Big Data to Enable Global Disruption of the Grapevine-powered Industries

D5.3 – Trust-aware Decision Support System

DELIVERABLE NUMBER	D5.3
DELIVERABLE TITLE	Trust-aware Decision Support System
RESPONSIBLE AUTHOR	Katrien Verbert (KU Leuven)

GRANT AGREEMENT N.	780751
PROJECT ACRONYM	BigDataGrapes
PROJECT FULL NAME	Big Data to Enable Global Disruption of the Grapevine-powered industries
STARTING DATE (DUR.)	01/01/2018 (36 months)
ENDING DATE	31/12/2020
PROJECT WEBSITE	http://www.bigdatagrapes.eu/
COORDINATOR	Pythagoras Karampiperis
ADDRESS	110 Pentelis Str., Marousi, GR15126, Greece
REPLY TO	pythk@agroknow.com
PHONE	+30 210 6897 905
EU PROJECT OFFICER	Mr. Riku Leppanen
WORKPACKAGE N. | TITLE	WP5 | Leveraging Data Value
WORKPACKAGE LEADER	KU Leuven
DELIVERABLE N. | TITLE	D5.3 | Trust-aware Decision Support System
RESPONSIBLE AUTHOR	Katrien Verbert (KU Leuven)
REPLY TO	katrien.verbert@cs.kuleuven.be
DOCUMENT URL	http://www.bigdatagrapes.eu/
DATE OF DELIVERY (CONTRACTUAL)	31 December 2019 (M24)
DATE OF DELIVERY (SUBMITTED)	20 December 2019 (M24)
VERSION | STATUS	2.0 | Final
NATURE	DEM (Demonstrator)
DISSEMINATION LEVEL	PU (Public)
AUTHORS (PARTNER)	Katrien Verbert (KU Leuven), Nyi Nyi Htun (KU Leuven), Diego Rojo Garcia (KU Leuven)
CONTRIBUTORS	
REVIEWER	Stefan Scherer (GEOCLEDIAN), Florian Schlenz (GEOCLEDIAN)

VERSION	MODIFICATION(S)	DATE	AUTHOR(S)
0.1	Initial ToC and document structure	27/11/2018	Nyi-Nyi Htun (KU Leuven)
0.2	Introduction and System Descriptions	05/12/2018	Nyi-Nyi Htun (KU Leuven)
0.3	Internal revision	07/12/2018	Katrien Verbert (KU Leuven)
0.4	A complete draft	07/12/2018	Nyi-Nyi Htun (KU Leuven), Diego Rojo Garcia (KU Leuven)
0.5	2nd Draft	10/12/2018	Diego Rojo Garcia (KU Leuven), Nyi-Nyi Htun (KU Leuven)
0.6	Internal revision	11/12/2018	Katrien Verbert (KU Leuven)
0.7	3rd Draft	11/12/2018	Nyi-Nyi Htun (KU Leuven), Diego Rojo Garcia (KU Leuven)
0.8	Partner revision	19/12/2018	Florian Schlenz (GEOCLEDIAN), Stefan Scherer (GEOCLEDIAN)
1.0	Final	20/12/2018	Nyi-Nyi Htun (KU Leuven)
1.1	System Description of the updated system.	29/11/2019	Diego Rojo Garcia (KU Leuven)
1.2	1st draft	06/12/2019	Diego Rojo Garcia (KU Leuven)
1.3	2nd draft	12/12/2019	Diego Rojo Garcia (KU Leuven), Nyi-Nyi Htun (KU Leuven)
1.4	Final draft	13/12/2019	Diego Rojo Garcia (KU Leuven), Nyi-Nyi Htun (KU Leuven)
1.5	Partner revision	18/12/2019	Florian Schlenz (GEOCLEDIAN), Stefan Scherer (GEOCLEDIAN)
2.0	Final	20/12/2019	Diego Rojo Garcia (KU Leuven), Nyi-Nyi Htun (KU Leuven)

PARTICIPANTS		CONTACT
Agroknow IKE (Agroknow, Greece)		Pythagoras Karampiperis Email: pythk@agroknow.com
Ontotext AD (ONTOTEXT, Bulgaria)		Todor Primov Email: todor.primov@ontotext.com
Consiglio Nazionale Delle Richerche (CNR, Italy)		Raffaele Perego Email: raffaele.perego@isti.cnr.it
Katholieke Universiteit Leuven (KULeuven, Belgium)		Katrien Verbert Email: katrien.verbert@cs.kuleuven.be
Geocledian GmbH (GEOCLEDIAN Germany)		Stefan Scherer Email: stefan.scherer@geocledian.com
Institut National de la Recherché Agronomique (INRA, France)		Pascal Neveu Email: pascal.neveu@inra.fr
Agricultural University of Athens (AUA, Greece)		Katerina Biniari Email: kbiniari@aua.gr
Abaco SpA (ABACO, Italy)		Simone Parisi Email: s.parisi@abacogroup.eu
APIGAIA (APIGEA, Greece)		Eleni Foufa Email: Foufa-e@apigea.com

ACRONYMS LIST

AHMoSe	Augmented by human model selection
DSS	Decision support system
ML	Machine learning
PA	Precision agriculture
WP	Work package

EXECUTIVE SUMMARY

This document presents an update of deliverable 5.3 where we design and evaluate a trust-aware decision support system that uses visualisation techniques to explain the influence of input (predictor) variables on prediction outcomes. Research has shown that prediction models currently employed in agricultural decision support systems (DSSs) remain opaque to users and hidden behind the software. This black-box nature can often lead to users not trusting the system's decisions especially when the system fails to provide meaningful explanations. Previous work has expressed that explaining a model's predictions is an important approach for earning users' trust. Visualisation is a powerful technique to address this problem and can effectively communicate uncertainty emerging from both data and prediction models.

