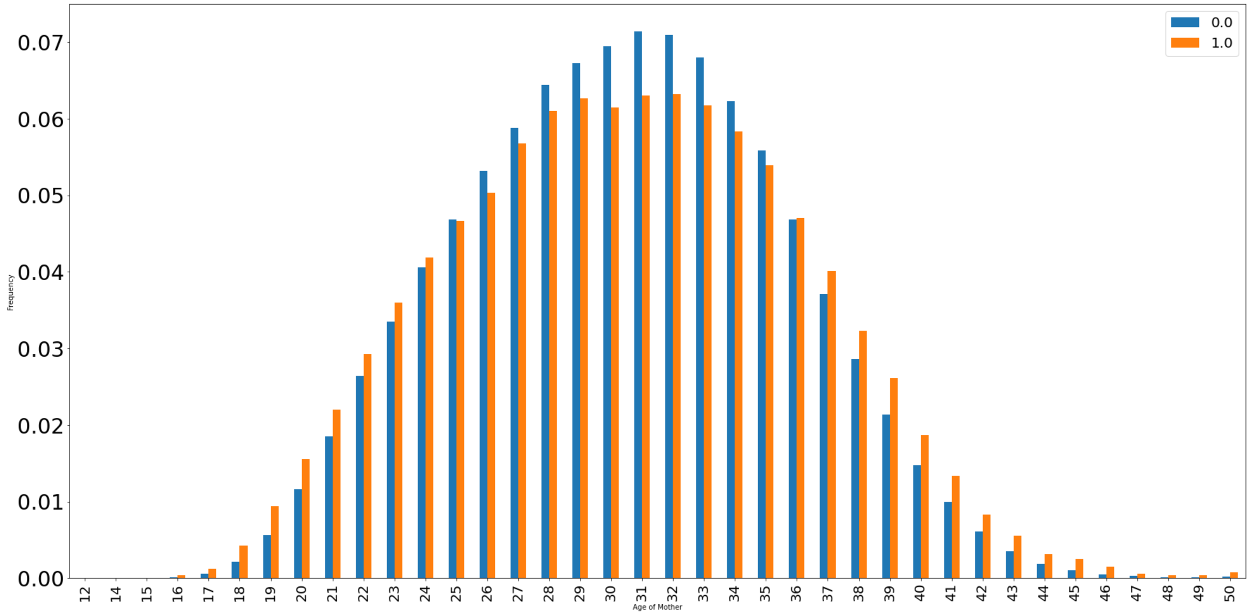


Figure 1.2

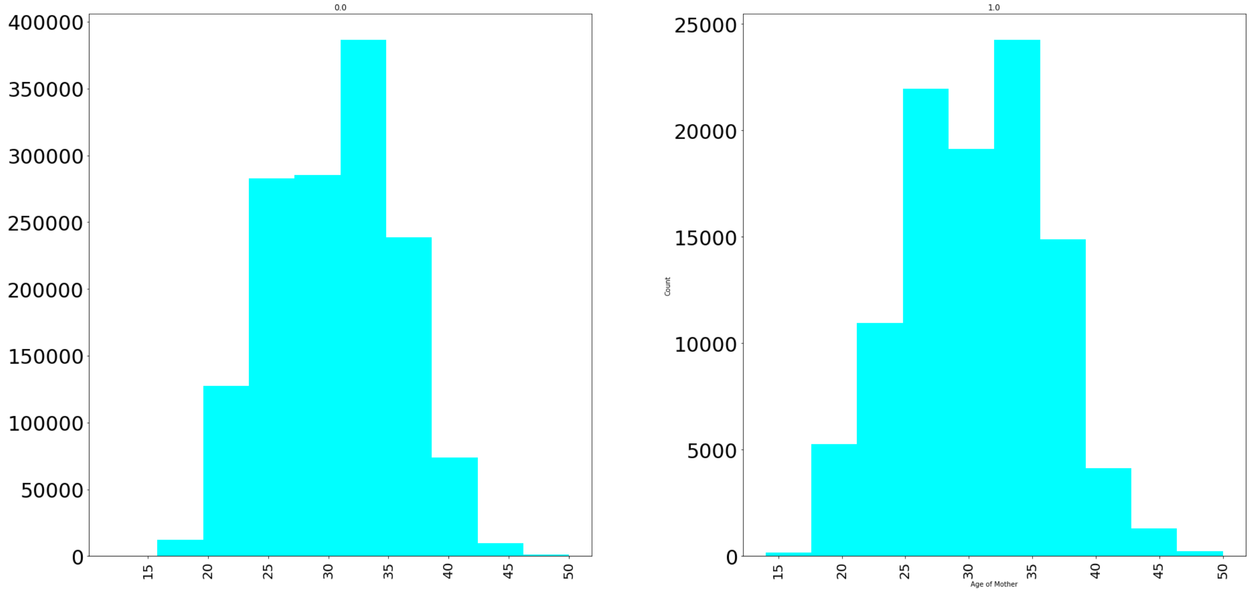


Figure 1.3

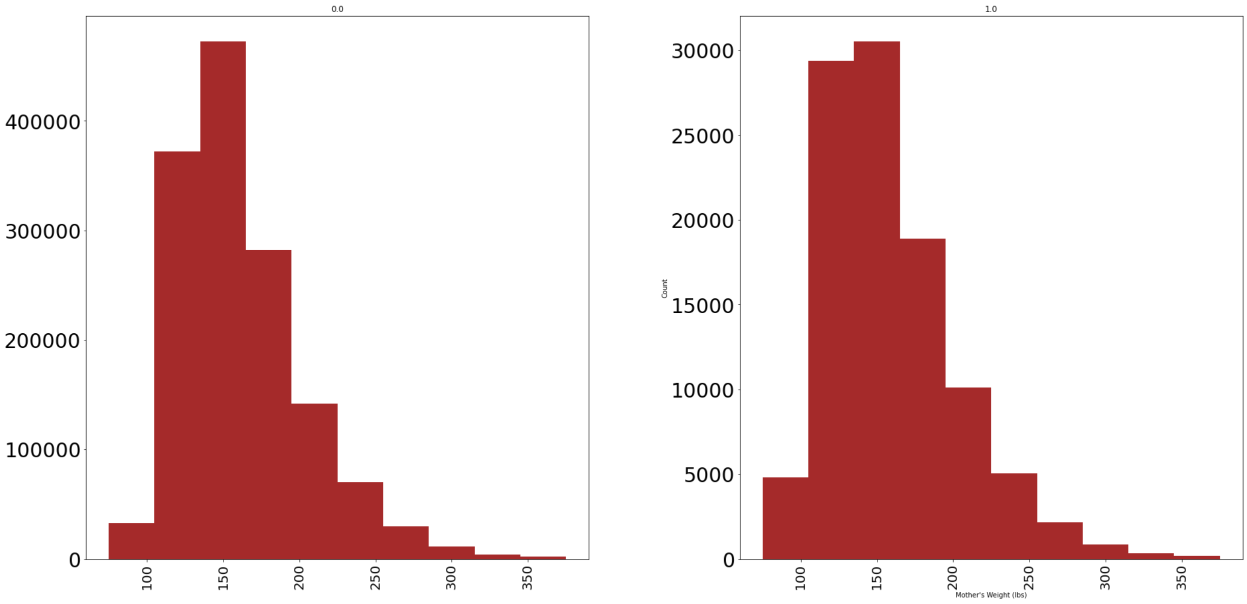


