Why Explain a decision made by AI? Explainable AI (XAI)

Rodrigo Almeida

Abstract—This paper aims to examine the literature on transparency for artificial intelligence (AI) decisions. With the increasing use of AI in various critical applications, it is essential to understand how these decisions are reached. Explainable Artificial Intelligence (XAI) has emerged as a response to this challenge, aiming to make AI decisions more transparent and interpretable. It offers a set of tools and techniques that can be used to explain the decisions and predictions made by AI systems with clarity and transparency. This is in contrast to ”Black Box” algorithms that output reults directly from data, without any explanation of how or why a specific decision was made. Even the developers of these models are often unable to explain the reasoning behind the model’s decision.

1 Introduction

ARtificial intelligence is increasingly becoming a part of our daily lives, being utilized in various critical applications, ranging from healthcare to finance and even in the military sector. AI models are relied upon for tasks such as medical diagnoses, self-driving cars, and decision-making in the military. As AI technology advances, our dependence on these models increases, making it more important to understand how they make their decisions. For critical applications, it is essential that AI systems are not only accurate and efficient but also transparent, interpretable, and explainable. This is because a wrong prediction can have severe consequences.[1] Decisions made by AI systems can significantly impact people’s lives, and therefore, it is crucial to comprehend how these decisions are reached.[3] This challenge is not only technical but also social. Policymakers, regulators, and the general public are increasingly concerned about the impact of AI systems on society.[6] The consequences of an incorrect prediction will vary depending on the field. In computing and business, a wrong prediction might lead to misleading recommendations, affecting the revenue of such businesses. In critical sectors like healthcare, however, such predictions could risk human life. It is therefore essential to open the ”black box” and understand how AI systems reach their conclusions. Doing so is crucial in ensuring accountability, transparency, and ethical use of these technologies, promoting public trust and regulatory compliance.[3]

In Figure 1, we can easily understand the main differences between the ”Black Box” and XAI.[2]

2 Definitions

Explainable AI (XAI) is a field focused on increasing human understanding of AI systems. Although terms like transparency, interpretability, and explainability are frequently used together, they can have different meanings, and there are differences between these theories.[4]

- Transparency: It is the ability to understand the inner workings of an AI system. It is the degree to which a model deemed transparent can be understood easily on its own. It is the opposite of a ”black box”, indicating that the models’s inner workings are understandable[1].

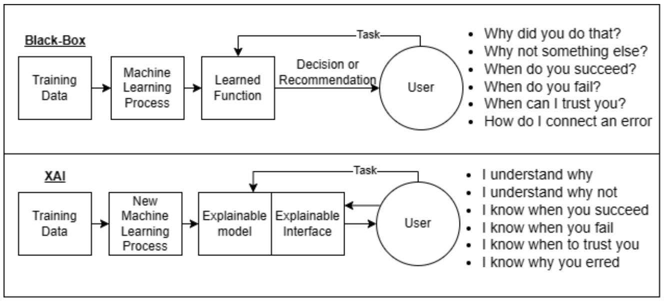

Fig. 1: Black box vs XAI

- Interpretability: It is the ability to explain complex concepts or results in a way humans can easily understand[7].
- Explainability: It is the concept that provides explanations as a bridge between AI systems and humans. It involves the ability to create AI systems that are not only accurate but also that people can understand.[7].
- Explainable AI (XAI): It is a field focused on increasing human understanding of AI systems. It is a set of tools and techniques that can be used to explain the decisions and predictions made by AI systems.[4]

3 Overview of Explainable AI Algorithms

Researchers have developed many algorithms to explain AI systems. These explanations could be categorized into two primary groups: self-interpretable models refer to the algorithm model itself or a representation of the algorithm that can be directly read and comprehended by a human. In this case, the model itself serves as the explanation. On the other hand, post-hoc explanations are descriptions, explanations, or models of the algorithm often produced by separate software tools. These tools aim to provide an understanding of how the algorithm operates. Post-hoc explanations are handy for algorithms whose inner workings are not entirely transparent, as they can be employed to generate insights without requiring deep knowledge of the algorithm’s internal mechanisms. Instead, they rely on querying the algorithm for outputs based on selected inputs.[8].

