CSE 6363 Machine Learning

Project I

*Submitted by*

Mohammed Ghouse Parvez Mohammed Akmal

1001625800

**1) LINEAR REGRESSION**

b) The following figures have the plots of Data points along with the functions of various degrees such as 1,2,3 and 4

1. **Order = 1**

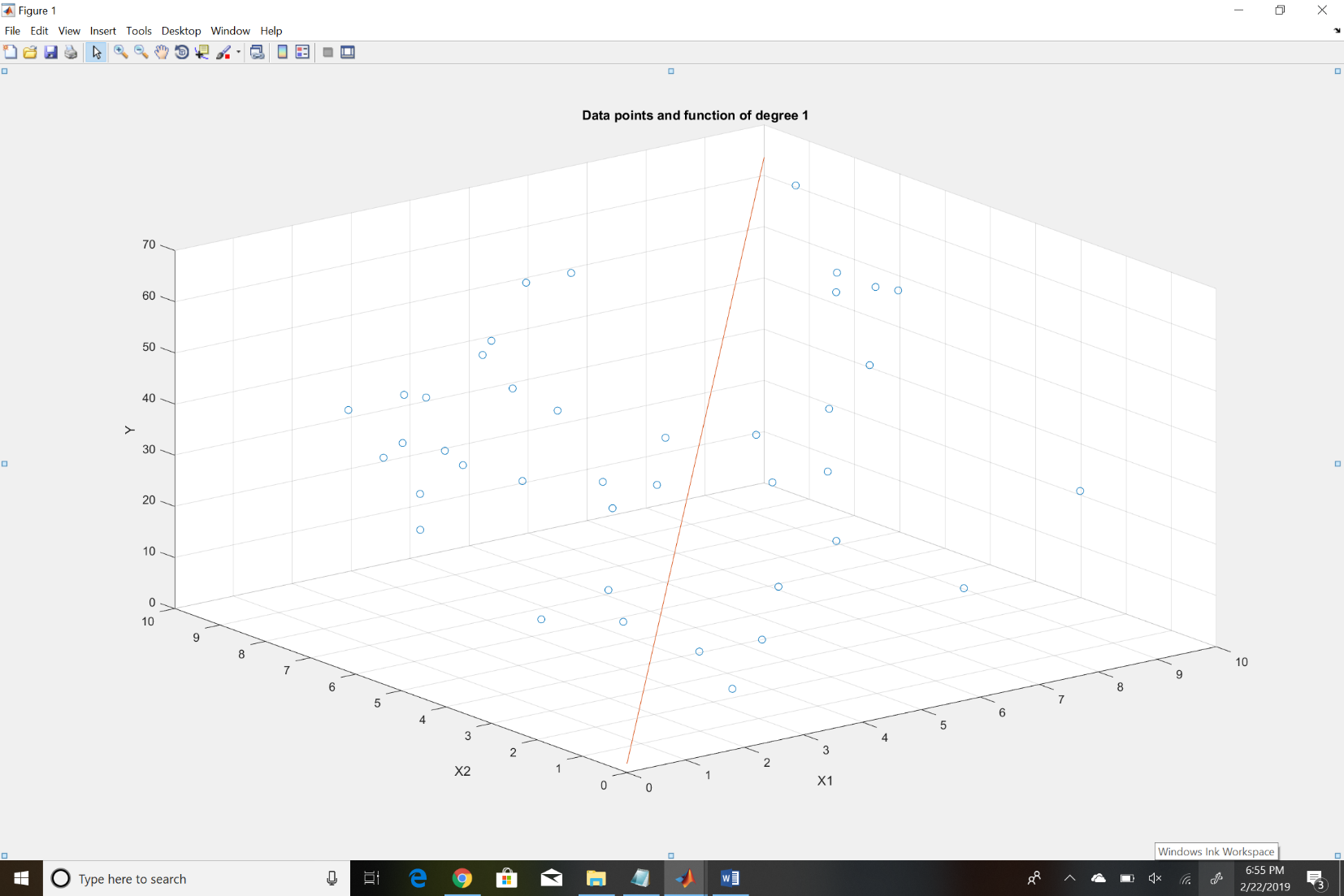


