Measuring Uncertainty and Explainability - Drug Consumption Classification Dataset

Manju Shettar, Tanay Khajanchi, Adarsh Pinjala

June 2024

1 Introduction

Machine Learning models have greatly improved in their predictive capabilities in the 21st century. Their functionality and use cases have evolved into conversational agents, strong decision making products, and independent applications themselves. However, the deployment of these models into high-level solutions not only requires these models to perform with highly accurate predictions, but also explainable and confident actions. As machine learning models have advanced rapidly, so have the requirements for model interpretability and uncertainty quantification.

This paper focuses on extracting uncertainty and explainability metrics from an ensemble of Decision Trees. By analyzing a Drug Consumption Classification dataset to predict cannabis usage on various demographical and personality attributes, we were able to cohesively measure the uncertainty and generate explanations from our model. In order to provide transparency and confidence in the model's predictions, we've graphed and plotted these metrics from our model ensemble.

Specifically, we used the Local Interpretable Model-agnostic Explanations (LIME) method to generate complete explanations of the model's decision making. These methods help visualize the individual contributions by each feature to the final classification output by the model. One of the focuses of this paper is a thorough analysis of the interactions between features during the model's training and inference stages. Evaluating the impact of using these features in this model uncovers how these features influence the likelihood of cannabis usage. By analyzing the prediction variance between the instances of the Decision Tree Ensemble, we quantify the uncertainty of the model's predictions. Uncertainty helps us quantify how confident the model's decision making is.

2 Motivation

Rapid advancements in machine learning and the adoption of advanced models in Artificial Intelligence systems has led to the need for transparent and explainable models. This is especially true for critical sectors such as healthcare, criminal justice and financial decision making, where normally

humans would go through data points to make their own classifications. All of these sectors deal with highly personal information for very important tasks. In order for users to confidently trust models with their data and the predictions the models make, users must understand why and how these predictions are reached.

Going specifically into the context of substance use prediction, models must be both accurate and interpretable to provide insight for the corporations and policy makers who utilize them. Understanding the model's interactions with different input features, like ethnicity, age and gender, is important to find inherent biases in datasets.

This paper aims to define methods that make models and the datasets they work with interpretable and confident, while being accurate. The main goals are as follows:

- **Enhancing Model Interpretability**: Decision Tree ensembles, also known as Random Forests, allow for gathering predictive and model weight distributions per instance. We employ interpretability techniques, such as LIME, to generate explanations of the model's actions and interactions with input features.
- **Quantifying Uncertainty**: Uncertainty definitions are split into two; aleatoric uncertainty, which is represented by uncertainty inherent in the dataset, and epistemic uncertainty, the model's lack of confidence. This metric not only helps identify which instances of the Decision Tree ensemble are less reliable, but it gives us a bigger picture of the overall variance between similar models and their classification of the same data.

3 Results

This section analyzes the results from our model's performance, along with a thorough dive into the metrics extracted from uncertainty and explainability measurements.

3.1 Feature Analysis

Understanding the interactions between features and the model's weights is one of our main goals. The dataset's features are as follows:

- **ID**: the number of record in original database. (used for reference only)
- **Age**: Age is the age of the participant and is represented by ranges; e.x. 18-24, 25-34.
- **Gender**: Gender is the gender of the participant, where 0.48246 represents female, and -0.48246 is male.
- **Education**: Education is the education level of the participant and is represented by numerical values representing university degrees, some college, left school, etc.
- **Country**: Country is country of current

residence and is represented by numerical values, e.x. Canada is 0.24923, USA is -0.57009.

- **Ethnicity**: Ethnicity is the ethnicity of the participant.
- **NScore**: Nscore represents Neuroticism. Individuals who score high on this personality score tend to experience feelings such as anxiety, worry and fear.
- **Escore**: Escore represents how outgoing and social a person is.
- **Oscore**: Oscore represents how open-minded a person is.
- **Ascore**: Ascore refers to agreeableness.
- **Cscore** Cscore refers to conscientiousness and self-discipline.
- **Impulsive** Impulsive refers to impulsiveness.
- **Sensation (SS)** SS is real world sensation measured by sensory receptors.

