UNIVERSITÉ LIBRE DE BRUXELLES
Faculté des Sciences
Département d'Informatique

Evaluation of DNA methylation-based and aging biomarkers resilience to noise and missing data

Akilleas Pappas

Promoter : Prof. Matthieu Defrance Master Thesis in Computer Sciences

Academic year 2019 - 2020

Acknowledgment

I just want to thank my thesis promoter, Mr. Matthieu Defrance, for giving me the opportunity to work on such a subject, for listening to my concerns and for the advices and explanations I received with a lot of patience from him. I would also like to thank the 2 other professor who accepted to join the jury : Mr. Tom Lenaerts and Mr. Maarten Jansen.

I'm very grateful to my entourage for their patience when I was struggling, their love when I was in a difficult situation and obviously their support during all this time.

I'm gonna finish by thanking the university and its secretariat for putting us in the right conditions to finish this thesis, especially in the current climate.

Contents

Chapter 1

Introduction

1.1 Thesis Objectives

The main objective of this master thesis is to use DNA methylation-based biomarkers and aging to evaluate the influence of noise and missing measurements on the prediction accuracy. We will implement different models to answer at those following questions : Which model is the best in term of accuracy and prediction ? Will models fall because of the lack of robustness and thus what is the most robust model? How much quantity of noise do we need to see a consequent diminution of accuracy on our predictions? Are missing values most destructive for data than noise ?

1.2 Background

The purpose of this section is to give the reader a scientific background so that he or she is able to understand what we are going to do and understand the concepts we use. The most important thing to understand is what a biomarker is and what it's used for so let's start with that.

Biomarkers

"A biomarker is a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention.[18]", this is definition of the biomarker since 1980, defined by the National Institutes of Health Biomarkers Definitions Working Group. To be more specified the United Nations and the International Labor Organization, has defined a biomarker as "any substance, structure, or process that can be measured in the body or its products and influence or predict the

incidence of outcome or disease[18]".

There is different types of biomarkers but here we will focus on biomarkers of aging. They are biomarkers that could predict functional capacity at some later age better than will "chronological age". We can deduce by the understanding of this sentence that biomarkers of aging would give the true "biological age", which may be different from the chronological age. As a little reminder , the "chronological age" is the calendar time that has passed since birth. Zero is the time at birth so if there is negative numbers it means that it indicates prenatal ages, whereas obviously positive numbers indicate postnatal ages. [8] "Biomarkers of aging based on DNA methylation data have enabled accurate age estimates for any tissue (or blood) during the entire life course [8]" according to J.Zhao.L and his colleagues. They explains that "These epigenetic clocks link developmental and maintenance processes to biological aging, giving rise to a unified theory of life course. And the point is that DNA methylation-biomarkers can be used to accurately estimate the age of tissues and cell types, forming an accurate epigenetic clock[8]."

"Anyway it has been recognized that valid and reliable biomarkers of aging will be needed to achieve the ancient goal of understanding, slowing, halting or even reversing aging."[15] This is why Baker and Sprott[15] proposed the identification of biomarkers that can accurately and rapidly predict the functional capability of a person or organ and how it changes with age, the goal being to identify markers of biological age instead of using chronological age because the last one is biased.

One of our task is to identify a restricted set of markers (e.g. genes) that allow an good estimation of the needed properties (theory of aging in our case). In fact it exists a lot of genes this is why we would have to select those who are the most correlate to age modification.

To do so we need power and useful tools to be able to use its biological data and to be able to extract interesting information from it. We will use machine learning combined to algorithms based on artificial intelligence and statistics. The main purpose as we said is to try different models and to use the combination of biomarkers and machine learning to predict something (e.g the age), then we will have to study the influence of noise on our set of data.

Chapter 2

State of the Art

In order to answer to all of this questions we have to introduce briefly in more details how Machine Leaning works because there is different types of Machine Learning depending of what our goal is.[19]

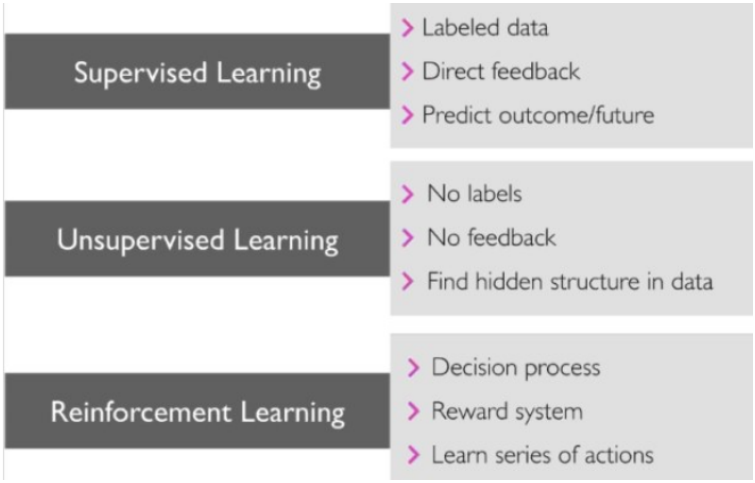


Figure 2.1: Sebastian Raschka explains 3 main types of machine learning.[19]

2.1 Supervised Learning

In supervised learning as you can see on the picture above the goal is to be able to accurately predict an outcome based on the available features.

Considering the example of filter email spam that you can found approximately everywhere on the internet because it is one of the most mainstream and understandable example, we can train a model using a supervised machine learning algorithm on a set of data of labeled emails, those emails are correctly marked as spam or not-spam, and the goal is to predict whether a new email belongs to

either of the two categories. This is just to illustrated that we can use Machine Learning for everything, in our case we will combined it with biomarkers in order to predict age so as mentioned supervised learning is used for prediction.

There is few things to do and few steps to follow so we can solve a given supervised learning problem. Ameet Talwalkar Mehryar Mohri [17] cites some example of an non-exhaustive list explaining what are those big steps, we will follow those later to implement our approach:

- "Determine training examples and training set."
- "Determine the number of input feature. The number of features should not be too big but should contain enough information to accurately predict the output."
- "Decide of the corresponding learning algorithm."
- "Run the learning algorithm on the gathered training set."
- "Evaluate the accuracy. The performance of the resulting function should be measured on a test set that is separate from the training set."

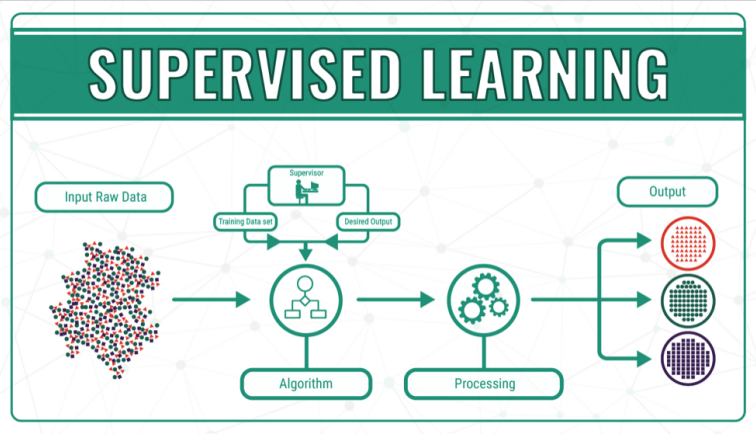


Figure 2.2: Supervised learning.[24]

2.2 DNA methylation biomarkers and aging

Now that we know the meaning of biomarkers and the goal of using Machine Learning which is in our case to predict age, we will have to focus on how we will accomplish that. From where will we extract the information needed to predict, how will we predict, will there be constraints and what are the consequences of that ?