Figure 1: Plot of data points along with function of degree 1

1. **Order = 2**

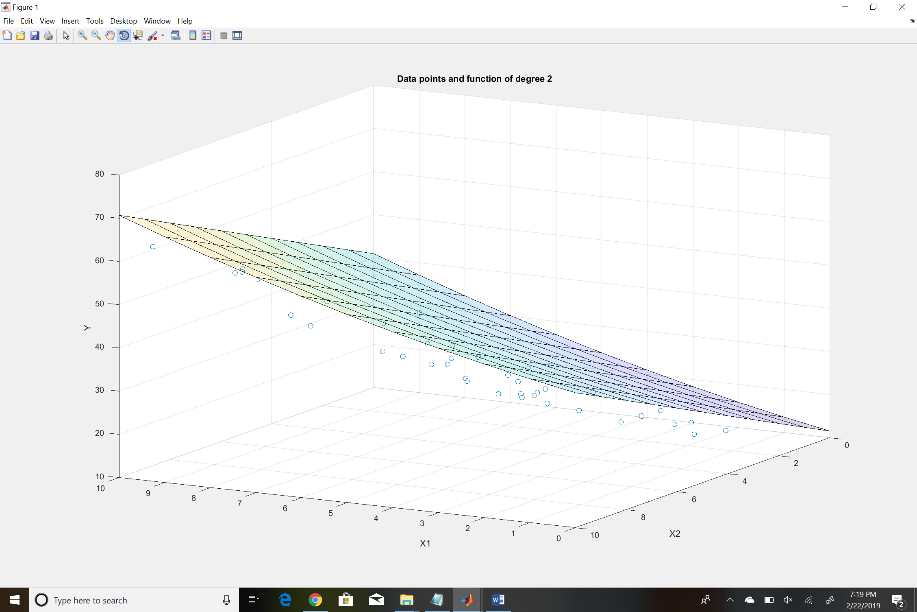


Figure 2a: Plot of data points along with function of degree 2 (view 1)

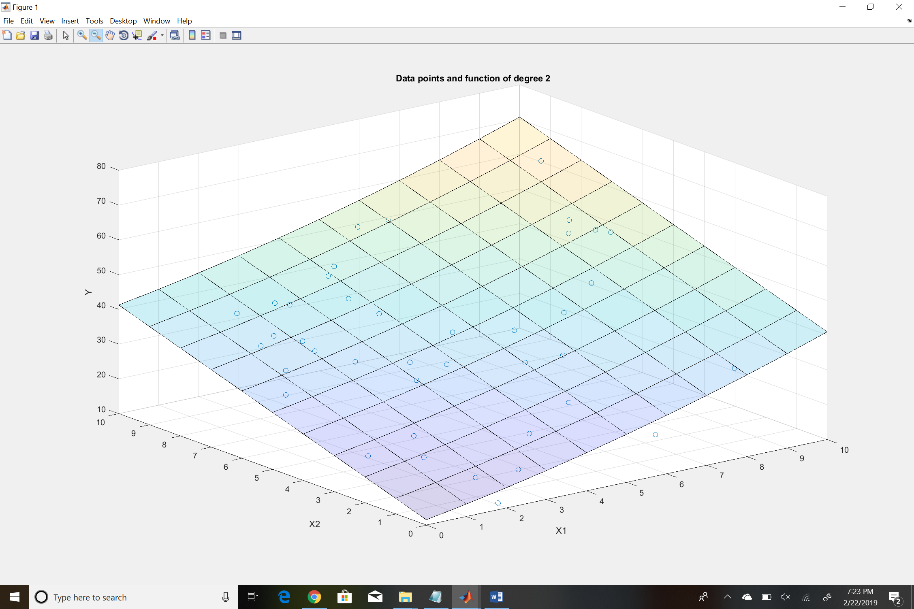
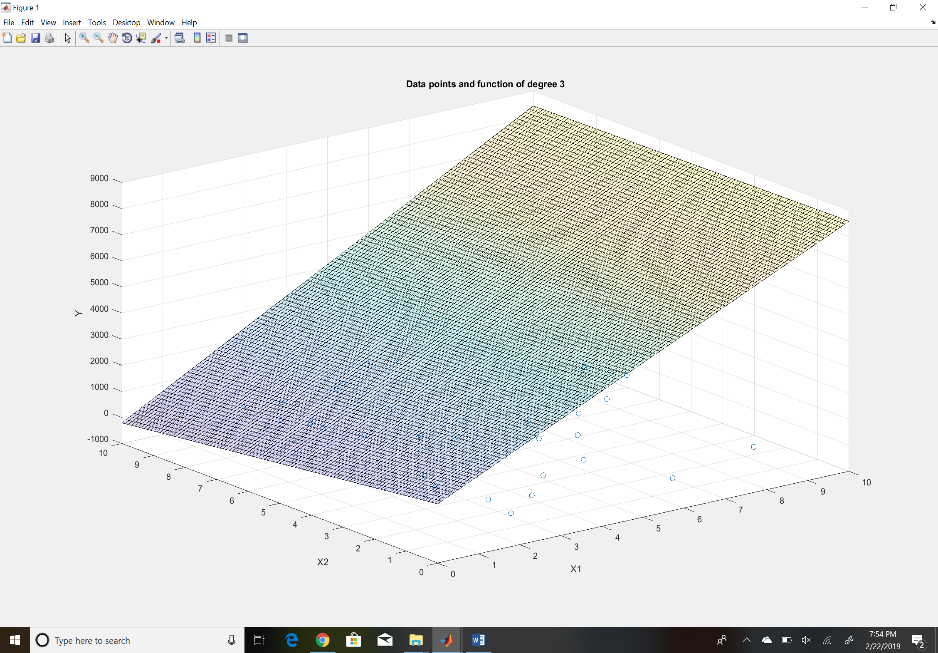