Straightaway, it becomes evident that it's difficult to analyze these raw feature values. Encoding these variables into values that our model can analyze is the first step of feature analysis. Our methodology included defining a set of floating point values to represent each feature value range and then encoding the original data using this set. For example, gender was transformed from the seemingly arbitrary floating point values to 0 and 1, for male and female. This encoding process makes the values more transparent to human users, leading to complete explanations.

Furthermore, personal attributes like ethnicity, education level, and age can be considered **protected attributes**. Ethically, these attributes are considered to require protection from discrimination or bias. Anti-discrimination laws and regulations exist to ensure fairness and equality in automated decision-making processes. We analyze the inclusion of these features in our discussion sections.

Along with these features, is a report of 18 drugs, including a false positive named Semeron, to identify over-claimers. For each participant, a measure for how often each drug is used was recorded. These raw classes are as follows:

- CL0 - Never Used
- CL1 - Used over a Decade Ago
- CL2 - Used in Last Decade
- CL3 - Used in Last Year
- CL4 - Used in Last Month
- CL5 - Used in Last Week
- CL6 - Used in Last Day

We focused on cannabis use in our project and constrained the data into two classes - frequent use and infrequent use. The input values that were transformed into class 0 or 'infrequent use' were CL0 (Never Used), CL1 (Used over a Decade Ago), CL2 (Used in Last Decade), and CL3 (Used in Last Year). The values that were transformed into frequent use were CL4 (Used in Last Month), CL5 (Used in Last Week), and CL6 (Used in Last Day). The 'use' attribute is now a single bit value, 0 for infrequent use and 1 for frequent use.

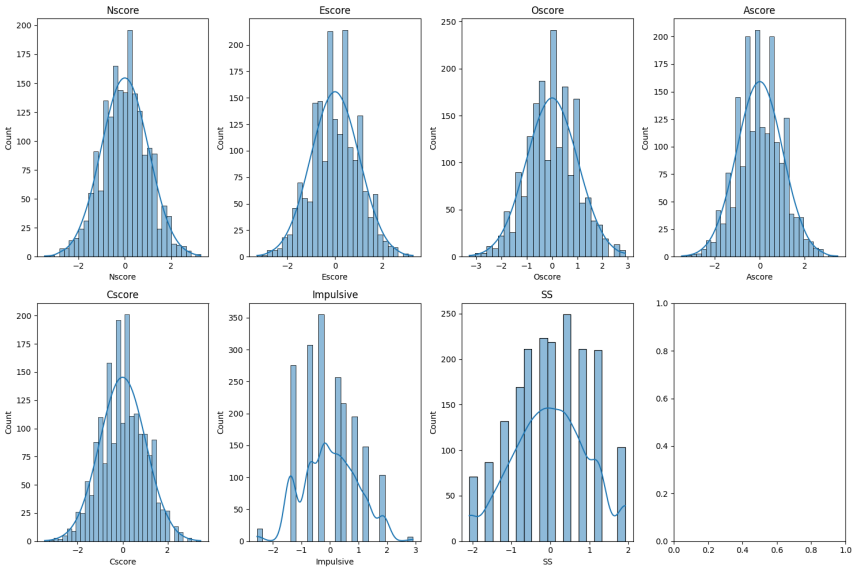
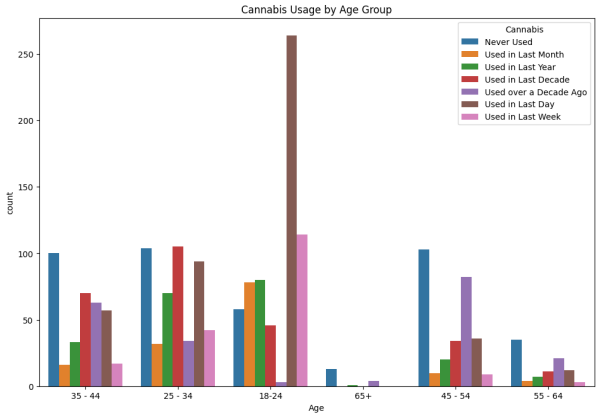


