metals

MDPI

Article

Data Driven Performance Prediction in Steel Making

Fernando Boto [1,*], Maialen Murua [1], Teresa Gutierrez [2], Sara Casado [2], Ana Carrillo [2] and Asier Arteaga [3]

[1] TECNALIA, Basque Research and Technology Alliance (BRTA), Mikeletegi Pasealekua 7, 20009 Donostia-San Sebastián, Spain; maialen.murua@tecnalia.com
[2] TECNALIA, Basque Research and Technology Alliance (BRTA), Astondo Bidea, Edificio 700, 48160 Derio, Spain; teresa.gutierrez@tecnalia.com (T.G.); sara.casado@tecnalia.com (S.C.); ana.carrillo@tecnalia.com (A.C.)
[3] Sidenor I+D, Barrio Ugarte s/n, 48970 Basauri, Spain; asier.arteaga@sidenor.com
* Correspondence: fernando.boto@tecnalia.com

Abstract: This work presents three data-driven models based on process data, to estimate different indicators related to process performance in a steel production process. The generated models allow the optimization of the process parameters to achieve optimal performance and quality levels. A new approach based on ensembles has been developed with feature selection methods and four state-of-the-art regression approximations (random forest, gradient boosting, xgboost and neural networks). The results show that the proposed approach makes the prediction more stable reducing the variance for all cases, even in one case, slightly reducing the bias. Furthermore, from the four machine learning paradigms presented, random forest is the one with the best results in a quantitative way, obtaining a coefficient of determination of 0.98 as a maximum, depending on the target sub-process.

Keywords: steel making; ensemble learning; feature selection; random forest; optimization

Citation: Boto, F.; Murua, M.; Gutierrez, T.; Casado, S.; Carrillo, A.; Arteaga, A. Data Driven Performance Prediction in Steel Making. *Metals* **2022**, *12*, 172. https://doi.org/10.3390/met12020172

Academic Editor: Chris Aldrich

Received: 30 November 2021
Accepted: 7 January 2022
Published: 18 January 2022

1. Introduction

In the industrial sector and in traditional industrial processes such as the steel industry, more and more elements of the chain are digitized. The availability of data is increasing but sometimes these data are difficult to structure and process, and it is also difficult to extract valuable information from them. In this context, efficiency in industrial processes is of great importance, especially when it is intended to reduce the ecological footprint while maintaining the production availability.

The quality of the final product is one of the most important indicators in any production process, but especially in iron and steel processes, where production conditions are very harsh and where there is hardly a theoretical framework and therefore disturbances dominate production.

Large amounts of data are available in these processes, but they are underused due to the difficulty in interpreting them, as well as the nature of the data corresponding to hostile, erratic and highly variable environments. Accordingly, data analysis tasks require a lot of expert knowledge in the domain to generate useful conclusions for business objectives. One of the biggest difficulties when working with data from an industrial process and in the steel industry in particular is uncertainty. We have to face a generalized uncertainty of different origins [1] due to the hostile and highly variable conditions.

Currently the production of high quality steel is supported by modern measuring systems gathering an increasing amount of high resolution quality and process data along the complete flat steel production chain [2].

The benefits of higher product quality, reduction of internal rejection and higher productivity will directly result in a reduction of the overall production costs. Moreover, it will

provide access to some very stringent product applications for which the quality requirements are very high. The competitiveness of steel companies will be significantly improved in productivity at the Finishing Shop and reduced raw materials and energy consumption.

But despite digitization, obtaining valuable conclusions through smart tools is not being easy task. Machine learning techniques have also burst into steel production as mentioned in the work by Laha et al. [3].

Although there have been several studies on the application of machine learning techniques, there are few works in the literature where these techniques are used especially in the prediction of the output performance of the steel manufacturing process.

There are some works on the use of neural networks to predict output parameters, such as the temperature of the liquid metal and the volume of oxygen blowing [4], metallurgical length in continuous casting (CC) where the steel solidifies, shell thickness at the end of the mold and the billet surface temperature [5], percentage of phosphorus in the final composition of the steel [6,7]. Mazumdar and Evans [8] provide a complete description of modern steelmaking processes together with physical and mathematical models and solution methodologies based on artificial intelligence. In the work presented in [9], prediction models based on production data are developed for casting surface defects. The authors show the importance of quality in the foundry industry by comparing 6 machine learning methods. Soft-sensing approaches are also proposed [10–12] to obtain a prediction of the content of silicon in the molten iron, hot metal temperature forecasting or slag amount prediction in EAF furnace. These works employ decomposition of the time series, neural networks, multivariate adaptive regression splines and ensemble learning approaches. The optimization of process parameters as a prescriptive strategy carried out in [13], is one of the most demanding tasks in the steel industry today.In this case using a surrogated model based on neural networks, within a multi-objective problem. Multi-output support vector regression and particle swarm optimization were used in [14] to optimize de process parameters in steel production.

This work belongs to a project framework in which activities towards a better product quality, productivity improvement and cost reduction of the steel making process have been developed. Strategies have been built via the modelling and control issues related to secondary metallurgy (SM), CC and hot rolling (HR) with the aim of optimizing these sub-processes. These models are the basis of an optimization methodology, in which the use of data-driven models, and therefore the use of machine learning techniques is a key point. This publication only presents the work developed with the data-driven models within this methodology.

The aim of this work is to develop process models to improve the steelmaking process and reduce the number of surface and sub-surface defects at the final product for micro-alloyed steels, ensuring a good performance of the three sub-processes that have an influence in the generation of surface defects: SM, CC and HR. These models are the basis of an optimization strategy for each sub-process, based on the relationship established between the control parameters and performance indicators.

Based on the growing demand of the microalloyed steels and the critical points for their correct manufacturing, the following data based models were developed and are presented in the following sections:

- SM model to predict the castability index of a heat. This index is a measure of the performance of the refining process, mainly influenced by the formation of solid micro-inclusions into liquid metal [15], that affects to the steel cleanliness.
- CC model to predict the temperature at the middle point of the upper face of the billet before the straightener during the continuous casting process.
- HR model to predict the minimum and average temperatures of the billet before the continuous rolling mill.

The description of the proposed approach goes from the data processing to the generation of the models through the analysis of the most relevant parameters. An important part of this approach is the comparison of feature selection strategies that were applied,

as well as the comparison of different regression paradigms. Within this strategy, it is worth highlighting the use of ensemble learning and a novel strategy based on different feature subspaces is presented. It combines four commonly employed feature selection techniques.

The article is structured as follows. Section 2 describes the key performance indicators in steel production process, and machine learning methodology employed. In Section 3, the data used is explained, the preprocessing results are provided, the performance metrics of the models, and the results are discussed in turn. Finally, in Section 4, some conclusions are drawn.

2. Materials and Methods

2.1. Steel Production and Quality Measurements

The steel can be produced via two routes: the integrated blast furnace (BF)/basic oxygen furnace (BOF) route from the primary raw material iron oxide and the electric arc furnace (EAF) route by using recycled steel (scrap). Figure 1 shows steel production from EAF furnace. Both routes have in common the secondary metallurgy and casting process. Castability index is a measurement developed from process data of the casting machine process. It is a number ranging from 0 to 10. The main concept of the index is that the steel flux in the casting machine nozzles is smooth if the heat is clean of oxide inclusions, and more peaky and has a different form if the heat is not clean and has many oxide inclusions. The calculation of the index is based in the time series data of the casting machine 6 stopper rods position. The index checks the slope and the peaks of those data and gets and overall "castability goodness value". So it is obtained by measured casting process data but, in practice, is measuring an effect accumulated from the secondary process. This approach has the disadvantage of indirect measurement but the advantage of having many more data, as classical cleanliness data are much more difficult to obtain.