Figure 1.4

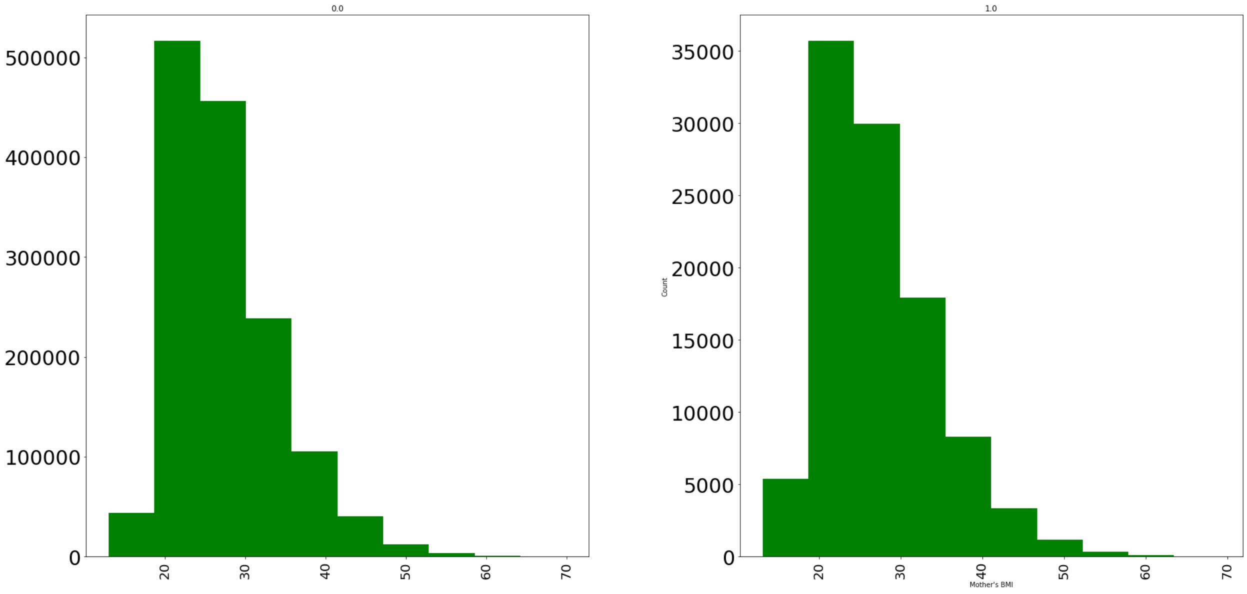


Figure 1.5

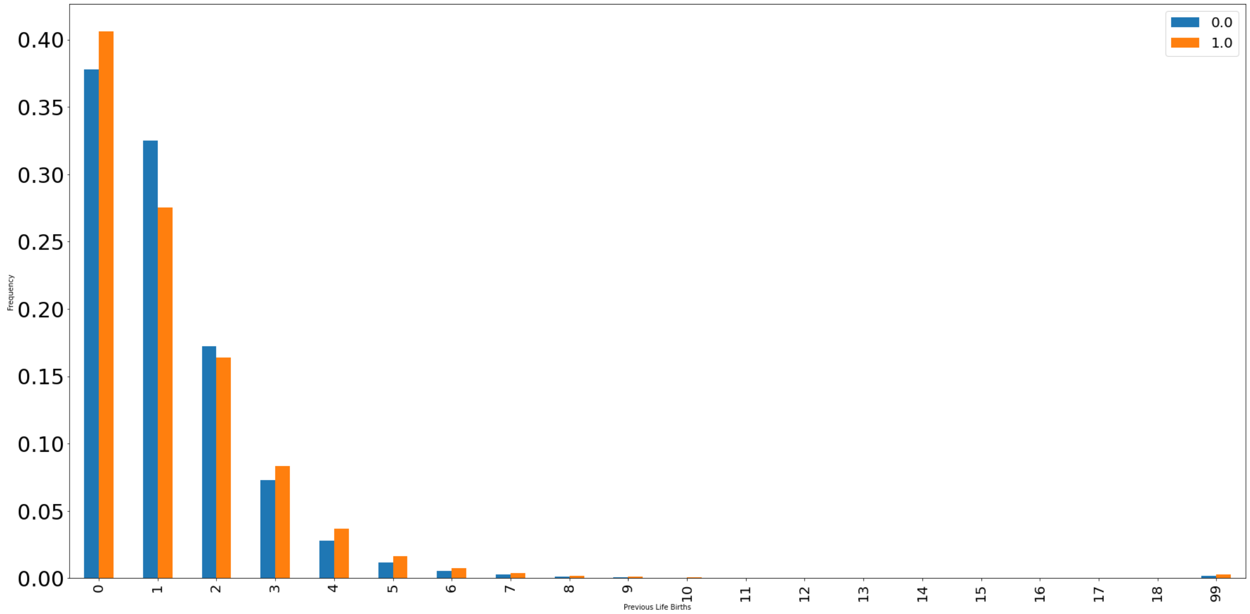


Figure 1.6

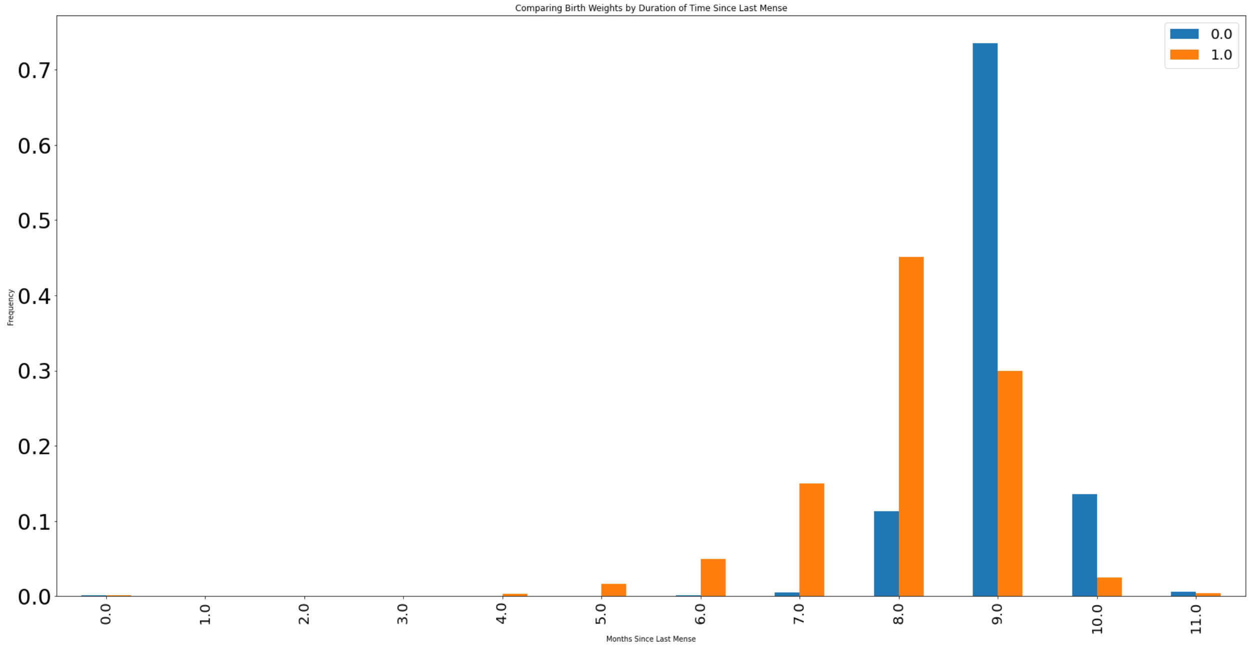