Figure 1: Feature distributions of Nscore, Escore, Oscore, Ascore, Cscore, Impulsive and SS personality features

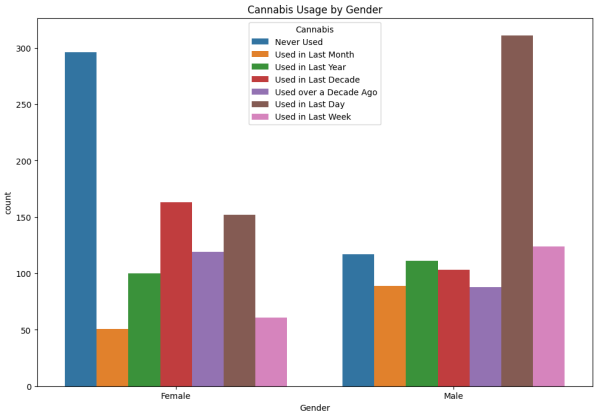
Our feature analysis begins with analyzing the distribution of the features. Figure 1 depicts the ranges of seven different features. From a top level analysis of the graph, we assume that the feature value distributions are fairly normal.

Since we are specifically focusing on predicting cannabis use, we generate the corresponding category distributions of participants who used cannabis in Figure 2. Continuing our analysis, several societal factors affect the distribution of cannabis use. Legalization laws, for example, explain the differences in cannabis use by age. Older people are less likely to have used cannabis, but there aren't that many participants in the last two age brackets as well. For the age graph, it seems like the female participants were recorded not to use cannabis as frequently as the male participants did. When these values are encoded, a lot more female participants will have their cannabis use feature as infrequent use than males, who are more likely to be classified as frequent use. Education level seems to be a balanced feature, the use ranges widely between education level, and it is uncertain whether or not this feature can be a decision maker in our model's final output just by looking at the distribution. We considered these feature value distributions to be important to the impact of

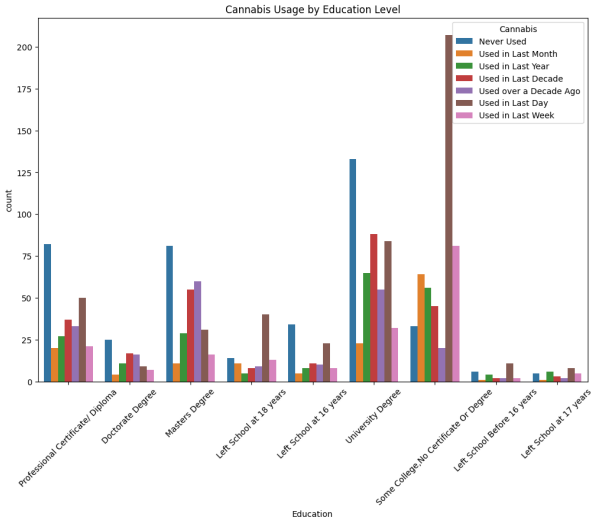
each feature on the model's final prediction.



(a) Use of cannabis by age.



(b) Use of cannabis by gender.



(c) Use of cannabis by education level.

Figure 2: Feature distributions based on cannabis use.

A thorough analysis on the features was completed once we finished training our model. Figure 3 defines a feature importance graph, with numerical values assigned to each feature to denote the impact of the feature's value on the final output of the model.

We made several notable observations from Figure 3. Firstly, we noticed the impact of protected features on the model's final prediction. The country of origin and age features make significant contributions to the model's classification and have high weights within the Decision Tree model. On the other hand, features like gender and ethnicity do not contribute as much importance. This makes for important discussion considering the inclusion on these features in machine learning applications, to ensure both data safety and user trust, while providing a model that is highly accurate.