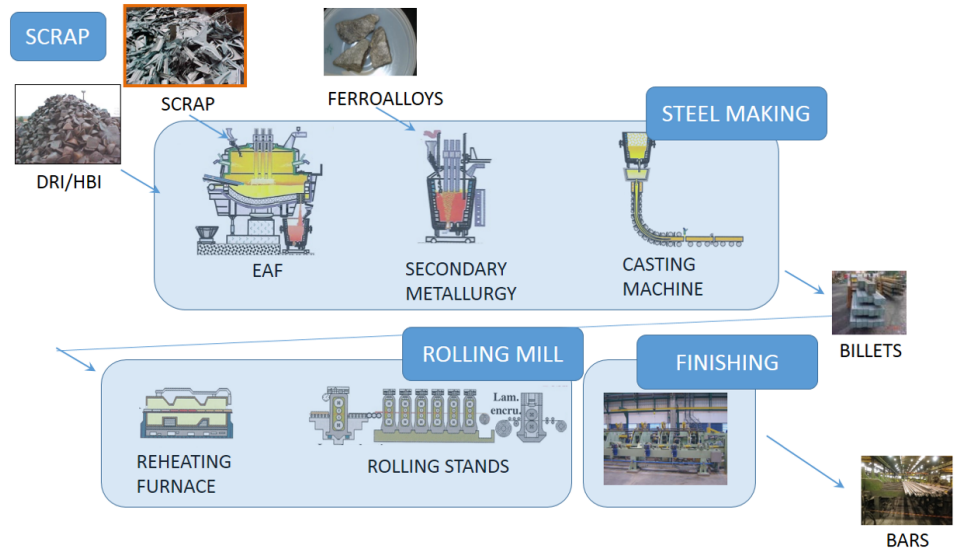

Figure 1. Steel making process description. From electric arc furnace to hot rolling and finishing (inspection) through secondary metallurgy and continuous casting.

Once temperature conditions and composition in the liquid steel are reached, the ladle is raised in a turret for casting liquid steel into a tundish with several strands. The CC is the process whereby molten steel is solidified into a "semi-finished" billet, by means of a watercooled open mold, secondary cooling combining water sprays and air, electromagnetic stirring, straightening and cutting the semi-product to the desired size.

After solidifying, the billets remain in any of the previous cooling beds transfer to the rolling mill. The billets are heated and hot rolled through several pairs of rolls to reduce the thickness and to obtain the desired shape and thickness of the bars. The steel quality is usually improved when it endures mechanical stresses in hot temperature.

Finally, the bars are inspected by quality control devices, such as Eddy Current technology based devices, to detect surface cracks. After this automated crack detection, there is a manual inspection of every crack-detected bar checking if the crack can be repaired or not.

Microalloyed steel grades are a family of steel composition characterized by the effect of alloying small amounts of elements like V, Nb, Ti, Al combined with C and/or N. Those elements exert an important effect on the steel properties by controlling the grain size evolution during different stages of the production process. In this way, it is possible to obtain reasonably good properties and avoid additional heat treatments. Heat treatments add additional costs to the product, additional energy use and emission generation. For these reasons, microalloyed steels are a growing trend in special steel industry for the global benefit they bring to different agents in the whole supply chain, but at the same time, they are quite demanding for the steel producer as one of their intrinsic properties is a low ductility at some critical temperatures. The low ductility makes them very prone to superficial cracks as the billet surface suffers the most important strains in different processes but mainly in the casting machine and HR. For this reason, it is important to control the temperatures before the straightener during the CC and before the continuous rolling mill during the HR process to avoid the cracks generation during these processes [16,17].

2.2. Ensemble Based on Feature Selection Approach

The most important aspect to take into account in this methodology is the ensemble learning strategy. Not only regarding the regression models used, but also the development of a novel approach, which will be explained in this section.

A supervised analysis starts with a database in which the predictor variables $x = \{x_j\}$ (where $1 \leq j \leq M$), M being the number of input variables and the dependent variable y are well defined for each use case or sub-process. So, we have for each regression problem a set of data defined as $\{X, y\}$.

Regression methods aim to predict a numerical value of a target variable given some input variables. To solve this task, the learning algorithm is asked to build a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$. In the metallurgical cases present in this paper different regression techniques have been used to predict a real value which is a performance measure of the process (such as the castability index, for secondary metallurgy). The selected regression modeling approaches are largely based on ensemble learning and a neural network has been included, as it is the most widely used strategy in the literature.

Ensemble methods train multiple learners to solve the same problem [18] with the ability to build a much stronger model than the base learners. One of the most important aspects when building an ensemble approach is diversity as well as reliability. One of the alternatives to provide this diversity is through generative methods, where the structure of the data set is modified [19]. Strategies based on this principle, as random forest and gradient boosting, are based on generative strategy where the data set is resampled randomly. In the approach presented in this work a generative point of view is also considered but with different non-random feature selection methods.

Figure 2 shows a common ensemble architecture. An ensemble contains L number of learners called base learners which are usually generated from training data by a base learning algorithm. A combination strategy has the responsibility to provide a unique decision.

By reducing the number of input features of the models, the effect of the curse of dimensionality problem that characterizes data with high number of features, can be alleviated. For example, by applying feature selection algorithms to construct subsets of features from the original data, diverse sets of models can be constructed. This is the option presented here, where the diversity is provided by a non random selection of features by applying different feature selection methods.

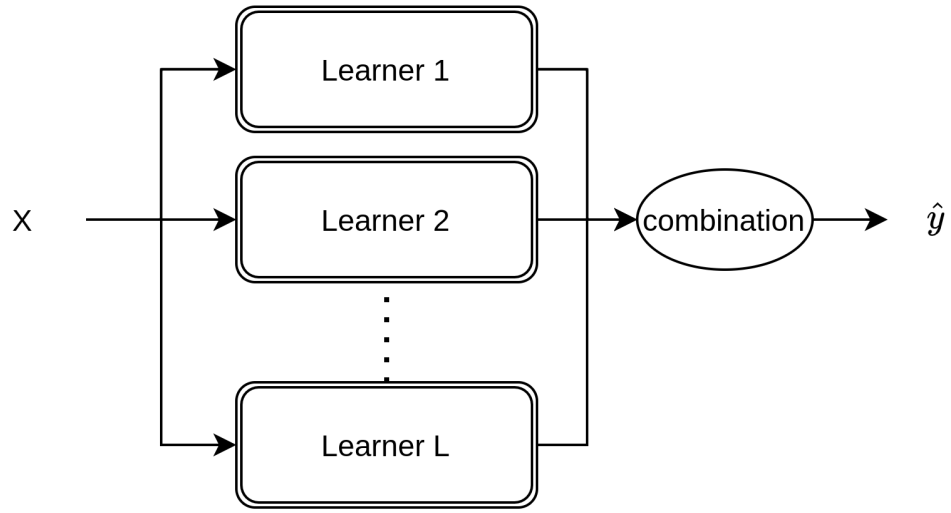

Figure 2. Ensemble strategy to combine multiple learners. Where X are the input features, L is the number of learners and $\hat{y}$ is the estimated output.

The approach proposed, called as meta-estimator, is defined as follows. Given a set of features defined as $X = \{x_i\}$, where $1 \leq i \leq n$ and n is the number of features, and given $S = \{S_l\}$, where $1 \leq l \leq L$ and L is the number feature selection methods.

Each S_l establishes a ranking and sort the set of features X creating L different sets X_l, which defines a different dimensional space to fit a regression model $f(X_l) \rightarrow \hat{y}_l$.

Having the prediction of each model $\hat{y}_l$, the final output is calculated by consensus as defined in Equation (1):

$$\hat{y} = \frac{\sum_{l=1}^{L} f(X_l)}{L} \tag{1}$$

Within this approach, three feature selection methods and four regressive paradigms have been selected. These methods are briefly explained below in the following Sections 2.3 and 2.4.

2.3. Feature Selection

The focus of feature selection is to select a subset of variables from the initial set of input variables which can efficiently describe it while reducing effects from noise or irrelevant variables and still provide good prediction results [20]. In this section, commonly used methods are revised and proposed to compare them with a real problem in the steel production where the number of the initial set of input variables is very high.