To demonstrate our first version of a decision support system, in the previous deliverable, we used an example wine quality dataset which was based on red variants of the Portuguese "Vinho Verde" wine and contains 1599 instances of 11 physicochemical (inputs) variables and a sensory (the output) variable which is wine quality. In this updated version, we have redesigned the system to integrate a dataset that has been collected by one of the pilot partners, AUA. The dataset contains quality assignments for one grape variety, which are based on four features of the grape: 1) total anthocyanin content, 2) berry fresh weight, 3) total soluble solids and 4) titratable acidity. The system aims to answer some of the most important questions in viticulture: 1) which machine learning (ML) model should I use with the data that is specific to this vineyard/grape variety, 2) how do various grape parameters affect the quality predictions of different ML models, and 3) which of the different ML models produces an output that is in-line with my knowledge? In this document, we describe the new version of our decision support system and the results of a qualitative evaluation which was conducted over several semi-structured interview sessions with the partners.

This document is structured as follows. Section 1 lays out an introduction to the deliverable describing our previous version and motivations. In Section 2, our DSS is described in detail together with the development technology we utilised. In Section 3, the results of the evaluation are presented. In Section 4, we provide a usage manual with instructions on how to obtain the source code. This document concludes with Section 5 where a summary of the deliverable is underlined.

TABLE OF CONTENTS

LIST OF FIGURES

1 INTRODUCTION

Decision Support System (DSSs) are designed to assist users with decision making activities while dealing with massive amounts of data. In the field of agriculture, different stakeholders such as farmers, advisers and policymakers use DSSs to often facilitate farm management and planning tasks [6]. Depending on the type of decision support required, data is first gathered from multiple sources including sensors, satellites and in-field observations, and analysed using a series of statistical models. The output is then presented to users in a number of ways such as tables and/or graphs.

Research has shown that user-centred development plays an essential role in fulfilling users' need for agricultural system modelling [8]. However, the models employed to date for agricultural DSSs remain opaque to users and hidden behind the software [1, 7, 16, 18]. This often leads to trust issues, notably when suggestions coming from a DSS fail to provide meaningful explanations [17]. If users do not trust a model or a prediction, they are unlikely to use it in making important business decisions. Trust in a system is based on competence, benevolence, and integrity, just like trust in a person [10]. Ribeiro et al. [15] expressed two different (but related) definitions of trust related to models and predictions. These are 1) *trusting a prediction* meaning a user trusts an individual prediction sufficiently to take some action based on it, and 2) *trusting a model* meaning the user trusts a model to behave in reasonable ways if deployed. Both are directly affected by how much the user understands a model's behaviour. In fact, Ribeiro et al. [15] argued that explaining a model's predictions is a much more important aspect for earning a user's trust. Moreover, explanations should be not just accurate, but also easily understandable.

Visualisation is a powerful technique to bridge this gap, and has demonstrated its usefulness in PA [19] to clearly communicate uncertainty emerging from both data and prediction models [5]. Visually explaining a model's predictions can be as simple as highlighting important input variables. Figure 1 (source: [15]) and Figure 2 (source: [2]) are two examples of such a case. Both figures show how colour coding and graphical approaches can explain a model's predictions and the influence of each input variable on the final output. As such, the use of visualisation techniques has been shown to provide essential insights and facilitates the understanding of machine learning models for complex problems [4, 11], and are especially useful for users with little machine learning knowledge [9, 21].

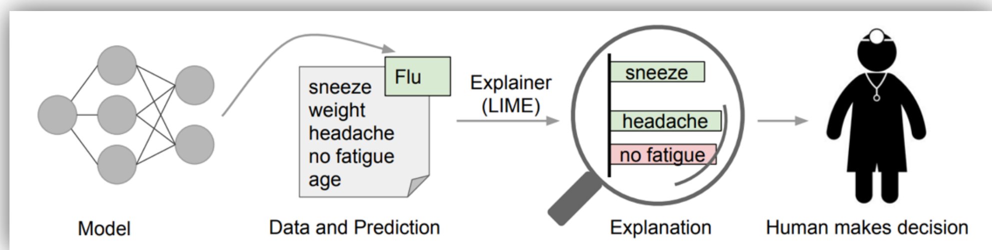


Figure 1: Explanation of a model which predicts flu based on the symptoms. Notice the use of simple colours: the green colour indicates that sneeze and headache are the primary indications of flu whereas the red colour indicates that having no fatigue is the evidence against flu [15].

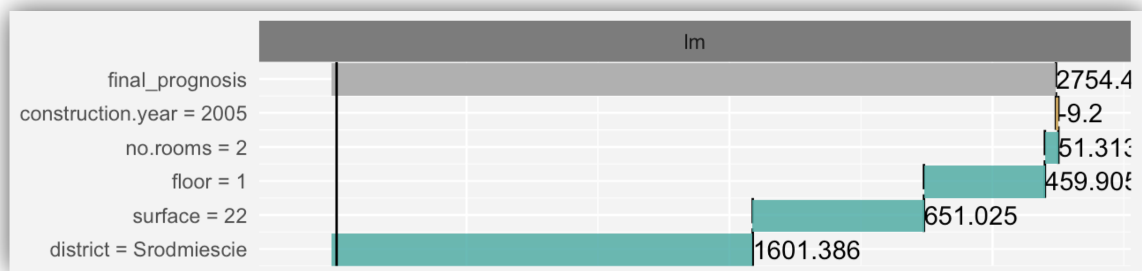


Figure 2: Explanation of a linear model that predicts house prices based on construction year, number of rooms, floor, surface area and district. The green bars indicate positive contributions of the input variables on the predicted price whereas the yellow bar (i.e. construction year) indicates negative contributions [2].