A comparative study of different methods for age estimation concluded that DNA methylation is the "most promising age-predictive biomarker". And a recent review of six types of potential biological age estimators concluded that the epigenetic clock is "the most promising molecular estimator of biological age."[10] But what is DNA methylation exactly ?

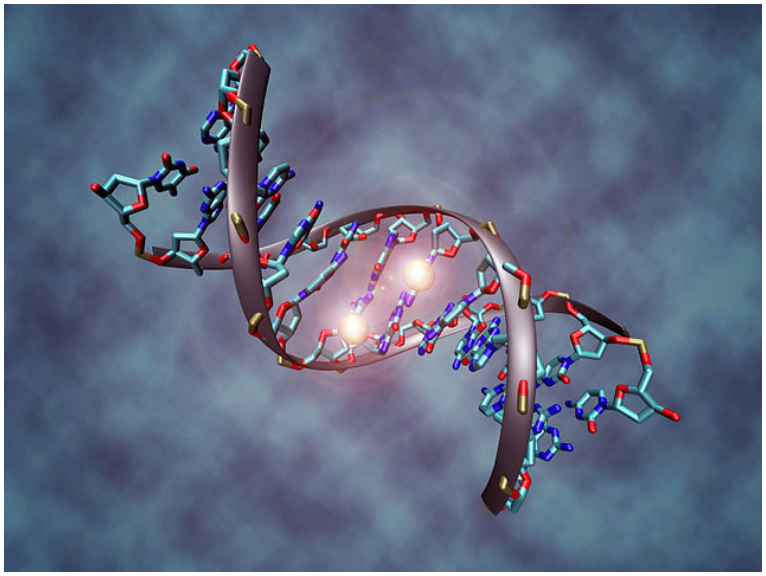


Figure 2.3: Representation of a DNA molecule that is methylated.

2.2.1 DNA methylation

"DNA methylation is an epigenetic mechanism used by cells to regulate gene expression. A number of mechanisms exist to regulate gene expression in eukaryotes, but DNA methylation is also a generally used epigenetic signaling tool that can fix genes in the 'OFF' position[14]" explains the Labclinics." In the above Figure 2.3, the two white spheres represent methyl groups. They are bound to two cytosine nucleotide molecules that make up the DNA sequence. Remember that a methyl group is an alkyl derived from methane, containing one carbon atom bonded to three hydrogen atoms. An example of methylation in Figure 2.4.

"Anyway, over the last decades, scientists have made varied discoveries about DNA methylation, the researchers linked the abnormal methylation of DNA to several harmful effects, including human diseases.[21]"."In humans and other mammals, DNA methylation levels can be used to accurately estimate the age of tissues and cell types, forming an accurate epigenetic clock.[22]" All those sentence prove us all the big discoveries which occurred and it make us thing that there is still work to do with DNA methylation in humans.

A longitudinal study of twin children showed that, between the ages of 5 and 10,

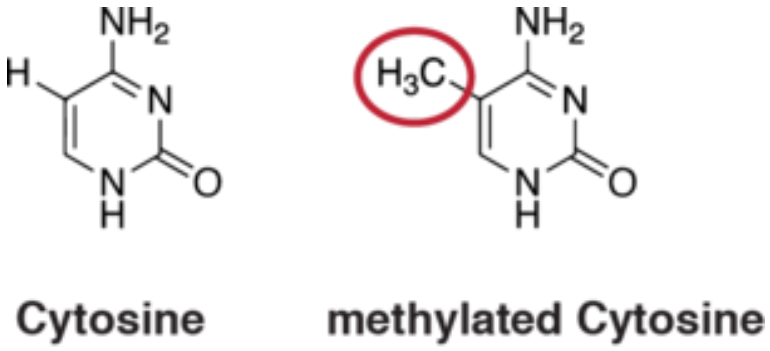


Figure 2.4: Example of molecule methylation.

there was divergence of methylation patterns due to environmental rather than genetic influences. **There is a global loss of DNA methylation during aging**. "In a study that analyzed the complete DNA cells in a newborn, a 26 years old individual and a 103 years old individual was observed that the loss of methylation is proportional to age which is very interesting to know.[2]"

2.2.2 Epigenetic

In fact, "Epigenetics is the study of heritable phenotype changes that do not involve alterations in the DNA sequence, it means that it is a mechanism of controlling genes.[7]" We have to control those genes because DNA is identical in all cells of an organism, but very different genes must be active in a brain cell than in a skin cell, by the way we will use tissues and blood cell in our implementation. Also try to remember that "For human being, one set of our genes is active while we are in the womb, another in our childhood, and yet others as we sexually mature and get older.[7]" which makes genes very delicate to analyze because it changes a lot of times during times life. Due to epigenetic,age estimators have been created as we will see below..

Epigenetic age estimators are "sets of CpGs (also known as clock CpGs) that are coupled with a mathematical algorithm to estimate the age (in units of years) of a DNA source, such as cells, tissues or organs[10]." Steve Horvath and Kenneth Raj described this estimated age, also referred as DNAm age, "not only as a reflection of chronological age but also as the biological age of the DNA source. Owing to their accuracy, DNAm age estimators are often referred to as epigenetic clocks.[10]" (Figure 2.5) .

Different epigenetic clock

"Epigenetic clock is a biochemical test that can be used to measure age. The test is based on DNA methylation levels as we saw before. At a simple level,

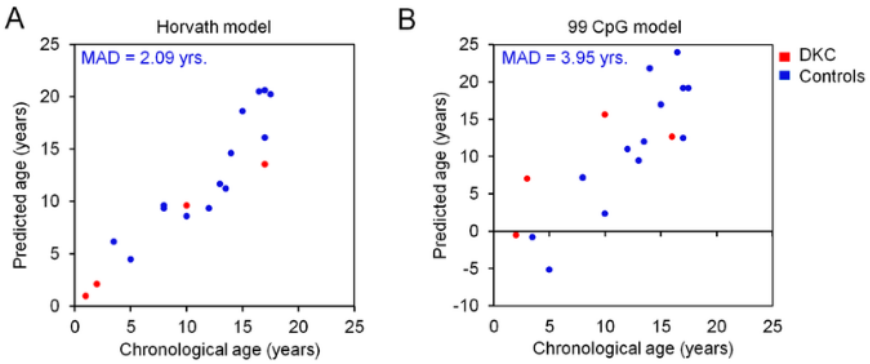


Figure 2.5: Example of epigenetic age-predictions of Dyskeratosis congenita samples following different models.[11]

the epigenetic clock is just a set of on/off gene switches that best correlates with age".[10] We can also see it like a process include in the body that continues forever, resulting in aging. There is different types of epigenetic clock depending of what we are analyse:

- Horvath's clock[22] : using Multi-tissue DNAm age estimator
- Hannum's clock[8] : using Single-tissue DNAm age estimator
- Levine's clock : using DNAm PhenoAge

A lot of factors can explain why biological age is higher than chronological age. Steve Horvath and Kenneth Raj again explains what is DNAm and why it is interesting to use this to make prediction : "The multi-tissue DNA methylation-based (DNAm) age estimator stands out because of its correlation with chronological age across multiple tissue types, its high accuracy in children, its strong correlation with gestational age (differentiation day) in neuronal cell culture models and the homogeneity of its age estimates across tissues. The phenotypic age estimator stands out in terms of its predictive accuracy for time to death, its association with smoking status etc. For example in smokers or people with Down syndrome (who age much faster), the biological age turned out to be higher than their chronological age. This is obviously not the case for animals because they do not drink or smoke. In general, DNAm PhenoAge and DNAm age as calculated by the single-tissue age estimator known as Hannum's clock outperform other blood-based biomarkers in regard to lifespan prediction[10]".