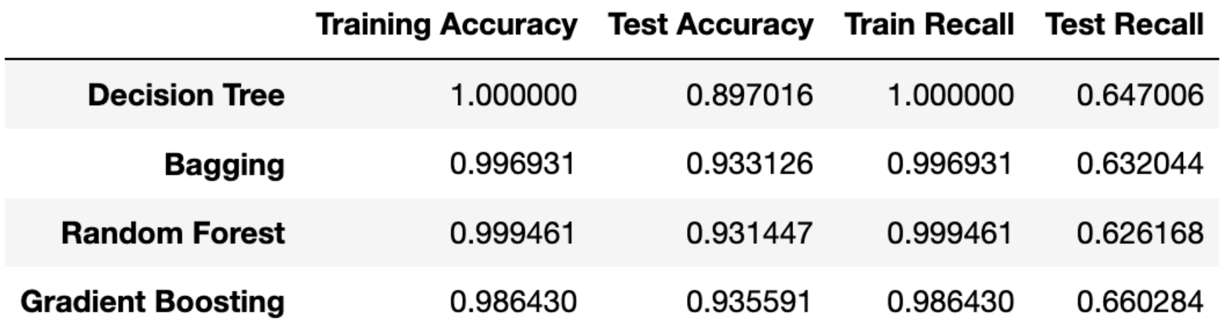


Figure 2.1



	Training Accuracy	Test Accuracy	Train Recall	Test Recall
Decision Tree	1.000000	0.897016	1.000000	0.647006
Bagging	0.996931	0.933126	0.996931	0.632044
Random Forest	0.999461	0.931447	0.999461	0.626168
Gradient Boosting	0.986430	0.935591	0.986430	0.660284