Furthermore, some personality scores also greatly impact the model's classification as well. Os-

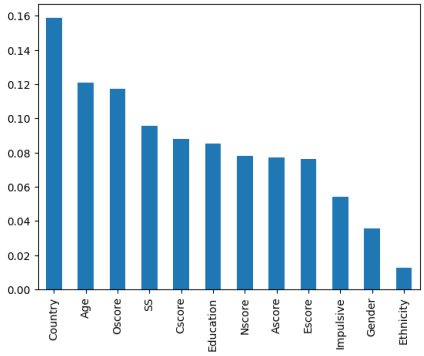


Figure 3: Feature importance graph

core, representing how open-minded a person is, and SS, referring to a perception and sensory test, have high weights assigned to them, while other personality scores seem to be weighted lower. We

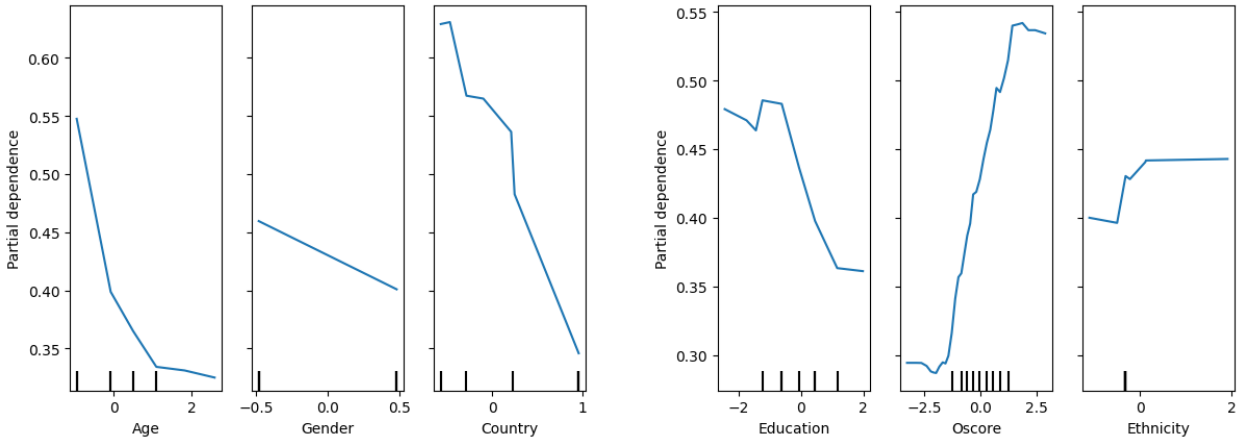


Figure 4: Partial dependency graphs for age, gender, country, education level, oscore, and ethnicity.

analyze Figure 4 as partial dependency graphs for six features; age, gender, country, education level, oscore, and ethnicity. This is an extension on the feature importance graph, with specific insight into the feature value range contribution on the final output. A graph with a positive slope, such as oscore represents that the feature has a positive correspondence with a positive classification that the participant uses cannabis. A negative score represents that higher values of the feature will result in a negative classification, like the partial dependency graph for age.

The steepness of the slopes is also important to consider. The higher the absolute value of the slope is, the higher contribution the feature has to the model's final output, in direct correlation with the feature importance graph in Figure 3. Features like gender and ethnicity have relatively gentle slopes and we analyze this representation to be that the feature values do not have much impact on the classification of the model.

3.2 Explainability

Explainability metrics lie in measuring the feature impact to predictions, essentially an expansion on our feature analysis. By using LIME, we generate a partial explanation, as depicted in Figure 5. LIME generates a composite explanation of a simplified model's results on a set of input data.

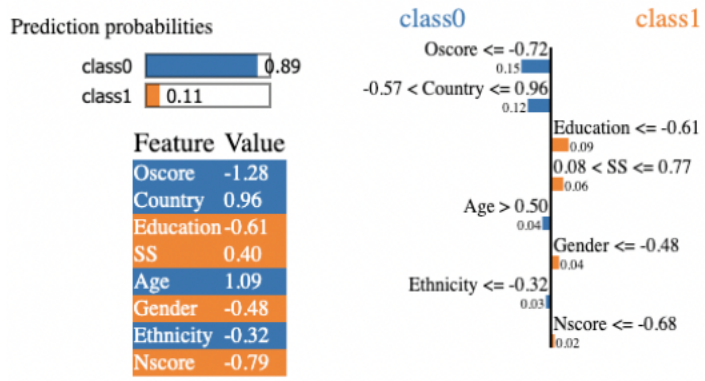


Figure 5: Feature importance graph

For the input data represented in Figure 5, we observe that the model classified the input as class 0 with an 89% prediction probability. LIME also describes how the model weighs each feature. The column of features depicts the value of each feature and the impact of each feature's value on the final class.