3.1 Self-interpretable Models

Models that are self-interpretable are those that serve as the explanation themselves. They not only describe the entire model but also go through each input, replicating it on the self-interpretable model, justifying each decision. Common self-interpretable models are decision trees and regression models. Ongoing research aims to create more interpretable models that surpass the accuracy of basic decision trees and regression models. These newer models include decision lists, decision sets, prototypes (representative samples of each class), feature combination rules, Bayesian Rule Lists, additive decision trees, and improved versions of decision trees.[8] Some sources suggest a trade-off between accuracy and interpretability, with self-interpretable models being less accurate than post-hoc models. In this case, the challenge is to balance the precision of the model with what it means to humans. However, scholars like Rudin[11] and Radin[9], in their research called ”Why Are We Using Black Box Models in AI When We Don’t Need To? A Lesson From an Explainable AI Competition”, disagree, explaining that a trade-off between accuracy and interpretability is not always necessary. They argue that interpretable models can be used without sacrificing accuracy. [8]

3.2 Post-hoc Explanations

Post-hoc explanations are descriptions, explanations, or models of the algorithm often produced by separate software tools. These tools aim to provide an understanding of how the algorithm operates. Post-hoc explanations are handy for algorithms whose inner workings are not entirely transparent, as they can be employed to generate insights without requiring deep knowledge of the algorithm’s internal mechanisms. Instead, they rely on querying the algorithm for outputs based on selected inputs.[8]. In other words, the explanations are generated after the model has made the decision or prediction and not during the training phase. Post-hoc explanations can be divided into two categories: local explanations and global explanations.

Table 1 shows a comparison between local and global explanations in XAI.

3.3 Local Explanation

Local explanations explain a subset of inputs. It focuses on explaining the output for a specific data point. The most common type is a per-decision or single-decision explanation, which explains the output or decision on a single input.[8] These explanations focus on a single prediction and aim to answer the following question: ”Why did this model make that particular prediction for that specific task?” There are some techniques used for local explanation. Some of the techniques used for local explanation are:

Aspect	Local Explanation	Global Explanation
Scope	Explains individual predictions.	Focus on understanding the model’s behaviour across the entire dataset.
Granularity	Aims to explain why a particular decision was made.	Higher level overview of the model’s behaviour, bringing up general trends and feature importance across the dataset.
Techniques	LIME, SHAP, individual instance inspection.	PDPs, ICE, SP-LIME, TCAV, decision sets, and summary of counterfactual rules.[8]
Use Cases	Focus on explaining why a specific prediction was made, especially in critical applications like healthcare or finance.	Focus on identifying biases, ensuring fairness, and gaining an overall understanding of the model’s behaviour for regulatory compliance and model improvement.
Interpretability	Easier to understand individual cases focusing on the specific explanation.	Wider view of the model makes it valuable for stakeholders and policymakers seeking a general understanding of the AI system.

TABLE 1: Comparison of Local and Global Explanations in XAI

3.3.1 LIME - Local Interpretable Model-agnostic Explainer

LIME takes a specific decision made by an ML model and examines nearby data points, creating a simplified and interpretable model representing a locally made decision[8]. The default model is logistic regression. It is often used to explain image classification models. LIME breaks down the image into smaller regions called superpixels when dealing with images. It then explores combinations of these superpixels, omitting some and replacing some with black. By doing this, LIME aims to understand and explain how the model’s decision is influenced by specific parts of the image.

3.3.2 SHAP - SHapley Additive exPlanations

SHAP is based on game theory concepts and can explain the predictions by calculating the contribution of every feature to the prediction. It explains how each feature impacts every model’s prediction by considering their interaction consistently and fairly. This helps increase the interpretability by explaining the logic behind every prediction.[10]

3.4 Global Explanations

Global explanations produce post-hoc explanations throughout the whole dataset. It outlines the process used

for the decision-making, mentioning tendencies,
features and possible biases that the model should have learned from the data. Global explanations are essential to understanding the model's behaviours and ensuring fairness, identifying biases, and increasing trust in AI systems.[8]
Such context in XAI helps stakeholders cope with potential issues by understanding the impact and implications of decisions made by a model. Some of the techniques used for global explanation are:

3.4.1 PDPs - Partial Dependence Plots

PDPs show the change in the response when a value changes in the feature, showing the relationship between the feature and the response. It is a global method, meaning that it considers all the data points in the dataset. The advantage of PDPs is that it is easy to understand and can be used for any model. Especially useful when trying to understand the relationship between a feature and the response and how the model behaves for individual features.[8]

3.4.2 ICE - Individual Conditional Expectation Curves

Individual conditional expectation curves are a more user-friendly way to understand how a feature influences a prediction for a single instance, by doing so, it makes it easier to understand how the model behaves for a specific case.

