

## 1. InterpreTabNet Synthetic Dataset Experiments

Table 1: AUC (classification task) mean and standard deviation (20 random seed trials) scores on 6 synthetic datasets from (Yoon et al., 2019). Results from the other models are reported from the TabNet paper which includes TabNet (Arik and Pfister, 2020) itself, INVASE (Yoon et al., 2019), L2X (Chen et al., 2018) and Tree Ensembles (Geurts et al., 2006). "Global" refers to the use of only globally salient features. "No Selection" refers to the use of all features without any feature selection.  $\uparrow$  indicates that the higher the score, the better the performance. The corresponding InterpreTabNet  $r_M$  values for each Syn1-Syn6 datasets are as follows: 5250, 4750, 4750, 3250, 5000, 3000. Red bolded numbers denote the best performance for each dataset.

Methods	Syn1	Syn2	Syn3	Syn4	Syn5	Syn6
	AUC $\uparrow$	AUC $\uparrow$	AUC $\uparrow$	AUC $\uparrow$	AUC $\uparrow$	AUC $\uparrow$
InterpreTabNet	<b>.696<math>\pm</math>.005</b>	.885 $\pm$ .003	.899 $\pm$ .004	<b>.790<math>\pm</math>.008</b>	<b>.791<math>\pm</math>.006</b>	<b>.880<math>\pm</math>.003</b>
TabNet	.682 $\pm$ .005	<b>.892<math>\pm</math>.004</b>	.897 $\pm$ .003	.776 $\pm$ .017	.789 $\pm$ .009	.878 $\pm$ .004
INVASE	.690 $\pm$ .006	.877 $\pm$ .003	<b>.902<math>\pm</math>.003</b>	.787 $\pm$ .004	.784 $\pm$ .005	.877 $\pm$ .003
L2X	.498 $\pm$ .005	.823 $\pm$ .029	.862 $\pm$ .009	.678 $\pm$ .024	.709 $\pm$ .008	.827 $\pm$ .017
Lasso-regularized	.498 $\pm$ .006	.555 $\pm$ .061	.886 $\pm$ .003	.512 $\pm$ .031	.691 $\pm$ .024	.727 $\pm$ .025
Tree	.574 $\pm$ .101	.872 $\pm$ .003	.899 $\pm$ .001	.684 $\pm$ .017	.741 $\pm$ .004	.771 $\pm$ .031
Global	.686 $\pm$ .005	.873 $\pm$ .003	.900 $\pm$ .003	.774 $\pm$ .006	.784 $\pm$ .005	.858 $\pm$ .004
No Selection	.578 $\pm$ .004	.789 $\pm$ .003	.854 $\pm$ .004	.558 $\pm$ .021	.662 $\pm$ .013	.692 $\pm$ .015

Synthetic datasets are crucial in determining whether the feature selection process is faithful towards the model's predictions. We use the same synthetic data generation model in TabNet (Arik and Pfister, 2020) and INVASE (Yoon et al., 2019), as well as a sample size of 10K training and 10K testing. The datasets are formulated where a subset of features determines the prediction. For Syn1-Syn3, they only depend on their specified salient features i.e. Syn1 only depends on features 0 and 1. For Syn4-Syn6, the salient features depend on instance-wise features i.e. the output of Syn4, relies on either features 0-1 or features 2-5 depending on the value of feature 10.

Table 1 indicates that InterpreTabNet with  $r_M$  values 5250, 4750, 4750, 3250, 5000, 3000 for each Syn1-Syn6 datasets outperforms existing methods in 4 out of 6 of the synthetic datasets. More notably in all Syn4-Syn6 while remaining competitive in the other two Syn2 and Syn3 datasets. With respect to Syn1-Syn3, InterpreTabNet's competitiveness indicates that it is able to accurately achieve global feature selection. Additionally, for Syn4-Syn6, the SOTA performance indicate that InterpreTabnet is the best at performing global feature selection when instance-wise redundant features are removed.

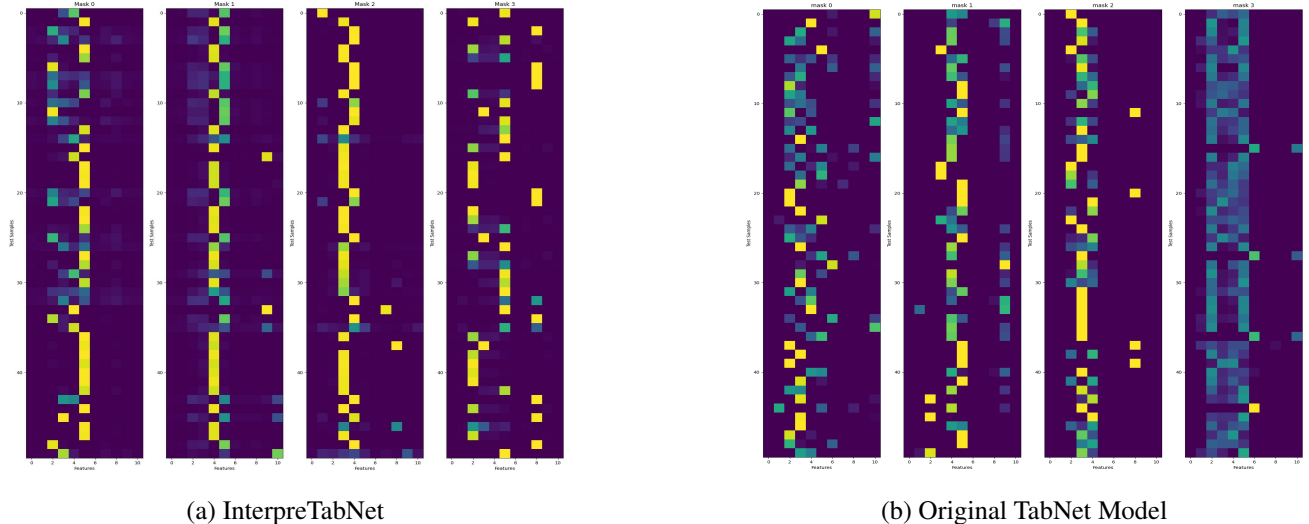


Figure 1: (a) Best performing model using InterpreTabNet  $r_M^* = 4,750$  with an AUC of 0.899 on Syn3 Dataset (Salient Features =  $[X_6 - X_9]$ ). (b) The baseline performance using the Original TabNet model, attaining an AUC of 0.897.