Figure 2.2

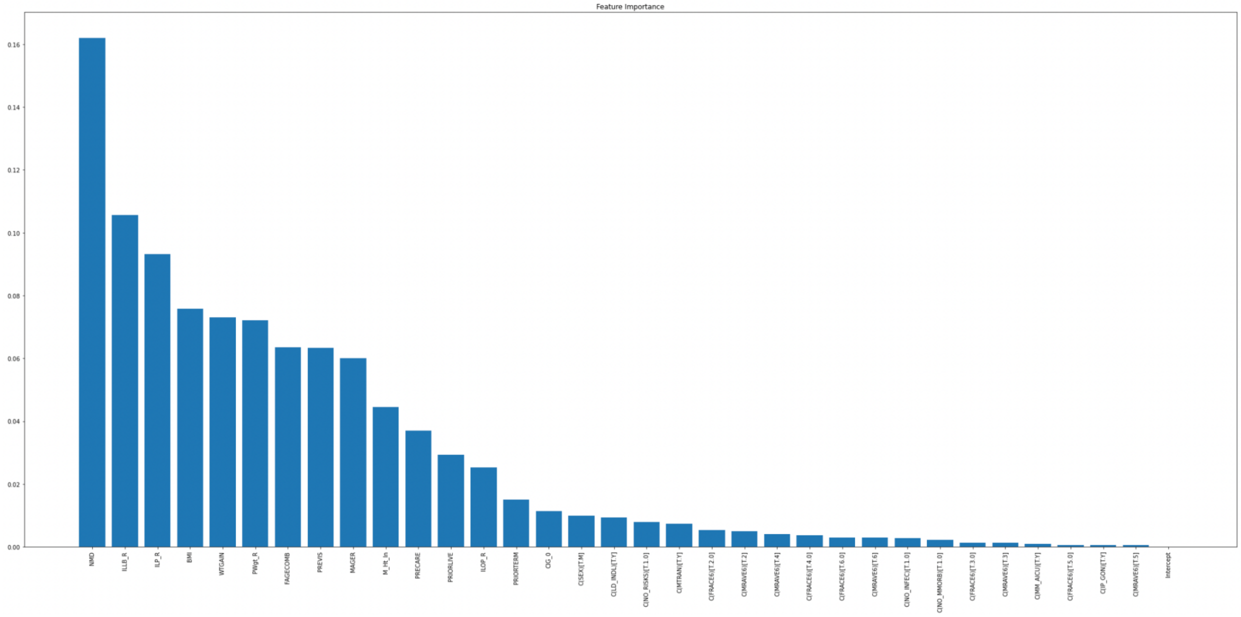


Figure 3.1

Mulitvariate Confusion Matrix (1000s)