In this deliverable, we aim to develop DSSs that nurture user trust. To achieve this, the system must be transparent, meaning it must be able to clearly communicate the prediction model with users and show differing effects of input variables on the model's output. Thus, in the previous version of the deliverable[1], we demonstrated an interface to assist with the wine quality prediction scenario, building on top of the prediction models presented by WP4 partners. In the new version of the deliverable, we focus on the scenario of grape quality prediction in collaboration with one of the pilot partners, AUA. The dataset collected by AUA contains quality assignments for one grape variety (*Vitis vinifera*), which are based on the following features: 1) total anthocyanin content, 2) berry fresh weight, 3) total soluble solids and 4) titratable acidity. The data came from 48 (10 × 20 meters) cells of a vineyard situated in Mikrothives, central Greece and is available for the harvest years of 2010, 2011 and 2012. The problem we discovered with data such as this is as follows:

> *Many of the observed data in the vineyards are yearly based (e.g., yield or grape quality at harvest), which makes it impossible to accelerate the gathering of data. Besides, the distribution of the data from year to year can vary a lot due to uncontrollable parameters like weather conditions. Thus, how can we address this gap and still create reliable machine learning models despite the lack of data from the farms?*

Previous work has tried to address this gap by designing knowledge-based decision support systems, such as fuzzy sets [14], that use the domain knowledge of viticulture experts. We followed a similar approach and designed AHMoSE (Augmented by human model selection) which compares and explains the predicted outcomes of various machine learning models and helps domain experts to select the models that fit their knowledge the most. Specifically, AHMoSE is designed to answer the following questions in viticulture: 1) which machine learning (ML) model should I use with the data that is specific to this vineyard/grape variety, 2) how do various grape parameters affect the quality predictions of different ML models, and 3) which of the different ML models produces an output that is in-line with my knowledge?

In the next section, Section 2, a detailed description of AHMoSE and the development technology is provided. We then present the results of a quantitative evaluation of AHMoSE in Section 3. The source code of AHMoSE has been uploaded to the Github repository of BigDataGrapes. Instructions on how to obtain the source code is described in Section 4.

[1] https://doi.org/10.5281/zenodo.2531615

2 SYSTEM DESCRIPTION

To aid domain experts with little machine learning (ML) knowledge to validate and compare different predictive regression models against each other, primarily using their domain knowledge, we have designed and developed AHMoSe (Augmented by Human Model Selection), a visual decision support system. The goals of this system are: (1) to be able to select models that generalize better to the data of interest and (2) to improve user understanding of the different models. AHMoSe can be used with any ML model as it is based on model-agnostic ML interpretation methods, and thus only needs access to the input (i.e., features) and the predicted output of the model.

2.1 DATA & KNOWLEDGE INPUTS

In this section we describe the two inputs required by AHMoSe: model explanations and domain knowledge. The concrete data file structure is detailed in Section 4.2.

2.1.1 Model Explanations Using SHAP

In order to not constrain ourselves with any present and future machine learning models, an interactive model analysis system should treat them as "black-box" models (i.e., no knowledge of models' internals) and use model-agnostic interpretation methods such as LIME [15], SHAP [12], DeepVid [20] or RuleMatrix [13]. In our system, we used the SHAP framework[2] to incorporate explanations as its methods show better consistency with human intuitions [12], which makes it a good fit for our system, as the end-users have to compare these explanations with their domain knowledge.

When obtaining explanations from the SHAP framework for a model and a dataset, one obtains, for each observation of the data, what effect the value of each input feature has on the predicted output (this effect is called the SHAP value). If the model is a regression model, as in our case, this SHAP value indicates how feature increments or decrements affect the output value for that observation. The SHAP framework also provides a base value which is the average model output over the dataset we passed. This value is used in our system to translate the SHAP value from an effect space to the dimensions of the predicted output, which makes the interpretation and the comparison with domain knowledge easier.

A way to understand how a single feature affects the predicted output of a model is to plot for each observation in our dataset the value of the feature versus the SHAP value of that feature (plus the base value if we want the predicted output dimension). An example of this can be seen in Figure 3. Note that vertical dispersion of the circles with the same feature value is possible due to the effects caused by the interactions with other features whose values are different across the observations.

[2] https://github.com/slundberg/shap

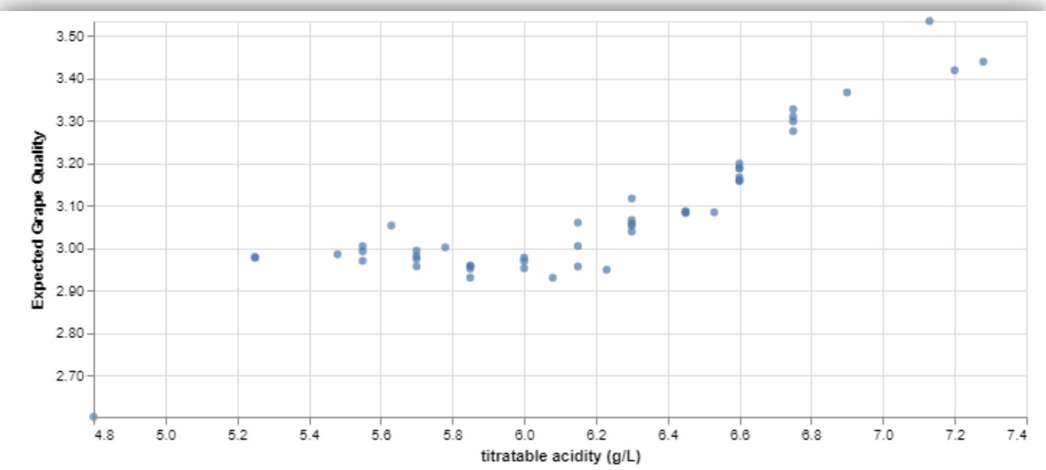


Figure 3: SHAP explanations of a model for titratable acidity. Each circle is an observation of our example scenario dataset. The y-axis shows, for each observation, the Expected Grape Quality, i.e. the addition of the SHAP value and the base value (expected grape quality of the average grape of our dataset).