Syn3 uses features 2-5 to generate its predictions. As observed in Figure 1, InterpreTabNet does indeed determine features 2-5 to be the most salient features as shown by the yellow bars in most of its masks. On the other hand, TabNet depicts lower salience levels when conducting its decision process for the prediction, making it harder for interpretation.

## 2. Verifying Correctness of GPT-4’s Analysis on Synthetic Dataset

We verify the correctness of GPT-4’s analysis by prompting it to interpret synthetic datasets generated in the manner from (Yoon et al., 2019). A synthetic dataset (*Syn1*) is considered where input features are generated from an 11-dimensional Gaussian distribution with no correlations across the features ( $X \sim N(0, I)$ ). The label  $Y$  is sampled as a Bernoulli random variable with  $\mathbb{P}(Y = 1|X) = \frac{1}{1+\logit(X)}$ , where  $\logit(X)$  is varied to create a synthetic dataset as follows:  $Syn1 = \exp(X_0, X_1)$ . As shown in the GPT-4 output below, the analysis is indeed robust and trustworthy since it does not show any signs of hallucinations or mistakes hence, verifying the integrity of GPT-4’s ability to interpret feature masks.

**Prompt** Conduct aggregate analysis on the description of the following feature mask. Please output ONLY a dictionary and no other natural language generation when generating the sentence as shown in the in-context example below. Please use single-word classification that encapsulates the meaning of the features if possible.

A synthetic dataset (*Syn1*) is considered where input features are generated from an 11-dimensional Gaussian distribution with no correlations across the features  $X \sim N(0, I)$ . The label  $Y$  is sampled as a Bernoulli random variable with  $\mathbb{P}(Y = 1|X) = \frac{1}{1+\logit(X)}$ , where  $\logit(X)$  is varied to create a synthetic dataset as follows:  $Syn1 = \exp(X_0, X_1)$ . Make sure to analyze the effect of  $Syn1 = \exp(X_0, X_1)$  in your aggregate analysis.

At the 0th step of feature selection, we observe mask 0 with the main feature highlighted as 0 which is  $X_0$ . At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 1 which is  $X_1$ . Lastly, at the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 0 and 1 which are  $X_1$  and  $X_2$ .

Here are in-context examples for few-shot learning.

The Poker Hand dataset is considered, which is a classification of poker hands into 10 distinct types based on attributes like rank of individual cards, suits, and card combinations. There are 50 test samples and 10 features.

At the 0th step of feature selection, we observe mask 0 with the main feature highlighted as 3 which is the rank of the card 2. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 7 which is the rank of card 3. At the 2nd step of feature selection, we observe mask 2 with the main feature highlighted as 6 which is the suit of card 3. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 3, 6, 7 and 10 which are the rank of cards 1 and 3 as well as the suits of cards 3 and 5.

Output: {"Mask 0": "Initially, the rank of card 2 is recognized as a significant predictor in the classification of poker hands. The rank of individual cards seems crucial at this stage.", "Mask 1": "At this step, the rank of card 3 is highlighted, further emphasizing the importance of individual card ranks in determining poker hand types.", "Mask 2": "Here, besides the rank of card 3, the suit of the same card is considered, suggesting that both rank and suit play roles in classifying poker hands.", "Mask 3": "This step combines multiple features, both ranks (cards 1 and 3) and suits (cards 3 and 5), indicating that a combination of these factors provides a broader understanding of poker hand classifications.", "Aggregate": "The rank of cards, specifically card 2 and card 3, appears consistently across different masks, indicating their primary role in poker hand classification. Suits, particularly of card 3, also have importance but are introduced later in the feature selection. The combined importance of both ranks and suits in the final step suggests that while individual card ranks are pivotal in initial classifications, understanding the relationship between card ranks and their respective suits provides a more comprehensive insight into the poker hand types. The consistent presence of card 3’s attributes (both rank and suit) underscores its pivotal role in determining poker hand categories."}

The Forest Cover Type dataset is considered which is a classification of pixels into 7 forest cover types based on attributes such as elevation, aspect, slope, and more. There are 50 test samples and 54 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 24, 29 and 34 which are different soil types. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 34 which is soil type. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 10, 34 and 37 which are wilderness area and soil types. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 0 and 29 which are elevation, and soil type.

---

Output: {"Mask 0": "The initial feature selection identifies that various soil types are significant predictors for the forest cover types. It appears that the characteristics of the soil play a vital role in determining the type of forest cover.", "Mask 1": "At this step, the model emphasizes even more on one particular soil type (34), suggesting it as a primary driver for the classification. This specific soil type may have unique characteristics making it a distinctive feature for certain forest cover types.", "Mask 2": "Here, besides the emphasis on soil type 34, another soil type (37) and wilderness area are considered. This suggests that both the nature of the terrain (wilderness) and the characteristics of the soil together contribute to the forest cover type.", "Mask 3": "This step introduces elevation as a significant predictor while still retaining a soil type. Elevation, combined with soil characteristics, might influence specific flora and subsequently determine forest cover types.", "Aggregate": "Soil type 34 is the most consistent feature across all the masks, appearing in every step, emphasizing its importance in the forest cover type classification. Another soil type, 29, is also consistent, appearing in both mask 0 and 3. The presence of different features across the steps, like wilderness area (mask 2) and elevation (mask 3), indicates that the landscape's nature and the physical height of the terrain are also important but might offer secondary or nuanced insights into forest cover types. The dominant predictor seems to be soil types, particularly soil type 34, suggesting that specific soil characteristics play a crucial role in determining forest cover types. However, aspects like elevation and the nature of the terrain (wilderness area) also contribute to the classification and shouldn't be overlooked."}

**GPT-4 Output** {"Mask 0": "Initially, the feature  $X_0$  is identified as a significant predictor for the dataset. This step suggests the individual effect of  $X_0$  on the label  $Y$  is crucial.",