Figure 2b: Plot of data points along with function of degree 2 (view 2)

**(iii) Order = 3**



*Figure 3a: Plot of data points along with function of degree 3 (View 1)*

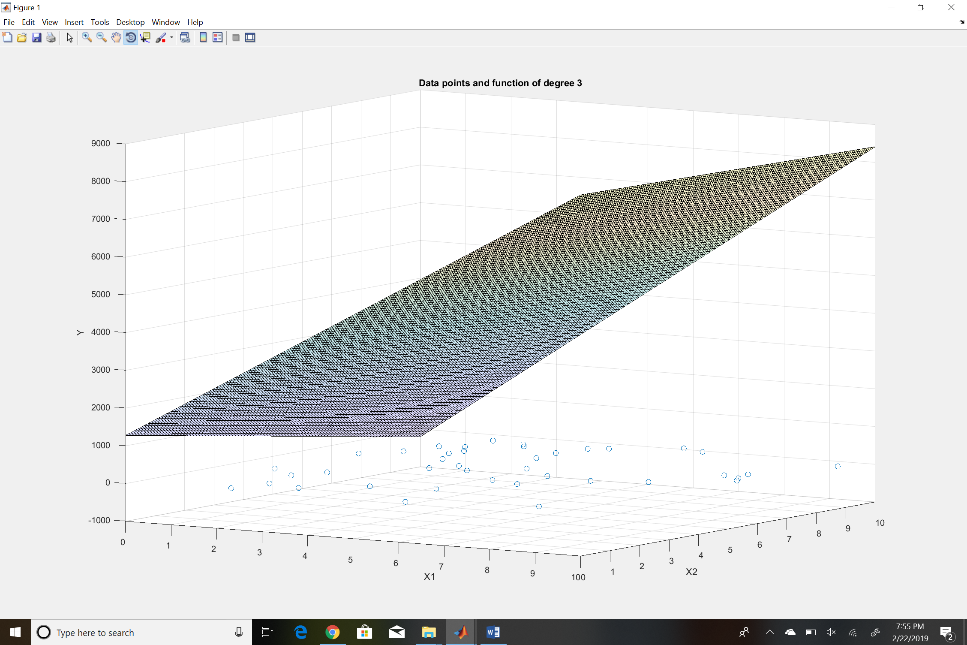
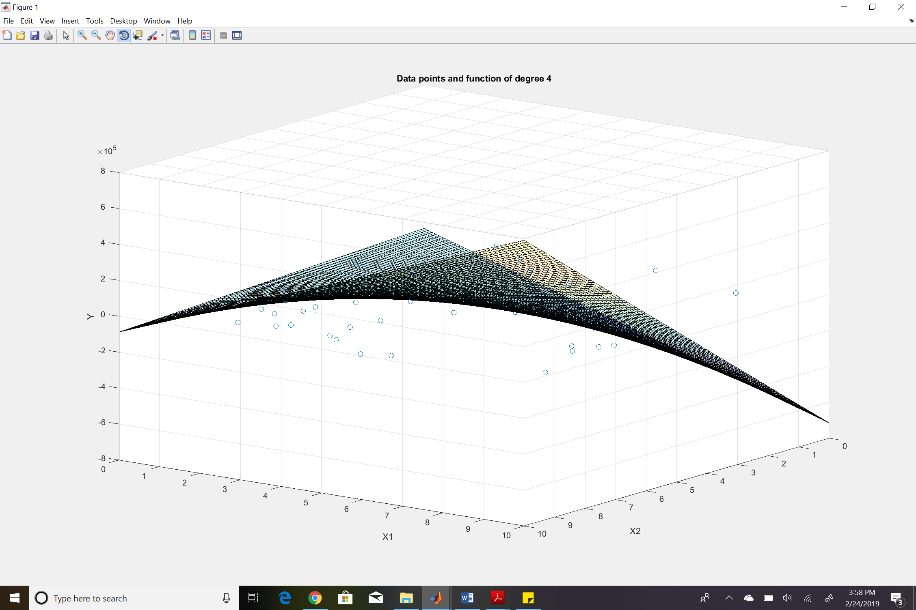
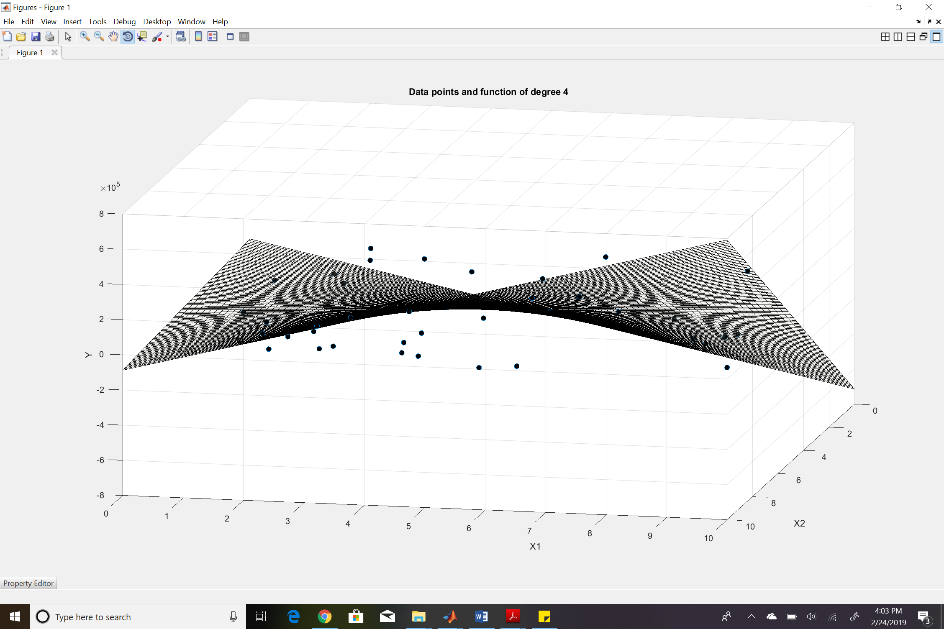


Figure 3b: Plot of data points along with function of degree 3 (View 2)

**(iv) Order = 4**



*Figure 4a: Plot of data points along with function of degree 4 (View 1)*



*Figure 4b: Plot of data points along with function of degree 4 (View 2)*

c) Table containing the order of polynomial versus it’s error percentage

|  |  |
| --- | --- |
| Order of Polynomial | Error (%) |
| 1 | 97.67 |
| 2 | 13.71 |
| 3 | 15.95 |
| 4 | 18.46 |

From the obtained data, we can see that polynomial of order 2 has the least error compared to other orders. Therefore, I will consider the second order polynomial to be the best prediction function as it has the least error and by Occam’s razor ‘The simpler one is better’. Furthermore, higher order polynomials (3,4, ..) may suffer from overfitting.

**2) LOGISTIC REGRESSION**

b) Table of Data versus the prediction by KNN, Naïve Bayes and Logistic Regression

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data | Class | Prediction (KNN)  K=5 | Prediction (Naïve Bayes) | Prediction (Logistic Regression) |
| (155, 40, 35) | W | W | W | W |
| (170, 70, 32) | M | M | M | M |
| (175, 70, 35) | W | M | M | M |
| (180, 90, 20) | M | M | M | M |

In this case, prediction by all three methods yield the same result. Therefore, we cannot conclude that logistic regression yields better results than other prediction methods. However, in general both KNN and Naïve Bayes assume that the features are conditionally independent from one another. But in our case of height, weight and age, this assumption does not hold true. Therefore, Logistic regression can yield better results because it accounts for the correlation between the features.