2.1.2 Domain Knowledge

To be able to carry out a knowledge-based validation of the regression models using the previously described model explanations, domain experts have to state their knowledge on how a target feature depends on each of the input features. To this end, experts should provide an expected range of the target feature value for as many different intervals of each input feature as considered. Experts could also be substituted by other sources of domain knowledge (e.g., research results). The intervals for the four features of the grape quality example scenario can be seen in Figure 4.

		Feature Value		Grape Quality	
Feature	Label	min	max	min	max
Anth	L	200	600	1.17	2.17
	M	600	800	2.23	3.23
	H	800	1000	3.35	4.35
	VH	1000	1400	4.25	5.00
BW	L	1	1.6	3.30	4.30
	M	1.6	2	2.44	3.44
	H	2	2.5	1.27	2.27
TA	L	3	5	1.64	2.64
	M	5	7	2.30	3.30
	H	7	12	2.79	3.79
TSS	L	15	21	1.36	2.36
	H	21	30	2.63	3.63

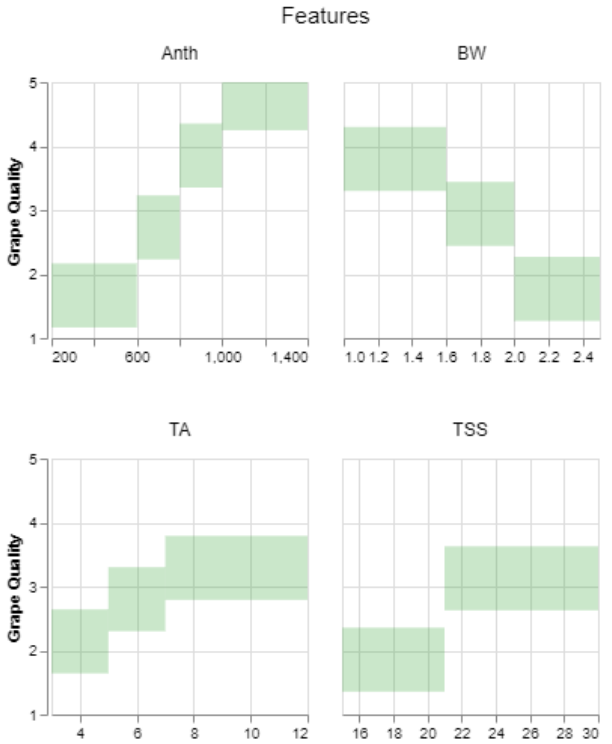


Figure 4: Table and visualisation of the domain knowledge intervals for the four input features of our example scenario: total soluble solids (TSS), titratable acidity (TA), total skin anthocyanins (Anth) and berry fresh weight (BW).

2.2 KNOWLEDGE-BASED VISUAL VALIDATION

To help domain experts in the validation and selection of the regression models using a knowledge-based approach, two different visualisations, described in the following sections, are used with each of the regression models: a knowledge-agreement dependence plot for each feature and a knowledge-agreement summary plot.

2.2.1 Knowledge-agreement Dependence Plot

Knowledge-agreement dependence plots (see Figure 5) help users understand, for each feature and model, the similarities and differences between the domain knowledge intervals (as shown in Figure 4) and the model explanations (as shown in Figure 3). This plot has two different layers: a knowledge layer and a model layer. While the model layer is different for each model and feature, the knowledge layer is different for each feature, but shared by all the models.

Knowledge layer: this layer uses a green rectangle mark for each of the knowledge intervals that correspond to the feature. The x-axis limits of each rectangle depict the minimum and maximum of each interval defined by a domain expert, whereas y-axis limits depict the range where the domain expert expects the target feature's (e.g., Grape Quality in our case) mean value to be.

Model layer: this layer uses a circle mark for each observation explanation corresponding to a certain model and feature. The x-position encodes the value of the feature on the observation the explanation was made on, whereas the y-position encodes the expected value of the target feature (e.g., Grape Quality) based on the model. Finally, the colour encodes the agreement (blue) or disagreement (orange) with the corresponding domain expert knowledge. If no knowledge interval covers the value of the feature for a circle, it is encoded in grey.