"Mask 1": "At this step, feature  $X_1$  is highlighted, indicating its individual significance in influencing the label  $Y$ , similar to  $X_0$ 's initial impact.",

"Mask 2": "Here, both features  $X_0$  and  $X_1$  are highlighted together, suggesting that the combined effect of these features significantly impacts the label  $Y$ , beyond their individual contributions.",

"Aggregate": "The progression of feature selection indicates a transition from individual to combined feature importance. Initially,  $X_0$  and then  $X_1$  are identified as significant on their own. However, by the final step, the combination of  $X_0$  and  $X_1$  is recognized as crucial, highlighting the importance of interactions between features in influencing the label  $Y$ . This suggests that while individual features have their significance, the interaction between  $X_0$  and  $X_1$ , especially considering the synthetic dataset's nature ( $\text{Syn1} = \exp(X_0, X_1)$ ), is vital for predicting  $Y$ . The synthetic dataset's structure emphasizes the multiplicative interaction between  $X_0$  and  $X_1$ , illustrating the complex relationship between features in modeling."}

### 3. InterpretTabNet Subgroup Feature Importance Analysis for Varying Subpopulation Characteristics

ADULT CENSUS INCOME MASK FIGURES

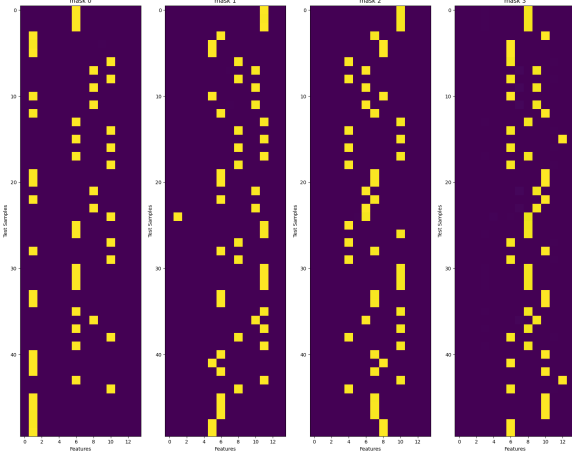


Figure 2: (a) InterpretTabNet (Sex Subgroup: Male)

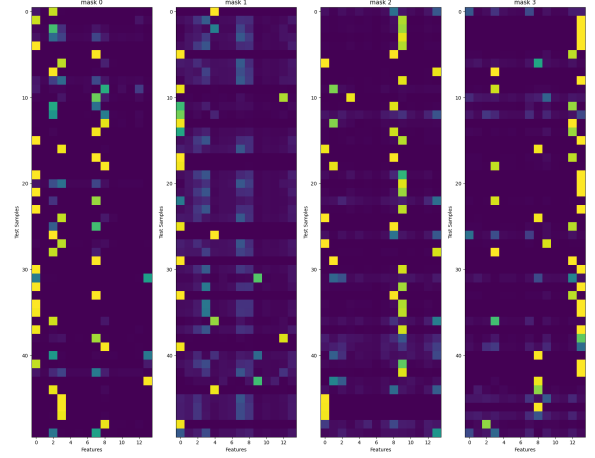


Figure 3: (b) InterpretTabNet (Sex Subgroup: Female)

Figure 4: Feature importance visualization across different education subgroups in InterpretTabNet.

To assess InterpretTabNet’s robustness and reliability, we conducted a subgroup analysis, examining if salient feature identification differs across subgroups. We divided the adult census income dataset into male and female subpopulations, with the feature importance visualization depicted in Figure 4. For males, mask 0 highlighted work class, occupation, race, and capital gain (features 1, 6, 8, and 10, respectively) as most influential for income prediction. Conversely, mask 0 for females pinpointed age, education, and marital status (features 0, 3, and 5) as salient features contributing to prediction.

The salient features match well with known socio-economic factors affecting income. For males, occupation and capital gains highlight the importance of job roles and investments. For females, marital status notably influences income, likely due to the socio-economic dynamics and potential household income sharing.

Subsequent masks for males consistently highlight features 6 (occupation), 8 (race), and 10 (capital gain). For females, mask 1 identifies feature 0 (age), mask 2 both feature 0 (age) and 8 (race), and mask 3 points feature 13 (native country). The minimum overlap in identified features suggests our model’s robustness in identifying differences in important features between different subpopulations of the data.

#### 4. Hyperparameter Guidelines

Hyperparameters such as  $N_d = N_a$ ,  $N_{steps}$ ,  $\gamma$ , and learning rate are tuned in the range per TabNet’s recommendations. In terms of the sparsity regularizer for InterpreTabNet,  $r_M$ , we recommend a smaller range e.g. [0,10000] for datasets with a low to moderate number of features and samples (Adult dataset), and a larger range e.g. [0, 1,000,000,000,000] for datasets with a larger number of features and samples (Higgs dataset). Within the sparsity algorithm itself, parameters such as the number of salient features and feature importance threshold can be adjusted to the user’s preference.

#### 5. Test Accuracy Scores with Mean and Standard Deviation

Table 2: Test Accuracy Scores (mean and standard deviations across 20 random seed trials in %) across Different Models and Datasets with Optimal Mask Regularizer Values ( $r_M$ ) for InterpreTabNet. Our InterpreTabNet achieved substantial improvements in interpretability across all the datasets and remains competitive in terms of accuracy in most datasets.