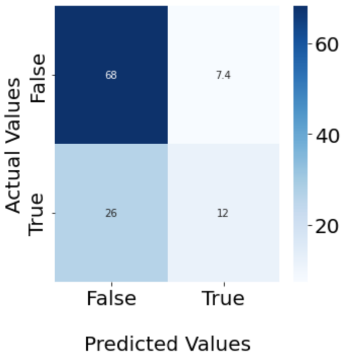


Figure 3.2

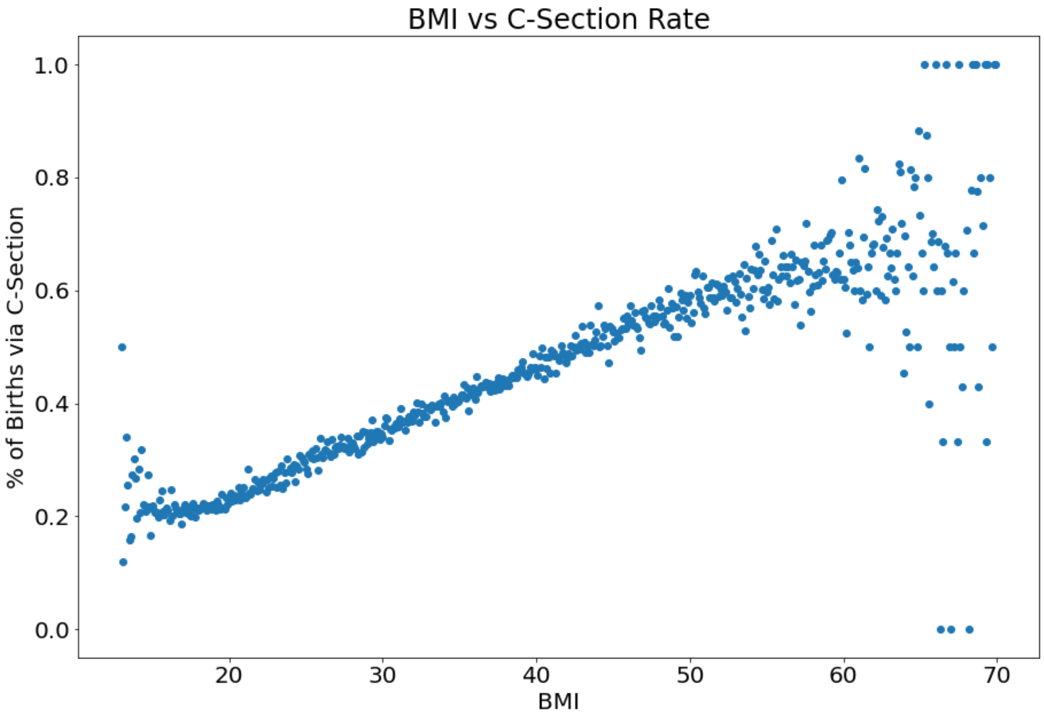


