1) $g(x_t | w_1, w_2, w_0) = w_2 (x_t)^2 + w_1 x_t + w_0$

a) We can solve this by converting our $Aw = y$ to

$A = D^T D$ and $y = D^T r$

Solve for w by using $w = (D^T D)^{-1} D^T r$.

$$\Rightarrow \begin{bmatrix} 0.2 & 0.05 & 0.3 \\ 0.07 & -0.3 & 0.15 \\ 0.12 & 0.6 & 0.3 \end{bmatrix} D^T \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}$$

$$D = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \end{bmatrix} = \begin{bmatrix} 1 & -2 & 4 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix} \Rightarrow D^T = \begin{bmatrix} 1 & 1 & 1 \\ -2 & 1 & 0 \\ 4 & 1 & 0 \end{bmatrix}$$

$$\Rightarrow w = \begin{bmatrix} 0.2 & 0.05 & 0.3 \\ 0.07 & -0.3 & 0.15 \\ 0.12 & 0.6 & 0.3 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ -2 & 1 & 0 \\ 4 & 1 & 0 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.2 & 0.05 & 0.3 \\ 0.07 & -0.3 & 0.15 \\ 0.12 & 0.6 & 0.3 \end{bmatrix} \begin{bmatrix} 6 \\ -1 \\ 11 \end{bmatrix}$$

$$= \begin{bmatrix} 4.45 \\ 2.37 \\ 3.42 \end{bmatrix}$$

$$\therefore w = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \begin{bmatrix} 4.45 \\ 2.37 \\ 3.42 \end{bmatrix} \Rightarrow \boxed{\begin{array}{l} w_1 = 4.45 \\ w_2 = 2.37 \\ w_3 = 3.42 \end{array}}$$

b) Using $w_1, w_2, w_3 \Rightarrow \boxed{g(x) = 3.42x^2 + 2.37x + 4.45}$

c) R^2 value $= 1 - E_{rse}$, we know that $E_{rse} = \dfrac{\sum_t \left[r^t - g(x^t | \theta)^2 \right]}{\sum_t (r^t - \bar{r})^2}$

we know $\bar{r} = 6/3 \Rightarrow \bar{r} = 2$

$$E_{rse} = \frac{\left[(2 - g(2|\theta))^2 + (3 - g(3|\theta))^2 + (1 - g(1|\theta))^2 \right]}{(2-2)^2 + (3-2)^2 + (1-2)^2}$$

$$= \frac{(-11.39)^2 + (-7.24)^2 + (-3.45)^2}{0^2 + (1)^2 + (-1)^2} = 97.02$$

$\therefore R^2 = 1 - 97.02 = \boxed{-96.02}$

Question 2

a) Two methods to select a good-fit and generalizable model are: Cross Validation and Structural Risk Minimization. Cross Validation is the process of dividing our data set into two parts(training and validation) and training them on models of varying complexities for the purpose of measuring the error on the validation set. Structural Risk Minimization is the process of using models on the basis of order. They are ordered by complexity which is directly related to the number of free parameters in our model.

b) The Bias-Variance tradeoff is one of the many trade offs that we have observed in the field of machine learning. When training a model, if we observe that it is unable to perform well on testing data due to its inability of catching patterns we say that there is a bias ie it is too simple of a model for our problem. Variance is when the model is incapable of producing useful results due the fact that it is now too complex and is considering all the noise from the testing data as well. The trade off here is that we do not want a model that is too simple to understand the data and make accurate predictions (underfitting) or too complex by capturing all the noise thus causing predictions supported by bad data (overfitting). So the best case is to find a model with the right complexity to interpret our data and this is called the Bias-Variance Trade off.

c) A high bias in our model shows that our model is unable to pick up on the underlying patterns in our data and make accurate predictions based on them solely on the fact that it is too simple of a model to catch them. This is called underfitting. Conversely, a high variance means that too much attention is paid to all the data to a point where it is no longer a prediction but just an attempt made from memorization. This is called overfitting.

3) a) $x = [5, 3, 2, 6]$

$y = [7, 5, 7, 5]$

Mean(x), $\bar{x} = \frac{5+3+2+6}{4} = 4$

Mean(y), $\bar{y} = \frac{7+5+7+5}{4} = 6$

$x - \bar{x} = [1, -1, -2, 2]$

$y - \bar{y} = [1, -1, 1, -1]$

$\text{Var}(x) = \frac{1^2 + (-1)^2 + (-2)^2 + (2)^2}{4} = \frac{10}{4}$

$\text{Var}(y) = \frac{1^2 + (-1)^2 + 1^2 + (-1)^2}{4} = \frac{4}{4} = 1$

$\text{Cov}(x, y) = -0.5$

$\therefore$ Covariance Matrix $= \begin{bmatrix} 2.5 & -0.5 \\ -0.5 & 1 \end{bmatrix}$

b) $z_1 = 1.5, \quad z_2 = 2.4 \quad r = \frac{-0.5}{\sqrt{2.5}\sqrt{1}} = 0.316$

$$p(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(z_1^2 - 2\rho z_1 z_2 + z_2^2\right)\right]$$

$$\Rightarrow \frac{1}{\pi(2.5)(1)(\sqrt{1-0.316})^2} \exp\left[-\frac{1}{2(1-(-0.316)^2)}\left(1.5^2 - \left(\frac{2(0.316)}{(1.5)(2.4)} + (2.4)^2\right)\right)\right]$$