Model / Dataset	Adult Census	Forest Cover	Poker Hand	Mushroom	Blastchar	Diabetes	Higgs
InterpreTabNet	<b>87.42 <math>\pm</math> 0.55</b>	<b>94.75 <math>\pm</math> 0.53</b>	99.50 $\pm$ 0.48	96.62 $\pm$ 0.35	72.96 $\pm$ 0.56	55.37 $\pm$ 0.47	53.08 $\pm$ 0.56
Original TabNet	85.55 $\pm$ 0.56	94.18 $\pm$ 0.63	99.00 $\pm$ 0.62	99.94 $\pm$ 0.31	76.22 $\pm$ 0.34	56.91 $\pm$ 0.53	52.94 $\pm$ 0.45
XGBoost	86.60 $\pm$ 0.64	92.30 $\pm$ 0.62	75.57 $\pm$ 0.47	99.69 $\pm$ 0.39	77.29 $\pm$ 0.53	<b>61.44 <math>\pm</math> 0.32</b>	<b>72.70 <math>\pm</math> 0.35</b>
LightGBM	86.20 $\pm$ 0.43	86.38 $\pm$ 0.64	78.47 $\pm$ 0.36	<b>100.00 <math>\pm</math> 0.37</b>	<b>77.86 <math>\pm</math> 0.46</b>	60.87 $\pm$ 0.39	72.62 $\pm$ 0.35
TabTransformer	85.09 $\pm$ 0.39	82.55 $\pm$ 0.39	<b>99.81 <math>\pm</math> 0.31</b>	<b>100.00 <math>\pm</math> 0.57</b>	73.17 $\pm$ 0.56	44.45 $\pm$ 0.34	51.97 $\pm$ 0.54
MLP	79.76 $\pm$ 0.65	84.89 $\pm$ 0.56	99.70 $\pm$ 0.56	99.82 $\pm$ 0.56	75.16 $\pm$ 0.61	53.99 $\pm$ 0.46	63.17 $\pm$ 0.36

#### 6. Training Stability of InterpreTabNet vs. TabNet

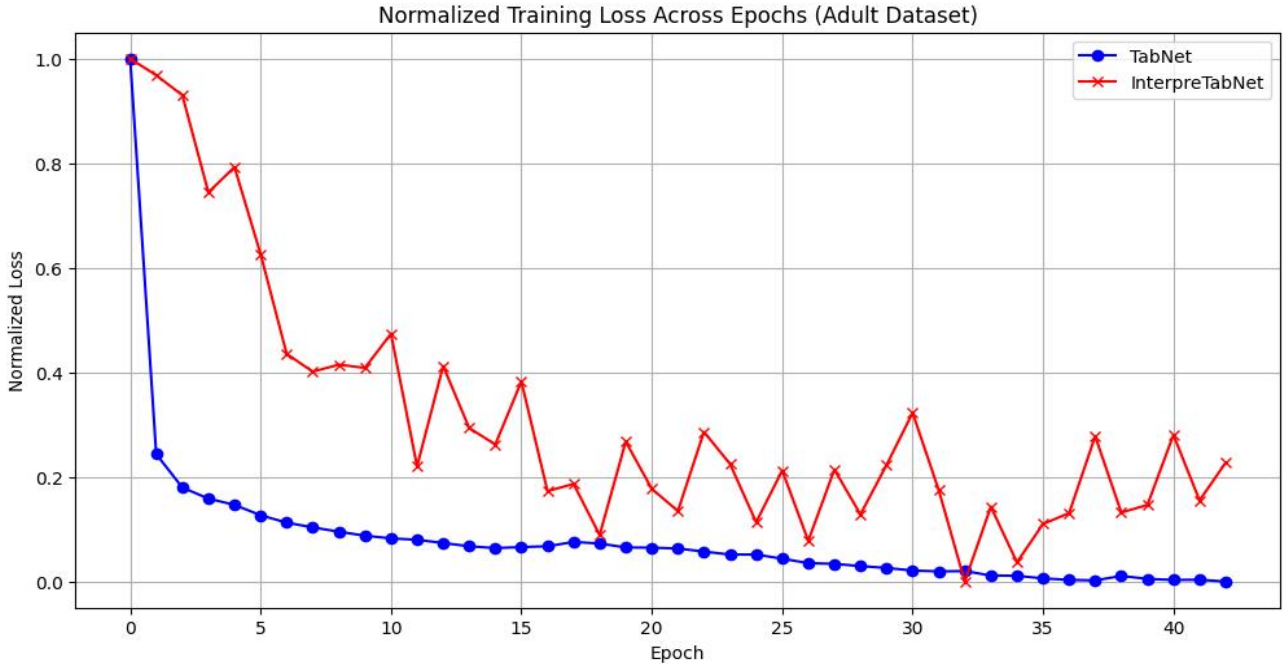


Figure 5: Normalized Training Loss of InterpreTabNet vs. TabNet for the Adult Income Dataset

Figure 5 demonstrates the normalized training loss across epochs for two models: TabNet and InterpreTabNet. TabNet’s loss decreases rapidly in the initial epochs and stabilizes quickly. In comparison, InterpreTabNet exhibits higher variability in loss reduction across epochs, which suggests a more exploratory learning process to determine the salient features. Despite the fluctuations, InterpreTabNet’s loss shows a general downward trend, which is a positive indicator of learning. Overall, while TabNet appears to have a smoother training process, InterpreTabNet remains competitive.

## 7. Survey: Comparing LLM-generated interpretations

Table 3: Human Evaluation Survey on LLM-Generated Interpretations from InterpreTabNet vs. TabNet

Model	Vote Percentage	Number of Votes
InterpreTabNet	60%	12
TabNet	40%	8
Total	100%	20

**Follow-up Human Evaluation Survey** We conducted a follow-up human evaluation survey on 20 Ph.D. and Masters students combined with a machine learning background. The survey was conducted in a blind format where the identities of the models were anonymous to ensure that our data was trusted and not biased. The survey asked: *"Which model provides a more relevant and reasonable interpretation?"*. 60% of respondents preferred InterpreTabnet as the model with a more interpretable output from GPT-4. This deeper evaluation of interpretability validated our motivation that sparse feature selection distilled the complexity of the data into a simpler and understandable form. This allowed GPT-4 to easily comprehend the information to provide a clear and concise output which explain the outcomes to the stakeholders. The survey's results can be found in Table 3.<sup>1</sup> The prompt and GPT-4 output for TabNet on the Adult dataset can be found in Section 7.1.

### 7.1. TabNet: Prompt and GPT-4 Output on Adult Dataset

**Prompt** Conduct aggregate analysis on the description of the following feature mask. Please output ONLY a dictionary and no other natural language generation when generating the sentence, as shown in the in-context example below. Please use single-word classification that encapsulates the meaning of the features if possible.

The Adult Census Income dataset is considered which predicts whether an individual's income exceeds \$50,000 per year. There are 50 test samples and 14 features.