In our case we will use the Horvath's clock as epigenetic clock so we can take multi tissues from different sources, like blood etc, it will be easier to show what we want to. However it would have been interesting to use only single-tissue DNA biomarkers to maybe try to compare both of clock's but It will not happen in this paper.

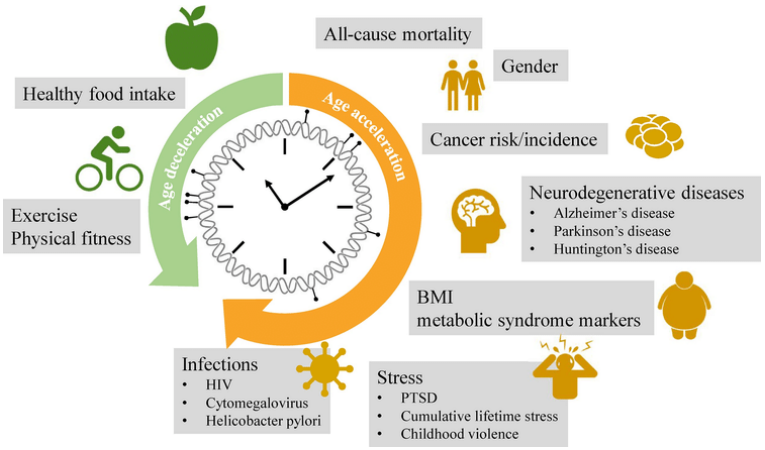


Figure 2.6: Epigenetic clock in life according Ken Declerck and Wim Vanden Berghe.[13]

What are the motivations of epigenetic clock ?

Overall, and **this is concerning us**, Yuri Deigin tell us that biological clocks are expected to be useful for studying what causes aging and what can be done against it. As he said "Remember that the main point of aging biomarkers is to use them to help us find the most effective ways to fight aging itself. Therefore, immediately after the methylation clock has established itself as such a potential biomarker, we have to check if it reflects something workable[6]".

Let's go a bit further in details and to describe the utility of biological clock we will give some interesting information. Just an example to understand and to introduce the next paragraph, in a large-scale study by Horvath et al.[5] on 13 000 people, it was found that "each year of difference between methylation age and chronological age (i.e. if you are 45 year old, and the methylation clock shows 46) increases one's mortality risk by 2% to 4%. Indeed methylations clock are highly correlated with mortality. This observation worked in both directions and had a cumulative effect[5]."

"Distinguishing several measures for assessing the accuracy of an age estimation method is useful because each measure has distinct advantages according to Horvath.[10]" We can compare measures like median error, which is the median absolute difference between the age estimate and chronological age. We can talk also about the root mean squared error and the mean absolute error. Finally the R square value is also important because it calculates the degree of correlation between predicted age and true age. We will use all those statistical measures to compare our different models and also to compare datasets with and with out noise.

Chapter 3

Approach considered

After detailing the state of art, the next step is to explain what we are going to do. We will start by describing what kind of data we will analyse to extract the information we need with the aim of comparing them. Then we will explain how we will do this, by using prediction algorithms included in libraries. Finally, we will add the notion of noise and explain why it is relevant to use it in our situation.

3.1 Data handling

This is a quick section to point out what kind of data we will use in a theoretical way, we will use blood data mainly but it is interesting to do is to take also other tissues data and then compare what are the changes of predictions depending of the type of data. Also we will take data from patients of every age range but we will detailed those data in our implementation chapter. Finally as we said we will insert noise in our data to try to understand the differences when we have noise and when we don't. All those things are going to be implemented using regression methods as algorithms of prediction, this is the purpose of our next section.

3.2 Prediction algorithms

3.2.1 Simple Linear Regression

Simple linear regression is a statistical method that makes us study relationships between two continuous variables and predict outcomes. The first variable is called the predictor which is the DNA in our case and the second variable is the outcome which is the age in our situation. It means that for each value of the predictor we will have a value for the outcome and at the end we can draw a line going through all the data point to approximate the dependency between the 2 variables. Obviously if we have more than 2 variables we called this regression a

Multiple Linear Regression. Unfortunately simple linear regression is not enough to prove causation, it proves that there is a correlation but it don't explain the cause-and-effect relationship. Also the redundancy information will force the linear regression to perform poorly because it will be unstable. Anyway in our case we can't predict age only with one DNA information of one gene because this is not relevant, we will have to use more features correlated to age. So we can forget about Simple Linear Regression but it forces us to introduces in a better way the Multiple Linear Regression because as we already said it is impossible to analyze the complex relationship between DNA methylation and age using a simple linear model.

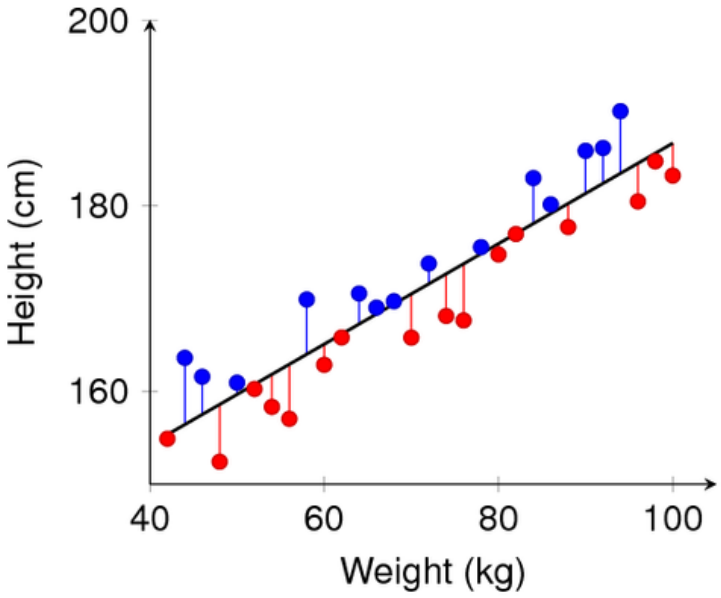


Figure 3.1: Example of linear regression.[12]

3.2.2 Multiple Linear Regression

Their is other types of regression as we said for example logistic regression but we are not interested because it outputs a binary answer true or false, yes or not , 0 or 1 and this is not what we want. However the Multiple Linear Regression is interesting because it estimates the relationship between two or more independent variables and one dependent variable, it is interesting for us because the prediction of the age given DNA data is computed using more than one variable otherwise it would be too easy. Furthermore Multiple Linear Regression is mainly used in biological domain so this is perfect for us. One example below to understand it better (Figure 3.2).

Here, Rebecca Bevans calculated the predicted values of the dependent variable

Figure 3.2: Example of multiple linear regression.[4]

(heart disease) across the full range of observed values for the percentage of people biking to work. She said that "to include the effect of smoking on the independent variable, we calculated these predicted values while holding smoking constant at the minimum, mean, and maximum observed rates of smoking.[4]" This example just to show that we can add multiple cause to a given consequence. In our case, for instance, if someone has a disease and we try to predict the age only with the disease parameter with out looking to DNA, we will found a biased result which is not what we want.

3.2.3 Support Vector Regression

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for regression challenges (in our case) but also for classification. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features we have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.[20] Let's focus now on the The Support Vector Regression (SVR) which uses the same principles as the SVM for classification, with only a few minor differences.