3.3 Uncertainty

We measure uncertainty metrics easily using our ensemble of Decision Trees. The most prominent quantification of uncertainty is variance. By recording the variance between model instances in our ensemble, we gain an estimate on how different the instances of the same model are performing when they are exposed to the same data.

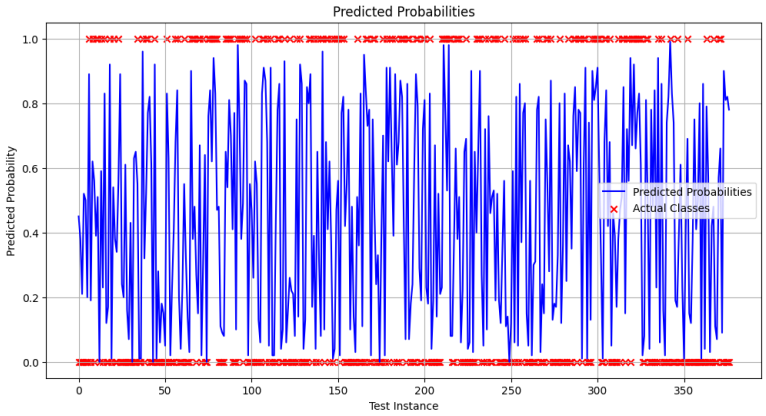


Figure 6: Predicted Probability graph

We analyze the predicted probabilities over 400 instances of our Decision Tree ensemble in Figure 6. The red x's represent the classes for predicted cannabis use. The lines in between represent the model's predicted probability. A predicted probability of including and over 50% will result in class 1, while a predicted probability of under 50% will result in class 0. With a perfectly confident model, we would see two blue flat lines at the decision boundaries represented by the classes. We observe that our model does not perform as well, we can even see some predicted probabilities being close to 50%, essentially the same thing as flipping a coin and getting it correct. The pre-

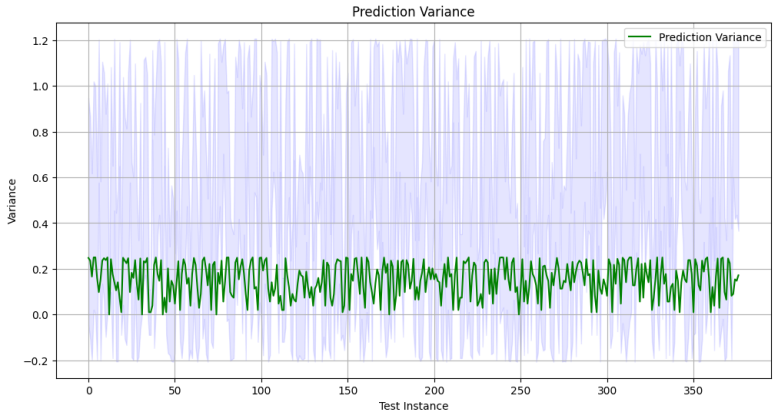


Figure 7: Prediction Variance graph

diction variance graph in Figure 7 depicts the variance within our model ensemble. Once again, with a perfectly confident model, we would see almost zero variance, as all model instances would confidently make predictions based on the same input.

The graph depicts the epistemic uncertainty inherent in the model ensemble, as the instances are predicting differently for the same input data. We analyze that this inherent uncertainty in the model is a representation of the model's confidence about the input data. Without a complete and accurate representation of the input, the model ensemble cannot perform consistently.

4 Discussion - Technological, Social and Policy Implications

With the gathering of uncertainty and explainability metrics, we move on to the discussion section of our project. Our project report defines methods for extracting metrics and in this section, we will expand on the significance of this quantification.

Our work on explainability and feature analysis defines a clear method that extracts the importance of each feature to the model's final predictions. Explainability is defined to be a complete, human-interpretable explanation of a model.