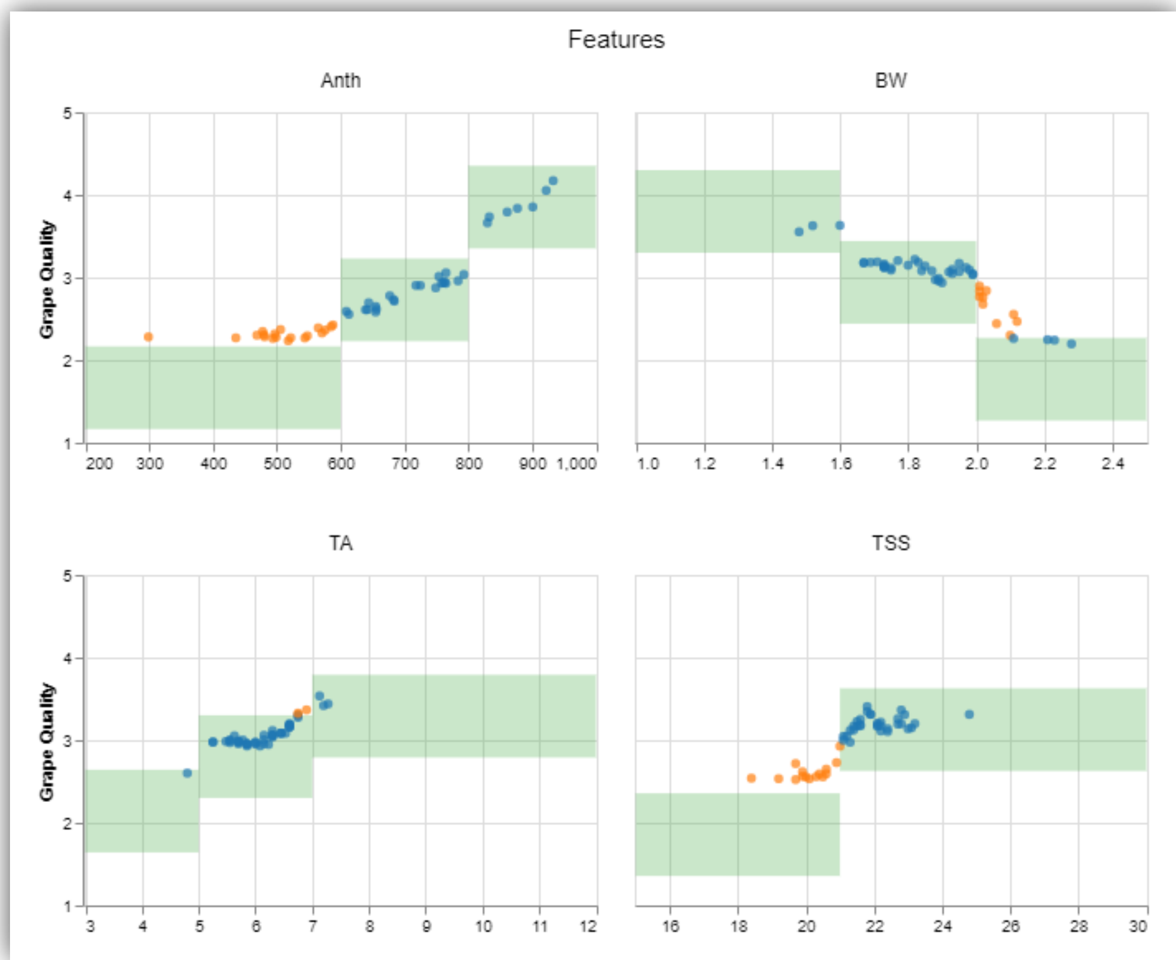


Figure 5: Example of knowledge-agreement dependence plots for four input features of our example scenario: total soluble solids (TSS), titratable acidity (TA), total skin anthocyanins (Anth) and berry fresh weight (BW).

2.2.2 Knowledge-agreement Summary Plot

The knowledge-agreement summary plot of each model (see Figure 6) has two objectives: (1) summarize the information of the knowledge-agreement dependence plots of all features and (2) show the importance that each feature has for the model. To this end, the knowledge-agreement summary plot uses a Marimekko chart.

The width of each group of stacked bars encodes the importance of a particular feature according to the given model explanations. For each model, the SHAP values show the effect of each feature in an observation. Thus, we calculate the importance of each feature on each model as the mean of the absolute value of the SHAP values.

The height of each stacked bar encodes the percentage of circles in the knowledge-agreement dependence plot of that feature that corresponds to each category: agreement (blue), disagreement (orange) or have no knowledge interval reference (grey).

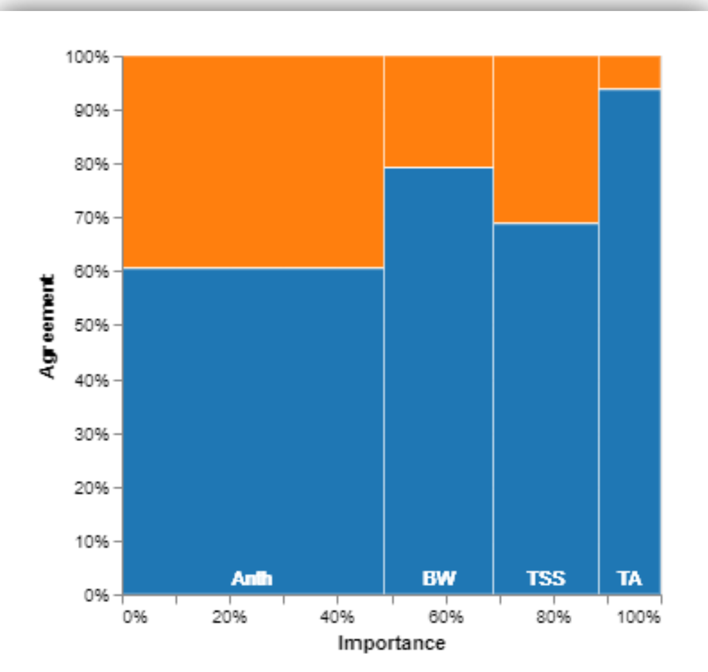


Figure 6: Example of knowledge-agreement summary plot.

2.3 AHMOSE INTERFACE

To facilitate interaction, we designed the interface of AHMoSe a sidebar and a visualisation panel. The sidebar controls (Figure 7-a) allow users to select and filter data to generate visualisations. The visualisation panel consists of a plot matrix that has one row for each of the selected models the user wants to analyse. Each row represents a model where two different visualisations are displayed: a knowledge-agreement dependence plot (Figure 7-b) for each feature and a knowledge-agreement summary plot (Figure 7-c).

Figure 7: The AHMoSe interface: a) the sidebar controls to select the use case, intervals, models and features to be visualized, b) the scatter plots highlight the comparisons between models' predictions (orange and blue dots) and domain expert knowledge (green rectangles), c) the Marimekko charts indicate the importance of each feature according to the given model, and the agreement between the model and domain expert knowledge.

2.4 DEVELOPMENT TECHNOLOGY

In this version of deliverable, we have used React[3] to build the system, Vega (together with Vega-Lite and Vega-Embed) for the development of the visualisations, and Material Components for React[4] for the user interface components. The previous version of the deliverable used Vue[5] as JavaScript Framework and Vue Material[6] components, which also follow the Material Design[7] specs, for the user interface elements. The decision to migrate to React has been taken, as in deliverable 5.1, because it is a more familiar technology for many of the development teams at many of the pilot partner institutions.

The use in this version of more commonly used visualisation in this version of the deliverable in order to ease the use and understanding of them, has led to the use of Vega[8] instead of d3.js[9] for the development of the visualisations.