3.4.3 TCAV - Testing with Concept Activation Vectors

This is a global algorithm designed to explain neural networks in a way that is easier to understand. It represents the neural network state using a linear interpretation of the internals of deep learning.

3.4.4 CAVs - Concept Activation Vectors

Concept Activation Vectors are user-friendly concepts used by TCAVs to explain neural networks. It is the numerical representation of a concept, making it more understandable to humans.

3.4.5 Decision sets

Unlike black-box models, decision sets capture the decision-making process by generating a set of rules outlining the conditions in which the model predicts specific outcomes, helping understand the model's decision boundaries.

4 Interpreting Predictions - A Case Study with the ”Adult Income” dataset

As a case study, the ”Adult Income” dataset[5] will be used to predict whether an individual earns more than €50,000/year based on 14 features. This dataset was extracted by Barry Becker [5] and is often cited in machine learning literature and research papers. Table 2 shows the 14 features listed by name, role type and demographic.

- **Age:** represents the age of a person, it is an important factor in predicting income because age can reflect an individual's level of experience and earning potential.

Variable Name	Role	Type	Demographic
age	Feature	Integer	Age
work class	Feature	Categorical	Income
fnlwgt	Feature	Integer	—
education	Feature	Categorical	Education Level
education-num	Feature	Integer	Education Level
marital-status	Feature	Categorical	Other
occupation	Feature	Categorical	Other
relationship	Feature	Categorical	Other
race	Feature	Categorical	Race
sex	Feature	Binary	Sex
capital-gain	Feature	Integer	—
capital-loss	Feature	Integer	—
hours-per-week	Feature	Integer	—
native-country	Feature	Categorical	Other
income	Target	Binary	Income

TABLE 2: Variable Descriptions

- **Work class:** can be categorized as Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, and Never-worked. It refers to the type of employment. Employment status is also essential, as income levels often vary according to employment status.
- **Fnlwgt:** final weight represents the number of people the census believes the entry represents.
- **Education Level:** represents the highest level of education an individual has obtained. It is divided into Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, and Preschool. The education level is a key predictor of income since individuals with higher education often have access to better-paying job opportunities.
- **Education Num:** a numerical representation of the educational level. This also helps understand the education obtained by an individual.
- **Marital Status:** describes the marital status of the individual. It is divided into Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, and Married-AF-spouse.
- **Occupation:** specifies the type of education the individual is involved with. It is divided into Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, and Armed-Forces. Occupation is also a critical factor that influences income, as income levels are different depending on occupation.
- **Relationship:** describes the relationship status of an individual, which can have an impact on household income and financial stability. It is divided into Wife, Own-child, Husband, Not-in-family, Other-relative, and Unmarried.
- **Race:** indicates the race of the individual. It should not directly affect income but could be linked to other socioeconomic factors that can impact earnings. It is divided into White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, and Black.
- **Sex:** specifies the gender of the individual. It can influence income for reasons such as wage gaps and

job segregation. It is divided into females and males.

- **Capital Gain:** represents capital gains for the individual. It contributes to overall income.
- **Capital Loss:** represents capital losses for the individual. It can impact the individual's financial situation.
- **Hours Per Week:** indicates the number of working hours per week. The number of hours worked in a week influences the income directly.
- **Native Country:** indicates the individual's country of origin. Depending on the country of origin, the individual may be impacted by fewer or more opportunities. It is divided into United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad & Tobago, Peru, Hong, and Holand-Netherlands.
- **Income:** the target variable indicating whether the individual's income exceeds $50,000 per year. **This is the variable we are trying to predict**.

4.1 Using Python to Interpret Predictions

In this section, the predictions made by a machine learning model using the "Adult Income" dataset will be interpreted using Python. The library `pandas` will be used for data handling, `sklearn` for building the model, and `shap` and `lime` for explainability.

4.1.1 Loading and Preparing the Dataset

Using `pandas`, the first step is to load the data and preprocess it in a way that handles missing values and splits the dataset into features `X` and target `y`.