Feature selection will be carried out using filtering methods. Filtering methods use variable ranking techniques as principle criteria for variable selection by ordering them. Ranking methods are used due to their simplicity and good performance which is reported for practical applications. A suitable ranking criterion is used to score the variables and a λ threshold is used that represents the percentage of variables that are filtered with each indicator. Ranking methods are filter methods since they are applied before classification to filter out the less relevant variables [21]. In this investigation, four ranking criteria are used: (1) Pearson's correlation; (2) mutual information (MI); (3) univariate linear regression; (4) recursive feature elimination. The employed λ was selected for each method independently to keep the 30% of the variables.

2.3.1. Correlation between Inputs and Target

Correlation between two variables is the most common criteria to filter a feature set. Pearson's correlation coefficient is probably the most widely used and simple measure for linear relationships between two normal distributed variables. Usually, Pearson's coefficient is obtained via a Least-Squares fit and a value of 1 represents a perfect positive

relationship, −1 a perfect negative relationship, and 0 indicates the absence of a relationship between variables. It is shown in Equation (2), where X and Y are two random variables, cov refers to the covariance measure and σ_x and σ_y are the standard deviations of the variables, respectively.

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} \tag{2}$$

Correlation ranking can only detect linear dependencies between input variables and the target.

2.3.2. Mutual Information

Mutual information between two random variables is a non-negative value, which measures the dependency between the variables. It is equal to zero if and only if two random variables are independent, and higher values mean higher dependency.

Usually the entropy is a key indicator to calculate the MI, in this method Nearest Neighbor estimator is used as in [22].

For two random variables, X and Y, let $P_{XY}(x, y)$ be the joint probability distribution and $P_X(x)$ and $P_Y(y)$ the marginal probability distributions. The MI between X and Y, denoted $I(X; Y)$, is defined as shown in Equation (3):

$$I(X; Y) = \sum_{x,y} P_{XY}(x, y) \log \frac{P_{XY}(x, y)}{P_X(x) P_Y(y)} \tag{3}$$

In information theory, MI is the amount of uncertainty in X due to the knowledge of Y [23].

2.3.3. Univariate Linear Regression

Univariate linear regression test ranks also the features depending on their correlation with the dependent variable. Being n the number of observations, the test is defined as shown in Equation (4):

$$\rho = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{std(x) std(y)} \tag{4}$$

where $\overline{x}$ and $\overline{y}$ are the mean values of X and Y and std refers to the standard deviation.

2.3.4. Recursive Feature Elimination

This method is a greedy optimization algorithm which aims to find the best performing feature subset. It repeatedly creates models and keeps aside the best or the worst performing feature at each iteration. It constructs the next model with the left features until all the features are exhausted. It then ranks the features based on the order of their elimination. This is an iterative procedure:

1. Train the classifier
2. Compute the ranking criterion for all features
3. Remove the feature with smallest ranking criterion

This iterative procedure is an instance of backward feature elimination [24] and references therein. For computational reasons, it may be more efficient to remove several features at a time, at the expense of possible classification performance degradation. In such a case, the method produces a feature subset ranking, as opposed to a feature ranking.

2.4. Regression Strategies

2.4.1. Boosting

The main idea of the boosting method is to add models to the set sequentially. In each iteration a "weak" model (base-learner) is trained with respect to the total error of the set generated up to the moment. To overcome the over-training problem of this kind of

algorithms, a statistical connection is established with the gradient-descent formulation to find local minimums. In Gradient Boosting models (GBM), the learning procedure is consecutively adjusted to new models to provide a more accurate estimate of the response variable. The main idea behind this algorithm is to build the new base models so that they correlate as much as possible with the negative gradient of the loss function, associated with the whole set. The loss functions applied can be arbitrary, but to give a better intuition, if the error function is the classic quadratic error, the learning procedure would result in an error adjustment. In general, the choice of the loss function depends on the problem, with a great variety of loss functions implemented up to now and with the possibility of implementing a specific one for the task. Normally, the choice of one of them is a consequence of trial and error.

A computationally flexible way of capturing the interactions between variables in a GBM is based on the use of classification trees. These models usually have the tree depth and divisions as parameters.

XGBoost or extreme gradient boosting is a scalable machine learning system for tree boosting [25]. Specifically, XGBoost uses a more regularized model formalization to control over-fitting, which users of this implementation say gives it better performance. Also provides an early stoping mechanism, parallel computing capabilities, custom loss functions, re-training and different base learners as linear regression.

2.4.2. Random Forest

Random forest is an algorithm developed by Breiman and Cutler [26] based on the idea of bagging ensemble technique by Breiman [27] himself and the random selection of attributes, introduced independently by Ho [28].

Random forest builds multiple decision trees and merges them together, usually with an average decision rule, to get a more accurate and stable prediction. One big advantage of random forest is, that it can be used for both classification and regression problems.

Thus, as Breiman defined [26], a random forest is a classifier consisting of a collection of tree-structured classifiers $h(x, \theta_k), k = 1, \ldots T$ where the $\{\theta_k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x and T is the number of trees.

Random forests for regression are formed by growing trees depending on a random vector θ such that the tree predictor $h(x, \theta)$ takes on numerical values as opposed to class labels. In this case, a random selection of features is used on top of bagging.

2.4.3. Neural Networks

Artificial neural networks (ANNs), also referred as neural networks (NNs), are basically a certain class of computational techniques for performing non-linear statistical modeling of data. Therefore, they can be used to model complex relationships between inputs and outputs, or to find patterns, including temporal patterns, in large amounts of data [29].

Nowadays, NNs are used in a large number of applications, but in all cases their composition is basically a certain kind of network of simple interconnected process elements so that they can generate a complex global behavior, this global behavior being determined by the connections between the elements of process (what is often referred to as the "weight of the interconnection"). The functioning of the ANNs is similar to that of the biological systems formed by the neurons (from which it takes its name), in such a way that the functions are developed collectively and in parallel through all the elements.

A NN consists of simple processing units, the neurons, and directed, weighted connections between those neurons. Here, the strength of a connection (or the connecting weight) between two neurons i and j is referred to as $w_{i,j}$. Thus, it is a sorted triple (NS, V, w) with two sets NS, V and a function w, where NS is the set of neurons and V a set $\{(i, j)|i, j \in \mathbb{N}\}$ whose elements are called connections between neuron i and neuron j. The function

$w : V \rightarrow \mathbb{R}$ defines the weights, where $w((i,j))$, the weight of the connection between neuron i and neuron j, is shortened to $w_{i,j}$.

3. Experimental Results and Discussion

3.1. Data Description and Pre-Processing

This section introduces the data based models implemented for each of the sub-processes mentioned above (secondary metallurgy, continuous casting and hot rolling), describing the most relevant aspects of the input and output variables of the models.

The variables involved in the models, are mainly of two types. First of all we have to take into account synchronous, cyclic or high frequency data, related to time series. Secondly, we refer to asynchronous data, acyclic or data related to specific events in the process, such as a casting event or the production batch. The different characteristics of these two types of variables require differentiated treatment and different data preprocessesing.

3.1.1. Secondary Metallurgy

The objective of the SM model is to provide a regression model that, given the process operating variables, which are the most important in the proper functioning of the process, provide a prediction of the castability index. As mentioned in the previous section, this is a critical parameter regarding the steel quality that is measured the day after performing the heat. In the secondary metallurgy sub-process data from the heats performed during 9 months were taken into account. Cyclic, acyclic, steel grade chemical compositions and ferroalloys additions are input variables for this model, as described in Table 1.

Table 1. Secondary metallurgy variables. Variable type as a taxonomy within the process and a short description in the case of Cyclic and Acyclic data.

