Low birth weight prediction description document.

Overview

This project is focused on predicting low birth weight in new born babies by considering data of the parents. Features like parent's age, race, education and habits like smoking and drinking are considered. Based on these features we can predict whether the baby to be born is under weight or not.
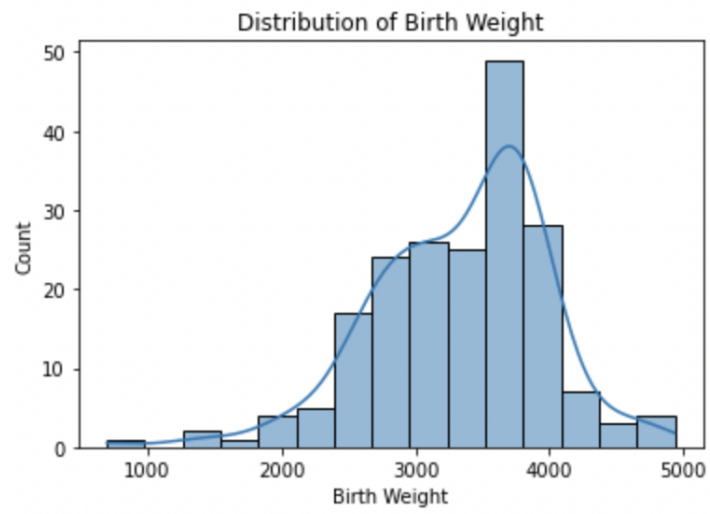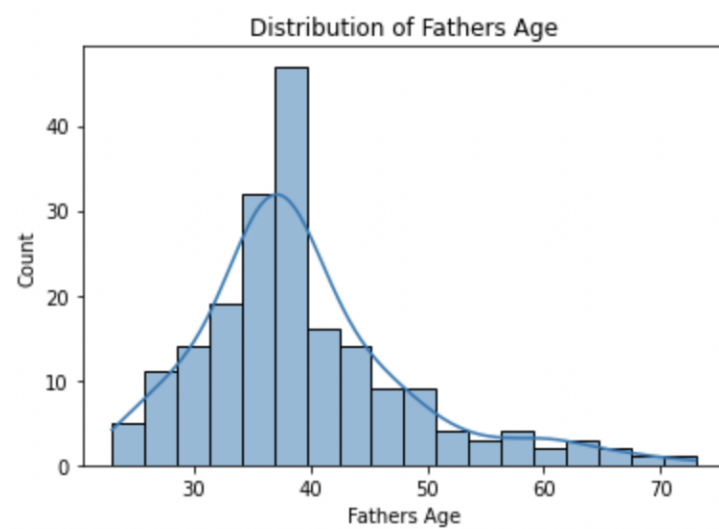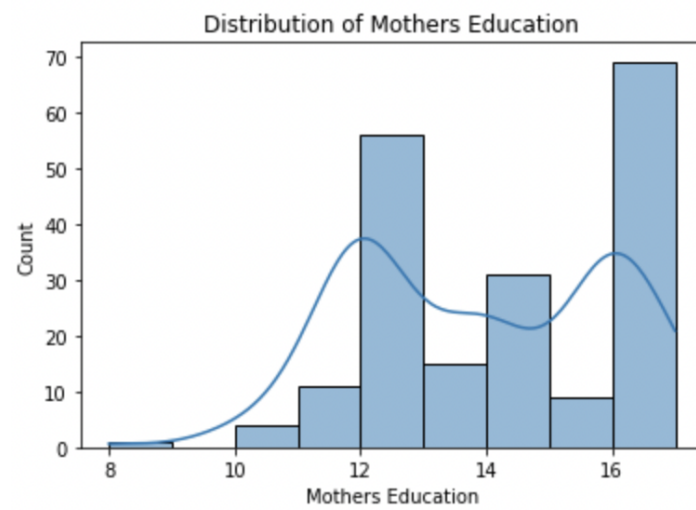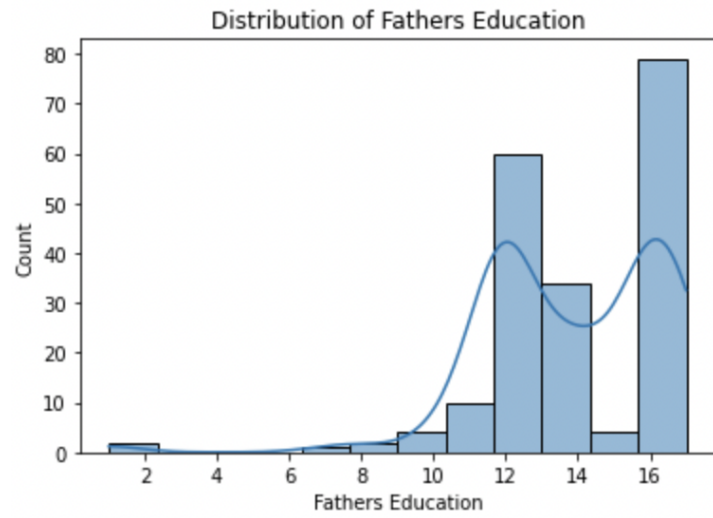
The Dataset