At the 0th step of feature selection, we observe mask 0 with barely any main features highlighted. At most feature 3 which is education is somewhat highlighted. At the 1st step of feature selection, we observe mask 1 with the main features highlighted as somewhat 7 and mainly 11 which are relationship and capital loss. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 3 and 5 which are education and marital-status. Lastly, at the 3rd step of feature selection, we observe mask 3 with difficulties in discerning the important features.

Here are in-context examples for few-shot learning.

The Poker Hand dataset is considered, which is a classification of poker hands into 10 distinct types based on attributes like rank of individual cards, suits, and card combinations. There are 50 test samples and 10 features.

At the 0th step of feature selection, we observe mask 0 with the main feature highlighted as 3 which is the rank of the card 2. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 7 which is the rank of card 3. At the 2nd step of feature selection, we observe mask 2 with the main feature highlighted as 6 which is the suit of card 3. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 3, 6, 7 and 10 which are the rank of cards 1 and 3 as well as the suits of cards 3 and 5.

Output: {"Mask 0": "Initially, the rank of card 2 is recognized as a significant predictor in the classification of poker hands. The rank of individual cards seems crucial at this stage.", "Mask 1": "At this step, the rank of card 3 is highlighted, further emphasizing the importance of individual card ranks in determining poker hand types.", "Mask 2": "Here, besides the rank of card 3, the suit of the same card is considered, suggesting that both rank and suit play roles in classifying poker hands.", "Mask 3": "This step combines multiple features, both ranks (cards 1 and 3) and suits (cards 3 and 5), indicating that a combination of these factors provides a broader understanding of poker hand classifications.", "Aggregate": "The rank of cards, specifically card 2 and card 3, appears consistently across different masks, indicating their primary role in poker hand classification. Suits, particularly of card 3, also have importance but are introduced later in the feature selection. The combined importance of both ranks and suits in the final step suggests that while individual card ranks are pivotal in initial classifications, understanding the relationship between card ranks and their respective suits provides a more comprehensive

<sup>1</sup>Link to the survey: <https://forms.gle/ZbGeXNF1HcSEYsNP7>

---

insight into the poker hand types. The consistent presence of card 3's attributes (both rank and suit) underscores its pivotal role in determining poker hand categories."}

The Forest Cover Type dataset is considered which is a classification of pixels into 7 forest cover types based on attributes such as elevation, aspect, slope, and more. There are 50 test samples and 54 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 24, 29 and 34 which are different soil types. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 34 which is soil type. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 10, 34 and 37 which are wilderness area and soil types. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 0 and 29 which are elevation, and soil type.

Output: {"Mask 0": "The initial feature selection identifies that various soil types are significant predictors for the forest cover types. It appears that the characteristics of the soil play a vital role in determining the type of forest cover.", "Mask 1": "At this step, the model emphasizes even more on one particular soil type (34), suggesting it as a primary driver for the classification. This specific soil type may have unique characteristics making it a distinctive feature for certain forest cover types.", "Mask 2": "Here, besides the emphasis on soil type 34, another soil type (37) and wilderness area are considered. This suggests that both the nature of the terrain (wilderness) and the characteristics of the soil together contribute to the forest cover type.", "Mask 3": "This step introduces elevation as a significant predictor while still retaining a soil type. Elevation, combined with soil characteristics, might influence specific flora and subsequently determine forest cover types.", "Aggregate": "Soil type 34 is the most consistent feature across all the masks, appearing in every step, emphasizing its importance in the forest cover type classification. Another soil type, 29, is also consistent, appearing in both mask 0 and 3. The presence of different features across the steps, like wilderness area (mask 2) and elevation (mask 3), indicates that the landscape's nature and the physical height of the terrain are also important but might offer secondary or nuanced insights into forest cover types. The dominant predictor seems to be soil types, particularly soil type 34, suggesting that specific soil characteristics play a crucial role in determining forest cover types. However, aspects like elevation and the nature of the terrain (wilderness area) also contribute to the classification and shouldn't be overlooked."}

**GPT-4 Output** {"Mask 0": "Initially, only education is somewhat highlighted, suggesting its early recognition as a potential factor in predicting income levels.",

"Mask 1": "This step brings relationship and capital loss into focus, indicating their significant roles in determining income above \$50,000.",

"Mask 2": "Education and marital-status are emphasized, reinforcing the importance of educational background and marital status in income prediction.",

"Mask 3": "Difficulties in discerning important features suggest a complexity in determining clear predictors by this stage.",

"Aggregate": "Across the steps, education appears consistently, underscoring its critical role in income prediction. Relationship and capital loss also emerge as key factors, with their introduction in mask 1 highlighting their importance. The repeated emphasis on marital status alongside education in mask 2 further indicates the significance of personal and social factors in determining income levels. However, the challenges faced in mask 3 reveal the complexities in identifying definitive predictors as feature selection progresses."}