[3] https://reactjs.org/
[4] https://github.com/material-components/material-components-web-react
[5] https://vuejs.org/
[6] https://vuematerial.io/
[7] https://material.io/design/
[8] https://vega.github.io/vega/
[9] https://d3js.org/

3 SYSTEM EVALUATION

We also conducted an evaluation with the partners to gather feedback. We interviewed a total of nine experts: five viticulture experts from AUA, INRA and Symbeeosis (V1 — V5) and four ML experts from CNR (ML1 — ML4). The experts were first presented with the tool and explained the individual components of the tool. They were then given a few minutes to interact with the tool and could ask questions for further clarification. If there were no further questions, we asked the experts the following 10 open-ended questions. Responses to each of the questions were recorded and later transcribed for analysis.

1. How do you imagine the tool being used in real life?
2. In which situations do you think this tool will be useful?
3. Which parts of the visualisations make you trust the Machine Learning models?
4. How did the tool affect your understanding of the influence of features on each model?
5. How easy was it to use the tool to select a model?
6. What other word choices (e.g., "agreement", "importance") would make the visualisation easier to understand?
7. Which aspects of the visualisations would you change to improve your understanding of the visualisations?
8. Which aspects of the visualisations would you change to improve your understanding of model explanations?
9. Which aspects of the visualisations would you change to ease the model selection task?
10. Do you have suggestions for improvement of the visualisations?

The transcribed data were coded and analysed following the thematic analysis approach [3], which resulted in four main themes: potential use cases, trust, usability, and understandability.

3.1 POTENTIAL USE CASES

This theme highlights potential use cases that the experts believe can benefit from our tool. All the experts recognized the advantages of using AHMoSE to compare the predictions of various ML models with one's knowledge. For example, one of the viticulture experts explained, "*If I can put my measurements in there and see how the model characterizes my grapes at hand, of my sample, and see for example if my anthocyanin content is 500, how does this relate with the model? Is my quality good or bad?*" (V1). Another interesting use case mentioned by the viticulture experts was detecting anomalies in data and unqualified products. For example, one expert mentioned, "*let's say if you have something unexpected happening on the field and you want to see how this correlates with the quality that you are going to get. So you do some measurements, and then you want to see how this is going to translate.*" (V2).

The ML experts, unsurprisingly, were more focused on the technical process behind each model. While they recognized the advantages of the tool for viticulture experts, they specifically pointed out the case which requires a viticulture expert to select a suitable model and better understand the features. For example, one expert explained, "*Of course, there are models that are clearly unsuitable but they show where the problems are, on which feature we have a problem in the ML algorithm. I don't know how to exploit this information but it is clear that the berry fresh weight is a difficult feature to use in all models.*" (ML2). This expert also pointed out the utility of this tool in anomaly detection such as, "*A domain expert could ask, `I didn't know that this feature, berry fresh weight, when between 2 and 2.4, the grape quality is difficult to predict. What is happening in that range?' In other words... anomaly detection.*" (ML2). Thus, it appears that anomaly detection is a potential use case the experts envision the tool being used the most, in addition to quality prediction. The experts also suggested to give the user more control, such as configuring different models for each threshold (ML4).

3.2 TRUST

The trust of users in a model often plays an important role for it to be adopted. Explaining a model's predictions, in a way that it is easily understandable, is perhaps the most important aspect for earning a user's trust [15]. This theme presents the aspects of our visualisations that may help users trust in the given ML models. All of the nine participants mentioned that the ability to see dis/agreements between models' predictions and an expert's knowledge can help them inspect further and thus promote trust. A viticulture expert explained, *"The thing that makes us trust the models is the fact that most of the time, there is a good agreement between the values predicted by the model and the ones obtained for the knowledge of the experts."* (V5). Another viticulture expert described, *"You actually help the model with some knowledge from the experts, so when I see for example that my basic knowledge agrees with the model, then we know that the model has some degree of certainty."* (V1). However, the same expert also emphasized that we *"cannot always take for granted and trust it with our eyes closed."* (V1). The ML experts also had a similar remark regarding trust. For example, one explained that *"[...] this visualisation is telling me an insight on what to look at. They are not going to take any decision or try to explain me why there is this problem; just telling me there could be something going on here"* (ML1). It appears that although the experts are aware of uncertainties, they recognize the potentials of visualisations and the ability to improve predictions with the knowledge of domain experts.

3.3 USABILITY

The usability theme focuses on the perceived ability of the tool to be used in real-life. As previously mentioned in Section 3.1, all the experts saw usability of the tool in various scenarios. However, they also expressed certain concerns and provided suggestions for improvement. For example, one viticulture expert mentioned *"it definitely needed a person there like [researcher name] to explain it a bit, the logic behind the models, what the green areas depict, the importance in the agreement of the model. It wasn't difficult to use it but definitely someone should be there to explain exactly what it is. But it wasn't difficult after that."* (V1). Adding more textual explanations and legends may help mitigate this limitation.

The ML experts also provided some interesting suggestions to improve the usability of the tool. These include sticky labels and using a mouse-over function to explain the models briefly and to show various metrics of the models. For example, one expert explained, *"When you slide over [hover] a model, give more details because I can probably imagine what the acronyms stand for but they are basically capital letters and numbers and also add some explanations about the metrics you use. Is RMSE 0.37 high or low, with respect to which reference value we are looking at?"* (ML1). A second ML expert suggested, *"What I would have on a specific plot, for example the berry fresh weight, is what the size of the error is of the orange dots. I mean the percentage of the total error you get, in terms of performance. What is the error that they bring to the total quality? Maybe you should give some insight of how big it is, how important it is to fix this behaviour for specific classes of visualisation."* (ML2). A third ML expert highlighted the importance of also showing the source of the data, whether is it model or human-generated. As these suggestions have highlighted, it appears that usability and understanding of the components may be correlated. In Section 3.4, we further present the understandability of the various components in our tool.