Figure 3.3

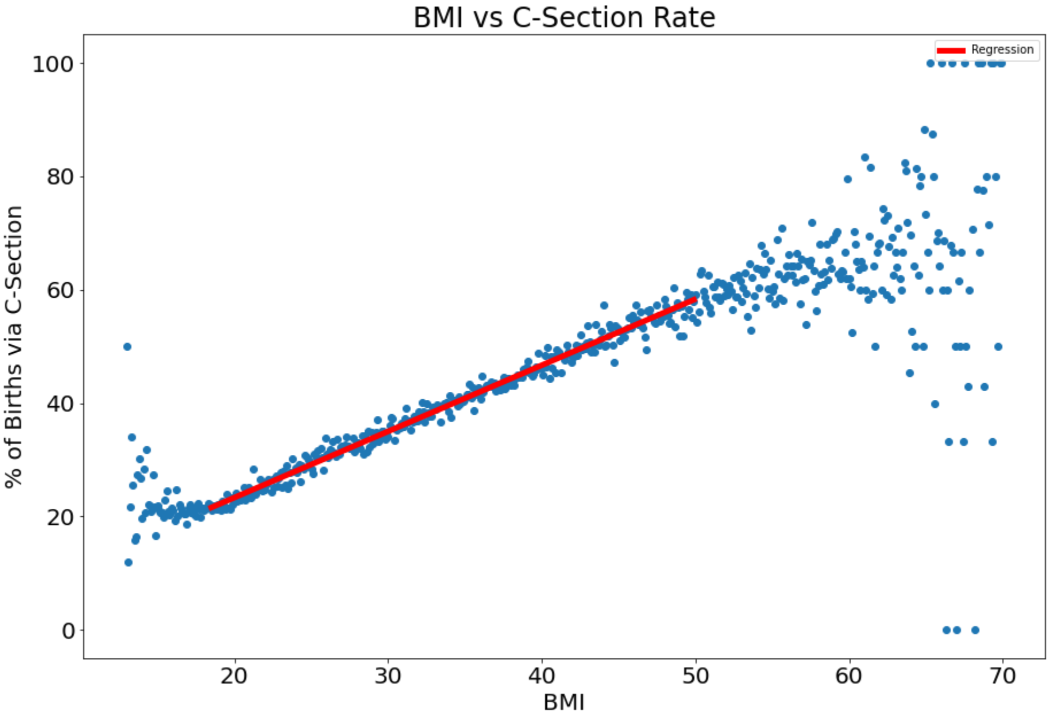
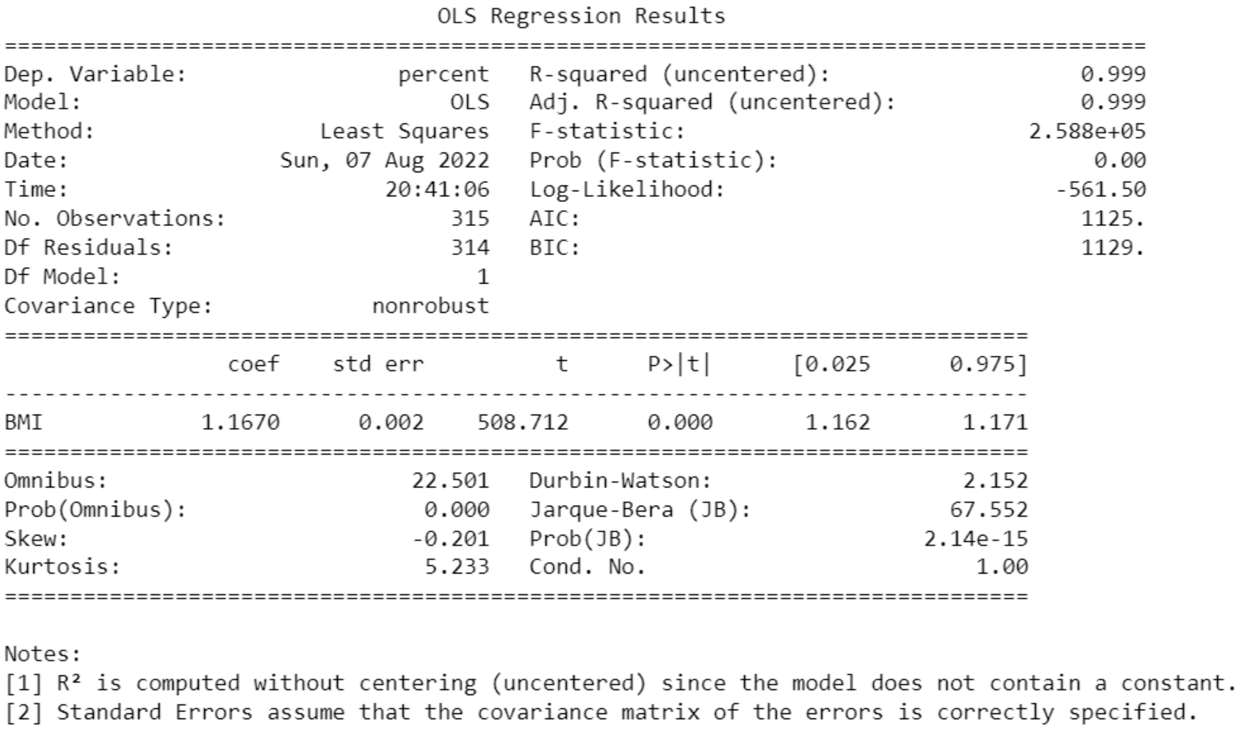