4.1.2 Training a Machine Learning Model

To train the model, `DecisionTreeClassifier` from scikit-learn will be used for the classification task. The model will be trained using the training set and then used to make predictions on the test set.

```
# Train the Model
decision_tree_model = DecisionTreeClassifier(
    random_state=42)
decision_tree_model.fit(X_train, y_train)
# Make predictions
y_pred = decision_tree_model.predict(X_test)
```

Listing 1: Training a Decision Tree Classifier

4.1.3 Evaluating the Model

```
# Evaluate the classifier
accuracy = accuracy_score(y_test, y_pred)
print(f"Decision Tree Accuracy: {accuracy}")
```

Listing 2: Evaluating the Model

To evaluate the model, `accuracy_score` from scikit-learn will be used. The accuracy score is the fraction of predictions the model got right. It returns the accuracy of the model, which is 0.81 in this case, meaning that the model got 81% of the predictions right.

4.1.4 Global Explanation with Feature Importance

To understand the overall model behaviour, feature importance can be used. It is a global explanation method that can be used to understand the model's behaviour across the entire dataset. It can be used to identify biases, ensure fairness, and gain an overall understanding of the model's behaviour for regulatory compliance and model improvement. [8] In this case, the feature importances will help understand which features are more critical for the model. The higher the value, the more important the feature is for the model.

```
# Get feature importance
importance = model.feature_importances_
# Plot feature importance
plt.barh(X.columns, importance)
plt.xlabel("Feature Importance")
plt.ylabel("Feature")
plt.show()
```

Listing 3: Feature Importance

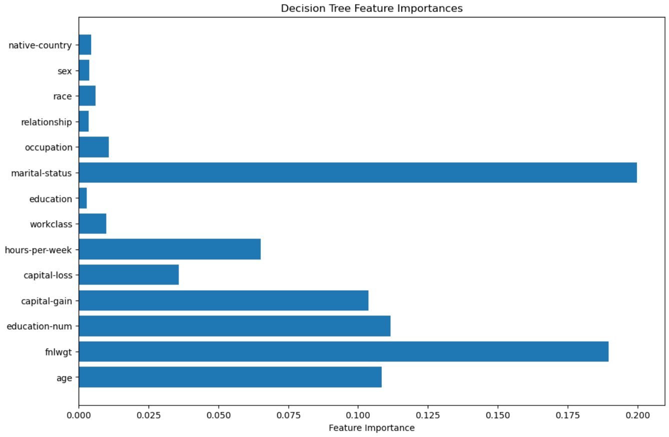

Fig. 2: Decision Tree Feature Importance

Figure 2 shows the output of code from Listing 3 for the "Adult Income" dataset. We can see that the most important features are `marital-status`, `fnlwgt`, `education-num`, `age` and `capital-gain`.

4.1.5 Local Explanation with SHAP

To understand the model's behaviour for a specific instance, SHAP (SHapley Additive exPlanations) can be used. It is a local explanation method that can be used to explain individual predictions. It explains how each feature impacts every model's prediction by considering their interaction consistently and fairly. This helps increase the interpretability by explaining the logic behind every prediction.[10] It explains each feature, showing how each feature contributes to the prediction.

```
shap_values = shap.TreeExplainer(
    decision_tree_model).shap_values(X)
shap.summary_plot(shap_values, X)
```