**3) LINEAR DISCRIMINANT ANALYSIS**

b) Table containing Data versus the prediction by LDA and Logistic Regression

|  |  |  |  |
| --- | --- | --- | --- |
| Data | Class | Prediction (LDA) | Prediction (Logistic Regression) |
| (155, 40, 35) | W | W | W |
| (170, 70, 32) | M | M | M |
| (175, 70, 35) | W | M | M |
| (180, 90, 20) | M | M | M |

From the table above, we can infer that prediction by LDA and Logistic Regression yields the same results. However, usually LDA is more sensitive to outliers unlike Logistic Regression which is not sensitive to outliers. Additionally, LDA has higher efficiency.

**Comparison of Boundaries**

**Plot for Linear Discriminant Analysis Boundary:**

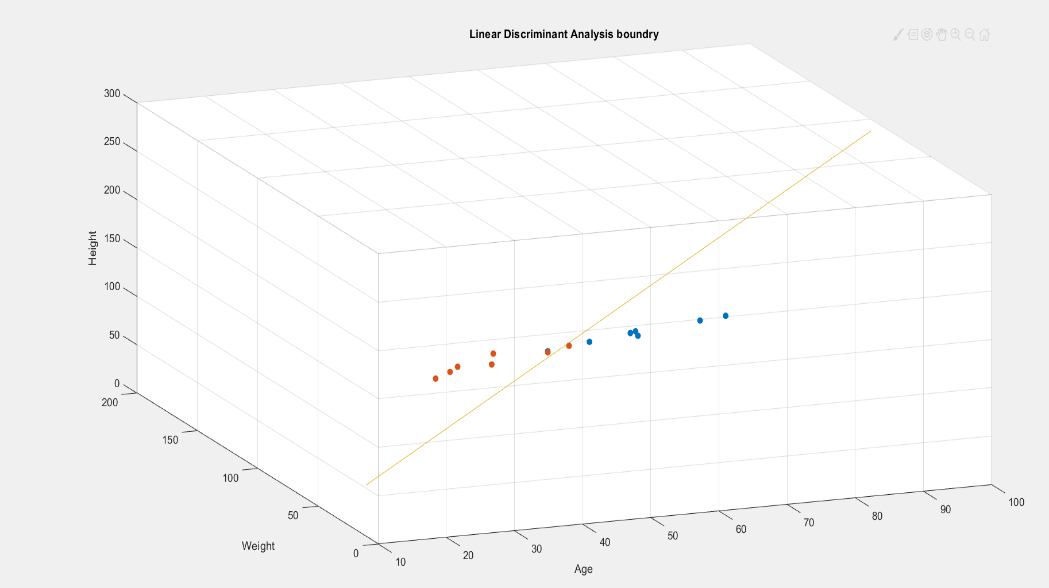


Figure 5: Linear Discriminant Analysis Boundary

**Plot for Logistic Regression Boundary:**

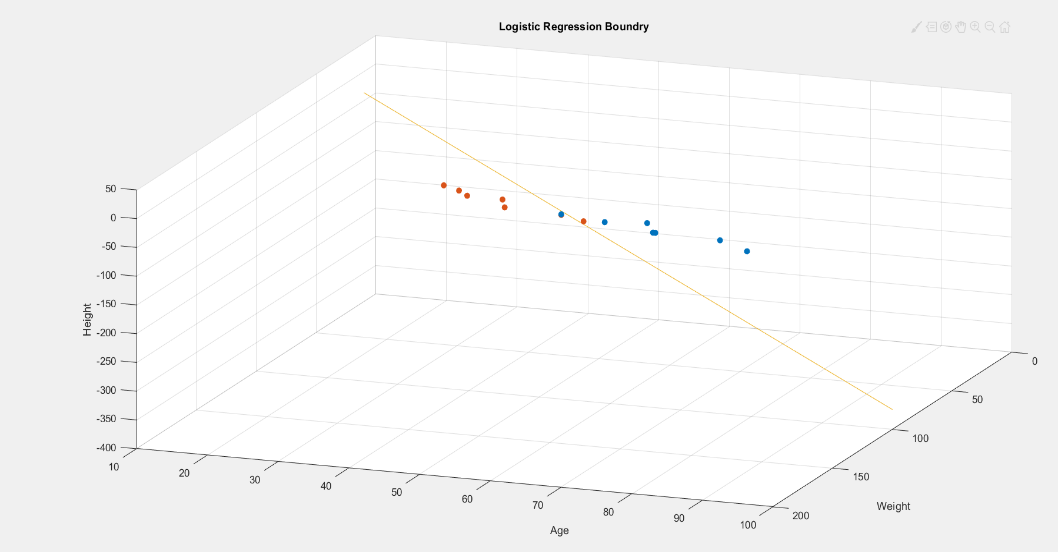
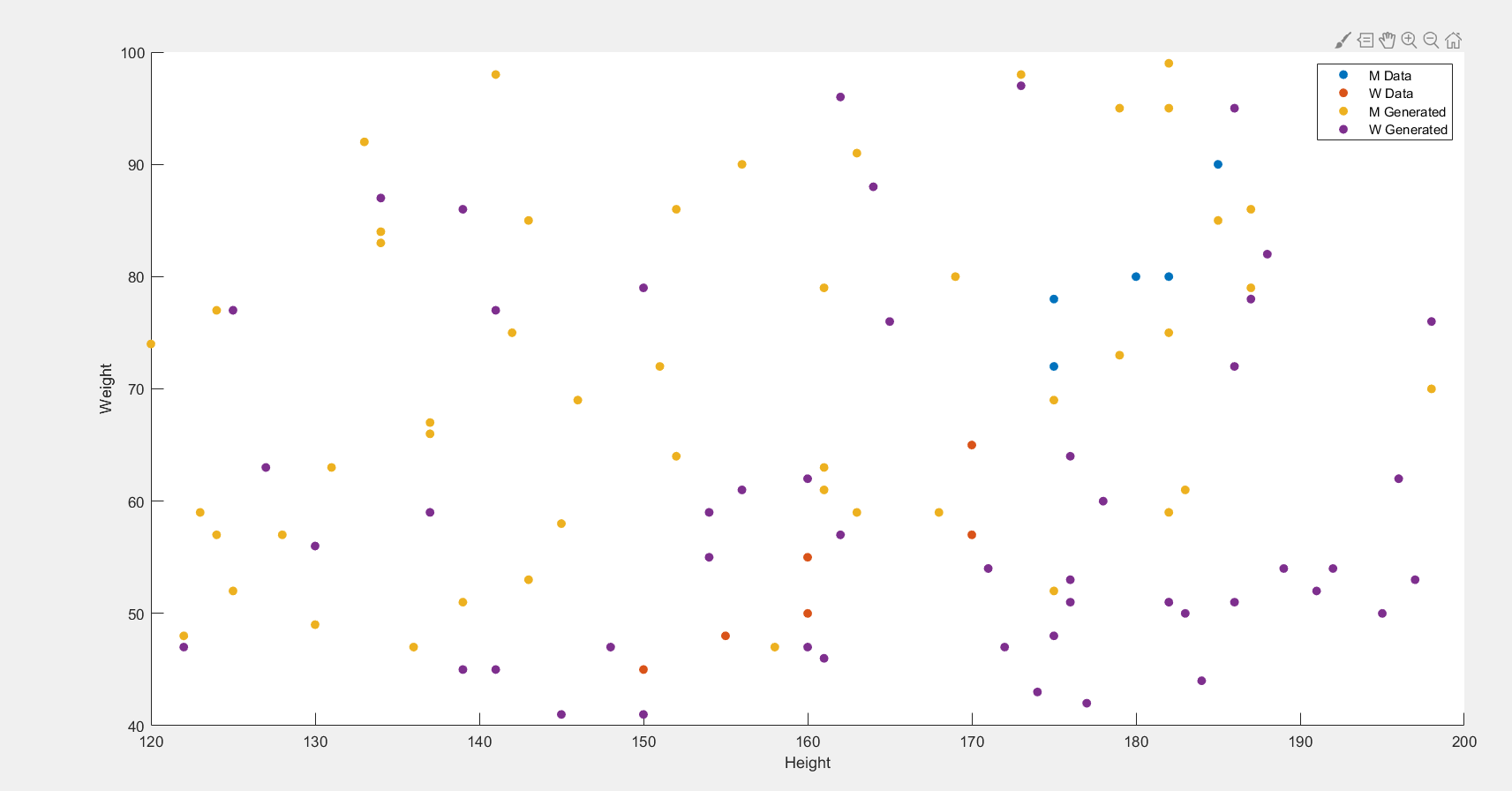