Variable Type	Description
Cyclic data	Argon consumption Argon vacuum consumption Nitrogen consumption Nitrogen vacuum consumption Initial and end temperature difference Temperature at the end of the process
Acyclic data	Total vacuum time Deep vacuum time Order inside sequence Casting temperature Temperature at the end of the refine process Wating time between refining and continuous casting Cap number of uses Warm-up time Total time
Steel grade Composition (at the initial and final stage of the process)	Carbon, Silicon, Manganese, Phosphor, Sulfur, Chromium, Nickel, Molybdenum, Copper, Aluminium, Nitrogen, Boron
Ferroalloys, deoxidants and slag formers	Steel lime, Pure alumina 100% + fluorspar, Manganese iron, Cok low nitrogen, Ferrosilicon, Silicium carbide, Silicon manganese, Charge Chrome, Ferrovanadium, Sulfur in thread, Niobium, Ingot Aluminum, Aluminum in wire, Ferro titanium wire, SI-CA 45 thread

Different pre-processing activities of the original data were carried out in order to refine the data, reduce dimensionality and improve prediction. One of the main tasks here was to correct certain inconsistencies and characterize the cyclical data. For example, the measurements of the flow of argon and nitrogen presented certain irregularities, as a consequence of the measuring system limitations and intrinsic process conditions, that

were corrected based on expert knowledge by using other variables measured by the equipment. The correction was based on the vacuum pressure and the back pressure as follows (Equation (5)):

$$\hat{f} = max(min((bp - c_1 + 1 - vp/100)/c_2, f), 0) \quad (5)$$

where f is the initial gas flow, $\hat{f}$ is the gas flow corrected, bp the back pressure, c_i are constants and vp vacuum pressure.

Figure 3 shows the effect of the correction on the gas flow with respect to the initial sequence.

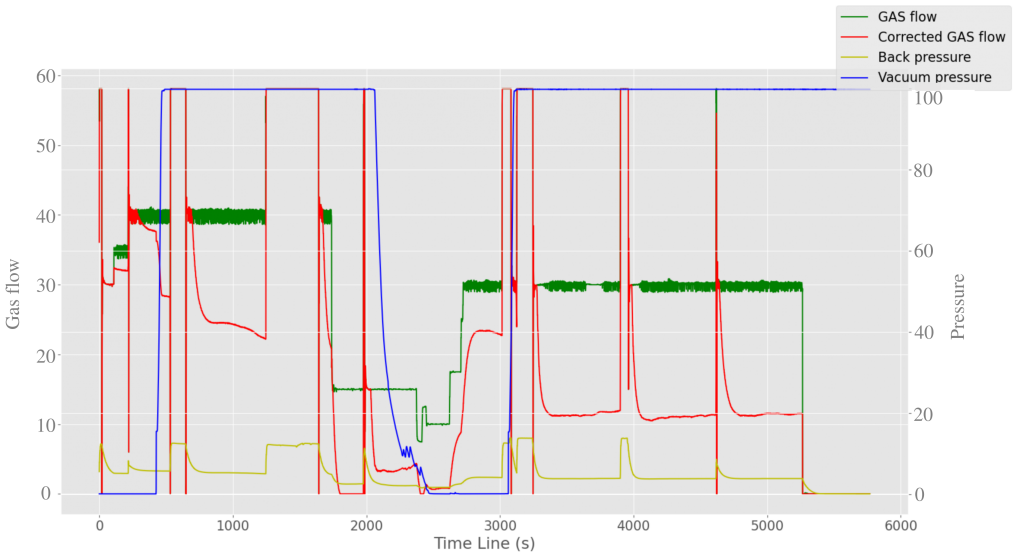

Figure 3. Gas flow correction, original gas flow (green) and corrected gas flow (red). Back (yellow) and Vacuum (blue) pressures are provided for a better understanding. Units are not displayed and scale is not real for confidentiality reasons.

After this correction, the extracted characteristic is the total consumption of argon and nitrogen (m^3/s) for the entire heat. In addition, the total consumption of each type of gas during the vacuum stage of the process was extracted as another key feature of secondary metallurgy.

The last two variables extracted from the cyclical data are those related to the process temperatures, which are trivially obtained from consulting the temperature curve throughout the process.

Acyclic data, steel grade composition and additions are extracted from the Manufacturing execution system (MES) for each heat. This information is important to successfully complete the data integration. In addition, a new variable, not existing in the initial data, was created with the alumina and the fluorspar to capture an important aspect of the ferroalloys.

In short, with these transformations and preprocessing stage, we got data to generate a predictive model to predict the castability index of the secondary metallurgy process, with a total of 54 process variables, including both process variables and the composition of the steel grade at the initial and the final phase of the process.

Then, following the general methodology for developing data-driven models, a study of the most relevant variables in this sub-process was made. The objectives of this task are mainly two: to reduce the number of variables in the model, to act on the precision of the model but above all its generalization, and secondly, to select the important variables of the process for a future optimization strategy.

To carry out the selection of features, the methods shown in Section 2.3 were experimented with. Each of the strategies was evaluated with the preprocessed data and in all

cases with a supervised point of view with the index of castability. Figure 4 shows the importance of most of the features with each selection method.

As it can be seen, there are many differences in the results obtained for each of the methods. This may be a consequence of the low absolute contribution of each variable in the process. The only variable that has a clear contribution is the sequence number, which is the order within a casting sequence. This conclusion is confirmed with expert knowledge demonstrating that it has a clear relationship with the operational process.

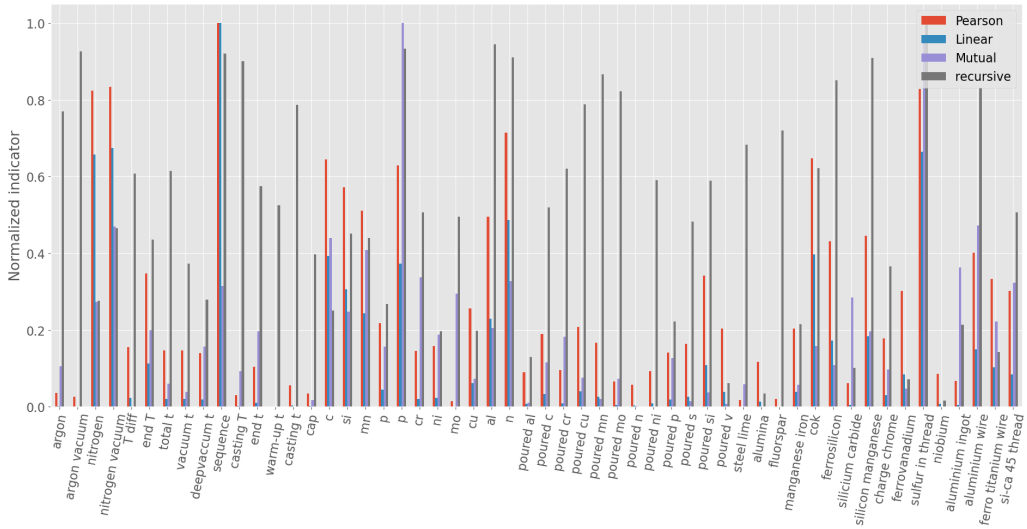

Figure 4. Feature selection normalized importance indicators per each variable with the performance variable in secondary metallurgy process.

3.1.2. Continuous Casting

The objective of the continuous casting model is to predict the temperature at the middle point of the upper face of the billet before the straightener during the continuous casting process, taking into account the process operating variables. As mentioned before, the control of this temperature is important to avoid the generation of cracks during the continuous casting process.