Listing 4: SHAP Explainer

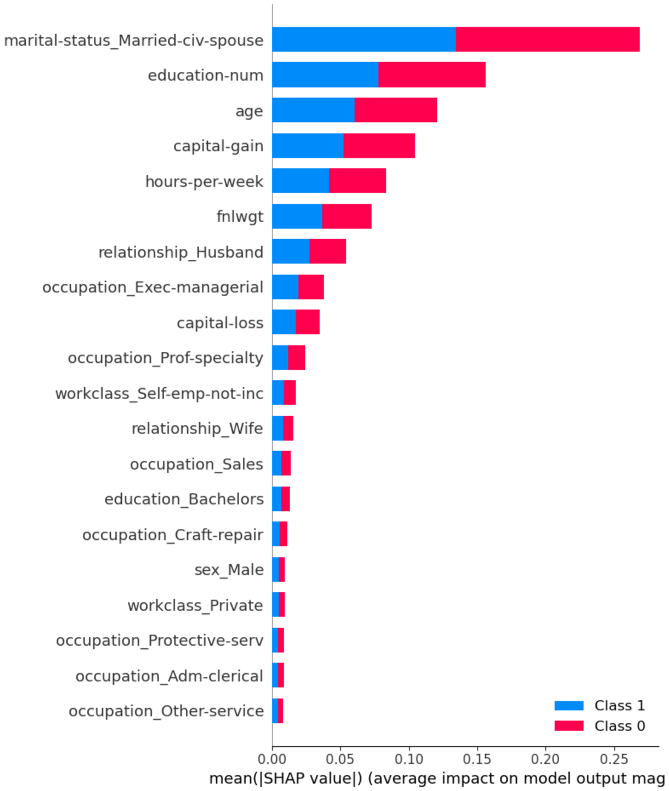

Fig. 3: SHAP Decision Plot

Note: In this SHAP summary plot, 'class 0' corresponds to incomes '≤€50,000', and 'class 1' corresponds to incomes '>€50,000'.

Figure 3 shows the output of code from Listing 4 for the "Adult Income" dataset. It demonstrates the key features influencing the model's predictions for the higher income bracket (class 1). The `marital-status` is shown as the feature with the highest mean SHAP value of 0.1345, suggesting a strong association between being married and predicting higher income. Regarding educational background, `education-num` with a mean SHAP value of 0.0779 suggests a considerable influence, meaning that the higher the education level, the higher the income. Similarly, `age` of the individual, with a mean SHAP value of 0.0603, indicates that older individuals tend to fall into the higher income category according to the model. Financial indicators like `capital-gain` with a mean SHAP of 0.0524, and `hours-per-week` with 0.0417 are also prominent. The more an individual earns from the capital gains, the more likely they are to fall into the higher income category. Similarly, the more hours an individual works per week, the more likely they are to fall into the higher income category. Other features like `fnlwgt`, `relationship_Husband` and various occupational roles such as `occupation_Exec-managerial`, `occupation_Prof-specialty` have smaller but significant positive SHAP values, indicating that these features also contribute to the model's predictions for the higher income category.

4.1.6 Local Explanation with LIME

LIME (Local Interpretable Model-agnostic Explainer) can also be used to explain individual predictions. It explains each feature, showing how each feature contributes to the prediction.

```
# Extracting feature names and weights from the LIME explainer
feature_names, weights = zip(*exp.as_list())
# Convert to list
feature_names = list(feature_names)
weights = list(weights)
# Check the data format
print(”The feature names are:”, feature_names)
print(”Weights:”, weights)
# Creating a bar plot
plt.figure(figsize=(8, 6))
sns.barplot(x=weights, y=feature_names, palette='viridis')
plt.title('Contribution of each feature to the Prediction')
plt.xlabel('Weight')
plt.ylabel('Feature')
plt.show()
```