The dataset can be found attached to the code file in the submission. The dataset contains 18 features and 196 tuples where the major features are age, education and race. Below listed are all the features in detail.
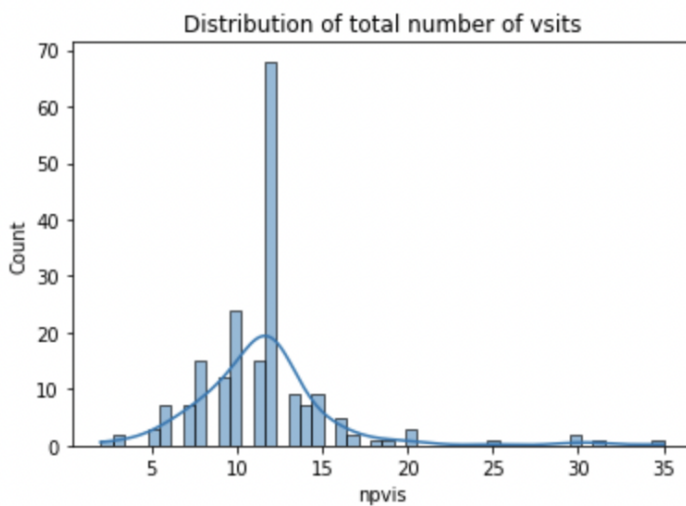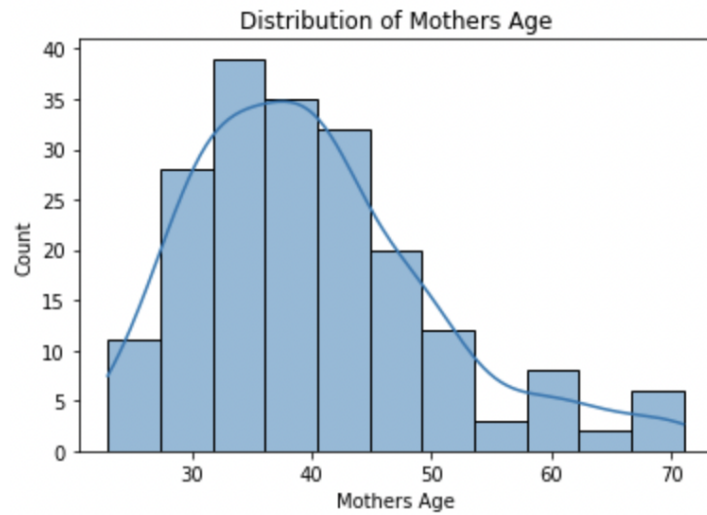
mage	meduc	monpre	npvis	fage	feduc	omaps	fmaps	cigs	drink	male	mwhte	mblck	moth	fwhte	fblck	foth	bwght
69		5	2	62		4	7	23	9	1	0	1	0	0	1	0	697
68	12	3	10	61	11	4	6	25	11	1	1	0	0	1	0	0	1290
71	12	3	6	46	12	2	7	21	12	1	0	1	0	0	1	0	1490
59	16	1	8	48	16	7	8	21	10	0	0	0	1	0	0	1	1720
48	12	4	6	39	12	2	9	17	13	0	1	0	0	1	0	0	1956
67	11	4	8	40	8	4	9	16	14	0	1	0	0	1	0	0	1984
54	12	2	12	46	12	9	9	17	12	1	0	1	0	0	1	0	2050
71	14	4	7	51	11	9	8	15	13	0	1	0	0	1	0	0	2068
56	12	1	9	53	14	8	9	14	9	1	1	0	0	1	0	0	2148
58	12	2	12	61	16	9	9	13	6	0	0	1	0	0	1	0	2180
60	11	7	8	58	1	7	9	9	10	0	1	0	0	1	0	0	2266
42	16	2	12	48	16	9	9	22	9	0	0	0	1	0	0	1	2310
29	12	3	11	32	13	9	9	13	10	0	1	0	0	1	0	0	2359
30	12	4	6	43	12	9	9	15	11	0	1	0	0	1	0	0	2438
34	16	2	12	39	16	9	9	23	9	0	0	0	1	0	0	1	2490
43	17	2	12	39	16	9	10	14	7	1	0	0	1	0	0	1	2500
33	12	2	16	36	16	8	9	13	6	0	0	0	1	0	0	1	2523
53	16	2	25	38	16	8	9	16	12	0	0	0	1	0	0	1	2523
42	14	2	12	36	16	9	9	18	7	0	0	0	1	0	0	1	2530
38	13	3	31	33	1	6	9	22	6	0	1	0	0	1	0	0	2544
34	17	2	10	39	17	9	9	14	5	0	0	0	1	0	0	1	2580
40	14	1	12	36	14	9	9	16	10	1	0	1	0	0	1	0	2580
44	16	2	12	44	16	9	9	13	9	1	0	0	1	0	0	1	2590
40	12	2	8	40	12	8	9	19	10	1	1	0	0	1	0	0	2608