First of all, because output is a real number and not classes it is very difficult to predict the information at hand, which can has infinite possibilities. Let's get in details now with Mariette Awad and Rahul Khanna explaining how SVM works : "SVR trains using a symmetrical loss function, which equally penalizes high and low misestimates. Using Vapnik's -insensitive approach, a flexible tube of mini-

mal radius is formed symmetrically around the estimated function, such that the absolute values of errors less than a certain threshold are ignored both above and below the estimate. In this manner, points outside the tube are penalized, but those within the tube, either above or below the function, receive no penalty. One of the main advantages of SVR is that its computational complexity does not depend on the dimensionality of the input space. Additionally, it has excellent generalization capability, with high prediction accuracy.[16]" Compared to multiple linear we will therefore have parameters to choose for our model so all those arguments make us think that this is an interesting model to implement.
An visual example of SVR is listed below on the Figure 3.3.

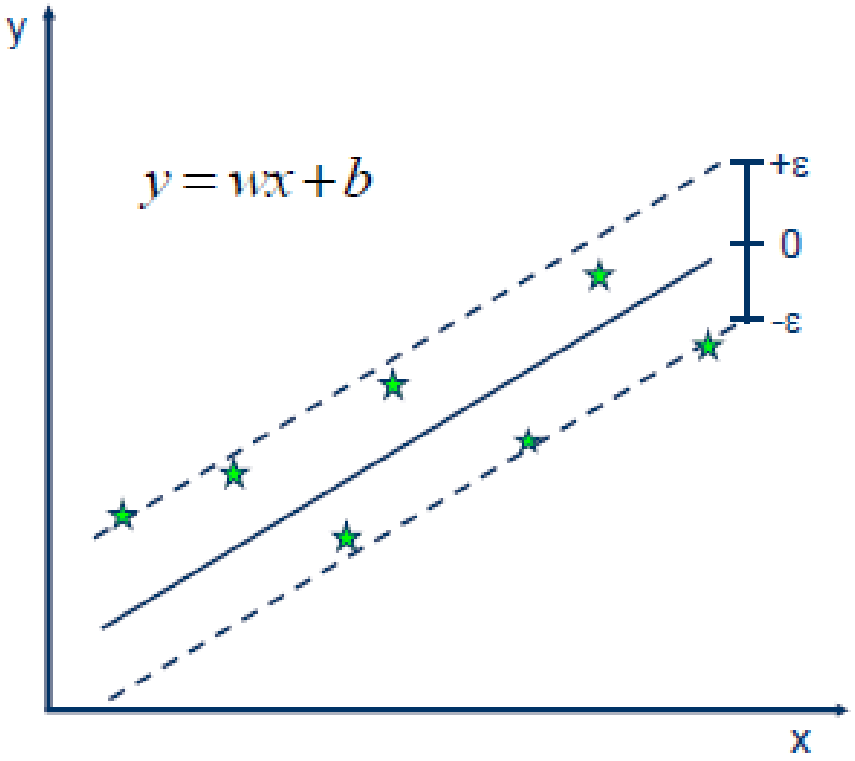


Figure 3.3: SVR example.[23]

3.2.4 Gradient Boosting Regression

To understand what Gradient boosting regression is you have first to understand what boosting is. Boosting is a method of converting weak learners into strong learners. This is what Gradient boosting regression is doing, it produces a prediction model in the form of an ensemble of weak prediction models (decision trees). DeepAi explain in more details how Gradient boosting works : "This technique builds a model in a stage-wise fashion and generalizes the model by allowing optimization of an arbitrary differentiable loss function. Gradient boosting basically combines weak learners into a single strong learner in an iterative fashion. As each

weak learner is added, a new model is fitted to provide a more accurate estimate of the response variable. The new weak learners are maximally correlated with the negative gradient of the loss function, associated with the whole ensemble. The idea of gradient boosting is that you can combine a group of relatively weak prediction models to build a stronger prediction model.[1]"In the same logic as SVR we will have parameters to handle such as the size of the tree, the learning speed, the number of steps, etc. We will talk about them in the Results section.

3.3 Noise

After outputting and after collecting the results of our prediction we have done nothing more than using algorithms and libraries. What we have to do is to play with our data set, use the notion of noise to see if our model will react well, if redundancy will appeared or not, and if our model will fall and predict really poorly, this is what we will look for.

3.3.1 What is noise ?

"Noise is a distortion in data, meaning that it makes random fluctuations on the data set, that is unwanted by the perceiver of data. Noise is anything that is spurious and extraneous to the original data, that is not intended to be present in the first place[3]" explains Amine Aoullay. Here in our case we will insert voluntarily noise in our data set to see how it reacts but we had to explain globally what noise was. It has been proved that by allowing some inaccuracy when training deep neural networks for example, not only the training performance but also the accuracy of the model can be improved if we add noise.

3.3.2 Generation of noise

Here our data are not supposed to be corrupted by noise but we want to inject some noise and see how our prediction will change. Obviously there is different ways to add noise but the most common one is the additive method but remember that noise can be anything. Let's suppose that a model is regulated by an equation

$$y = S.x$$

where y, x are Real numbers and S is a matrix of Real numbers.

As we said most of the time it is logical and practical that the noise is additive, meaning our model will give something like that :

$$y = S.x + r$$

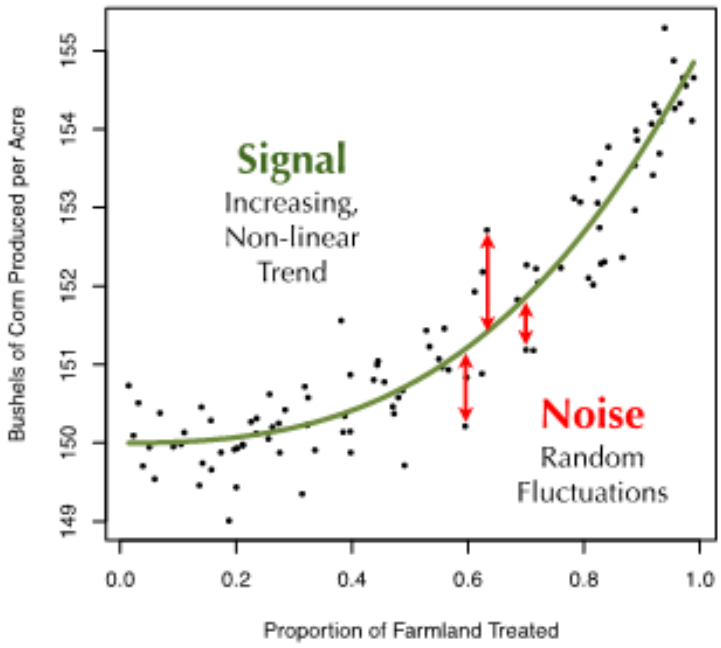


Figure 3.4: Example of noise causing deviation of data.[9]

where r is sampled from a random variable of any distribution. The most logical one would be a Gaussian distribution , making then a Gaussian noise following this : r $\sim \mathbb{N}(\mu,\sigma^2)$. To make it easier we will set up the mean to 0 and only keep the variance moving.

"Adding gaussian noise to inputs randomly is like telling our model to not change the output in a ball around our exact input. It adds robustness against overfitting to the model.[3]" argues Amine Aoullay. Important thing to know is that if we perturb our input by a very small amount, our intended output doesn't change in almost all cases. We will have to test gradually till how much amount of noise our model will lose a lot of accuracy. It will in fact depends of the robustness and redundancy of our model, because the more redundancy there is the more the noise will be insignificant until it become uncontrollable and we lost too much accuracy..