Since the value of this temperature was not measured in real time by any hardware sensor, it was modeled by using the information provided by a thermal model. This thermal model predicts the evolution of the temperature distribution during the solidification process of the billet in the continuous casting machine (see Figure 5). It is a mathematical model solving the heat transfer equations. Several cross sections of the billet, distributed along the strand, moves through the continuous caster. The model considers the exchanging heat with the mold wall and afterward with the secondary cooling system, rolls and ambient. The differential equation that governs the heat transfer and is solved by the model is expressed as Equation (6):

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{Q}{k(T)} = \frac{\rho(T) \cdot Cp(T)}{k(T)} \frac{\partial T}{\partial t} \tag{6}$$

where ρ, Cp and k and are the density, specific heat and thermal conductivity of the steel grade depending on the temperature, T is the Temperature and Q is the transferred heat from the exterior. The thermophysical properties of the steel grade, such as conductivity, specific heat, latent heat, density, liquidus and solidus temperature are calculated by a commercial software, depending on steel composition. The model was used to predict the temperature at the middle point of the upper face of the billet before the straightener depending on the process parameters. The model was validated with temperature measurements of the billet before the straightener that were performed with a scanner located on top of the billet. The results of this validation showed that the temperatures calculated

by the model were in a good agreement with the measured temperatures, with an average difference in the temperatures of about 11 ºC (considering that the temperature in this zone is around 1100 ºC, it means an error about 1%).

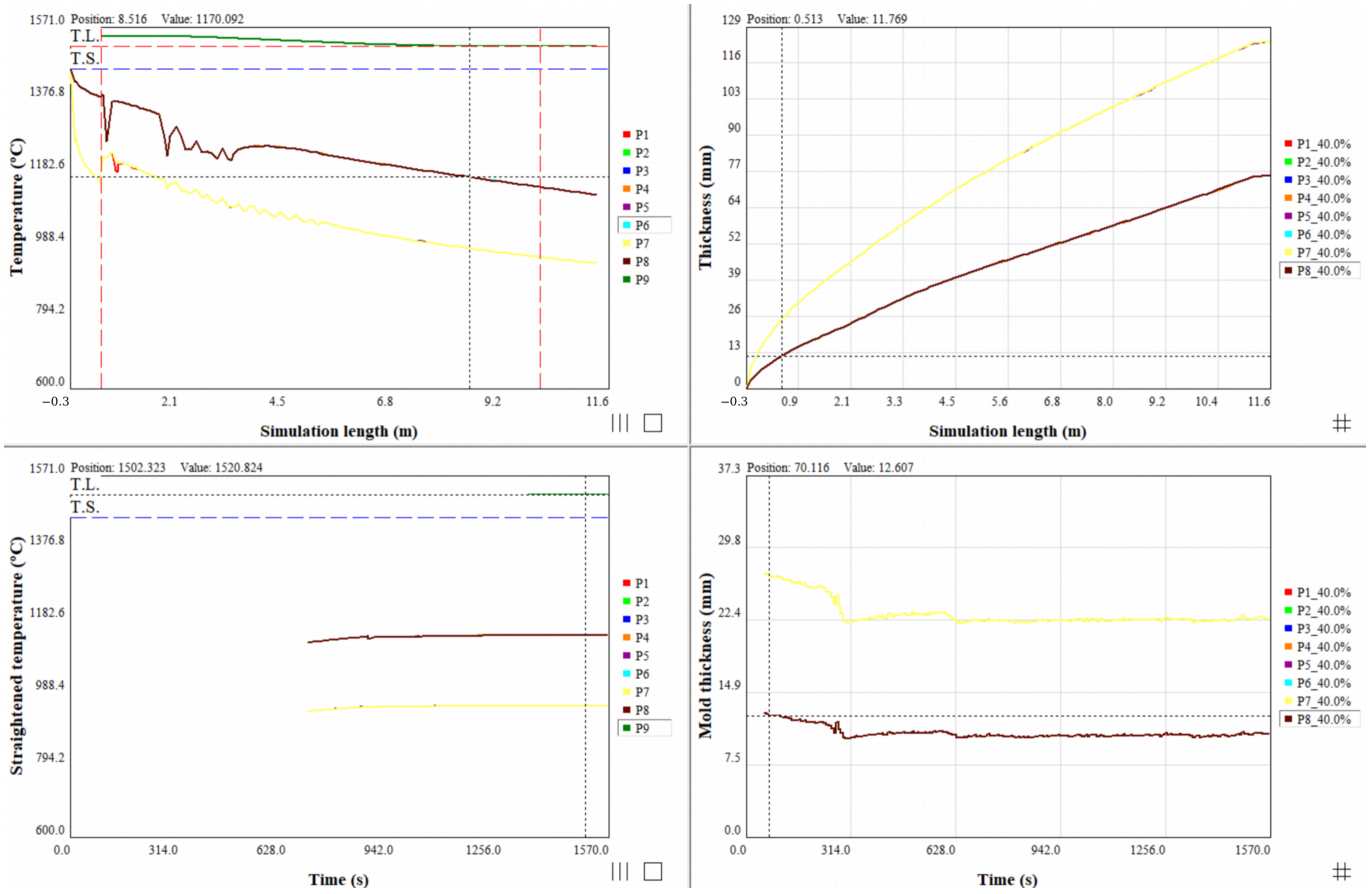

Figure 5. Continuous Casting thermal model output. On the top: evolution of the temperatures (**left**) and thickness (**right**) along the strand in corners, central points of the faces and center of the billet. At the bottom: evolution of the temperatures before the straightener (**left**) and the thickness at the end of the mould (**right**).

Once the model was validated, the process parameters of the heats performed during 9 months were analysed to identify process parameters sets corresponding to steady conditions. As a result, almost 3000 parameters sets (data set) were obtained, covering the usual range of the process window for the different parameters. Then, the thermal model was executed for each element of the data set in order to get the value of the temperature at the middle point of the upper face of the billet before the straightener. Finally, these temperatures were used for the development of the data driven model. The reason for implementing a data-driven model, instead of using the thermal model, lies in the differences in computation times required by both types of models (the data model is much faster than the thermal model), a key aspect for an efficient integration with optimization algorithms.

The Table 2 shows the final input variables for the regression model implemented to predict the temperature at the middle point of the upper face of the billet before the straightener. It includes both process parameters (mainly parameters related to the primary and secondary cooling) and the composition of the steel grade.

Table 2. Continuous casting variables with a short description for specific variables and compositions.

Variable Type	Description
Specific variables	Lance temperature Mold input temperature Casting speed Steel level (mold) Mold flow rate Flow rate in the zones of the secondary cooling Preassure in the zones of the secondary cooling zones
Compositions	Carbon, Silicon, Manganese, Phosphor, Sulfur, Chromium, Nickel, Molybdenum, Copper, Aluminium, Nitrogen, Niobium, Titanium, Vanadium Boron

3.1.3. Reheating Furnace and Hot Rolling

The objective of the hot rolling process modelling was to predict the temperature of the billet before the rolling mill taking into account the process parameters. As mentioned before, the control of this temperature is important to avoid the generation of cracks during the hot rolling As this temperature is acquired continuously, the average and minimum temperatures of the billet were extracted to correctly characterize the process and they were taken as response variables or objective of the two models to be developed. In this process, two models have been developed to predict the temperature before the rolling mill. As this temperature is acquired continuously, the average and minimum temperature have been extracted to correctly characterize the process. These two variables are very relevant for the correct rolling of the billet, so both are taken as response variables of the models to be developed.

Data integration played a relevant role in generating the necessary data for the hot rolling modeling. The data source for the process parameters was compressed files with XML format.

These files store the whole processing information of each billet hot rolled.

The information that was extracted from these files is detailed in the Table 3 and can be structured in the following types: general data, dimensions, speeds, times, furnace times/temperatures and other temperatures.

In addition, it was necessary to integrate other important data for the model to be developed, accessible from other data sources. Specifically, we are referring to the steel grade compositions obtained from the secondary metallurgy data base (composition described in Table 1). Finally, with the information of the billets hot rolled during 6 months and after a feature selection stage, we got a data set with 22 predictor variables and two objective variables, with which two independent models were generated and analysed.