## 8. TabNet Feature Selection with Minimal Decision Steps

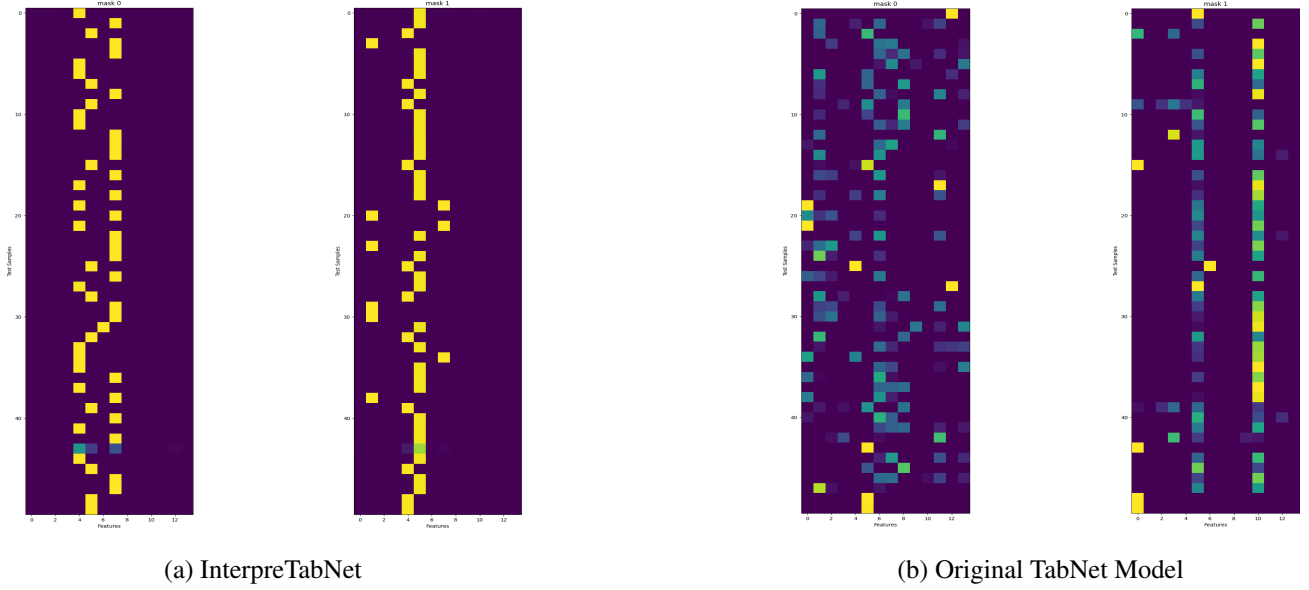


Figure 6: (a) Best performing model using  $r_M^* = 4$  for InterpreTabNet has an accuracy of 87.48 on the Adult Dataset. (b) The baseline performance using the Original TabNet model, attains an accuracy of 86.87%.

We observe in Figure 6 that even when we select the least possible decision steps  $N_{steps} = 2$ , TabNet utilizes every single feature aggregated across the two masks. On the other hand, InterpreTabNet selects clear salient features in its decision making process.



---

## 9. Prompting GPT-4 with Persona of a Domain Expert

We conditioned GPT-4 on a prompt where the model was asked to comment on the feature masks as if they were an expert in the broad topic that the predictive model was built for i.e. "Assume that you are an Economist - Adult Dataset", "Assume that you are a Physician - Diabetes Dataset". This goal of this experiment is to give the reader a qualitative assessment of the variability in the interpretations based on what an expert in the domain might care about.

We summarize the variation of results in Sections 9.1 and 9.2 — what we find is that the expert interpretations do not vary significantly compared to the original interpretations given different "expert" prompts. Therefore, we can anticipate that the interpretations are "robust".

### 9.1. Adult Dataset - Economist

#### GPT-4 Output

Dataset & Expert Persona	Output from GPT-4
Adult Dataset, Economist	{ "Mask 0": "Initial selection highlights workclass, marital status, and relationship as key predictors of income levels, indicating that employment type, marital situation, and family roles are crucial for income classification.", "Mask 1": "This step maintains focus on workclass and marital status, reaffirming their importance in predicting income levels and suggesting a consistent pattern in the significance of employment type and marital situation.", "Mask 2": "The focus shifts to race, capital gain, and hours-per-week, introducing new dimensions such as ethnic background, financial investment returns, and work hours as significant factors influencing income.", "Mask 3": "Reintroduces workclass and relationship along with capital gain, indicating that while employment type and family roles remain pertinent, financial investments emerge as a critical predictor.", "Aggregate": "Workclass and marital status appear as consistent predictors across multiple masks, emphasizing the role of employment type and marital situation in income determination. Relationship and capital gain also emerge as significant, with the latter highlighting financial investments' importance. The inclusion of race and hours-per-week in one mask suggests a broader perspective, considering ethnic background and work dedication. Overall, employment type, marital status, financial investments, and to a lesser extent, family roles, ethnic background, and work hours, are key to predicting income levels." }

---

Adult Dataset, No Persona	{"Mask 0": "This suggests that early in the feature selection process, the model finds that work-related factors (like the type of work) and personal relationships (both marital status and other relationships) are significant predictors for income exceeding \$50,000.", "Mask 1": "Here, the model emphasizes even more on work nature and marital status, suggesting these two as the primary drivers for the prediction. Relationship as a feature has been excluded in this step, which might indicate a possible correlation with marital status.", "Mask 2": "At this step, the model seems to be looking beyond just personal and work nature to more financial and demographic aspects. Capital gain, which is directly related to financial health, is considered along with hours-per-week, potentially pointing to a relationship between longer working hours and higher income. The inclusion of race might indicate underlying societal or economic factors at play.", "Mask 3": "This combination brings back features from both the initial and second masks. Workclass and relationship are revisited, which further underscores their importance. Capital gain remains consistent from the second step, emphasizing its importance as a predictor for income.", "Aggregate": "The features that consistently appear across multiple masks are: workclass (masks 0, 1, and 3), relationship (masks 0 and 3), and capital gain (masks 2 and 3). This highlights that these features play a pivotal role in predicting whether an individual's income exceeds \$50,000 per year. Marital status is seen in the initial stages but disappears later, while hours-per-week and race come into play at the second step. This could indicate that while some features provide foundational predictive power, others may provide more nuanced insight or refinement to the model's predictions. Work-related aspects, personal relationships, and financial metrics seem to be the main predictors of an individual's income level in this dataset. It would be crucial to further analyze these features' exact impacts and understand any potential correlations among them for a more comprehensive understanding."}
---------------------------	--

---

**Prompt** Assume that you are an Economist. Conduct aggregate analysis on the description of the following feature mask. Please output ONLY a dictionary and no other natural language generation when generating the sentence as shown in the in-context example below. Please use single-word classification that encapsulates the meaning of the features if possible.

The Adult Census Income dataset is considered which predicts whether an individual's income exceeds \$50,000 per year. There are 50 test samples and 14 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 1, 5 and 7 which are workclass, marital status and relationship. At the 1st step of feature selection, we observe mask 1 with the main features highlighted as 1 and 5 which are workclass and marital status. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 8, 10 and 12 which are race, capital gain and hours-per-week. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 1, 7 and 10 which are workclass, relationship and capital-gain.

Here are in-context examples for few-shot learning.

The Poker Hand dataset is considered, which is a classification of poker hands into 10 distinct types based on attributes like rank of individual cards, suits, and card combinations. There are 50 test samples and 10 features.