3.3.3 Missing Data

We can't talk about noise if we don't talk about missing data , both are closely linked. In fact missing data can be interpret like a zero value, meaning we have no information so we can imagine that a strong nice could deviate some data pretending it is a missing measurement. The important thing is that we can take advantage about those missing data to make our model robust, it is not always a negative aspect. However we can make our prediction of our model worst by removing those missing data because you reduce the amount of data you have available which can be dramatic. This is why we have to use them as much as we

we will use the noise, but how ?

Drop out

In our dataset we had no missing values but we decided to add some to see if our model is robust. There is a lot of different techniques to handle missing values, for instance Deductive Imputation, Mean Imputation, Regression Imputation but in our case we decided to us the Dropout method. As explained above missing values can be interpret like a zero value so we decided to modified value to 0 in a gradual way, i.e. 10% of the dataset, 30% and finally 50% of modified data to see how well our models resist. Note that we will use the same system for the adding noise first of all 10% then 30 and finally 50%. Doing it in this way will provide to us a better comprehension of the modification that noise and missing values are making.

Chapter 4

Implementation of our considered approach

It is now time in this section to apply in practice our theoretical concepts we saw above. Note that our the implementation has been done using Python 3.6 version.

4.1 Collection of data

In this thesis all the blood dataset comes from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) (**https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi**). We took all of these DNA methylation data from the HumanMethylation27 BeadChip platform (27k array) because this is where we had the most pertinent results. We tried and look for different and various dataset, excluding those with out age or with too much missing values. We finally arrived to the conclusion of keeping 3 big datasets (GSE20242 , GSE41037 and GSE19711) for a total of 1308 different patients of different races background and a diversifcation of age going from 16 to 91 years. To show the different performances of our models we decide to divided this number in a 7:3 ratio for respectively the training and testing part. You can find all the datasets in CSV format and the 3 python files for each different model in this github link. **https://github.com/Akiltour/thesis**

4.2 Features selection

With a total of more than 27 000 features (genes) we can't make a good and powerful prediction with this big number because we are not sure that every feature as a correlation with age, some features could and would have biased the prediction results because they are not pertinent so we decide to proceed a feature selection of the CPG sites (genes). Our regression models has to find the most optimal

coefficients for all the attributes and these are the results we get :

```
[-32.23243958  99.19302924 -30.46768113 -39.96048004  78.71815122
 -59.87534911]
```

Figure 4.1: Matrix of coefficient

Theses are the most age-correlated features we call them CPG sites and they are 6: cg22736354, cg06493994, cg02228185, cg09809672, cg19761273 and cg01820374. Those value means that, let's take the first value as example, for a unit increase in "age", there is a decrease of 31.51 units in the methylation level of the CPG site cg22736354. It has been proved by Horvath and other scientist that the combination of these six markers had the highest accuracy. As you can see some of them are positively correlated and other negatively.

Here is below a more detailed table of what are those CpG sites and where they can be found.

Age-related CPG sites		
CPG ID	Gene ID	Full name
cg01820374	CSNK1D	Casein kinase 1 delta
cg19761273	SCGN	secretagogin, EF-hand calcium binding protein
cg09809672	LAG3	Lymphocyte activating 3
cg0222818	ASPA	Aspartoacylase
cg06493994	NHLRC1	NHL repeat containing E3 ubiquitin protein ligase 1
cg22736354	EDARADD	EDAR associated death domain

4.3 Models implementation

We decide to implement and to try 3 models with the aim of comparing them and seeing which one performs best either without noise or with. Those 3 models as explained above in our considered approach are the following : Multiple Linear Regression, Support Vector Regression and finally the Gradient Boosting Regression.

4.3.1 Statistical Measurements of the performances

To analyze the performance of each model we will of course display the graphs representing the predictions, but to be sharper and more precise we will also use some statistical measurements. We will use the R^2 measurement, the Mean Absolute Error (MAE), the Mean Squared Error (MSE) and finally the Root Mean Squared Error (RMSE) to do so so it will be easier to us to make comparison.

Statistical Measurements	
Name	Signification
R^2	Degree of correlation between predicted age and true age
MAE	Measure of errors between paired observations
MSE	Average squared difference between the estimated values and the actual value
RMSE	Quadratic mean of the differences between estimated values and the actual value

4.3.2 Comparison of the 3 models with out adding noise

For **the multiple linear regression**, no parameters to take into consideration, we use the sklearn package from Python to implement this algorithm and for all others too and it gives us those results when we apply it to our dataset. (Figure 4.2). For **the support vector regression** all the parameters are those by default : kernel used = 'rbf', degree of the polynomial kernel function = 3, tolerance for stopping criterion = 0.001, C (Regularization parameter) = 1.0, epsilon (tube) = 0.1. Finally the **gradient boosting regression** parameters are the follow : loss (loss function to be optimized) = 'lad' (least absolute deviation), learning rate = 0.03, n-estimators (number of boosting stages to perform) = 300, all other parameters like the min-sample, lambda, the max-depth, the verbose and others are set by default.

The first thing we can see with our eyes is that the Multiple Linear Regression model is less precise and less accurate, many more points are far away from the right equation x=y. It seems that Gradient Boosting and Support Vector regressions are quite similar conspicuously. Remember that our goal is to see which model is the most robust when noise is added so the results we get here gives us some interesting information but it doesn't answer to our questions. Anyway we are going to display a little table with the statistical measurement of each model to make the difference sharper. (Table below) All the values were identified on the same CpG sites mentioned above. The results by the R^2 values showed that the prediction accuracy of the gradient boosting regressor and the support vector regressor are superior to the multiple linear regressor. However the 3 models are making good enough prediction but both of them stand out a little bit more from the pack.

Regarding the other statistical measurement we can deduce that the smaller the more accurate our model is. For instance, if we take the RMSE measurement we can see that his value is 7.2653 years which is more than 10% of the mean value of the patient's age (4.9 years). This is the lower value of the 3 models so we can interpret by saying that this model makes the fewest mistakes. Note that RMSE penalize large error because it this a root square so it can be really appropiate in our case unlike other measurements

Statistical measurements of each model				
Model	R^2	MAE	MSE	RMSE
MLR	0.8259	5.9114	62.2048	7.8870
SVR	0.8523	5.5970	52.7840	7.2653
GBR	0.8450	5.6383	55.3952	7.4428

4.4 Noised addition to our models

Now this is when we enter in our interesting phrase which consists ,for the first part, to add noise gradually to our dataset (10,30 and 50 percent) for each model to finally arrive to the conclusion of which one is the most robust and why so. The noise is randomly add following a gaussian with parameters mu = 0 and sigma = 0.1. The second part will be the same process but with missing values (drop out of 10, 30 and 50 percent).

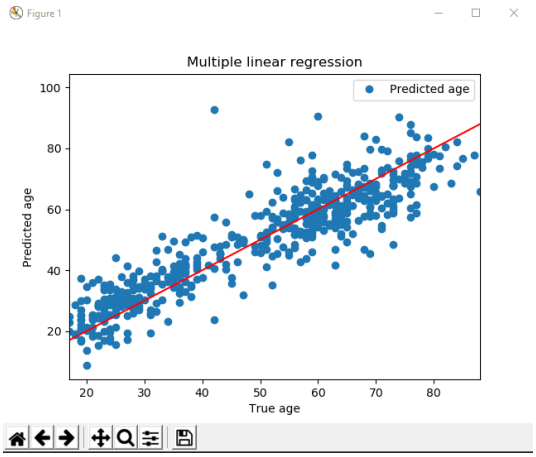
4.4.1 Multiple linear regression with noise

Let's get start with MLR , we have 3 graphics where noise is added in each but not with the same ratio. As we explained above we begin with 10% then 30 and finally 50. (Figure 4.8) We can notice that in general, the more noise we increase in our data, the more the values and predictions made deviate from the true values, which is logic . We add a disturbance to the data which implies a decrease in accuracy. We will see with the help of the statistical measurements that from 50% of noise added the model becomes much less robust but note that the model does not flow even if its efficiency decreases. The table follows with more detail. However we will discuss about all of the results at the end of the presentation of all models.