Listing 5: LIME Explainer

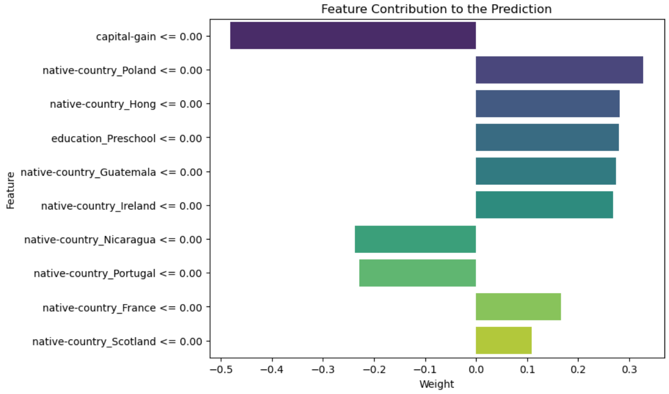

Fig. 4: LIME Decision Plot

Figure 4 shows the output of the code from Listing 5 for the "Adult Income" dataset. According to the model's prediction, this output highlights the top ten features that significantly influence an individual's income category. Each feature is listed alongside a numerical weight, which quantifies the feature's influence on the prediction. The feature `capital-gain ≤ 0.00` has the most significant negative weight (-0.4817176518811231), suggesting that a strong link between having no or minimal capital gain and earning an income that is less than €50,000 annually. On the other hand, features such as `native-country_Poland ≤ 0.00`, `native-country_Hong ≤ 0.00`, and `education_Preschool ≤ 0.00` has positive weights (0.3280648148846505, 0.2812235627400507, and 0.27962517534951975, respectively), suggesting that the absence of these conditions (e.g., not being from Poland, not being from Hong, not having preschool education) tends to be associated with earning a higher income. This LIME analysis provides a localized interpretation of a specific instance, offering insights into why the model might have predicted a particular income category for an individual based on their feature values. Other features, such as `native-country_Guatemala ≤ 0.00` and `native-country_Ireland ≤ 0.00`, also positively influence predictions towards a higher income, but their impact is somewhat less when compared to the other features. The negative weights of `native-country_Nicaragua` $\leq$ `0.00` and `native-country_Portugal` $\leq$ `0.00` indicates a tendency towards predicting lower income levels for individuals from these countries.

5 Conclusion

This literature review emphasizes the significance of Explainable Artificial Intelligence (XAI) and its importance in various sectors. The accuracy and efficiency of AI systems alone are not enough; they also need to be transparent, interpretable, and explainable. This is particularly important for critical applications where the consequences of a wrong prediction can be severe. The review discusses the challenges posed by black box models that are difficult to understand and explain. It highlights the importance of explainability in AI systems, stressing the need for transparency and interpretability. XAI has emerged as a response to these challenges, aiming to make AI decisions more transparent and interpretable. The review further differentiates between transparency, interpretability, and explainability, explaining how each contributes to making AI more understandable and accountable. The case study using the "Adult Income" dataset demonstrates how XAI can be used to interpret predictions made by a machine learning model. Tools like SHAP and LIME provide insights into the model's behavior, helping us understand why the model made a particular prediction for a specific individual.

As AI becomes more integrated into our lives, the need for explainability will only increase. It is not only a technical challenge but also a social one. "Why Explain a Decision Made by AI?" is not only a question but also a call to action for the AI community to make AI more transparent, interpretable, and explainable. The future of AI, as discussed in this review, is one where AI systems are not only accurate and efficient but also transparent, interpretable, and explainable, aligning with society's needs, values, and ethical considerations.

References

[1] Amina Adadi and Mohammed Berrada. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)". In: *IEEE Access* 6 (2018), pp. 52138–52160. DOI: 10.1109/ACCESS.2018.2870052. URL: https://doi.org/10.1109/ACCESS.2018.2870052.

[2] *AI concept vs. XAI Concept*. Image depicting the difference between AI and XAI concepts. URL: https://www.darpa.mil/ddm_gallery/xai-figure2-inline-graphic.png.

[3] P. Angelov et al. "Explainable artificial intelligence: an analytical review". In: *WIREs Data Mining and Knowledge Discovery* 11 (5 2021). DOI: 10.1002/widm.1424.

[4] Anna Markella Antoniadi et al. "Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review". In: *Applied Sciences* 11.11 (2021), p. 5088. URL: https://doi.org/10.3390/app11115088.

[5] Barry Becker and Ronny Kohavi. *Adult*. UCI Machine Learning Repository. 1996. URL: https://archive.ics.uci.edu/dataset/2/adult.

[6] Finale Doshi-Velez and Been Kim. *Towards A Rigorous Science of Interpretable Machine Learning*. 2017. arXiv: 1702.08608 `[stat.ML]`.

[7] Leilani H. Gilpin et al. *Explaining Explanations: An Overview of Interpretability of Machine Learning*. 2019. arXiv: 1806.00069 `[cs.AI]`.

[8] P Jonathon Phillips et al. *Four principles of explainable artificial intelligence*. NIST Interagency or Internal Report (NISTIR) 8312. 2020. URL: https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8312.pdf.

[9] Joanna Radin. *Joanna Radin*. MIT press profile. URL: https://hdsr.mitpress.mit.edu/user/joanna-radin.

[10] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 1135–1144. ISBN: 9781450342322. DOI: 10.1145/2939672.2939778. URL: https://doi.org/10.1145/2939672.2939778.

[11] Cynthia Rudin. *Cynthia Rudin - Google Scholar Citations*. Google Scholar profile. URL: https://scholar.google.com/citations?user=mezKJyoAAAAJ&hl=en.