Figure 6: Logistic Regression Boundary

c) Sample Data generated

[123, 53, 54, 'M']  
[187, 49, 72, 'W']  
[131, 71, 70, 'M']  
[170, 41, 66, 'W']  
[183, 75, 6, 'W']  
[181, 94, 51, 'M']  
[135, 88, 38, 'M']  
[125, 53, 16, 'M']  
[143, 76, 11, 'M']  
[169, 63, 68, 'M']  
[127, 80, 37, 'M']  
[122, 44, 82, 'M']  
[127, 53, 20, 'M']  
[164, 91, 72, 'M']  
[173, 86, 54, 'M']  
[198, 70, 52, 'W']  
[180, 45, 22, 'W']  
[152, 52, 85, 'M']  
[177, 49, 15, 'W']  
[159, 47, 78, 'M']  
[154, 54, 26, 'W']  
[120, 71, 98, 'M']  
[191, 68, 18, 'W']  
[149, 64, 27, 'M']  
[175, 58, 78, 'M']  
[198, 47, 71, 'W']  
[144, 54, 29, 'M']  
[191, 82, 11, 'W']  
[127, 85, 80, 'M']  
[170, 54, 25, 'W']  
[172, 51, 78, 'M']  
[163, 97, 82, 'M']  
[132, 98, 38, 'M']  
[169, 74, 27, 'M']  
[171, 84, 93, 'M']  
[128, 96, 11, 'M']  
[122, 78, 61, 'M']  
[146, 42, 43, 'W']  
[187, 44, 28, 'W']  
[167, 97, 58, 'M']  
[190, 41, 45, 'W']  
[172, 85, 62, 'M']  
[134, 50, 82, 'M']  
[174, 51, 13, 'W']  
[157, 59, 83, 'M']  
[199, 68, 98, 'M']  
[186, 65, 12, 'W']  
[140, 40, 86, 'M']  
[159, 74, 97, 'M']  
[146, 41, 5, 'W']  
[133, 94, 53, 'M']  
[181, 54, 17, 'W']  
[127, 59, 97, 'M']  
[185, 73, 69, 'M']  
[126, 67, 64, 'M']  
[178, 80, 13, 'M']  
[131, 85, 85, 'M']  
[190, 75, 7, 'W']  
[152, 45, 77, 'M']  
[152, 87, 53, 'M']  
[131, 46, 33, 'M']  
[130, 47, 39, 'M']  
[128, 73, 70, 'M']  
[147, 86, 98, 'M']  
[172, 91, 86, 'M']  
[143, 77, 46, 'M']  
[187, 89, 91, 'M']  
[167, 47, 85, 'M']  
[168, 50, 99, 'M']  
[178, 84, 71, 'W']  
[153, 63, 52, 'W']  
[170, 71, 50, 'W']  
[197, 54, 15, 'W']  
[172, 90, 96, 'W']  
[125, 89, 10, 'W']  
[193, 78, 1, 'W']  
[164, 44, 68, 'W']  
[151, 53, 94, 'W']  
[163, 65, 91, 'W']  
[192, 87, 91, 'W']  
[169, 90, 48, 'W']  
[168, 40, 48, 'W']  
[153, 64, 8, 'W']  
[168, 81, 30, 'W']  
[140, 46, 63, 'W']  
[145, 96, 64, 'W']  
[164, 94, 94, 'W']  
[150, 94, 88, 'W']  
[141, 65, 86, 'W']  
[132, 71, 2, 'W']  
[166, 59, 50, 'W']  
[168, 69, 44, 'W']  
[136, 67, 11, 'W']  
[124, 53, 36, 'W']  
[132, 79, 69, 'W']  
[198, 43, 60, 'W']  
[126, 61, 36, 'W']  
[169, 58, 43, 'W']  
[144, 40, 14, 'W']  
[154, 93, 10, 'W']

**Plot of generated data points as well as training data points:**



There is not much similarity between generated and training data points as the amount of data in the training set was not sufficient enough to train the model accurately. However we can see that few data points of the same class from the generated and training data are close to each other.