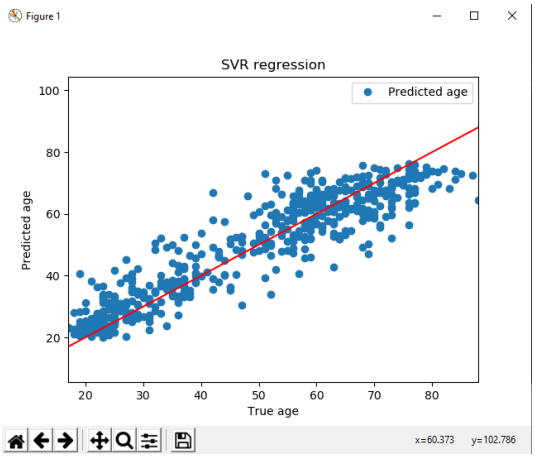
Statistical measurements of MLR with noise				
Model	R^2	MAE	MSE	RMSE
MLR-10	0.7788	6.8368	79.0388	8.8903
MLR-30	0.7021	8.1670	106.4158	10.3158
MLR-50	0.6431	9.0428	127.4986	11.2915

4.4.2 Support Vector Regression with noise

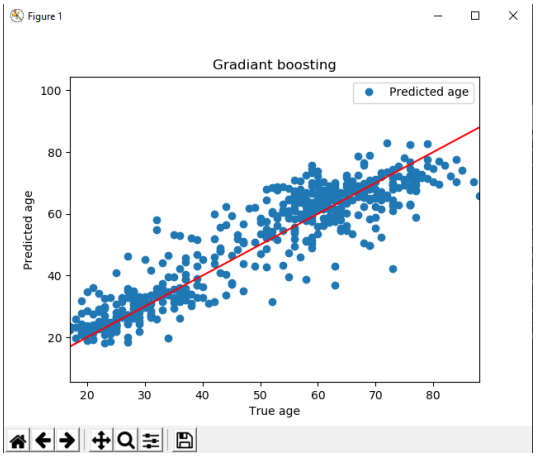
Same principle for the support vector regression model, note that we keep the same parameters. Here below the graphics and the table with the statistical measurements. (Figure 4.4)



(a) Multiple Linear Regression



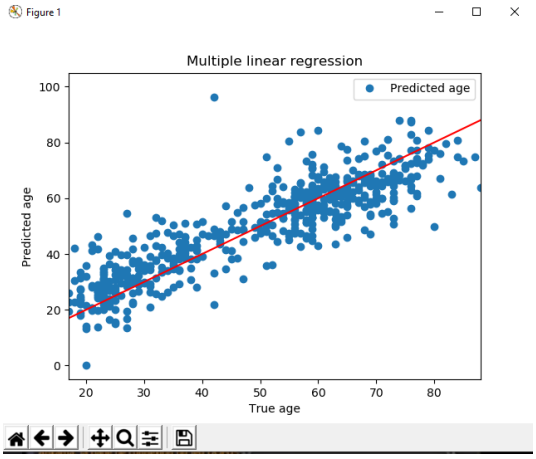
(b) Support Vector Regression



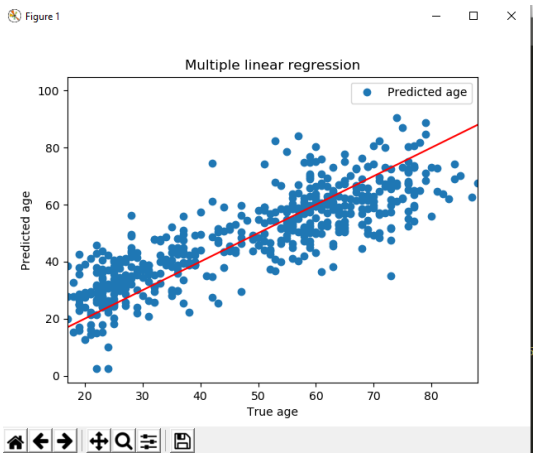
(c) Gradient Boosting Regression

Figure 4.2: Comparison between true age and predicted age predicted by the 3 models

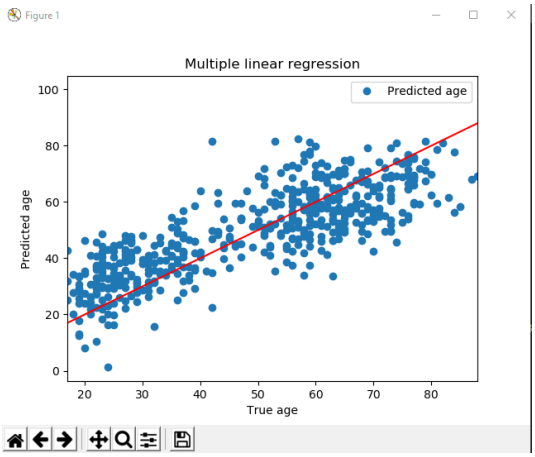
Statistical measurements of SVR with noise				
Model	R^2	MAE	MSE	RMSE
SVR-10	0.8058	6.4668	69.3843	8.3297
SVR-30	0.7386	7.5937	93.3911	9.6639
SVR-50	0.6813	8.5069	113.8609	10.6705



(a) MLR with 10% of noise



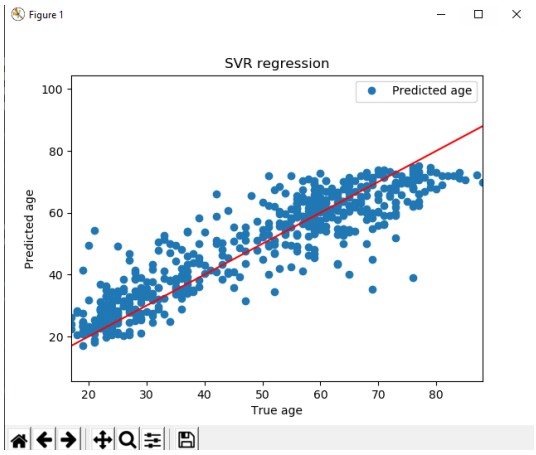
(b) MLR with 30% of noise



(c) MLR with 50% of noise

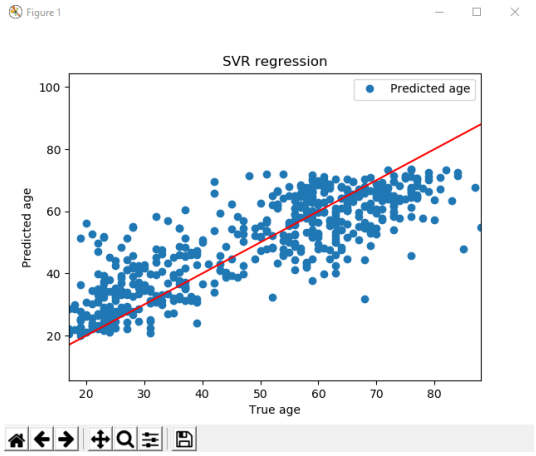
Figure 4.3: Comparison between true age and predicted age predicted by MLR with added noise

A quick glance allows us to see that we get better results and a better robustness with SVR rather than with MLR but be will discuss of the reason later.