At the 0th step of feature selection, we observe mask 0 with the main feature highlighted as 3 which is the rank of the card 2. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 7 which is the rank of card 3.

---

At the 2nd step of feature selection, we observe mask 2 with the main feature highlighted as 6 which is the suit of card 3. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 3, 6, 7 and 10 which are the rank of cards 1 and 3 as well as the suits of cards 3 and 5.

Output: {"Mask 0": "Initially, the rank of card 2 is recognized as a significant predictor in the classification of poker hands. The rank of individual cards seems crucial at this stage.", "Mask 1": "At this step, the rank of card 3 is highlighted, further emphasizing the importance of individual card ranks in determining poker hand types.", "Mask 2": "Here, besides the rank of card 3, the suit of the same card is considered, suggesting that both rank and suit play roles in classifying poker hands.", "Mask 3": "This step combines multiple features, both ranks (cards 1 and 3) and suits (cards 3 and 5), indicating that a combination of these factors provides a broader understanding of poker hand classifications.", "Aggregate": "The rank of cards, specifically card 2 and card 3, appears consistently across different masks, indicating their primary role in poker hand classification. Suits, particularly of card 3, also have importance but are introduced later in the feature selection. The combined importance of both ranks and suits in the final step suggests that while individual card ranks are pivotal in initial classifications, understanding the relationship between card ranks and their respective suits provides a more comprehensive insight into the poker hand types. The consistent presence of card 3's attributes (both rank and suit) underscores its pivotal role in determining poker hand categories."}

The Forest Cover Type dataset is considered which is a classification of pixels into 7 forest cover types based on attributes such as elevation, aspect, slope, and more. There are 50 test samples and 54 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 24, 29 and 34 which are different soil types. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 34 which is soil type. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 10, 34 and 37 which are wilderness area and soil types. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 0 and 29 which are elevation, and soil type.

Output: {"Mask 0": "The initial feature selection identifies that various soil types are significant predictors for the forest cover types. It appears that the characteristics of the soil play a vital role in determining the type of forest cover.", "Mask 1": "At this step, the model emphasizes even more on one particular soil type (34), suggesting it as a primary driver for the classification. This specific soil type may have unique characteristics making it a distinctive feature for certain forest cover types.", "Mask 2": "Here, besides the emphasis on soil type 34, another soil type (37) and wilderness area are considered. This suggests that both the nature of the terrain (wilderness) and the characteristics of the soil together contribute to the forest cover type.", "Mask 3": "This step introduces elevation as a significant predictor while still retaining a soil type. Elevation, combined with soil characteristics, might influence specific flora and subsequently determine forest cover types.", "Aggregate": "Soil type 34 is the most consistent feature across all the masks, appearing in every step, emphasizing its importance in the forest cover type classification. Another soil type, 29, is also consistent, appearing in both mask 0 and 3. The presence of different features across the steps, like wilderness area (mask 2) and elevation (mask 3), indicates that the landscape's nature and the physical height of the terrain are also important but might offer secondary or nuanced insights into forest cover types. The dominant predictor seems to be soil types, particularly soil type 34, suggesting that specific soil characteristics play a crucial role in determining forest cover types. However, aspects like elevation and the nature of the terrain (wilderness area) also contribute to the classification and shouldn't be overlooked."}

---

## 9.2. Diabetes Dataset - Physician

### GPT-4 Output

Dataset & Expert Persona	Output from GPT-4
Diabetes Dataset, Physician	<p>{ "Mask 0": "This initial selection highlights the importance of discharge disposition, the number of diagnoses, and specific medications (glyburide and citoglipton) in predicting readmissions. It suggests that patient outcomes and the specific treatment protocols, including medication use, play a crucial role in the likelihood of a patient being readmitted.",</p> <p>"Mask 1": "At this stage, the model focuses on the admission type and discharge disposition, indicating the significance of the nature of admission and the patient's discharge status in predicting readmissions. This could reflect how initial hospitalization circumstances and post-discharge plans impact patient outcomes.",</p> <p>"Mask 2": "Here, the inclusion of admission type, number of diagnoses, and metformin (another medication) suggests a continued emphasis on the circumstances of admission and the complexity of the patient's health condition, as well as the role of specific diabetes medication in influencing readmission risks.",</p> <p>"Mask 3": "By focusing on the number of diagnoses and medications (glipizide and metformin), this step underscores the critical nature of the patient's health status and the treatment of diabetes as key factors in predicting hospital readmissions. It highlights the significance of managing complex health conditions and diabetes treatment.",</p> <p>"Aggregate": "Across the selection process, the consistent features are the number of diagnoses and the emphasis on specific diabetes medications (metformin appearing in masks 2 and 3, and other diabetes-related medications in masks 0 and 3). This indicates a strong relationship between the complexity of a patient's health (as represented by the number of diagnoses) and the treatment of diabetes in predicting readmissions. Discharge disposition and admission type are also significant but might offer insights into the logistical aspects of care that affect readmission risks. The aggregate analysis underscores the complexity of managing diabetes within the hospital setting and its impact on readmission rates, highlighting the importance of both clinical and administrative factors." }</p>

---

Diabetes Dataset, No Persona	{ "Mask 0": "In the initial step, the model identifies discharge disposition id, number of diagnoses, and specific medications (glyburide, and citoglipton) as key features. This suggests a combination of administrative data (discharge process), clinical data (diagnoses), and prescribed medications play a significant role in predicting patient readmission.", "Mask 1": "Here, the model focuses on the patient's admission type and discharge disposition, emphasizing the importance of the hospital admission and discharge process in determining the likelihood of readmission. The absence of clinical and medication features in this step indicates a potential higher influence of administrative processes at this stage of feature selection.", "Mask 2": "The model reintroduces clinical data (number of diagnoses) and adds a new medication (metformin), while retaining admission type id. This combination underscores a nuanced interplay between the nature of hospital admission, the complexity of patient's health conditions, and specific treatments in predicting readmissions.", "Mask 3": "This step focuses on the number of diagnoses and two medications (glipizide and metformin). The consistency of number of diagnoses across masks highlights its pivotal role, while the introduction of glipizide and retention of metformin point towards an evolving emphasis on medication in conjunction with clinical data.", "Aggregate": "Number diagnoses (masks 0, 2, and 3) and discharge disposition id (masks 0 and 1) are recurrent, underscoring their persistent importance in predicting readmissions. Admission type id also appears consistently (masks 1 and 2), linking the administrative process tightly with readmission predictions. The changing set of medications across the masks (glyburide, citoglipton, metformin, glipizide) points to the dynamic nature of treatment's role in the prediction, warranting a closer look into specific medication combinations and their interactions with clinical and administrative data. The aggregate analysis reveals a complex interplay between administrative processes, clinical data, and medication in predicting patient readmissions, suggesting a multifaceted approach for model refinement and interpretation." }
------------------------------	--

