1 Tables

Event Log	Accuracy		
	Base	Modified	Enriched
−age, −gender	.916 ± .002	.915 ± .002	.915 ± .002
−age, +gender	.916 ± .002	.920 ± .004	.926 ± .003
+age, −gender	.916 ± .002	.920 ± .003	.928 ± .004
+age, +gender	.916 ± .002	.923 ± .003	.932 ± .003

Event Log	ΔDP (Age)			ΔDP (Gender)		
	Base	Modified	Enriched	Base	Modified	Enriched
−age, −gender	.014 ± .012	.014 ± .011	.014 ± .011	.009 ± .003	.021 ± .018	.021 ± .018
−age, +gender	.010 ± .005	.008 ± .001	.042 ± .028	.008 ± .005	.064 ± .041	.837 ± .081
+age, −gender	.014 ± .015	.030 ± .058	.903 ± .033	.037 ± .026	.033 ± .031	.052 ± .033
+age, +gender	.018 ± .013	.005 ± .003	.917 ± .017	.007 ± .006	.004 ± .002	.836 ± .044

Tab. 1.1: Evaluation of accuracy and ΔDP for the four versions of the *Hospital Billing* log. The attributes *age* and *gender* are annotated based on whether they introduce a bias (+) or not (−). The reported values represent the mean (μ) and standard deviation (σ) across validation folds, expressed as $\mu \pm \sigma$.

Event Log	Accuracy			ΔDP		
	Base	Modified	Enriched	Base	Modified	Enriched
cs	.816 ± .005	.848 ± .003	.890 ± .001	.004 ± .004	.002 ± .003	.996 ± .002
hb (−age −gender)	.916 ± .002	.915 ± .002	.915 ± .002	.011 ± .004	.017 ± .005	.017 ± .005
hb (−age +gender)	.916 ± .002	.920 ± .004	.926 ± .003	.009 ± .001	.036 ± .039	.439 ± .563
hb (+age −gender)	.916 ± .002	.920 ± .003	.928 ± .004	.026 ± .016	.032 ± .002	.478 ± .602
hb (+age +gender)	.916 ± .002	.923 ± .003	.932 ± .003	.012 ± .008	.005 ± .001	.877 ± .057
bpi	.817 ± .003	.822 ± .001	.829 ± .004	.058 ± .004	.053 ± .006	.510 ± .094

Tab. 1.2: Evaluation of accuracy and ΔDP for the *Cancer Screening (cs)* log, the *BPI Challenge 2012 (bpi)* log, and four versions of the *Hospital Billing (hb)* log, where the attributes *age* and *gender* are annotated based on whether they introduce a bias (+) or not (−). Since the *hb* event log uses two sensitive attributes, we report their average ΔDP. The reported values represent the mean (μ) and standard deviation (σ) across five validation folds, expressed as $\mu \pm \sigma$.

Event Log	Number of Nodes	Number of removed Nodes	Depth
cs	26.6 ± 4.7	1 ± 0	10.8 ± 0.4
hb (−age, −gender)	33.4 ± 2.6	0 ± 0	11.0 ± 0.7
hb (−age, +gender)	45.4 ± 3.2	1 ± 0	12.2 ± 0.8
hb (+age, −gender)	42.2 ± 4.3	1 ± 0	12.8 ± 1.0
hb (+age, +gender)	47.0 ± 3.4	2 ± 0	12.8 ± 0.4
bpi	38.2 ± 3.0	1 ± 0	11.2 ± 0.4

Tab. 1.3: Evaluation of the characteristics of the distilled decision tree for the *Cancer Screening (cs)* log, the *BPI Challgenge 2012 (bpi)* log and all versions of the *Hospital Billing (hb)* log. For *hb*, the attributes *age* and *gender* are annotated based on whether they introduce bias (+) or not (−). The reported values represent the mean (μ) and standard deviation (σ) across validation folds, expressed as $\mu \pm \sigma$.

Bias Strength	Accuracy			ΔDP		
	Base	Modified	Enriched	Base	Modified	Enriched
0.50	.797 ± .004	.851 ± .003	.851 ± .003	.005 ± .008	.001 ± .001	.001 ± .001
0.55	.805 ± .003	.848 ± .003	.861 ± .006	.001 ± .002	.002 ± .002	.800 ± .444
0.60	.808 ± .003	.849 ± .003	.869 ± .001	.000 ± .001	.003 ± .003	.996 ± .002
0.65	.813 ± .003	.851 ± .003	.882 ± .002	.002 ± .002	.001 ± .001	.973 ± .055
0.70	.817 ± .003	.850 ± .003	.892 ± .003	.002 ± .002	.000 ± .000	.999 ± .001
0.75	.825 ± .003	.851 ± .004	.904 ± .002	.002 ± .002	.000 ± .001	.999 ± .001
0.80	.825 ± .005	.848 ± .007	.912 ± .004	.001 ± .001	.001 ± .001	.998 ± .001
0.85	.838 ± .002	.853 ± .002	.925 ± .003	.001 ± .001	.001 ± .001	.999 ± .002
0.90	.841 ± .002	.852 ± .004	.936 ± .002	.001 ± .002	.000 ± .001	.999 ± .001
0.95	.844 ± .004	.850 ± .005	.945 ± .003	.004 ± .004	.000 ± .001	.000 ± .000
1.00	.854 ± .003	.854 ± .003	.958 ± .001	.004 ± .002	.001 ± .001	.000 ± .001