3.2. Process Modeling Results

In this section, the results obtained for all models presented in the previous section will be shown and discussed. The validation method used was hold out with repetitions since there is a lot of deviation between different simulations. Likewise, the coefficient of determination (a.k.a R^2) was used as a metric to evaluate the goodness of each approximation, as well as statistical tests to see significant differences between them.

Holdout evaluation is a cross-validation approach whereby the available data are partitioned into a training set and a test set. The purpose of holdout evaluation is to test a model on different data to those from which it is learned. This provides an unbiased estimate of learning performance, in contrast to in-sample evaluation. Repeated holdout evaluation experiments are performed, each time with a different partition of the data, to create a distribution of training and test sets with which an algorithm is assessed [30].

Table 3. Raw data information from hot rolling process.

Variable Type	Description
General information	Heat number Manufacturing order Billet
Dimensions	Bar diameter Billet lengths
Speeds	Roughing speed Continuous mill
Times	Furnace time Download time Shelling time Roughing times Rolling mill time
Furnace times	Pre-heating zone time Heating zone time Maintenance zone time
Furnace temperatures	Surface temperatures Average temperatures Core temperatures
Other billet temperatures	Before roughing Before rolling mill Finishing exit block

The coefficient of determination (R^2) depicts the quality of the model to replicate the results, and the proportion of variation of the results that can be explained by the model. The metric is defined in Equation (7).

$$R^2 = \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2} \tag{7}$$

where m are number of instances, y_i and $\hat{y}_i$ are real and predicted values, and $\bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$.

In the previous section, especially in the case of SM, it was seen that each feature selection strategy provided a different dimensional space. The hypothesis that we handle in this work is that this result is due to the complexity of the problem, so that no selection method has the correct answer but all together can offer a better solution.

The possibility of being able to generate diverse spaces fits with an ensemble learning strategy, in which, taking into account a regressive paradigm, different models are generated for each of the selection strategies. These models are part of an ensemble and by means of a model fusion rule for each dimensional space, a single output is provided, which is the global estimate.

The meta-estimator (Section 2.2) using the three selection strategies explained in Section 2.3 was tested in SM, CC and HR, giving positive results in SM as can be seen in Table 4. To validate this strategy in the different processes, gradient boosting has been used as a base learner, which is the best base line without the meta-estimator approximation. This table summarizes the obtained R^2 using all features and the different considered feature selection techniques for the three models. The results with all features and the best method are indicated in bold. In CC and in the rolling process there are hardly any differences, although looking at the results in a global way, the methodology of an ensemble of models with different non-random dimensional spaces seems to be a good strategy for this type of problem where establishing the main characteristics of the process

is complicated. In addition, this form of combination of models gives a stability to the prediction that none of the other strategies has.

Table 4. Feature selection strategy comparison, number of features used and R^2 score for each sub-process SM, CC and HR. The best results per model are indicated in bold.

Selection Method	SM		CC		HR	
	#Features	R^2	#Features	R^2	#Features	R^2
All features	54	**0.42**	26	**0.93**	22	**0.87**
Pearson correlation	17	0.42	8	0.88	7	0.82
Univariate linear regression	17	0.42	12	0.9	9	0.85
Mutual information	17	0.41	8	0.89	7	0.85
Recursive selection	17	0.42	8	0.91	7	0.86
Meta-estimator	30	**0.43**	14	**0.92**	14	**0.86**

Considering that the ensemble based on feature selection approach can be a valid solution to produce a reliable and robust model for the performance prediction of a metallurgical process, a study was carried out using each of those proposed in Section 2.4 as a basic learning method.

Table 5 shows the obtained results using the meta-estimator strategy changing the base regression learner in each case. The best obtained R^2 values are indicated in bold. A clear conclusion that can be drawn is that a base model with a random forest produces the best results. In the case of SM, for example, random forest has quite better results than a NN with a much lower computational cost. However, the NN has competitive results compared to random forest or gradient boosting for both continuous casting and hot rolling. Also, it is noteworthy the low reliability of the XG Boost method in CC compared to other algorithms, which is an improvement on the state-of-the-art of a boosting strategy. However, in SM, and HR their results are similar and the computational cost of this approach is very low comparing especially with neural networks. A scatter plot for real measurements and predicted values with regression estimate is shown in Figure 6. This graph shows how the estimate for the Random forest strategy fits quit well in CC and HR. The high reliability provided by this strategy is visually appreciated for the entire magnitude range.

Table 5. Benchmark of applied regression methods where the R^2 is indicated for each sub-process: SM, CC, and HR. The best results per model are indicated in bold.

Regression Model	Secondary Metallurgy	Continuous Casting	Hot Rolling
Gradient Boosting	0.43	0.92	0.86
Random Forest	**0.44**	**0.94**	**0.98**
Neural Network	0.29	0.92	0.94
XG Boost	0.43	0.87	0.85

The Friedman statistical test is a non-parametric statistical procedure to compare more than two related samples [31]. The default assumption, or null hypothesis, is that the multiple paired samples have the same distribution. A rejection of the null hypothesis indicates that one or more of the matched samples have a different distribution. Taking a significance value of 0.05, if the p-value is below the significance level, then the test says that there is enough evidence to reject the null hypothesis and that the samples were probably drawn from populations with different distributions.

On the other hand, the Wilcoxon signed-rank test is a nonparametric statistical procedure to compare two samples that are paired or related. The parametric equivalent of the Wilcoxon signed rank test is known by names such as Student's t-test, t-test for matched pairs, t-test for paired samples, or t-test for dependent samples. The default assumption for the test, the null hypothesis, is that the two samples have the same distribution. In this case, in the same way, a significance of 0.05 is taken as the threshold.

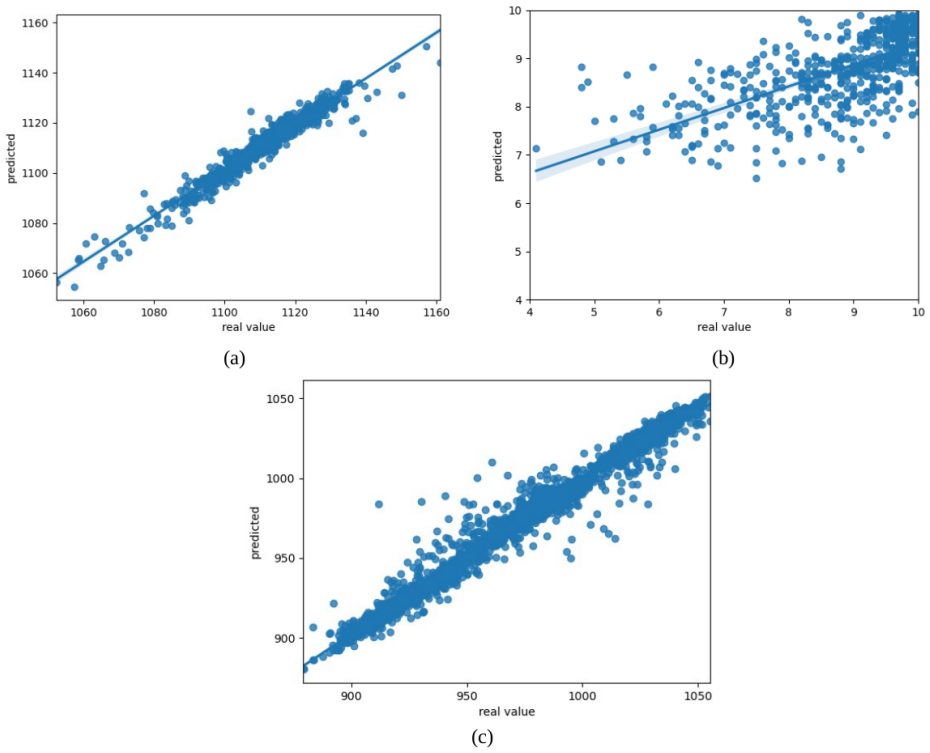

Figure 6. Scatter plots real vs. predicted values for Random forest strategy in case of CC (**a**) , MS (**b**) and HR (**c**). A linear regression line shows the trend line of the scatter plot result set. MS unit is the castability index, the CC and HR units are both temperature (ºC).