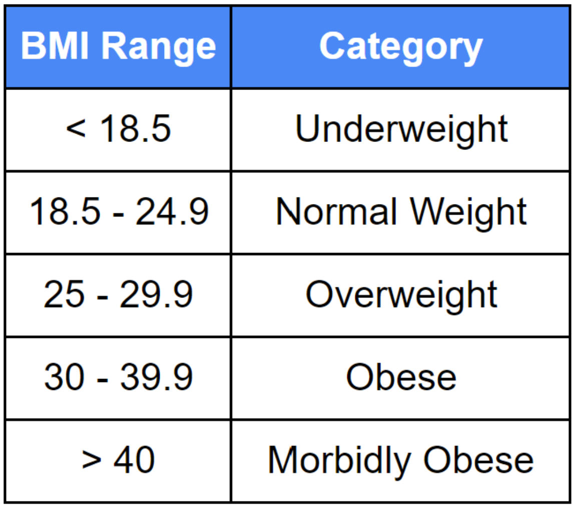
Figure 3.4



```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:                percent   R-squared (uncentered):                   0.999
Model:                            OLS   Adj. R-squared (uncentered):              0.999
Method:                 Least Squares   F-statistic:                          2.588e+05
Date:                Sun, 07 Aug 2022   Prob (F-statistic):                        0.00
Time:                        20:41:06   Log-Likelihood:                         -561.50
No. Observations:                 315   AIC:                                      1125.
Df Residuals:                     314   BIC:                                      1129.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
BMI            1.1670      0.002    508.712      0.000       1.162       1.171
==============================================================================
Omnibus:                       22.501   Durbin-Watson:                   2.152
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               67.552
Skew:                          -0.201   Prob(JB):                     2.14e-15
Kurtosis:                       5.233   Cond. No.                         1.00
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

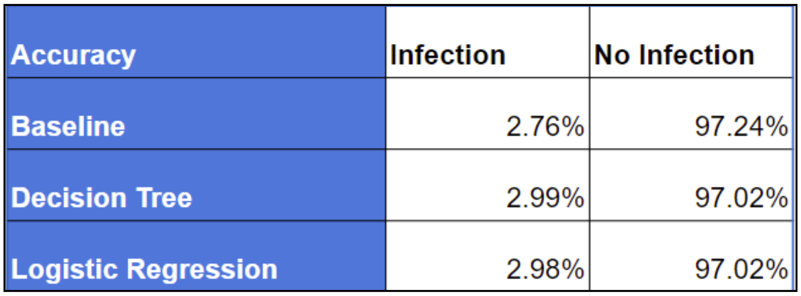
Figure 3.5*



BMI Range	Category
< 18.5	Underweight
18.5 - 24.9	Normal Weight
25 - 29.9	Overweight
30 - 39.9	Obese
> 40	Morbidly Obese

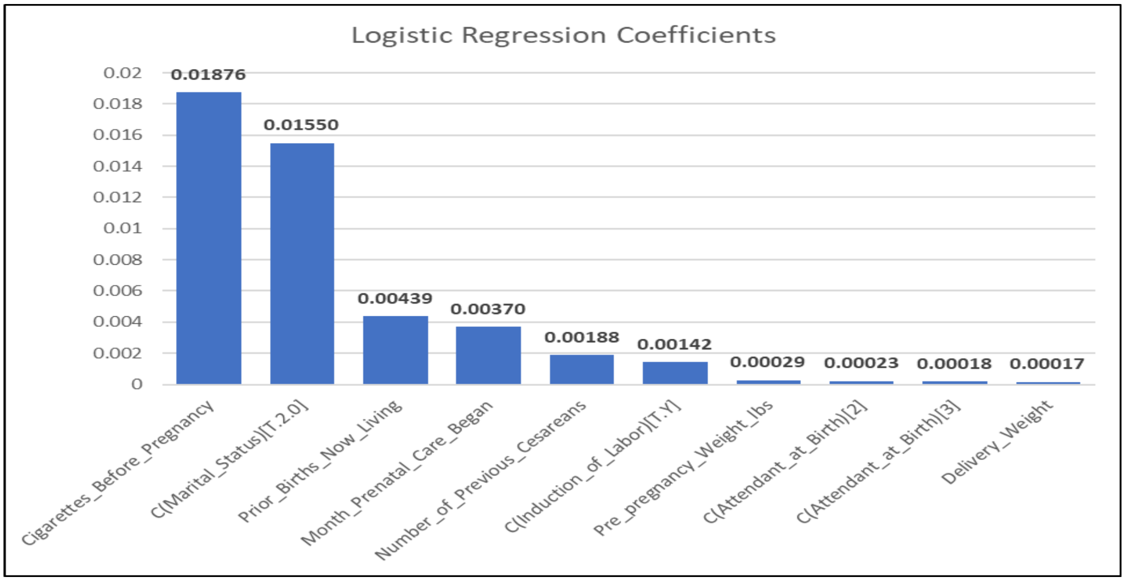
*Adult women BMI categorization

Figure 4.1



Accuracy	Infection	No Infection
Baseline	2.76%	97.24%
Decision Tree	2.99%	97.02%
Logistic Regression	2.98%	97.02%

Figure 4.2



References

Al-Shawwa, Mohammed & Abu-Naser, Samy S. (2019). Predicting Birth Weight Using Artificial Neural Network. *International Journal of Academic Health and Medical Research (IJAHMR)* 3 (1):9-14.

https://philpapers.org/rec/ALSPBW

Berg CJ, Callaghan WM, Syverson C, Henderson Z. Pregnancy-related mortality in the United States, 1998 to 2005. Obstet Gynecol. 2010 Dec;116(6):1302-1309. doi: 10.1097/AOG.0b013e3181fdfb11. PMID: 21099595.

https://pubmed.ncbi.nlm.nih.gov/21099595/

Bever AM, et al. Fetal growth patterns in pregnancies with first trimester bleeding: the NICHD Fetal Growth Studies. *Obstetrics and Gynecology*. 2018;

DOI: 10.1097/AOG.0000000000002616

https://www.nichd.nih.gov/newsroom/news/050918-early-pregnancy-bleeding#

Centers for Disease Control and Prevention. (n.d.). *Natality - Birth Records (expanded) documentation*. Centers for Disease Control and Prevention. Retrieved August 7, 2022, from https://wonder.cdc.gov/wonder/help/Natality-expanded.html

Fafard St-Germain A-A, Kirby RS, Urquia ML (2022) Reproductive health among married and unmarried mothers aged less than 18, 18–19, and 20–24 years in the United States, 2014–2019: A population-based cross-sectional study. PLoS Med 19(3): e1003929.

https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1003929

F.O. Dare, A.S. Ademowore, O.O. Ifaturoti, A. Nganwuchu,

The value of symphysio-fundal height/abdominal girth measurements in predicting fetal weight,

International Journal of Gynecology & Obstetrics, ISSN 0020-7292, https://doi.org/10.1016/0020-7292(90)91018-L.

https://www.sciencedirect.com/science/article/pii/002072929091018L

Mayo Foundation for Medical Education and Research. (2022, June 16). *C-section*. Mayo Clinic. Retrieved August 7, 2022, from https://www.mayoclinic.org/tests-procedures/c-section/about/pac-20393655

US births (2018)

https://www.kaggle.com/datasets/des137/us-births-2018