---

**Prompt** Assume you are a Physician. Conduct aggregate analysis on the description of the following feature masks. Start off with an analysis of the individual masks, followed by an aggregate analysis of all masks combined. Please format the output into a dictionary as shown in the in-context examples. The output should only contain the formatted output, no other natural language generation is required.

The Diabetes 130-US hospitals for years 1999-2008 dataset is considered, which is a classification of patient encounters into readmitted or not readmitted categories based on attributes like the number of laboratory tests performed, the number of medications prescribed, diagnoses, and other clinical and administrative data. There are 50 test samples and 50 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 7, 21, 31, and 40 which are discharge disposition id, number diagnoses, glyburide, and citoglipton. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 6 and 7 which are admission type id and discharge disposition id. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 6, 21, and 45 which are admission type id, number diagnoses, and metformin. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 21, 30, 45 which are number diagnoses, glipizide, and metformin.

Here are in-context examples for few-shot learning.

The Adult Census Income dataset is considered which predicts whether an individual's income exceeds \$50,000 per year.

---

There are 50 test samples and 14 features. At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 1, 5 and 7 which are workclass, marital status and relationship. At the 1st step of feature selection, we observe mask 0 with the main features highlighted as 1 and 5 which are workclass and marital status. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 8, 10 and 12 which are race, capital gain and hours-per-week. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 1, 7 and 10 which are workclass, relationship and capital-gain.

Output: {"Mask 0": "This suggests that early in the feature selection process, the model finds that work-related factors (like the type of work) and personal relationships (both marital status and other relationships) are significant predictors for income exceeding \$50,000.", "Mask 1": "Here, the model emphasizes even more on work nature and marital status, suggesting these two as the primary drivers for the prediction. Relationship as a feature has been excluded in this step, which might indicate a possible correlation with marital status.", "Mask 2": "At this step, the model seems to be looking beyond just personal and work nature to more financial and demographic aspects. Capital gain, which is directly related to financial health, is considered along with hours-per-week, potentially pointing to a relationship between longer working hours and higher income. The inclusion of race might indicate underlying societal or economic factors at play.", "Mask 3": "This combination brings back features from both the initial and second masks. Workclass and relationship are revisited, which further underscores their importance. Capital gain remains consistent from the second step, emphasizing its importance as a predictor for income.", "Aggregate": "The features that consistently appear across multiple masks are: workclass (masks 0, 1, and 3), relationship (masks 0 and 3), and capital gain (masks 2 and 3). This highlights that these features play a pivotal role in predicting whether an individual's income exceeds \$50,000 per year. Marital status is seen in the initial stages but disappears later, while hours-per-week and race come into play at the second step. This could indicate that while some features provide foundational predictive power, others may provide more nuanced insight or refinement to the model's predictions. Work-related aspects, personal relationships, and financial metrics seem to be the main predictors of an individual's income level in this dataset. It would be crucial to further analyze these features' exact impacts and understand any potential correlations among them for a more comprehensive understanding."}

The Forest Cover Type dataset is considered which is a classification of pixels into 7 forest cover types based on attributes such as elevation, aspect, slope, and more. There are 50 test samples and 54 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 24, 29 and 34 which are different soil types. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 34 which is soil type. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 10, 34 and 37 which are wilderness area and soil types. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 0 and 29 which are elevation, and soil type.

Output: {"Mask 0": "The initial feature selection identifies that various soil types are significant predictors for the forest cover types. It appears that the characteristics of the soil play a vital role in determining the type of forest cover.", "Mask 1": "At this step, the model emphasizes even more on one particular soil type (34), suggesting it as a primary driver for the classification. This specific soil type may have unique characteristics making it a distinctive feature for certain forest cover types.", "Mask 2": "Here, besides the emphasis on soil type 34, another soil type (37) and wilderness area are considered. This suggests that both the nature of the terrain (wilderness) and the characteristics of the soil together contribute to the forest cover type.", "Mask 3": "This step introduces elevation as a significant predictor while still retaining a soil type. Elevation, combined with soil characteristics, might influence specific flora and subsequently determine forest cover types.", "Aggregate": "Soil type 34 is the most consistent feature across all the masks, appearing in every step, emphasizing its importance in the forest cover type classification. Another soil type, 29, is also consistent, appearing in both mask 0 and 3. The presence of different features across the steps, like wilderness area (mask 2) and elevation (mask 3), indicates that the landscape's nature and the physical height of the terrain are also important but might offer secondary or nuanced insights into forest cover types. The dominant predictor seems to be soil types, particularly soil type 34, suggesting that specific soil characteristics play a crucial role in determining forest cover types. However, aspects like elevation and the nature of the terrain (wilderness area) also contribute to the classification and shouldn't be overlooked."}

---

## References

- S. O. Arik and T. Pfister. Tabnet: Attentive interpretable tabular learning, Feb 2020. URL <https://arxiv.org/abs/1908.07442v4>.
- J. Chen, L. Song, M. J. Wainwright, and M. I. Jordan. Learning to explain: An information-theoretic perspective on model interpretation, 2018.
- P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Mach. Learn.*, 63(1):3–42, apr 2006. ISSN 0885-6125. doi: 10.1007/s10994-006-6226-1. URL <https://doi.org/10.1007/s10994-006-6226-1>.
- J. Yoon, J. Jordon, and M. van der Schaar. INVASE: Instance-wise variable selection using neural networks. In *International Conference on Learning Representations*, 2019. URL [https://openreview.net/forum?id=BJg\\_roAcK7](https://openreview.net/forum?id=BJg_roAcK7).