$\Rightarrow 0.106 \exp[-5.71] = 0.0035$

$\therefore$ Joint Bivariate Density $= 0.0035$

Question 4

a) Mean Imputation is the process of inputting missing values in a dataset with the mean of the observed values. This process helps in filling up a data set while ensuring there are no massive fluctuations.

b) Average of x1 = $\sum$x1 / N = 6.44

   This means that All the empty values in x1 are now **6.44**

   Average of x1 = $\sum$x2 / N = 7.22

   This means that All the empty values in x1 are now **7.22**

Question 5

a) The Euclidean Distance is a well known mathematical concept that lets you calculate the shortest distance between two points in a space. An application of this in AI could be to find the similarity between two vectors(data) by seeing how far apart they are from each other or its implementation in the core of KNN. Unlike the Euclidean Distance, the Mahalanobis distance is the distance between a point and a distribution. This can be used in clustering methods where we can determine which group a point belongs to by calculating the distance between the point and the mean of each cluster. Using the Euclidean Distance might prove to be more useful in a case where the data points might be independent so a calculation of distances might be sufficient for a classification. Using the Mahalanobis distance might prove to be more accurate in a case where there is more correlation between the points and we can classify our data points with reference to our given correlations.

b) For the given multivariate distribution, I expect the contour map to look like Elliptical. The center of the Ellipse would be the mean vector and the covariance vector would provide the shape and size.

6) $\mu_A = [2, 3, -1]^T$ $\quad \Sigma_A = \begin{bmatrix} 1 & 0.5 & 0.3 \\ 0.5 & 2 & 0.2 \\ 0.3 & 0.2 & 1 \end{bmatrix}$

$\mu_B = [0, 1, 4]^T$ $\quad \Sigma_B = \begin{bmatrix} 1.5 & 0.1 & 0.4 \\ 0.1 & 1.8 & 0.6 \\ 0.4 & 0.6 & 2 \end{bmatrix}$

$\det(\Sigma_A) = 1((2\times1) - (0.2\times0.2)) - 0.5(0.5\times1) - (0.2\times0.3)) + 0.3((0.5\times0.2) - (2\times0.3))$

$= 1.59$

$\det(\Sigma_B) = 1.5((1.8\times2) - (0.6\times0.4)) - 0.1((0.1\times0.6) - (0.1\times0.4)) + 0.4((0.1\times0.6) - (1.8\times0.4))$

$= 4.6$

$\det(\Sigma^{-1}A) = 1/5.9 = 0.628$

$\det(\Sigma^{-1}B) = 1/4.6 = 0.217$

Multivariate Normal distribution:

$$P(x|C_i) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_i)^T(\Sigma_i^{-1}(x-\mu_i)\right]$$

For class A:

$$\frac{1}{(2\pi)^{1.5}\times(1.59)^{1/2}} = 0.053$$

$$P(x|C_1) = 0.053 \exp\left[-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)\right.$$

$$(x-\mu) = \begin{bmatrix} -1 \\ -1 \\ +1 \end{bmatrix}$$

$$\Rightarrow (x-\mu_1)^T \cdot \Sigma_A \cdot (x-\mu) = [-1\ -1\ 1]\ (\Sigma_A)\ \cdot \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix} = 3.04$$

(Given.)

$$\Rightarrow (0.05)\exp\left(\frac{-1}{2}\times 3.04\right) = 0.0109$$

Class B:

$$P(x|C_2) = \frac{1}{(2\pi)^{1.5}(4.6)^{0.5}} \exp\left[\frac{-1}{2}(x-\mu_2)\Sigma_2^{-1}(x-\mu_2)\right]$$

$$\frac{1}{(2\pi)^{1.5}(4.6)^{0.5}} = 0.0037$$

$$(x-\mu_2)^T \Sigma_B^{-1} (x-\mu) = \begin{bmatrix}1 & 1 & -4\end{bmatrix} \Sigma_B^{-1} \begin{bmatrix}1\\1\\-4\end{bmatrix} = 13.336$$

$$P(x|C_2) = 0.0037 \exp\left(\frac{-1}{2} \times 13.336\right)$$

$$= 0.00004$$

$\therefore P(x|C_1) > P(x|C_2) \Rightarrow P(x|C_A) > P(x|C_B)$

$\Rightarrow$ Predicted Class A

b) $P(x|C_A) = 0.0107$ (The covariance does not change)

$$P(x|C_2) = 0.05 \times \exp \times \left[\frac{-1}{2}(x-\mu)^T (\Sigma_A^{-1})(x-\mu)\right]$$

$$\downarrow$$

$$\frac{-1}{2}\begin{bmatrix}2.2136 & 0.4122 & -4.4852\end{bmatrix}\begin{bmatrix}1\\1\\-4\end{bmatrix}$$

$\Rightarrow$ $0.05 \times \exp(-1/2 \times 21.627)$

$\Rightarrow$ $0.05 \exp 0.000001006$

$\therefore P(x|C_1) > P(x|C_2)$

The predicted class would still be A