(a) SVR with 10% of noise

(b) SVR with 30% of noise



(c) SVR with 50% of noise

Figure 4.4: Comparison between true age and predicted age predicted by SVR with added noise

4.4.3 Gradient Boosting Regression with noise

Finally the last model, the same parameters are kept for gradient boosting regression and we will apply the same process than the 2 others models. (Figure 4.5) We see that gradient boosting regression is less efficient than SVR but still better than MLR and we will see why in the next section.

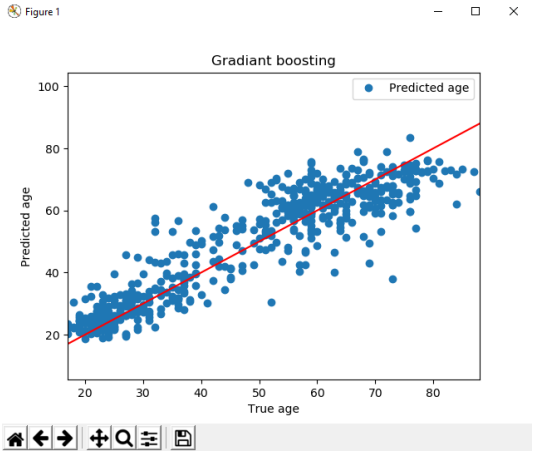
Statistical measurements of GBR with noise				
Model	R^2	MAE	MSE	RMSE
GBR-10	0.7826	6.6115	72.9306	9.0118
GBR-30	0.7144	7.8763	100.0439	10.9789
GBR-50	0.6657	8.9012	119.3784	11.5578

4.4.4 Discussion

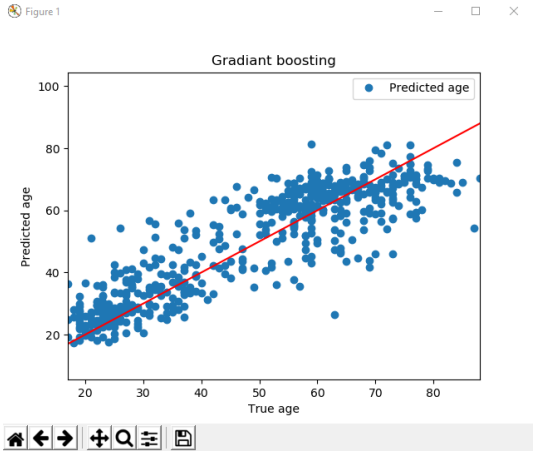
We did notice that even with noise integration the 2 models SVR and GBR performs better than the MLR and this is normal because the Multiple Linear Regression will analyze the relationships between several independent variables and one dependent variable but if the dataset is too complicated it will tend to make bad predictions and we see it in the statistical measurements, this model is the clearly weakest compared to the other 2. In its defense it is the simplest model to implement and understand but that's precisely what makes it a little too easy to solve the problem. The last 2 models are very similar with similar accuracies even if SVR is a little more precise.

The 2 methods are adequate and it should not be forgotten that in Machine Learning no algorithm is better than another in general, each one is just more adapted to the situation and the stated problem. In our case we only have 6 features to worry about and SVR starts to have problems when the number of features increases and gets closer to the number of samples and in the same way when the dataset is too large, which is not really our case so SVR is interesting for us. The 2 models are weakening and are overfitting when there is a lot of noise in the data which is logical we prove it clearly in the results.

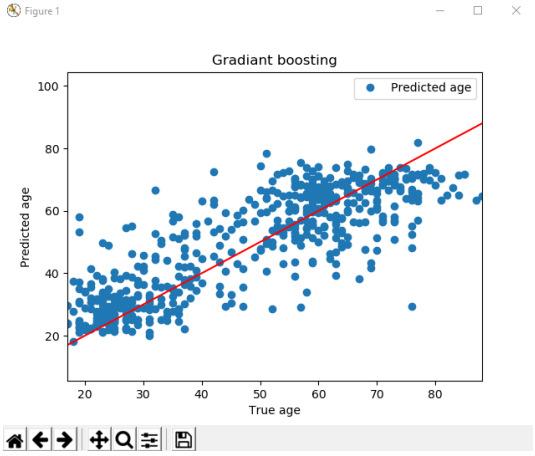
Finally, it is true that GBR is more precise and more rigorous when a problem is difficult because it uses precisely the boosting which allows it to facilitate the task and to be more precise. However with its tree process GBR takes much longer during the execution. Generally speaking, each of the two models has its positive and negative points but in our situation and with our dataset SVR performs a little bit better than GBR when we add some noise even if both models have an undeniable robustness.



(a) GBR with 10% of noise



(b) GBR with 30% of noise



(c) GBR with 50% of noise

Figure 4.5: Comparison between true age and predicted age predicted by GBR with added noise

4.5 Missing values added to our models (drop out)

The other interesting thing to analyse is how each model reacts to missing value, a lot of regression model has some ability to withstand missing values. We will modified our dataset to add in random way missing data in the same way as we did with noise. We will use the dropout method, meaning that values will be set to 0 randomly in out dataset (10% and 30% only) just to see which model is the most robust, our goal is not to manage THOSE missing values because we have a clean dataset without missing value so we just want to add some ourselves to see how our models react to it. All parameters and assumptions done previously are unchanged in the results below.

4.5.1 Multiple Linear Regression with missing values

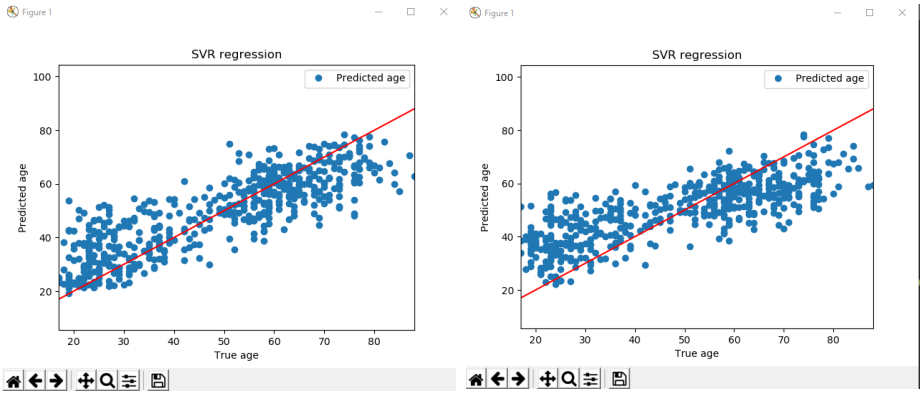
Degree of correlation between predicted age and true age with missing values	
Model	R^2
MLR-missing-10	0.5476
MLR-missing-30	0.3141

(a) MLR with 10% of missing values (b) MLR with 30% of missing values

Figure 4.6: Comparison between true age and predicted age predicted by MLR with missing values