Bias Strength	Number of Nodes	Number of removed Nodes	Depth
0.50	27.0 ± 4.7	0.0 ± 0.0	10.8 ± 0.4
0.55	25.0 ± 1.4	0.8 ± 0.4	10.6 ± 0.5
0.60	24.6 ± 1.7	1.0 ± 0.0	10.4 ± 0.5
0.65	27.8 ± 4.4	1.0 ± 0.0	10.2 ± 0.4
0.70	27.0 ± 4.7	1.0 ± 0.0	10.2 ± 0.4
0.75	26.2 ± 5.0	1.0 ± 0.0	10.6 ± 0.5
0.80	29.4 ± 5.9	1.0 ± 0.0	11.0 ± 0.0
0.85	25.4 ± 1.7	1.0 ± 0.0	10.8 ± 0.4
0.90	25.8 ± 1.1	1.0 ± 0.0	10.8 ± 0.4
0.95	27.8 ± 5.4	1.0 ± 0.0	10.8 ± 0.4
1.00	23.0 ± 0.0	1.0 ± 0.0	10.0 ± 0.0

Tab. 1.4: Evaluation of the accuracy, ΔDP and characteristics of the distilled decision tree for varying bias strengths. The reported values represent the mean (μ) and standard deviation (σ) across validation folds, expressed as $\mu \pm \sigma$.

Num. Attributes	Accuracy			ΔDP		
	Base	Modified	Enriched	Base	Modified	Enriched
1	.807 ± .003	.820 ± .003	.862 ± .003	.000 ± .000	.000 ± .000	.999 ± .001
2	.796 ± .002	.807 ± .003	.830 ± .002	.001 ± .001	.045 ± .099	.501 ± .009
4	.799 ± .001	.807 ± .002	.822 ± .004	.001 ± .001	.005 ± .010	.374 ± .005
6	.799 ± .005	.805 ± .005	.816 ± .005	.002 ± .002	.028 ± .014	.267 ± .077
8	.799 ± .004	.801 ± .001	.812 ± .002	.001 ± .001	.020 ± .013	.238 ± .029
10	.801 ± .004	.803 ± .003	.811 ± .005	.004 ± .007	.033 ± .021	.168 ± .094

Num. Attributes	Number of Nodes	Number of removed Nodes	Depth
1	24.6 ± 2.6	1.0 ± 0.0	10.4 ± 0.9
2	31.0 ± 4.2	1.8 ± 0.4	10.4 ± 0.9
4	47.8 ± 4.1	7.4 ± 0.9	10.8 ± 0.4
6	81.8 ± 19.7	21.0 ± 6.0	11.6 ± 0.9
8	77.8 ± 15.7	22.4 ± 5.2	10.8 ± 0.4
10	60.2 ± 16.0	13.2 ± 8.8	10.8 ± 0.4

Tab. 1.5: Evaluation of the accuracy, ΔDP and characteristics of the distilled decision tree for a varying amount of sensitive attributes. The reported values represent the mean (μ) and standard deviation (σ) across validation folds, expressed as $\mu \pm \sigma$.

Num. Decisions	Accuracy			ΔDP		
	Base	Modified	Enriched	Base	Modified	Enriched
2	.765 ± .001	.793 ± .002	.843 ± .002	.003 ± .003	.001 ± .001	.998 ± .003
4	.684 ± .006	.710 ± .019	.791 ± .003	.000 ± .001	.007 ± .002	.997 ± .002
8	.627 ± .004	.648 ± .009	.749 ± .003	.003 ± .003	.009 ± .009	.983 ± .021
12	.595 ± .003	.648 ± .001	.729 ± .003	.003 ± .002	.005 ± .004	.968 ± .019
16	.582 ± .004	.638 ± .004	.716 ± .002	.004 ± .003	.002 ± .001	.939 ± .028
20	.570 ± .002	.625 ± .005	.702 ± .006	.003 ± .002	.004 ± .001	.921 ± .035

Num. Decisions	Number of Nodes	Number of removed Nodes	Depth
2	19.8 ± 2.3	1.0 ± 0.0	6.8 ± 0.4
4	33.0 ± 3.5	2.0 ± 0.0	10.4 ± 0.5
8	61.8 ± 3.6	5.8 ± 1.1	16.4 ± 0.9
12	90.2 ± 1.8	9.6 ± 0.9	25.0 ± 0.0
16	124.6 ± 0.9	14.6 ± 0.9	33.0 ± 0.0
20	154.2 ± 1.8	17.6 ± 0.9	41.0 ± 0.0

Tab. 1.6: Evaluation of the accuracy, ΔDP and characteristics of the distilled decision tree for a varying amount of biased decisions. The reported values represent the mean (μ) and standard deviation (σ) across validation folds, expressed as $\mu \pm \sigma$.

Event Log	Accuracy (NN)	Accuracy (DT)
cs	.848 ± .003	.847 ± .003
hb (−age, −gender)	.915 ± .002	.913 ± .002
hb (−age, +gender)	.920 ± .004	.918 ± .004
hb (+age, −gender)	.920 ± .003	.917 ± .003
hb (+age, +gender)	.923 ± .003	.919 ± .002
bpi	.822 ± .001	.823 ± .002

Tab. 1.7: Comparison of the accuracy of the modified neural network (*NN*) against the accuracy of the modified decision tree (*DT*) for the *Cancer Screening (cs)* log, the *BPI Challgenge 2011 (bpi)* log and all versions of the *Hospital Billing (hb)* log. For *hb*, the attributes *age* and *gender* are annotated based on whether they introduce bias (+) or not (−). The reported values represent the mean (μ) and standard deviation (σ) across validation folds, expressed as $\mu \pm \sigma$.