We utilize LIME to complete the first two steps of explainability; data representation and process-

ing. The third step of explainability is explanation generation, which involves language generation. We simplify this process by simply listing the numerical contributions of each feature and their encoded value to the model's class prediction.

We conclude that the inclusion of explainability metrics in feature analysis is crucial to the development of Artificial Intelligence systems. By designing systems that are explainable, corporations are able to design systems with more clarity and give users trust in data safety. Machine Learning models are still 'black box-esque', even to experienced developers and even with explainability methods, it's difficult to analyze why a model makes its decisions. Mathematical models represent data in abstract ways, from forming pixel clusters to defining value ranges. It's crucial for technology to explain itself to increase user trust and confidence.

We even predict that in the future, as the reach for Artificial General Intelligence gets closer, policymakers will begin to petition for systems that explain their thinking. In fact, with the rise of GPT-4, many researchers are calling for a halt on AI development for the next five years in a new field called AI safety. The White House has created an AI Safety and Security Board to advance AI's responsible development and deployment, led by executives like Sam Altman and Sundar Pichai.

As for social implications, our feature analysis resulted in several notable observations. The biggest one is the impact of including protected attributes like education level and age on the model's classification. We observe that the inclusion of some protected features greatly impacted the final classification, sometimes even being decision makers for the model. This is a very gray area of data analysis, as several legal policies like the Civil Rights Act actively protect against discrimination because of race, national origin, age, or genetic information.

We consider uncertainty a separate metric, but the inclusion of uncertainty allows users to view not only the model's final classification, but the amount of uncertainty that was associated with the classification as well. This is very significant for a wide variety of sectors that use machine learning, like medical scanning.

One popular task in this sector is breast cancer screening, where a model is trained on images that vary greatly. The training images often have very little in common and the inclusion of input data for inference only makes this worse. In the real world, data and data formats transform and can contain noise from faulty sensors. If the model considers the noise as normal input, it can lead to incorrect or even out-of-sample classifications, where the model thinks the input data is something entirely different from what it trained on.

Quantifying uncertainty becomes extremely important in this context. It is important to have

a model that is highly accurate, but it is important at the same time to have a model that knows when it is wrong. Our decision tree ensemble was making classifications on near 49-51 prediction probability splits. To the user, this classification appears as a confident prediction when it fact it wasn't.

It becomes crucial to include uncertainty metrics in Artificial Intelligence systems. As autonomous technology advances, math-based decisions receive input from a variety of different sensors in inconsistent data formats. A model that is trained with a specified sensor and data format will perform confidently on similar data recorded on that sensor, but may face complications when exposed to different data representations. This is why it is important to quantify how uncertain a model is when it's looking at some input data.

5 Conclusion

We have defined several methods to extract explainability and uncertainty from a dataset and a model's classifications in this project. By gaining insight into why and how the models classify input data, we are able to make machine learning models more transparent and interpretable to human users.

References

[1] YouTube, "Intro to Deep Learning - Uncertainty in Deep Learning," 2021. [Online].https://www.youtube.com/watch?v=veYq6EWZyVc.

[2] "Over 20 Technology and Critical Infrastructure Executives, Civil Rights Leaders, Academics, and Policymakers Join New DHS Artificial Intelligence Safety and Security Board to Advance AI's Responsible Development and Deployment" Homeland Security, 26 Apr. 2024, [Online]. www.dhs.gov/news/2024/04/26/over-20-technology-and-critical-infrastructure-executives-civil-rights-leaders

[3] "The Role of Protected Attributes in Ai Fairness." Trust Science, 27 Mar. 2024, [Online].www.trustscience.com/resource/blog/archive/the-role-of-protected-attributes-in-ai-fairness.

[4] Gilpin, Leilani H., et al. "Explaining Explanations: An Overview of Interpretability of Machine Learning." IEEE, arXiv.Org, 3 Feb. 2019, [Online].arxiv.org/abs/1806.00069

[5] Begoli, Edmon, et al. "The Need for Uncertainty Quantification in Machine-Assisted Medical Decision Making." Research Gate, Jan. 2019, [Online].www.researchgate.net/publication/330203342_The_need_for_uncertainty_quantification_in_machine-assisted_medical_decision_making