Friedman's non-parametric statistical significance test was performed to analyze whether there are significant differences between the error distributions of the regression models implemented. For the three sub-processes treated in this work, the statistic exceeds the threshold of significance, so there is at least one model that has a different error distribution in all cases. Table 6 shows the *p*-value obtained for each sub-process.

Table 6. The obtained *p*-values from the Friedman's test for each sub-process.

Secondary Metallurgy	Continuous Casting	Hot Rolling
1.88×10^{-55}	0.0017	9.77×10^{-20}

Taking this general conclusion, a pair-wise comparison has been made between the regression methods for each sub-process with the Wilcoxon test as can be seen in Figure 7. The objective was to extract the underlying differences between the error distributions of the different models. The main conclusion is that the NN is the model with an error distribution that has a significant difference comparing with the rest of the regression paradigms, except XGBoost in HR. This shows that in cases where the random forest method exceeds the rest in R^2 score, it is statistically the method with the highest reliability. Figure 8 shows the differences between the error distributions of the random forest and neural network of the three sub-processes.

Regarding the error distribution of the ensemble methods, there are some significant differences in the Wilcoxon test for CC. Any way, the use of one method or another in this case, taking into account the R^2 is a matter of reliability. In the case of SM there is not a significant difference between gradient boosting and ensemble approaches.

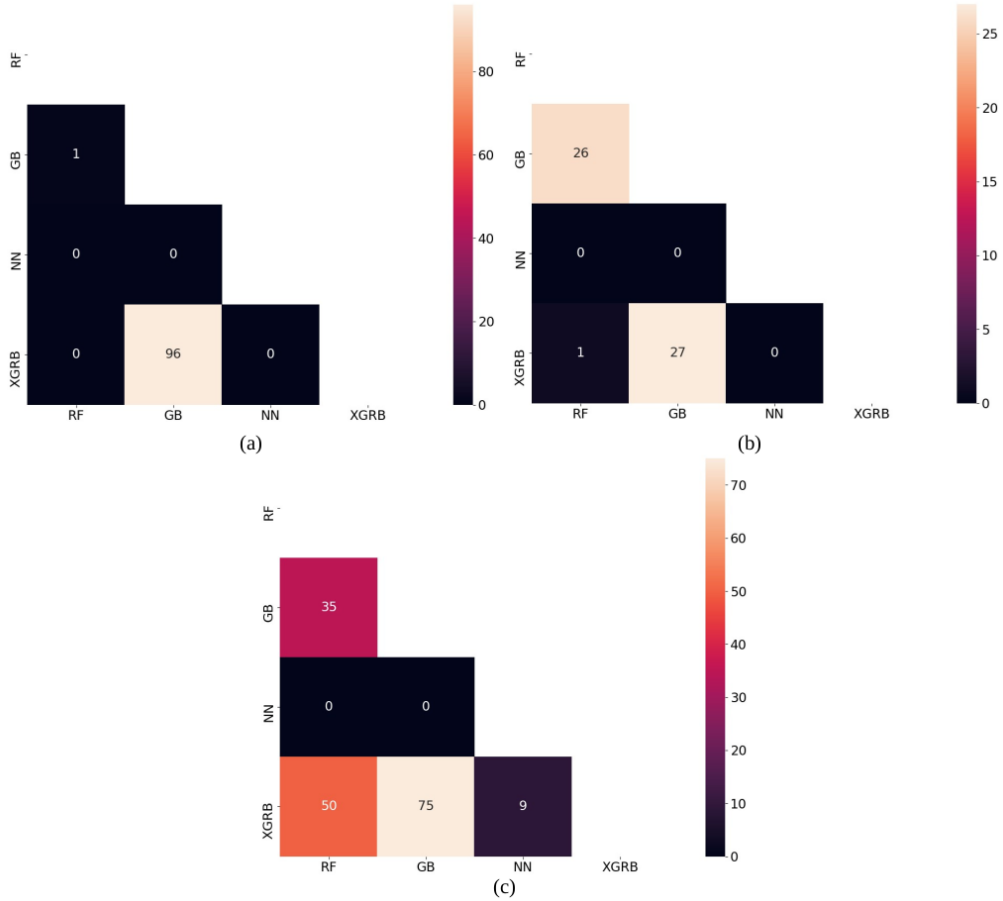

Figure 7. Wilcoxon signed rank test *p*-values (%) (**a**) CC, (**b**) MS and (**c**) HR. RF stands for Random Forest, GB Gradient Boosting, NN Neural Network and XGRB Extreme Gradient Boosting regressor.

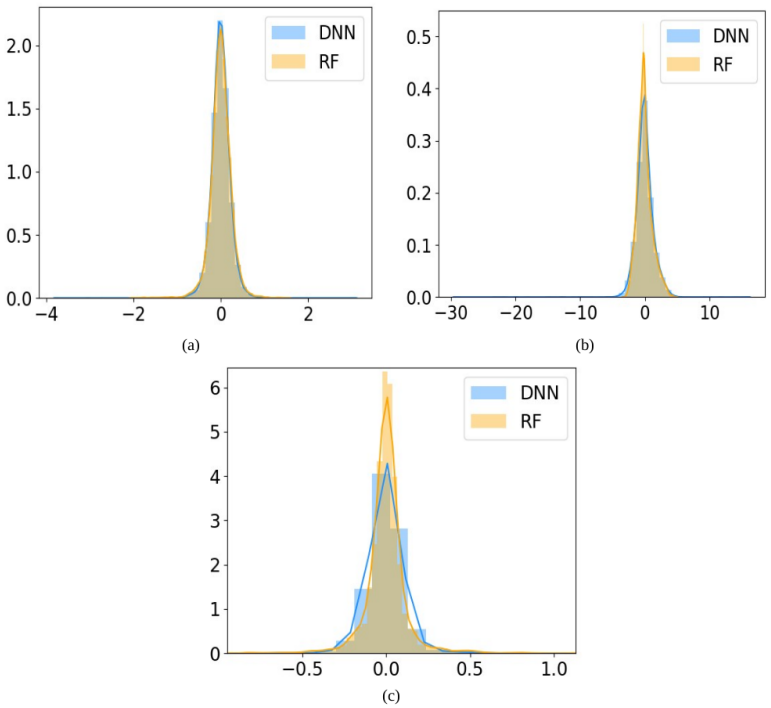

Figure 8. Error distribution for the random forest and neuronal network regression methods in CC (**a**), MS (**b**) and HR (**c**). *x*-axis is error produced by the model and y-axis is the density.

Lastly, the Wilcoxon paired test for HR draws slightly different conclusion. There are only significant differences between NN, random forest and gradient boosting. This enhances the hypothesis of the random forest method as an approximation with greater reliability for this sub-process and for the rest.

4. Conclusions

The defects in the final product in a steel making process are related to inefficiencies in the different sub-processes involved. SM, CC and HR are the ones that have the greatest impact, and each one has its own indicators to determine these inefficiencies. The estimation of these indicators through data-driven models is an utility that allows establishing the causes of poor production and thus optimizing operations to move towards efficient production. For these models to be used in plant operation, two visual decision-making tools are being developed. First, a simulator of the indicators with some inputs to the process, in such a way that the operator himself can explore the parameter space of the process. Second, a search tool for optimal process parameters, where the cost function is based on the generated model itself.

In this work, various data-driven models of a steel production process have been developed. These models developed for each sub-process independently estimate a relevant indicator such as the castability index for secondary metallurgy, the temperature in the billet before straightening in the continuous casting machine and temperatures of the billet before the continuous rolling mill.

In the methodology used in this study, different methods of feature selection and different regression strategies come into play. A novel approach based on ensemble strategy with a generative approach is also presented, using several selection methods to generate different base learners in order to obtain greater diversity in prediction.