3.4 UNDERSTANDABILITY

This theme focuses on how easy it was to understand the various components of the tool and how we can improve them. While all the experts reported that they understood the intentions of the tool, they also mentioned that the Marimekko charts (Figure 7-c) could be improved. For example, one viticulture expert explained that he wanted to see whether *"[...] this model is in the agreement with the basic bibliography. Is it in the agreement with the expert and his/her knowledge?"* (V1). One suggestion to show this agreement was to introduce confidence intervals (ML4). Two of the viticulture experts suggested that explicitly showing the success rate of each model may be helpful. For example, *"The right part may be a little bit confusing. And it would be useful to have a final number [of] the percentage of success."* (V2). The tool currently shows the Root-Mean-Square Error (RMSE) for each model, which may be confusing for the experts who do not have an ML background.

One ML expert stated that the Marimekko charts could be better explained by showing the importance and agreement in relation to a particular baseline, for example, *"importance of a feature, and agreement with human assessor or some description about that."* (ML1). Another ML expert suggested that showing more statistics or even the underlying model (although this could get too technical) may help understand the models better. For example, *"[...] what is the fraction of the error per feature. Given the wrong answer on each feature, the fraction of the error computed on that specific dimension but maybe this is too technical."* (ML2). Surely, understandability of the models, visualisations and features may differ between individual experts depending on their background. However, we believe that the suggestions provided by the ML experts could also help improve understandability for viticulture experts.

4 USAGE INSTRUCTIONS

4.1 DOWNLOADING THE CODE

A demo of this version of deliverable 5.3 is available at: https://bigdatagrapes-eu.github.io/d5.3-ahmose. The source code of the system has been uploaded to Github: https://github.com/BigDataGrapes-EU/d5.3-ahmose. Please follow the following steps to download the source code and run locally.

Step 1. Install npm with Node.js from https://www.npmjs.com/get-npm

Step 2. Download or clone the project:

```
$ git clone https://github.com/BigDataGrapes-EU/d5.3-ahmose.git
```

Step 3. Navigate to the cloned/downloaded folder:

```
$ cd d5.3-ahmose
```

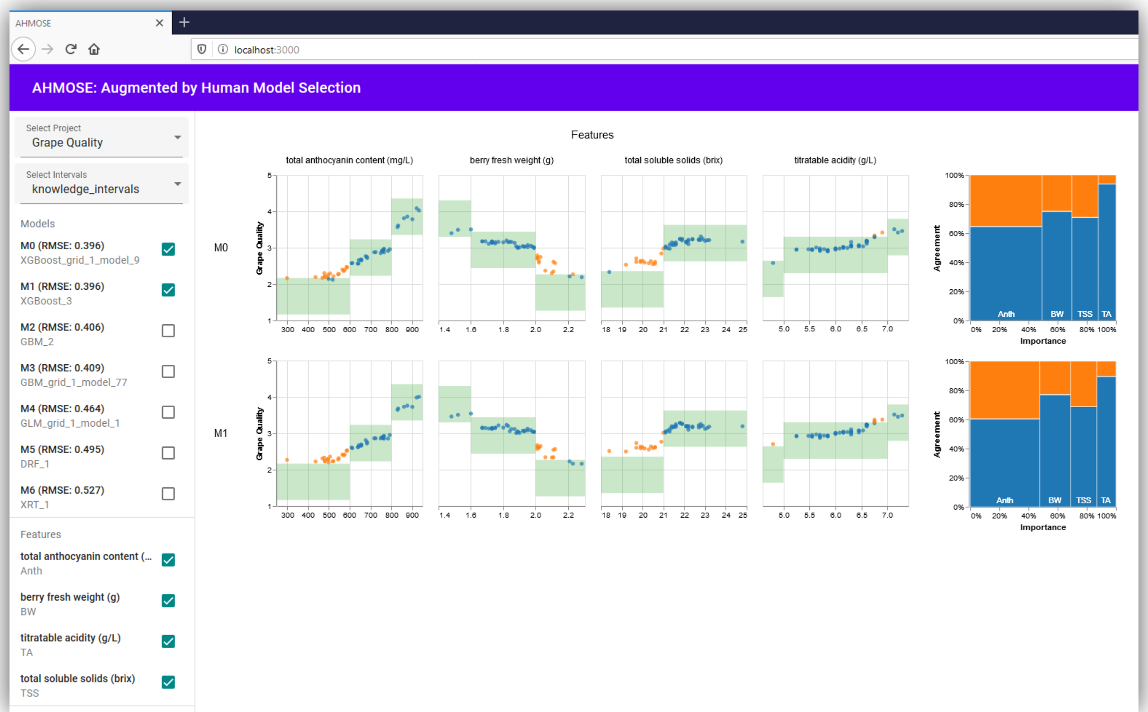
Step 4. Install dependencies:

```
$ npm install
```

Step 5. Run the app:

```
$ npm start
```

Step 6. Open http://localhost:3000 in browser



4.2 ADDING NEW SCENARIOS

To add new scenarios, create a new folder in `public/data/` with the scenario name and the following structure before running the app:

```
\---Scenario Name
    |   features_info.csv
    |   models_data.csv
    |   models_info.csv
    |
    \---intervals
    |   knowledge_intervals_1.csv
    |   knowledge_intervals_2.csv
    |   ...
    |   knowledge_intervals_N.csv
```

Where the structure of each file is the following:

- `models_info.csv` contains the name of the models and its root mean square error (RMSE) score:

```
model,RMSE
XGBoost_grid_1_model_9,0.395577489097816
XGBoost_3,0.395667516092918
...
```

- `models_data.csv` contains for each model, item (i.e. observation) and feature, its value, SHAP value and expected value (SHAP value + base value).

```
model,item,feature,value,shap_value,expected_value
DRF_1,1,Anth,299,-0.5201137609431272,2.5603869701381665
DRF_1,2,Anth,544,-0.563813505095174,2.5166872259861197
...
```

- `features_info.csv` contains the feature name on the other files and a (longer) label that would be use on the charts title.

```
feature,feature_label
Anth,total anthocyanin content (mg/L)
BW,berry fresh weight (g)
...
```