4.5.2 Support Vector Regression with missing values

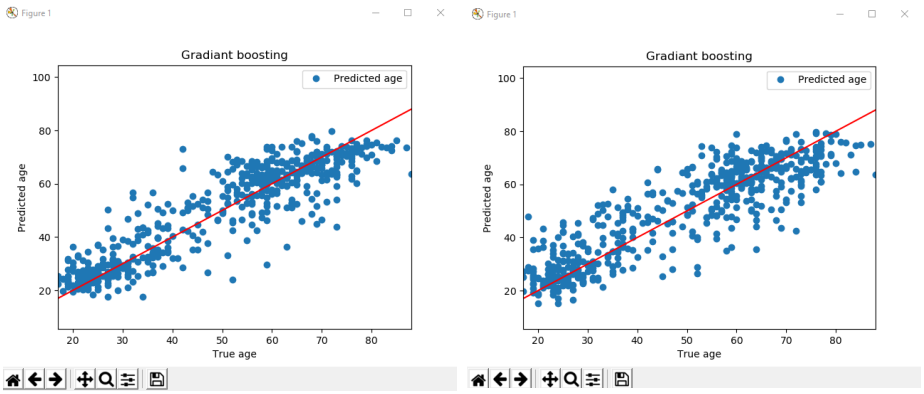
Degree of correlation between predicted age and true age with missing values	
Model	R^2
SVR-missing-10	0.7333
SVR-missing-30	0.5686



(a) SVR with 10% of missing values (b) SVR with 30% of missing values

Figure 4.7: Comparison between true age and predicted age predicted by SVR with missing values

4.5.3 Gradient Boosting Regression with missing values



(a) GBR with 10% of missing values (b) GBR with 30% of missing values

Figure 4.8: Comparison between true age and predicted age predicted by GBR with missing values

Degree of correlation between predicted age and true age with missing values	
Model	R^2
GBR-missing-10	0.8231
GBR-missing-30	0.7741

4.5.4 Discussion

The results are striking but following the logic established above with noise, the Multiple Linear Regression model fall completely apart after adding only 10% of missing value. You can see on the graph that the points deviates greatly from the point. Indeed no technique is used by the regressor to handle missing values, they will simply make the predictions worse because the model doesn't have the tools

to handle them, which makes MLR a very bad model to handle missing values. One of the solution would be to fill in the missing data manually which is a loss of time and a bad way to handle missing values where SVR and GBR does it automatically. As we can see the winner is GBR we can thank his technique of boosting and steps tree, indeed after each step and each tree created, if a missing value appears there will be an automatic update with the median of each sample, if the values are not replaced then effectively we will have disastrous results. The results proves it and the logic follows, GBR is a more complex model but it manages and allows us to withstand more problems such as missing values. As we have already repeated, there is no better model to follow, but an model adapted to different situations. MLR is good when we're in a simple problem without noise or missing value, when the complexity of the dataset increases we'll lean towards SVR and especially GBR which is much more flexible.

Chapter 5

Conclusion

To conclude this thesis and to answer again to our problematic which is the evaluation of our biomarkers resilience to noise and missing data it is good to remember that age prediction with the help of biomarkers and DNA methylation in computer science culture has come a long way and has had a meteoric rise, we can predict and analyze things through the human body that we couldn't do before. And here in our case by comparing different prediction models we came to the conclusion that with noise and missing data the most suitable model was the Gradient Boosting Regression. Thanks to DNA methylation-based and aging biomarkers, it allowed us to predict a person's age in a fairly robust and accurate way without weakening when faced with incomplete or erroneous data. And this is important because this is what can happen when we collect DNA-methylation on humans. Sometimes we don't have all the information but it is essential to be able to predict the age of the person with a good accuracy even if our database is not perfect. However, science is not always accurate, assumptions and results can be biased if we do not take everything into account. Indeed we could go much further in this thesis and in future research because this was only the visible surface. It is good to remember that methylation changes with age and can be drastically different in each of us, it would also be interesting to compare these predictions by separating gender or creating age ranges. One of the last things that can be done to go even further would be to collect DNA-methylation from somewhere other than the blood, such as saliva, skin tissue etc. In our specific case we stopped collecting DNA-methylation in the blood and without separating age and gender. Of course it is not a novelty what has been demonstrated in this paper but it has been done in our own way and on a small scale to finally come to a conclusion that other scientists have found but on a larger scale.

Bibliography

[1] Gradient boosting. **https://deepai.org/machine-learning-glossary-and-terms/gradient-boosting**, 2019.

[2] Cristiano Antonio Capurso, Gaetano Crepaldi. *Benefits of the Mediterranean Diet in the Elderly Patient*. Springer, 2018.

[3] Amine Aoullay. How to use noise to your advantage ? **https://towardsdatascience.com/how-to-use-noise-to-your-advantage-5301071d9dc3**, 2018.

[4] Rebecca Bevans. An introduction to multiple linear regression. **https://www.scribbr.com/statistics/multiple-linear-regression/**, 2020.

[5] Colicino E et al. Chen BH, Marioni RE. Dna methylation-based measures of biological age: meta-analysis predicting time to death. *Aging (Albany NY)*, 8(9):1844–1865, 2018.

[6] Yuri Deigin. What is a biomarker of aging and why do we need it? **https://medium.com/@yurideigin/epigenetic-clock-of-aging-709c1fe1e554**, 2019.

[7] Brenner CA Dupont C, Armant DR. Epigenetics: definition, mechanisms and clinical perspective. *Seminars in Reproductive Medicine.*, 27(5):351–357, 2009.

[8] J; Zhao L; Zhang L; Hughes G; Sadda S; Klotzle B; Bibikova M; Fan JB; Gao Y; Deconde R; Chen M; Rajapakse I; Friend S; Ideker T; Zhang K Hannum, G; Guinney. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell.*, 49(2):359–367, 2013.

[9] Joseph W. Richards Henrik Brink and Mark Fetherolf. *Real-World Machine Learning: Model Evaluation Optimization*. Hanning, 2016.

[10] Steve Horvath and Kenneth Raj. Dna methylation-based biomarkers and the epigenetic clock theory of ageing. *Nature reviews*, 19:371–384, 2018.

[11] Qiong Birkhofer Carina Gerstenmaier I Weidner, Carola Lin. Dna methylation in prdm8 is indicative for dyskeratosis congenita. *Oncotarget*, 7(10), 2016.

[12] Jake. Example: Regression with residuals. **http://pgfplots.net/tikz/examples/regression-residuals/**, 2020.

[13] Wim Vanden Berghe Ken Declerck. Back to the future: Epigenetic clock plasticity towards healthy aging. *Elsevier*, 2018.

[14] Labclinics. Role of dna methylation in disease. **https://www.labclinics.com/en/role-dna-methylation-disease/**, 2018.

[15] Lee S. D. Shin K.-J. Lee, H. Y. Forensic dna methylation profiling from evidence material for investigative leads. *BMB Rep.*, (49):359–369, 2016.

[16] Rahul Khanna Mariette Awad. *Efficient Learning Machines*. Apress Open, 2015.

[17] Ameet Talwalkar Mehryar Mohri, Afshin Rostamizadeh. *Foundations of Machine Learning*. The MIT Press ISBN 9780262018258, 2012.

[18] WHO International Programme on Chemical Safety Biomarkers in Risk Assessment. Validity and validation. **http://www.inchem.org/documents/ehc/ehc/ehc222.htm.**, 2001.

[19] Sebastian Raschka. *Python Machine Learning*. Packt Publishing Limited, 2015.

[20] Sunil Ray. Understanding support vector machine(svm) algorithm from examples. **https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/**, 2015.

[21] Sally Robertson. What is dna methylation? **https://www.news-medical.net/life-sciences/What-is-DNA-Methylation.aspx#**, 2018.

[22] Horvath S. Dna methylation age of human tissues and cell types. *Genome Biology.*, 14(10), 2013.

[23] Dr. Saed Sayad. Support vector machine - regression (svr). **https://www.saedsayad.com/support_vector_machine_reg.htm**, 2010-2020.

[24] Ronald van Loon. Machine learning explained: Understanding supervised, unsupervised, and reinforcement learning. **https://bigdata-madesimple.com/machine-learning-explained-understanding-supervised-unsupervised-and-reinforce** 2018.