In the tests carried out to validate the methodology, experimental results are presented for the models of the three sub-processes. Firstly, it has been verified that the selection of features with the presented methods allows maintaining the reliability of the models in most cases. Secondly, the approximation with ensemble techniques by means of non-random subsampling of features, maintains in the same way the reliability, giving it more stability and reducing the variance of the prediction. Even in the case of SM we see a slight improvement.

A study has been carried out with different regression strategies, based on the proposed ensemble approximation. Analyzing the coefficient of determination, the most clear conclusion is that the Random Forest method obtains the best results, even above NNs. In a more statistical way, the non-parametric tests of Friedman and Wilcoxon reveal significant differences between boosting and bagging paradigms and NNs. With the results of the models alone (not meta-estimator), the boosting strategy produces better results, which is not the case in the meta-estimator. This fact requires further analysis to be carried out in future work.

The direct prediction of the defects in the final product is a very challenging task, which is further complicated by the difficulties of the product traceability along the process, anyway a data-driven model for direct prediction of defects is being worked on. At any case, as mentioned before, the indicators of the sub-processes can be understood as indirect measurements of that quality.

Finally, all these process models are within a process optimization methodology/platform, using global methods such as evolutionary algorithms, being of great interest to the steel industry.

Author Contributions: Conceptualization, methodology and investigation, F.B., M.M. and T.G.; software, F.B., M.M., S.C. and A.C.; writing, review and editing, F.B., M.M., T.G., S.C., A.C. and A.A.; project administration, T.G.; resources, A.A.; funding acquisition, T.G. and A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the European Union's Horizon 2020 Research and Innovation Framework Programme [grant agreement No 723661; COCOP; http://www.cocop-spire.eu (accessed on 6 January 2022)]. The authors want to acknowledge the work of the whole COCOP consortium. This article reflects only the author's views and the Commission is not responsible for any use that may be made of the information contained therein.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Grzegorzewski, P.; Kochanski, A.; Kacprzyk, J. (Eds.) Data Preprocessing in Industrial Manufacturing. In *Soft Modeling in Industrial Manufacturing*; Springer International Publishing: Cham, Switzerland, 2019; pp. 27–41. [CrossRef]
2. Brandenburger, J.; Colla, V.; Nastasi, G.; Ferro, F.; Schirm, C.; Melcher, J. Big Data Solution for Quality Monitoring and Improvement on Flat Steel Production. *IFAC-PapersOnLine* **2016**, *49*, 55–60. [CrossRef]
3. Laha, D.; Ren, Y.; Suganthan, P.N. Modeling of steelmaking process with effective machine learning techniques. *Expert Syst. Appl.* **2015**, *42*, 4687–4696. [CrossRef]
4. Falkus, J.; Pietrzkiewicz, P.; Pietrzyk, W.; Kusiak, J. Artificial neural network predictive system for oxygen steelmaking converter. In *Neural Networks and Soft Computing*; Springer: Berlin, Germany, 2003; pp. 825–830.
5. Grešovnik, I.; Kodelja, T.; Vertnik, R.; Šarler, B. Application of artificial neural networks to improve steel production process. In Proceedings of the IASTED International Conference on Artificial Intelligence and Soft Computing, ASC 2012, Napoli, Italy, 25–27 June 2012.
6. Monteiro, L.V.; Sant'Anna, A. Application of Neural network for modeling steelmaking process. In Proceedings of the Congreso Latino-Iberoameroamericano de Investigación Operativa, Rio de Janeiro, Brasil, 24–28 September 2012.
7. Shukla, A.K.; Deo, B. Mathematical modeling of phosphorus prediction in BOF steelmaking process: A fundamental approach to produce low phosphorus steels and ensure direct tap practices. In Proceedings of the International Conference on Metal and Alloys, METALLO 2007, Kanpur, India, 7–10 December 2007.
8. Mazumdar, D.; Evans, J.W. *Modeling of Steelmaking Processes*; CRC Press: Boca Raton, FL, USA, 2009.
9. Chen, S.; Kaufmann, T. Development of Data-Driven Machine Learning Models for the Prediction of Casting Surface Defects. *Metals* **2022**, *12*, 1. [CrossRef]
10. Diniz, A.P.M.; Côco, K.F.; Gomes, F.S.V.; Salles, J.L.F. Forecasting Model of Silicon Content in Molten Iron Using Wavelet Decomposition and Artificial Neural Networks. *Metals* **2021**, *11*, 1001. [CrossRef]
11. Díaz, J.; Fernández, F.J.; Prieto, M.M. Hot metal temperature forecasting at steel plant using multivariate adaptive regression splines. *Metals* **2020**, *10*, 41. [CrossRef]
12. Murua, M.; Boto, F.; Anglada, E.; Cabero, J.M.; Fernandez, L. A slag prediction model in an electric arc furnace process for special steel production. *Procedia Manuf.* **2021**, *54*, 178–183. [CrossRef]
13. Miriyala, S.S.; Subramanian, V.R.; Mitra, K. TRANSFORM-ANN for online optimization of complex industrial processes: Casting process as case study. *Eur. J. Oper. Res.* **2018**, *264*, 294–309. [CrossRef]
14. Yan, Y.; Lv, Z. A Novel Multi-Objective Process Parameter Interval Optimization Method for Steel Production. *Metals* **2021**, *11*, 1642. [CrossRef]
15. Lino, R.E.; Marins, Â.M.F.; Marchi, L.A.; Mendes, J.A.; Penna, L.V.; Neto, J.G.C.; Caldeira, J.H.P.; da Costa e Silva, A.L.V. Influence of the chemical composition on steel casting performance. *J. Mater. Res. Technol.* **2017**, *6*, 50–56. [CrossRef]
16. Riaz, S.; de Toledo Bandeira, G.A.; Arteaga, A.; Komenda, J.; Zamberger, S.; Triolet, N.; Erdem, E. *Precipitation: Behaviour of Microalloyed Steels during Solidification and Cooling*; Technical Report; European Union: Luxembourg, 2010.
17. Pohu, B.; Collet, J.L.; Nguyen, T.; Lannoo, G.; Husain, Z.; Lan, Y.; Latz, A.; Schreiber, S.; Calvillo, G.P.; Theuwissen, K.; et al. *Control of Precipitation Sequences during Hot Rolling to Improve Product Uniformity of Titanium Containing High Strength Steels (PRETICONTROL)*; Technical Report; European Union: Luxembourg, 2021.
18. Zhou, Z.H. *Ensemble Methods: Foundations and Algorithms*; CRC Press: Boca Raton, FL, USA, 2012.
19. Re, M.; Valentini, G. Ensemble Methods: A Review. In *Advances in Machine Learning and Data Mining for Astronomy*; Chapman & Hall: London, UK, 2012; pp. 563–594.
20. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
21. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **2014**, *40*, 16–28. [CrossRef]
22. Ross, B.C. Mutual information between discrete and continuous data sets. *PLoS ONE* **2014**, *9*, e87357. [CrossRef] [PubMed]
23. Hoque, N.; Bhattacharyya, D.; Kalita, J.K. MIFS-ND: A mutual information-based feature selection method. *Expert Syst. Appl.* **2014**, *41*, 6371–6385. [CrossRef]
24. Kohavi, R.; John, G.H. Wrappers for feature subset selection. *Artif. Intell.* **1997**, *97*, 273–324. [CrossRef]
25. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
26. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
27. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]

28. Ho, T.K. Random decision forests. In Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 14–16 August 1995; Volume 1, pp. 278–282.
29. Smith, M. *Neural Networks for Statistical Modeling*; Thomson Learning: Belmont, CA, USA, 1993.
30. Sammut, C.; Webb, G.I. (Eds.) Holdout Evaluation. In *Encyclopedia of Machine Learning*; Springer: Boston, MA, USA, 2010; pp. 506–507. [CrossRef]
31. Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.* **1937**, *32*, 675–701. [CrossRef]