51	16	3	11	37	14	9	9	16	9	0	0	0	1	0	0	1	2630
29	14	1	15	37	15	8	10	13	9	1	1	0	0	1	0	0	2633
31	16	1	15	34	12	9	9	11	10	1	0	1	0	1	0	0	2637
30	13	3	10	39	12	3	6	23	6	1	1	0	0	1	0	0	2637
42	13	2	10	42	11	2	5	18	8	0	1	0	0	1	0	0	2639
33	16	2	10	42	16	9	9	17	8	1	0	1	0	1	0	0	2658
47	16	1	10	39	16	8	9	20	10	1	0	0	1	0	0	1	2700

mage - mother's age, meduc - mother's education, monpre - mother's prenatal condition, npvis - no of visit to the clinic, fage - father's age, feduc - father's education, omaps, maps, cigs - smoking, drinks - drinking, male - gender, (mwhite, mblack, moth, fwhite, fblack, foth)- race of parents, bwght - birth weight.

Feature Engineering

Distribution of Birth Weight

Distribution of Fathers Education

Distribution of Mothers Education

Distribution of Fathers Age

Distribution of Mothers Age

Distribution of total number of vsits

Models and Results

Three models were fitted on the dataset to perform regression analysis. The models are OLS Regression, Lasso Regression and ARD Regression. The results of the models are as shown below:

```
Model      Train Score      Test Score        Test Gap
-----      -----------      ----------        ------------
OLS        0.7238           0.6667            0.0571
*Lasso*    0.7228           0.6758            0.047 <--- Selected model
ARD        0.7233           0.6749            0.0484
```

The selected model is Lasso regression because of the least train-test gap.