- Each file on the `intervals` folder contains an expected range (y1-y2) of the target feature value for as many different intervals (x1-x2) of each input feature as considered (see also Figure 4).

```
feature,label,x1,x2,y1,y2
Anth,Anth_L,200,600,1.1666666666666667,2.166666666666667
Anth,Anth_M,600,800,2.2301587301587302,3.2301587301587302
...
```

5 CONCLUSIONS

In this document, we presented an updated version of deliverable 5.3: trust-aware decision support system. In collaboration with a pilot partner, AUA, we have designed AHMoSE, a decision support system that assists viticulture experts with little machine learning knowledge to compare and understand grape quality predictions of various machine learning models, and to select the models that fit their knowledge the most. AHMoSE is designed to answer some of the most important questions in viticulture, such as: 1) which machine learning (ML) model should I use with the data that is specific to this vineyard/grape variety, 2) how do various grape parameters affect the quality predictions of different ML models, and 3) which of the different ML models produces an output that is in-line with my knowledge?

Detailed explanations of the system and development technology have been presented in Section 2. We also conducted a qualitative evaluation of the system by administering semi-structured interviews with the partners. The results of this evaluation are presented in Section 3. The code of the system has been published at the Github repository of BigDataGrapes[10]. Instructions on how to obtain the code have also been provided in Section 4.

[10] https://github.com/BigDataGrapes-EU/d5.3-ahmose

6 REFERENCES

[1] Accorsi, P. et al. 2014. HydroQual: Visual analysis of river water quality. *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)* (Oct. 2014), 123–132.

[2] Biecek, P. 2018. DALEX : Descriptive mAchine Learning EXplanations. (2018).

[3] Braun, V. and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*. 3, 2 (2006), 77–101. DOI:https://doi.org/10.1191/1478088706qp063oa.

[4] Donoso-Guzmán, I. and Parra, D. 2018. An interactive relevance feedback interface for Evidence-Based Health Care. *International Conference on Intelligent User Interfaces, Proceedings IUI* (Mar. 2018), 103–114.

[5] Frías, M.D. et al. 2018. An R package to visualize and communicate uncertainty in seasonal climate prediction. *Environmental Modelling & Software*. 99, (Jan. 2018), 101–110. DOI:https://doi.org/10.1016/J.ENVSOFT.2017.09.008.

[6] Gutiérrez, F. et al. 2019. A review of visualisations in agricultural decision support systems: An HCI perspective. *Computers and Electronics in Agriculture*. Elsevier B.V.

[7] Han, E. et al. 2017. Climate-Agriculture-Modeling and Decision Tool (CAMDT): A software framework for climate risk management in agriculture. *Environmental Modelling & Software*. 95, (2017), 102–114. DOI:https://doi.org/10.1016/j.envsoft.2017.06.024.

[8] Jones, J.W. et al. 2017. Brief history of agricultural systems modeling. *Agricultural Systems*. 155, (Jul. 2017), 240–254. DOI:https://doi.org/10.1016/j.agsy.2016.05.014.

[9] Kahng, M. et al. 2019. GAN Lab: Understanding Complex Deep Generative Models using Interactive Visual Experimentation. *IEEE Transactions on Visualization and Computer Graphics*. 25, 1 (Jan. 2019), 310–320. DOI:https://doi.org/10.1109/TVCG.2018.2864500.

[10] Komiak, S. et al. Comparing Customer Trust in Virtual Salespersons With Customer Trust in Human Salespersons. *Proceedings of the 38th Annual Hawaii International Conference on System Sciences* 175a-175a.

[11] Kulesza, T. et al. 2015. Principles of Explanatory Debugging to personalize interactive machine learning. *International Conference on Intelligent User Interfaces, Proceedings IUI* (Mar. 2015), 126–137.

[12] Lundberg, S. and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. (2017). DOI:https://doi.org/10.3321/j.issn:0529-6579.2007.z1.029.

[13] Ming, Y. et al. 2019. RuleMatrix: Visualizing and Understanding Classifiers with Rules. *IEEE Transactions on Visualization and Computer Graphics*. 25, 1 (Jan. 2019), 342–352. DOI:https://doi.org/10.1109/TVCG.2018.2864812.

[14] Papageorgiou, E.I. et al. 2016. Fuzzy sets in agriculture. *Studies in Fuzziness and Soft Computing*. Springer Verlag. 211–233.

[15] Ribeiro, M.T. et al. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2016), 1135–1144.

[16] Rossi, V. et al. 2014. Addressing the implementation problem in agricultural decision support systems: The example of vite.net®. *Computers and Electronics in Agriculture*. 100, (2014), 88–99. DOI:https://doi.org/10.1016/j.compag.2013.10.011.

[17] Sinha, R. and Swearingen, K. 2002. The role of transparency in recommender systems. *CHI'02 extended abstracts on Human factors in computing systems* (2002), 830–831.

[18] Tayyebi, A. et al. 2016. SmartScape™: A web-based decision support system for assessing the tradeoffs among multiple ecosystem services under crop-change scenarios. *Computers and Electronics in Agriculture*. 121, (2016), 108–121. DOI:https://doi.org/10.1016/j.compag.2015.12.003.

[19] Wachowiak, M.P. et al. 2017. Visual analytics and remote sensing imagery to support community-based research for precision agriculture in emerging areas. *Computers and electronics in agriculture*. 143, (2017), 149–164.

[20] Wang, J. et al. 2019. DeepVID: Deep Visual Interpretation and Diagnosis for Image Classifiers via Knowledge Distillation. *IEEE Transactions on Visualization and Computer Graphics*. 25, 6 (Jun. 2019), 2168–2180. DOI:https://doi.org/10.1109/TVCG.2019.2903943.
[21] Yosinski, J. et al. 2015. Understanding Neural Networks Through Deep Visualization. (Jun. 2015).