# FYS-STK4155 Project 2

Bendik Steinsvåg Dalen & Gabriel Sigurd Cabrera

November 1, 2019

**Abstract**

## Introduction

## Data

### Credit Card Data

Our first dataset contains real credit card metadata for 30,000 people, in the form of a `.xls` file; each given datapoint (or person) has 23 features and one *binary output* denoting whether or not they've defaulted on their credit card debt. These features can be summarized as follows:

| Feature No. | Description | Data Type |
|---|---|---|
| 1 | Total Credit Given | Continuous |
| 2 | Gender | Categorical |
| 3 | Education | Categorical |
| 4 | Marital Status | Categorical |
| 5 | Age | Continuous |
| 6-11 | Month-Wise Repayment Status | Categorical |
| 12-17 | Month-Wise Bill Statement | Continuous |
| 18-23 | Month-Wise Amount Paid | Continuous |

For more detailed information regarding this dataset, and the file itself, visit https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

### The Franke Function

The second dataset will be given by the *Franke function*, which is defined as follows:

$$
\begin{aligned}
f(x, y) = &\frac{3}{4}\exp\left(-\frac{(9x-2)^2}{4} - \frac{(9y-2)^2}{4}\right) \\
&+ \frac{3}{4}\exp\left(-\frac{9x+1}{49} - \frac{9y+1}{10}\right) \\
&+ \frac{1}{2}\exp\left(-\frac{(9x-7)^2}{4} - \frac{(9y-3)^2}{4}\right) \\
&- \frac{1}{5}\exp\left(-(9x-4)^2 - (9y-7)^2\right)
\end{aligned}
$$

We will be solving the Franke function for 100 $x$-values and 100 $y$-values in the range $[0, 1]$, leaving us with a grid containing a total of 10000 $xy$ coordinate pairs. This leaves us with the values plotted in Figure 1.

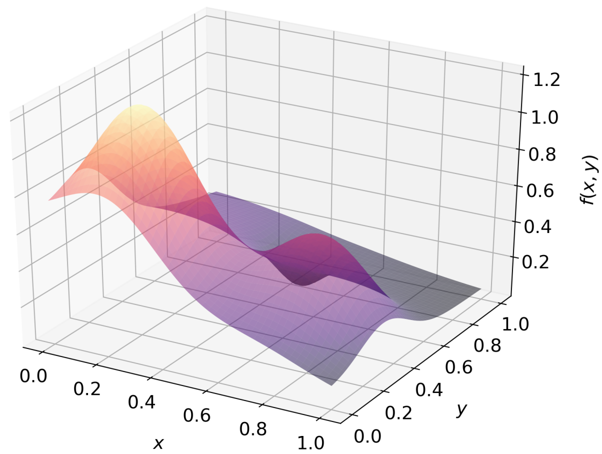


Figure 1: The *Franke function* for $x$ and $y$ values ranging from zero to one.

In addition, we will also be adding *Gaussian noise* to each value $f(x, y)$, such that we are left with values as seen in Figure 2.

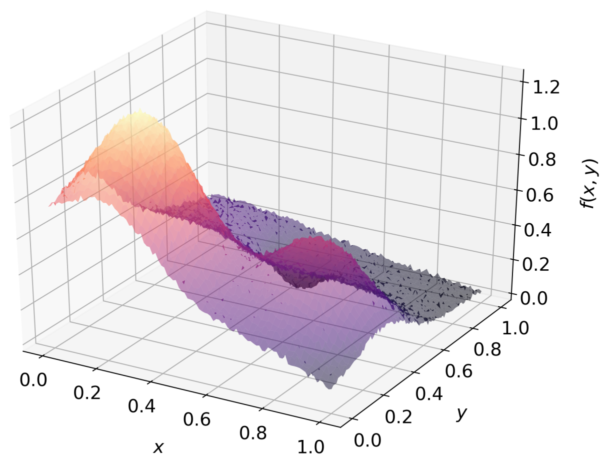


Figure 2: The *Franke function* for $x$ and $y$ values ranging from zero to one, with a Gaussian noise $N(0, 0.01)$

# Method

## Mean Squared Error

To get a measure of success with respect to the implemented method and parameters, we can calculate the mean difference in the squares of each measured output $y_i$ and their respective predicted outputs $\hat{y}_i$:

$$MSE(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \mathbb{E}\left[(\mathbf{y} - \hat{\mathbf{y}})^2\right]$$

The lower the $MSE$, the closer the polynomial approximation is to the original dataset. If it is too low, however, we run the risk of overfitting our dataset, which is not desireable either – fortunately, this not an issue within the scope of this report.

## $\mathbf{R^2}$ Score

Another measure of success is the *coefficient of determination*, colloquially known as the $R^2$ score, is given by the following expression:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y}_i)^2}$$

The closer $R^2$ is to one, the closer the polynomial approximation is to the input/output dataset, although a perfect score can once again arise due to overfitting just as in the case of the $MSE$.

# Results

# Discussion

# Conclusion